วารสารสมาคมวิศวกรรมเกษตรแห่งประเทศไทย
ปีที่ 25 ฉบับที่ 2 (2562) 19-27
Available online at www.tci-thaijo.org/index.php/TSAEJ

บทความวิจัย

ISSN 1685-408X

Informative selection of spectra obtained from an online sugar content prediction system of sugarcane by using statistical index

Kittisak Phetpan[1]*, Vasu Udompetaikul[1], Panmanas Sirisomboon[1]

[1]Department of Agricultural Engineering, Faculty of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand

*Corresponding author: Tel: +66-99-324-2127, E-mail: kphetpan@gmail.com

Abstract

The main objective of this research was to evaluate the applications of principal component analysis (PCA), cluster analysis and statistical index to be used for informative selection of spectra obtained from an online soluble solids content (SSC) measuring system of sugarcane. Three studying steps were conducted. The assessment of these three methods for the classification between sugarcane and slat detecting signals from the ideal data set (static scanning case) were firstly performed. Secondly, the applications of these methods in the classification for the testing data set of the dynamic case were performed and then the best method was selected to be used for filtering out the slat detecting signals from all dynamic data set. Finally, the partial least square (PLS) modelling by using the output spectra from the best filtration method against their SSC values was done in order to compare with the model obtained from the previous work. The initial assessment of these three methods showed clear grouping between sugarcane and non-sugarcane (slat) spectra for ideal scanning case. Seventy-four spectra were randomly selected from the dynamic data set to be the testing data set for determining the best methods in filtering out the slat spectra. Based on this test, the statistical index was the best one, showing 89.19% for overall accuracy, compared to the PCA and cluster analysis reflecting the accuracy between 48.65-56.76%. So, the statistical index was used to filter out slat spectra from all dynamic data set. Based on this step, the output spectra and their SSC values were used for modelling, displaying a coefficient of determination of calibration (R^2) of 0.768. Its internal validation showed a root mean square error of cross-validation (RMSECV) of 0.43 °Brix, whereas that of 0.33 °Brix was shown in previous work. However, no statistically significant difference was observed at a 95% confidence interval. Therefore, the statistical index proposed in this study is the informative selection of spectra obtained from an SSC prediction system of sugarcane. This information would be useful for further work in developing this online system for installing on the conveyor of sugarcane harvester.

Keywords: Spectral classification, Soluble solids content, Elevator conveyor, Sugarcane quality monitoring system

1 Introduction

Agricultural products are one of the necessary resources for all living things around the world. Sugarcane is one of the main agricultural products, used in sugar production worldwide and used to produce alternative fuels in the form of ethanol in some countries as well (Cookson, 2012). With the increasing world population, it is expected to reach 8.6, 9.8 and 11.2 billion in 2030, 2050 and 2100, respectively (United Nations, 2017). So, to increase the yield and maintain the quality in the production of sugarcane without enlargement of growing areas should be focused to support the increasing population.

However, the world of sugarcane production is still confronted with the problem of variation in the yield and especially quality within the fields (Kingston and

Received: January 14, 2019
Revised: March 22, 2019
Accepted: April 24, 2019
Available online: May 18, 2019

Hyde 1995; Bramley and Quabba 2001). This has prompted efforts in accessing the variation within the fields to achieve improvement in the production such as field input.

Several means such as electronic refractometers (Mccarthy and Billingsley 2002), microwaves (Klute 2007; Nelson 1987; Shah and Joshi 2010) and spectroscopic techniques (Nawi et al. 2013a; Nawi et al. 2013b; Phetpan et al. 2018) have been studied to develop as a rapid sugarcane quality measurement. Interestingly, Phetpan et al. (2018) developed the non-destructive, visible and near-infrared spectroscopic technique for online measuring soluble solids content (SSC) of sugarcane billets on an elevated conveyor. With their results reporting a root mean squares error of prediction (RMSEP) of 0.30 °Brix, it shows the possibility in measuring the SSC of moving sugarcane billets using the spectroscopic technique. For the continuous online spectral measurement, a spectral screening process is needed to filter out unwanted spectra because many sources are detected and recorded. However, the spectral filtration proposed in Phetpan et al. (2018) study was just the basic screening process by eliminating all low signal reflections despite being the sugarcane responses.

Pattern recognition and classification are typically divided into two main groups, i.e. unsupervised and supervised methods. Grouping of any objects with no supervisor in the sense of known membership is called unsupervised manner (Otto, 2017). Principal component analysis (PCA) and cluster analysis are some of the unsupervised learning methods, widely used in the chemometric analysis. If the membership of objects to classify is known, the methods of the supervision called pattern recognition can be used (Otto, 2017). Linear learning machine (LLM), discriminant analysis, the soft independent modeling of class analogies (SIMCA), and support vector machines (SVMs) are some of this one for the application in recognizing patterns (Otto, 2017). In the real world, not only these two methods were applied for the object classification, but also statistical index such as Normalized Difference Vegetation Index (NDVI) and Green-Red Vegetation Index (GRVI) widely used in the remote sensing area. These are the well-known indexes used for tracking phenological changes which are probably used in distinguishing between the green vegetation and the other types of the ground covers by measuring the difference between two bands, i.e. near-infrared and red light for the NDVI and green and red light for the GRVI (Motohka et al. 2010).

