Multivariate product adapted grading of Scots Pine sawn timber for an industrial customer, part 2: Robustness to disturbances

Linus Olofsson a, b, Olof Broman a, Johan Skog b, Magnus Fredriksson a and Dick Sandberg a, b

aWood Science and Engineering, Luleå University of Technology, Skellefteå, Sweden; bRISE Bioeconomy, Research Institutes of Sweden, Skellefteå, Sweden

ABSTRACT

Holistic-subjective automatic grading (HSAG) of sawn timber by an industrial customer’s product outcome is possible through the use of multivariate partial least squares discriminant analysis (PLS-DA), shown by part one of this two-part study. This second part of the study aimed at testing the robustness to disturbances of such an HSAG system when grading Scots Pine sawn timber partially covered in dust. The set of 308 clean planks from part one of this study, and a set of 310 dusty planks, that by being stored inside a sawmill accumulated a layer of dust, were used. Cameras scanned each plank in a sawmill’s automatic sorting system that detected selected feature variables. The planks were then split and processed at a planing mill, and the product grade was correlated to the measured feature variables by partial least squares regression. Prediction models were tested using 5-fold cross-validation in four tests and compared to the reference result of part one of this study. The tests showed that the product adapted HSAG could grade dusty planks with similar or lower grading accuracy compared to grading clean planks. In tests grading dusty planks, the disturbing effect of the dust was difficult to capture through training.

Introduction

Part one of this two-part study showed that it was possible to predict a customer’s product grade outcome by holistically and subjectively grade sawn timber using partial least squares discriminant analysis. However, given a sawmill’s dusty environment, it is important to know how a visual grading system is affected by disturbances. This second part of the study aimed at testing if a holistic-subjective automatic grading system is robust to disturbance in the form of a layer of dusty on top of the scanned sawn timber.

Method development (part 1 summary)

Automatic dry sorting stations at sawmills in Scandinavia use mainly objective rule-based automatic grading (RBAG) to sort sawn timber into standardized visual quality grades, e.g. (NTGR, 1994). RBAG is objective by nature, i.e. it uses different measurement rules (limits) to define grades. The problem with using rules to separate grades that motivated part one of this study (Olofsson et al., 2019), as well as previous work by Lycken and Oja (2006); Berglund et al. (2015), and Olofsson et al. (2017), was that a large set of correlated grading rules had to be manually created by some expert for a coherent grading. Sawmill customers often have a holistic-subjective view of sawn timber quality, meaning they judge the whole piece at the same time and rules can be overridden based on the overall appearance of the piece. The large set of grading rules required, and the difficulty in defining them to holistically describe a customer’s subjective definition of desirable quality sawn timber means that attempts to customize grading rules are seldom made. This problem is prominent when a sawmill is trying to make customized quality grades for customers whose needs are not at all in line with the standardized grades, as this requires that big changes to the standardized grading rules have to be made. The collaborating industrial planing mill in this study is an example of this. The problem with customizing an RBAG system for a new holistic-subjective grade manifests itself primarily in two ways defined by Lycken and Oja (2006) as: (1) it is difficult for a customer to describe their subjective view of the desired sawn timber quality in a way that can easily be defined in objective grading rules, (2) the number of variables that can be controlled to specify a grade is often large enough to make customization complicated.

The problems with RBAG for a customer were addressed in part one of this study by implementing a holistic-subjective automatic grading (HSAG) method of Scots pine sawn timber with a collaborating sawmill and planing mill. Using the current sawmill hardware, an automatic scanning, grading, and sorting system was used to show that it is possible to grade sawn timber with HSAG according to the planing mills quality grade outcome. The grading method used was multivariate partial least squares discriminant analysis (PLS-DA), where prediction models were trained on aggregated feature measurements of each plank and the manually determined quality grade yield. This was done even though the planing mill split, planed, and milled the sawn timber.
The tested grading models gave an overview of the grading accuracy level, and sensitivity to the random selection of test sets. The prediction models showed a stable behaviour, indicating that the PLS-DA can be suitable for HSAG. The HSAG was in part tested by 5-fold cross-validation to correctly grade 74% of the planks, in a data set of 308 planks. This grading accuracy can be compared to the 76–85% grading accuracy achieved in re-substitution tests in previous work by Berglund et al. (2015) and Lycken and Oja (2006) when grading sawn timber according to visual standard grades from NTGR (1994).

Grading results from a PLS-based prediction model are highly dependent on how accurately the test set is represented by the training set. Since the input to a sawmill is naturally heterogeneous wood, it is highly probable that the input material to the sorting station has never been experienced by the currently active grading model. The problem of having a training set that is representative of the test set is especially important in the context of systematic differences, due to e.g. difference in visual scanning conditions. This gave rise to the question of how robust a PLS-based HSAG grading method is to unforeseen disturbances (changes) of the input. This was tested in part two of this study project.

Robustness to disturbances

Some automatic dry sorting stations in Scandinavian sawmills grade sawn timber by visual appearance by objective rule-based automatic grading (RBAG). This is done by cameras, advanced feature detection algorithms, and a set of well-defined grading rules. Such grading rules defines feature limits, e.g. the maximum size of dead knots or maximum number of sound knots, and are often created to follow standards like NTGR (1994). Lycken and Oja (2006) stated that the hit-rate for defect classification in Nordic visual automatic grading systems was around 70–80%, which resulted in approximately 80–90% plank-grade hit-rate, depending on grading rules and material. Part one of this study graded sawn timber using the same hardware and feature detection grading rules and material. Part one of this study graded sawn timber using the same software and feature detection grading rules and material. Part one of this study graded sawn timber using the same software according to visual standard grades from NTGR (1994).

Apart from grading accuracy, the selected thresholds were compared to the respective optimal threshold values. Since the threshold was chosen based on the training set, the selected threshold might be far from its optimal value if the training set does not represent the test set due to the disturbance of the dust.

Materials and methods

The collection of scanner data from the sawmill, and customer response from the planing mill; the construction of regression components; the implementation of PLS prediction models; and the testing methodology followed the same procedure as in part one of this study (Olofsson et al., 2019). The scanner, grader, and sorting system used was the same Boardmaster by FinScan (Anon, 2018) at Kåge Sawmill. Dialogue with FinScan ensured that the scanner data of the two batches were comparable. Lundgren’s planing mill was again the industrial customer with the same product and production process as in part one of this study.
**Material**

Two data sets were used: (1) the 308 planks used in part one of this study, referred to as the clean data set; and (2) a new set of 310 planks with the unforeseen dusty disturbance, referred to as the dusty data set. Both sets of planks originated from the same large sawing batch, so that there were no differences in sawing or handling conditions, except for that the second, dusty, data set had been stored uncovered indoors at the sawmill for one year before being scanned.

The dusty data set consisted of 310 Scots pine (*Pinus sylvestris* L.) planks, sawn from top-logs from the sawmill’s log yard sorting station. Each log was cant sawn, and the centre yield of two planks was used for the study. The planks measured 50 by 150 mm in cross-section and varied between 3.4 and 5.5 m in length. The planks were dried to 14% moisture content before being stored and eventually scanned at the sawmill.

During storage, the planks were exposed to the dusty environment which left dust stains especially on the ends of the planks, and with visible marks from the spacers. Example of planks from the two data sets are shown in Figure 1. Almost every plank in the second data set had one dusty side and one clean side, as the dust settled on top of each plank, but some few planks were clean altogether or dusty on both sides. The plank orientation was considered random, i.e. the dust settled randomly on either plank face. The Boardmaster was calibrated as in part one of the study as if the dust was not present.