Based on spectral data set in this study gathering sugarcane and non-sugarcane detecting signals, adopting some statistical procedures in order to obtain the required spectra of sugarcane during an online scanning is necessary. The unsupervised and statistical index methods are suitable for this data set because of their unknown memberships of the spectra during classification. Therefore, the objectives of this research are threefold: 1) to initial assess the application of principal component analysis (PCA), cluster analysis and statistical index in filtering out non-sugarcane spectra for static data set, 2) to select the best one for classifying the sugarcane and non-sugarcane spectra in dynamic data set and 3) to establish the partial least square (PLS) model by using the output spectra from the best filtration method against their SSC values.

2 Materials and methods

2.1 Samples

Fifty clumps of sugarcane were used in this study. In each clump, the sugarcane was chopped into billets with an approximate length of 20 cm (see in Phetpan et al., 2018 for more details). So, there were 50 groups of sugarcane billets in total.

2.2 Spectral detection

An online measurement system used in this study consisted of two main parts, a cane billet elevator and a spectral acquisition system. A vis/NIR spectrometer (AvaSpec-2048-USB2, Avantes BV, Netherlands) was installed for the spectral acquisition system, operating in the spectral range of 350-1100 nm with the spectral resolution of 2.4 nm. It was set the integration time for 14 ms, yielding approximately 90% full-scale Analog-to-Digital Converter (ADC) of the reference material reflectance (see in Phetpan et al., 2018 for more details). Static and dynamic spectral detections were performed.

2.2.1 Static spectral detection

To understand spectral behavior obtained from various sources such as sugarcane, the floor of the

elevator, and slat for conveying cane, spectral measurement of the static detection was necessary. Figure 1 shows two types of spectral detection, i.e. single and paired cane scans and slat scans. In the case of floor detection, it was the spectral measurement without the presentation of any objects to the optic fiber not presented in the Figure. Overall, five repetitions for each detection were performed. So, twenty spectra in total were obtained.

2.2.2 Dynamic spectral detection

Each group of sugarcane billets was put on the elevator and was conveyed by the slat, mounted to the conveyor chain, in order to collect the spectra. Two replications with two repetitions for spectral detection were performed (see in Phetpan et al., 2018 for more details). Four spectral sets (19 spectra each) were obtained from each group and two hundred spectral sets were the number in total. However, one hundred and seventy-six sets remained after the visual observation. This number was used to be the data for optimizing the filtration process. Note that another twenty-four sets identified some problems showing none of sugarcane absorption even if the cane was scanned.

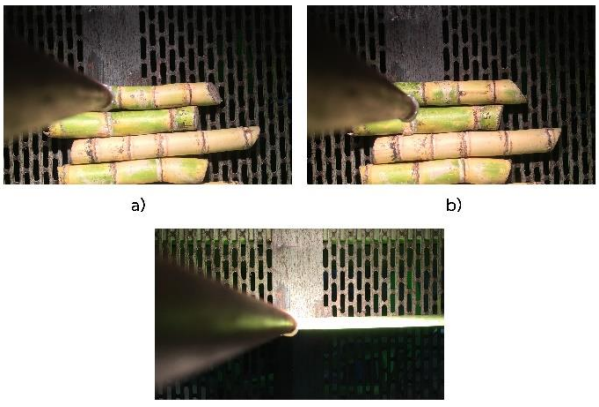

Figure 1 Types of static spectral detection for a) single cane detection, b) paired cane detection and c) slat detection.

2.3 Data analysis

Grouping of analytical data according to their similar elemental pattern is possible by either means of clustering methods or by projecting the high dimensional data onto lower dimensional space (Otto, 2017). These methods are performed with no supervisor in the sense of known membership of objects to classify. So, principal component analysis (PCA), one of the projection methods, and cluster analysis were used for classifying sugarcane and non-sugarcane samples.

To decompose the original matrix (X) by a product of the score (T) and loading (P) matrices, is the key idea of PCA which can be formulated as follows:

$$X = TP^T \quad (1)$$

With this concept, to project matrix (X) or spectral matrix onto the lower dimensional space (T), the equation (1) is to be converted by:

$$T = XP \quad (2)$$

To interpret the results of PCA for the discrimination between the sample groups, visualizing by plotting the elements of two scored vectors of the matrix (T), especially the first two vectors explaining the most variance in the original matrix, could be done.