**Data collection**

The ID marked sawn timber was scanned by the Boardmaster and feature variables regarding knots and bark were saved before the timber was delivered to the planing mill. Other feature variables detected by the Boardmaster, such as wane or cracks, were ignored in this study as knots and bark features are critical for the customer and suitable for holistic-subjective grading. The ignored variables were ignored by the planing mill as well. Each plank was split into three boards which were each planed, milled, and manually graded – each board given grade A for desirable or grade B for undesirable for the planing mill. Each plank ID was associated with a digital label according to the majority of board grades produced, to study borderline cases, i.e. planks were labelled AAA, AA, BB, or BBB – omitting a B or A from the mixed labels AAB and BBA to keep labelling clean. Each plank was given a grade A or grade B based on the majority of grade A or grade B boards produced, i.e. AAA and AA was given the grade A for desirable, and BB and BBB planks the grade B for undesirable.

The quality distributions of the two data sets are given in Table 1, which showed that the two data sets had very similar proportions of grade A planks, with a slightly larger proportion labelled AAA in the dusty data set.

**Applying PLS regression**

Below is a brief overview of the implementation of PLS described, for a more thorough explanation, see Olofsson et al. (2019) – especially Table 1 and Figure 1. Using the knot and bark measurements from the Boardmaster, which provides size, position, and defect type, a detailed set of 3564 aggregated variables were created to expand the ability of the HSAG system to capture, in objective measurements, the subjective quality traits desired by the customer. Each of the 22 created variables was measured for 6 defect types, separately for each of the plank’s 3 faces (edges together), and separately measured in 9 different sections of each plank which was divided into 1, 3, and 5 sections in the lengthwise direction of the plank. In total, $22 \cdot 6 \cdot 3 \cdot 9 = 3564$ variables were created.

Using the aggregated feature variables from the Boardmaster and the planing mill quality grade assessment, both data sets could be used to create an explanatory matrix $X$ (up to 618 by 3564) and a response matrix $Y$ (up to 618 by 1). Using the SIMCA 14 software (Anon, 2019), these matrices were used to train PLS regression models based on different training sets, which were then used to predict the quality outcome of a series of test sets.

Based on new measurements of a test plank, a PLS prediction model predicts a value approximately between 0 and 1

---

**Table 1.** The number of clean and dusty planks in each data set, shown with proportion of each label and plank grade.

<table>
<thead>
<tr>
<th>Plank grade</th>
<th>Clean set</th>
<th>Dusty set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Label</td>
<td>Number</td>
</tr>
<tr>
<td>A</td>
<td>AAA</td>
<td>126</td>
</tr>
<tr>
<td></td>
<td>AA</td>
<td>73</td>
</tr>
<tr>
<td>B</td>
<td>BB</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>BBB</td>
<td>66</td>
</tr>
<tr>
<td>Totals</td>
<td></td>
<td>308</td>
</tr>
</tbody>
</table>

---

**Figure 1.** Examples of three clean planks of data set 1 (top), and three dusty sides of planks in data set 2 (bottom). The bottom planks show clear dust shades between the clear marks caused by the spacers. The number written on each plank is the unique ID-number of each data set and the numbers 8 and 9 repeats by chance.
which can be thought of as a probability of that plank being of grade A. A threshold value decides what the prediction model classifies as of grade A or not. The class-separating threshold of each prediction model is determined by choosing a threshold that optimally separates the classes A and B in the training set of that prediction model.

**Testing specifications**

From part one of this study, the average grading results of the 5-fold cross-validation of the clean data set served as the reference case. The grading accuracy of the PLS-based HSAG method was tested according to four user-case scenarios, and compared to the reference case (test 0), detailed in Table 2.

Six new PLS prediction models are created in part two of this study. One model was trained on the complete clean data set, and based on the training set the threshold 0.56 was selected. This model was used in test (1) to test the discriminant model’s robustness to unforeseen disturbances (research question (1) presented in the introduction). Five more prediction models were created for 5-fold cross-validation of both the clean and the dusty data sets together. The benefit of cross-validation over a single training and test set is that the average behaviour is less test set sensitive, which is necessary due to the high variability and complexity of the data sets used in this study. Each of the five models was tested on a unique fifth of the combined data set, and trained on the rest. The test sets were proportionally selected from both the clean and dusty data sets, as well as proportionally selected from each label class. For each model a threshold was determined from the training set, which on average was 0.56 and the average prediction result was calculated. The same average prediction results of the 5-fold cross-validation was used in tests (2–4), where tests (3–4) are a split of test sets from test (2). Tests (2–4) share the same prediction results and the same 5 prediction models for the cross-validation, the results of test (2) are separated into tests (3) and (4) to show differences in grading outcome between the different test sets. These tests were performed to test the average grading accuracy of the discriminant models when taking the dust into account by training the prediction models on a mixed set of planks (research question (2) in the introduction).

The robustness of a grading model to unforeseen disturbances was, apart from grading accuracy, also investigated by studying the used threshold compared to the optimal threshold value for the different tests. Selecting a threshold based on a training set that is not is not representative of the intended test set could result in a lowered grading accuracy and was therefore also investigated. The maximum grading accuracy of e.g. test (1) could be the same as for the reference model (test 0) but at a different threshold than for the reference case. In part one of this study, the average grading accuracy of the 5-fold cross-validation was 74% at the average class-separating threshold 0.55, compared to the average peak grading accuracy of 76% at the optimal threshold 0.52. Compared to the optimal, this is a loss of 2 percentage points of grading accuracy and a 0.03 points miss of the optimal threshold value. Similar comparisons will be made for each test in part two of this study. This loss of grading accuracy and miss of optimal threshold is due to the fact that the tested model was not optimized for the current test set. Furthermore, the dust layer was expected to lower the grading accuracy, especially in test (1).

**Results**

To evaluate the effects of unforeseen disturbances on the accuracy of PLS-based HSAG, the results of several prediction models were tested in tests 1–4. Prediction results were compared with the reference 5-fold cross-validation analysis from part one of this study, presented as test (0) in Table 3 for easy reference. The test specifications are detailed in Table 2 and the results of each test are presented in Table 3 using the corresponding selected threshold.

In test (1), a single prediction model was trained on the entire clean data set from part one of this study and was tested by grading the entire dusty data set from part 2 of this study. The prediction results are visualized in Figure 2 where the prediction results are shown for the selected class-separating threshold 0.56, and as test (1) in Table 3. In test (1), 73% of the dusty planks were correctly graded in comparison to the 74% correctly graded planks according to the reference case.

**Optimal grading accuracy and threshold**

The optimal grading accuracy of a prediction model could in hindsight be found at some optimal class-separating threshold for each test. A small loss of grading accuracy

---

### Table 2. Test setups and objectives.

<table>
<thead>
<tr>
<th>#</th>
<th>Training set</th>
<th>Testing set</th>
<th>Test objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Clean</td>
<td>Clean</td>
<td>The reference use-case where the sawmill graded clean planks by a prediction model trained on clean planks.</td>
</tr>
<tr>
<td>1</td>
<td>Clean</td>
<td>Dusty</td>
<td>To test the effect of the dust by grading dusty plank with a prediction model trained on clean planks.</td>
</tr>
<tr>
<td>2</td>
<td>Clean and dusty</td>
<td>Clean and dusty</td>
<td>To test the behaviour of a prediction model that was trained to take unforeseen disturbances into account by training the prediction model on both clean and dusty planks, here grading a set of mixed clean and dusty planks.</td>
</tr>
<tr>
<td>3</td>
<td>Clean and dusty</td>
<td>Clean</td>
<td>To test the behaviour of a prediction model that was trained unnecessarily to take unforeseen disturbances into account when grading only clean planks.</td>
</tr>
<tr>
<td>4</td>
<td>Clean and dusty</td>
<td>Dusty</td>
<td>To show the difference in grading outcome compared to test (1) when grading dusty planks when the prediction models were trained on a training set including dusty planks.</td>
</tr>
</tbody>
</table>
compared to the optimal was expected as the threshold selection was not optimized for the test sets. The grading accuracy of the reference test (0) is shown in Table 4 to miss out on 2 percentage points of grading accuracy by using a threshold 0.03 points away from the optimal class-separating threshold. The grading accuracy of the prediction model(s) of each test, and their respective (average) thresholds, are compared to the optimal setting in Table 4.