For the cluster analysis, the similarity of the objects is decided based on the distance measures. The shorter the distance between objects the more similar they are (Otto, 2017). In this study, the Euclidean distance, widely used, was applied according to:

$$d_{ij} = \left[\sum_{k=1}^{K} \left|x_{ik} - x_{jk}\right|^2\right]^{1/2} \quad (3)$$

where K is the number of variables; i, j are the indices for object i and j.

In addition, the Normalized Difference Vegetation Index (NDVI) and the Green-Red Vegetation Index (GRVI), which are a normalized ratio of near-infrared and red reflectance and that of green and red reflectance, respectively, were initially tested to assess the possibilities for classifying the different spectra in this work. No such approaches were possible.

With the behavior of spectra scanned under the static state, it seems that the spectra had their own pattern according to their sources (Figure 2). The own pattern of sugarcane responses presents the low and high variation in the range from 560 to 640 nm and

from 680 to 750 nm, respectively. High variation in both ranges is the characteristic of slat detection, whereas low variation throughout the spectral range is that of floor detection. With very low %reflection and variation of floor spectra, these original floor spectra can be easily cut off from other raw spectra. The standard deviation (SD) throughout the spectral line for 1 was used as a threshold for removing the floor spectra.

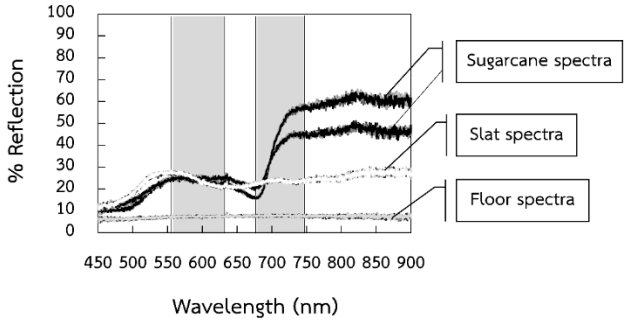

Figure 2 Spectral characteristics based on the detection of the different objects.

Based on these two leftover patterns, the concept of the normalized ratio mentioned above was modified to be an alternative channel for this work. So, a statistical index relying on a normalized ratio of SD of reflection value in the range from 560 to 640 nm and 680 to 750 nm was proposed in this study. The index is defined as follows:

$$\text{Statistical index} = \frac{SD_{560\ to\ 640} - SD_{680\ to\ 750}}{SD_{560\ to\ 640} + SD_{680\ to\ 750}} \quad (4)$$

For the data analysis, the spectral range of 450-900 nm, defined as a full range, and that of 560-640 and 680-750 nm were used for the application of PCA and cluster analysis. PCA and cluster analysis were firstly applied to the static data set after the floor spectra removal in order to test the filtering out slat spectra based on both the full range and the 560-640 and 680-750 nm. In the case of PCA, estimation of the loading matrices P^T from the scaled-matrices (X) based on two spectral ranges and computation of new score matrices (T) according to the equation (2) were performed, respectively. With this computation, two matrices (T) based on these two ranges were the output of the PCA and were used for the discrimination between the sugarcane and slat spectra. For the classification with cluster analysis, the Euclidean distance was applied based on two spectral ranges resulting in different K according to the equation (3). Based on these two means, four outputs of the classification were obtained, 2 outputs (450-900 nm and 560-640, 680-750 nm) from the PCA and that from the cluster analysis. For the statistical index, it was also applied to the data of static data set. The specific index value was then obtained to be an indicator to keep up the sugarcane responses. In order to evaluate the performance of these three methods in filtering the slat spectra out, they were applied to dynamic spectral data set which was already cut the floor spectra off. Ten sample groups (74 spectra) out of one-hundred and seventy-six groups were randomized and used as a testing set for the evaluation. The results showing the percentage of correct classification of those compared among the three methods were presented. Based on these, the best one was applied to screen non-sugarcane (slat) spectra out from all the dynamic data set. The spectra survived from this process were mathematically pretreated by moving average (MA) smoothing with segment size of 21 points and standard normal variate (SNV) in order to minimize spectral noises and to diminish the offset effect in the spectra, respectively. At the end of this study, these spectra were modeled against their SSC values with partial least square (PLS) regression in order to compare with the results obtained from previous work (Phetpan et al., 2018). So, the PLS model, resulting standard error of cross validation ($SECV$) was compared using a Fisher's test (F value) (Molla et al., 2016) to see a significant difference in statistics. The F value was calculated as

$$F = \frac{SECV2}{SECV1}, \text{ where } SECV1 < SECV2 \quad (5)$$

The calculated F value was compared with the confidence limit F critical (1 – α, n_1 – 1, n_2 – 2), obtained from the distribution F table, where α is the test significance level (α = 0.05), n_1 for the sample number modeled at the previous work, and n_2 for that in this work (n_1 and n_2 = 100 for this test) (Molla et al., 2016). The differences between these two models are significant when F > F limit.