Discussion

The present study showed that the HSAG system used was robust to, but not unaffected by, unforeseen disturbances in the form of a dust layer on the sawn timber. As in part one of this study, it was evident that when the sawmill and customer discuss which threshold to use to reach satisfactory grading outcome, they need to take grading accuracy as well as grading outcome into account. For example, the grading and sorting outcome from the sawmill using a very high class-separating threshold, say 0.8, would be a very small batch of almost entirely grade A planks with almost no incorrectly graded B-grade planks in the delivered batch to the planing mill, but at a great cost for the customer due to the low volume (see Figure 5 in part one of this study). For this reason, a threshold close to 0.5 is preferable for a balance of sawmill and customer satisfaction, as well as being close to the optimal grading accuracy. The prediction results of all prediction models are conceptually visualized by Figure 2.

Table 3. Misclassification table (confusion matrix) of the reference prediction results (test 0, Olofsson et al., 2019), and for all the tests performed (tests 1–4), showing the grading outcome of each respective test for each plank-label class, using the selected thresholds. Test (0) used the average class-separating threshold 0.55, test (1) used the class-separating threshold 0.56, and tests 2–4 used the average class-separating threshold 0.56. Predicted plank grades were determined at the sawmill, and the observed grades were the grade outcome of each plank at the planing mill. Tests 0, 2, 3, and 4 are based on 5-fold cross-validation and shows the average values. Predicted number of planks are rounded to whole number of planks.

<table>
<thead>
<tr>
<th>Label</th>
<th>Test 0</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed</td>
<td>Predicted</td>
<td>Observed</td>
<td>Predicted</td>
<td>Observed</td>
</tr>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>AAA</td>
<td>24</td>
<td>21</td>
<td>147</td>
<td>137</td>
<td>53</td>
</tr>
<tr>
<td>AA</td>
<td>14</td>
<td>11</td>
<td>55</td>
<td>31</td>
<td>25</td>
</tr>
<tr>
<td>BB</td>
<td>8</td>
<td>5</td>
<td>39</td>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>BBB</td>
<td>13</td>
<td>9</td>
<td>69</td>
<td>29</td>
<td>26</td>
</tr>
<tr>
<td>Totals</td>
<td>59</td>
<td>41</td>
<td>310</td>
<td>218</td>
<td>120</td>
</tr>
</tbody>
</table>

Table 3: Observed vs predicted plot for test (1), where the dusty planks were predicted by the prediction model trained on the clean data set. The observed axis shows the plank grade from the planing mill while the predicted axis shows the predicted plank grade value from the prediction model used at the sawmill. Grade A planks are represented by 1 and grade B planks by 0. The grade separating threshold 0.56 is shown as a vertical line which defines planks with a predicted value above the threshold as of grade A, otherwise B. These prediction results are conceptually similar for all models used.

Figure 2. Observed vs predicted plot for test (1), where the dusty planks were predicted by the prediction model trained on the clean data set. The observed axis shows the plank grade from the planing mill while the predicted axis shows the predicted plank grade value from the prediction model used at the sawmill. Grade A planks are represented by 1 and grade B planks by 0. The grade separating threshold 0.56 is shown as a vertical line which defines planks with a predicted value above the threshold as of grade A, otherwise B. These prediction results are conceptually similar for all models used.

Figure 3. The grading accuracy as a function of threshold for the prediction models in tests; (0): trained on clean data and predicting clean data, from part one of this study; (1): trained on the entire clean data set and grading the entire dusty data set; (2): trained both data sets and predicting both data sets; (3): trained on both data sets and grading the clean data set; and (4): trained on both data sets and grading the dusty data set.
Table 4. Comparison of the used threshold and optimal threshold and the corresponding grading accuracy of each test, also showing the loss of grading accuracy and difference of threshold compared to the optimal setting. Tests 0, 2, 3, and 4 are based on 5-fold cross-validation and shows the average values.

<table>
<thead>
<tr>
<th>Test</th>
<th>Accuracy</th>
<th>Optimal Accuracy</th>
<th>Used threshold</th>
<th>Optimal threshold</th>
<th>Accuracy loss (points)</th>
<th>Threshold difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>74%</td>
<td>76%</td>
<td>0.55</td>
<td>0.52</td>
<td>−2%</td>
<td>0.03</td>
</tr>
<tr>
<td>1</td>
<td>73%</td>
<td>73%</td>
<td>0.56</td>
<td>0.56</td>
<td>0%</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>74%</td>
<td>75%</td>
<td>0.56</td>
<td>0.50</td>
<td>−1%</td>
<td>0.06</td>
</tr>
<tr>
<td>3</td>
<td>77%</td>
<td>78%</td>
<td>0.56</td>
<td>0.54</td>
<td>−1%</td>
<td>0.02</td>
</tr>
<tr>
<td>4</td>
<td>70%</td>
<td>74%</td>
<td>0.56</td>
<td>0.50</td>
<td>−4%</td>
<td>0.06</td>
</tr>
</tbody>
</table>

In tests 1–4 (Table 3), the grading accuracy was similar to that of the reference-case test (0), but there was a tendency for the dust layer to affect the PLS-based HSAG results negatively. As in part one of this study, the label classes with a higher number of planks in the data sets, shown in Table 1, were overall graded with a higher grading accuracy, which can be seen by comparing the number of planks in each label class with the corresponding grading accuracy in Table 3. Grade B planks were graded with the lowest grading accuracy, which could be attributed to the disproportionate number of grade B planks in the data sets. Table 1 shows that 35% (approximately 110 pieces in each data set) of the planks are of grade B. A data set with a proportional number of planks per label-class might be desired, unless an increased grading accuracy of grade B planks comes with an undesirable trade-off of reduced grading accuracy of grade A planks. The lower grade B grading accuracy could also be a consequence of using a data set with few grade B planks, as Lycken and Oja (2006) estimated that a minimum of 100 planks per plank grade are required for PLS-based grading of sawn timber by visual grades, indicating that more grade B planks might be required for PLS-based product adapted grading.

In test (1), the prediction model had been trained only on clean planks, and the model graded dusty planks with similar accuracy to grading clean planks. The grading accuracy dropped 1% point from 74% to 73% at the selected thresholds 0.55 and 0.56, respectively (Table 3). This test simulated a real-world scenario where, based on the clean data from part one of this study, a prediction model was created and used at the sawmill. The model was used to grade dusty plank and the PLS-based HSAG system would have functioned as expected but with a 1% point lower grading accuracy than the reference test (0). Due to the fact that the prediction model was used to grade planks with unforeseen disturbances, which it had not been trained for, this test was assumed prior to testing to show the worst grading performance, which was not the case. Test (1) showed the lowest grading accuracy for thresholds close to 0.5 but showed much higher grading accuracy than in the other tests for thresholds above 0.7. This could be because this is the only test that completely separates the two data sets for training and testing, making the separation of the classes more overlapping than in the reference test (0) (see Figure 2 in Olofsson et al. (2019)). The grading accuracy of test (1) means that the sawmill and customer could have decided on a very high threshold, up to say 0.8, without great loss of grading accuracy as in the other tests, however this test did not represent a common use-case.

In test (2), the prediction model was trained on a mixed set of both clean and dusty planks and graded a mixed data set by 5-fold cross-validation. The grading scenario in test (2) was close to the reference case, where the prediction model was trained and tested on similar data, i.e. clean, and mixed planks respectively. The grading accuracy curve for test (2) shown in Figure 3 was very similar to the curve for the reference test (0). This indicated that the sawmill should train the prediction model on data that is as representative of the intended grading batch as possible, dusty or not, as a mixed-trained model retained the grading accuracy compared to the reference test (0) when dusty planks are part of the test set. At the selected threshold 0.56, the grading accuracy increased from 73% to 74% compared to the reference, and since the prediction results are based on cross-validation on both the clean and the dusty planks, it was expected to see similar behaviour as the reference cross-validation of the clean data set.

In tests (3) and (4) the same grading models as in test (2) graded clean planks and dusty planks separately. Table 3 and Figure 3 showed that the grading accuracy was higher when grading clean planks than dusty planks. At the threshold 0.56, the mixed-trained model graded 77% of clean planks correctly, compared to 70% when grading dusty planks, compared to the 74% of correctly graded clean planks of the reference test (0). As expected, the model graded clean planks better than dusty planks. In Figure 3, the grading accuracy curves of test (3) and (4) behaved very much like expected, as they are the separation of test (2) into clean and dusty test sets separately; test (3) showed a higher grading accuracy than test (2) for almost all thresholds, while the opposite is true for test (4). Comparing tests (0) and (3) indicated that the mixed-trained model can be expected to perform equally or slightly better when grading clean planks than a model trained only on clean planks, and the mixed-trained model managed the highest prediction accuracy of all tests in test (3) (Table 4). One argument for the higher grading accuracy of the mixed-trained model when grading clean planks was the benefit of the larger training set, despite half of the training set being the dusty data set. However, as the dusty data set was not completely covered in dust, as almost all planks were only exposed to the settling dust on the top-face, there was a net positive effect of training on the dusty data set as well when grading clean planks. Comparing tests (1) and (4) indicated that the use of a mixed-trained model performed slightly worse when predicting the dusty planks, i.e. it was not possible in these tests to train a prediction model to take the dust into account for a higher grading accuracy of dusty planks. Comparing tests (0) and (3) showed that it was possible to train a prediction model on a larger data set and achieve higher grading accuracy when grading clean planks.

The grading accuracies were slightly higher for the models grading clean planks (tests 0 and 3) than for the models...
grading dusty planks (tests 1 and 4). Furthermore, prediction model trained on the combined data sets (test 2) did not grade planks more accurately than the reference model. This indicates that regardless of what data set(s) the prediction model is trained on, the dust on the planks introduces some difficulties. This was expected as the Boardmaster was calibrated for clean planks, and the layer of dust was bound to affect the feature detection in some way. For the sawmill, these tests indicated that the prediction model used should be trained on a data set that is as large as possible and as representative of the intended grading batch as possible. This is especially true when grading batches of sawn timber that is not expected to be dusty.

According to Table 4 the threshold for test (1) was by chance selected optimally, which can be explained by the flat grading accuracy curve in Figure 3 for thresholds between 0.4–0.6. In Figure 3 the maximum difference of the grading accuracy between test (0) and (1) was 4% points at the threshold 0.51, indicating that the loss of prediction accuracy due to the dusty could have been larger than the measured 1% point.

The robustness of PLS-based HSAG was, apart from the grading accuracy, investigated by measuring the optimal threshold stability, seen in Table 4. The threshold 0.56 was in both test (1) and tests (2–4) purposely chosen to imitate the use of a PLS-based HSAG system at the sawmill, with the goal of maximizing grading accuracy. However, as the prediction models tested were not optimized for the tests set there is an optimal threshold that can only be determined with known grade outcome at the planing mill. If the training data is representative of the test set, the threshold selected based on the training data should be close to the optimal threshold value. Any large changes in grading accuracy or large miss of optimal threshold might indicate that the prediction model is not well trained for the current test set, i.e. the training data is not fully representing the test data. Table 4 shows that in the reference case (test 0) the average threshold value of the 5-fold cross-validation is 0.03 points away from the optimal threshold, which resulted in a 2% points loss of grading accuracy. Given the size of the clean data set and complexity of the data used, this kind of variation was expected and is the reference case for optimal threshold stability. In test (1) the clean training data was not expected to be completely representative of the dusty testing data, but the threshold selected based purely on the clean training data was (surprisingly) the optimal threshold (Table 4), which can be explained by the approximately flat grading accuracy curve in Figure 3 for thresholds between 0.3 and 0.8. Anecdotally, in Figure 3 the maximum difference of the grading accuracy between test (0) and (1) was 4% points at the threshold 0.51, indicating that the loss of prediction accuracy due to the dusty could have been larger than the measured 1% point with a slightly different test set.

The threshold value difference between the used and optimal value for tests with dusty planks in the test set (tests 1, 2, and 4) showed higher differences (0.06 points), or the above mentioned optimal threshold ambiguity in test (1), compared to the tests with only clean plank in the test set. The larger threshold difference indicated that when predicting dusty planks the training data was not fully representative of the test set, no matter the training set. In tests (0) and (3) the threshold value difference was lower, at 0.03 and 0.02 points, which indicated that when grading clean planks the training sets were more representative of the test sets, no matter the training set. For a user of a PLS prediction model, it is very important that the training data is representative of the test data, otherwise grading outcome estimated from the training data might not transfer well to the actual grading outcome of a test or when grading for a customer. The optimal threshold stability measurements showed that when grading dusty planks, slight changes in grading outcome might occur unexpectedly.

Conclusions

The PLS-based HSAG of Scots Pine sawn timber investigated in this study was robust to unforeseen disturbances in the form of a layer of dusty on top of the sawn timber when grading planks for an industrial customer. Prediction models trained and tested on clean planks served as reference and predicted on average the plank grade of new clean planks correctly 74% of the time. The prediction accuracy dropped 1% point to 73% when a similar model was used to grade dusty planks, which indicated that PLS-based product adapted HSAG was robust to the disturbances of the dust. However, the class-separating threshold used (0.56) was seemingly selected optimally by chance for this test, and the loss of prediction accuracy could have been up to 4% points for a different threshold in the range 0.4–0.6.

Prediction models trained on both clean and dusty planks performed on average as good or better than the reference models when grading only clean planks, with on average 77% correctly graded clean planks. This indicated that the prediction model should be trained on a data set as large as possible consisting of planks as representative of the planks to be graded as possible, even if they are dusty on one side. It was, however, not possible in this study to improve the grading accuracy of dusty planks by training the prediction model on dusty planks as well as clean planks. Further research should investigate if separate grading methods can detect disturbances of the sawn timber measurements and handle the identified pieces accordingly.

The class-separating thresholds used in the study were selected based on the assumption that the training data was representative of the test data of each test. When the dusty planks were graded in any of the tests, the estimated grading outcome determined by the training data did not completely translate to the grading outcome of the tests, which indicated that the dust introduced some difficulty when selecting a class-separating threshold, as well as lowering the overall grading accuracy. Further research is required to fully understand how, and if, unexpected input to the PLS-based HSAG system should be handled. A larger data set with more extreme disturbances would be preferred for such a study.

Note

1. The threshold 0.56 is the same for test (1) and tests (2–4) by chance.
Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Financial support from the Swedish Innovation Agency (Vinnova), project Sawmill 4.0 Customized flexible sawmill production by integrating data driven models and decisions tools 2018-02749, is gratefully acknowledged.

ORCID

Linus Olofsson  http://orcid.org/0000-0002-5562-5142
Dick Sandberg  http://orcid.org/0000-0002-4526-9391

References