In this study, the statistical index computations were conducted by R software (R Core Team, 2018), using the dplyr package (Wickham et al., 2017), whereas the PCA, cluster analysis, mathematical pretreatments as well as the PLS modelling were conducted by the software for multivariate analysis (Unscrambler X 10.3, Camo, Norway).

3 Results and discussion

The application of PCA and statistical index employed with the static data set to filter out slat spectra are shown in Figure 3 and 4, respectively, whereas that of cluster analysis is shown in Table 1. The PCA outcome with the first principle component (PC) explained 99% of the total variation in the spectra identifies the tendency in classifying between sugarcane and slat spectra. The results of PCA based on the two spectral ranges show the same distribution patterns. With this result, the spectral range of 560-640 and 680-750 was used for further analysis, whereas zero was used as a threshold for the classification.

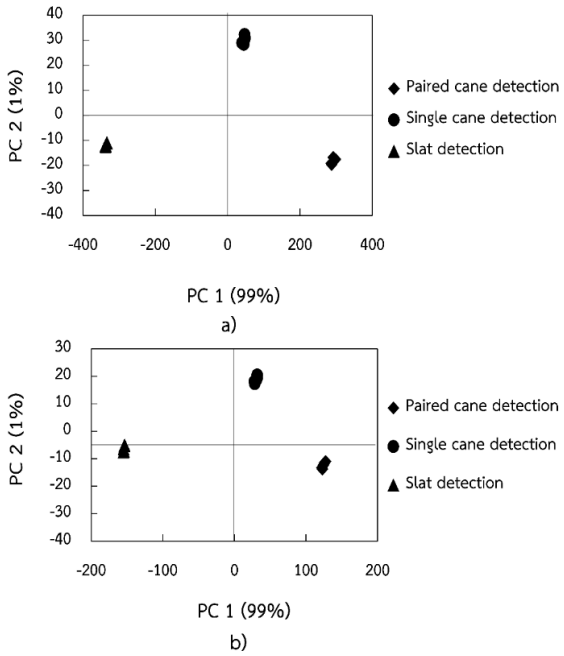

Figure 3 Spectral filtration output of the PCA applied to static data set a) outcome based on full spectral range (450-900 nm) and b) outcome based on 560-640 and 680-750 nm.

For the statistical index test, the scatter plot clearly groups between two different objects (cane and slat). Sugarcane detection has a very high value in negative, whereas slat detection has very high in positive. Therefore, the value of -0.6 was selected as a threshold to distinguish between sugarcane and slat detections. With this threshold, the spectra could be filtered out if the value was higher than -0.6. In the case of cluster analysis, the results shown in the Table 1 identify that this method probably works well in the future for filtering out the slat detecting signals. However, these tests were done based on ideal spectral data set (static spectral detection). So with the very good results they are, these applications were also tested with the dynamic data set.

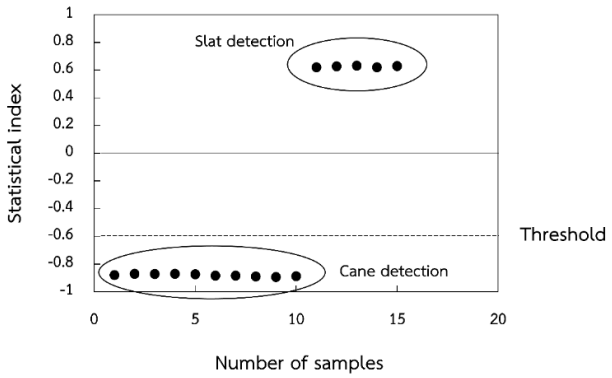

Figure 4 Spectral filtration output of statistical index applied to the static data set.

Table 1 Application of cluster analysis in filtering out slat spectra for data set of static spectral detection.

Objects	450-900 nm	560-640, 680-750 nm
Single cane detection	0	0
	0	0
	0	0
	0	0
	0	0
Paired cane detection	0	0
	0	0
	0	0
	0	0
	0	0
Slat detection	1	1
	1	1
	1	1
	1	1
	1	1
Overall accuracy	100%	100%

Note: 0 is the symptom of sugarcane identification and 1 is that of slat identification

Table 2 shows the evaluation results of the performance of the three methods in classifying between sugarcane and slat spectra from the testing data set (74 spectra previously mentioned). Percentage of overall accuracy obtained from PCA and cluster analysis are between 48.65-56.76%. There is a bit difference of overall accuracy in PCA employed based on two different spectral ranges, whereas the same result is obtained from the cluster analysis. This identifies that the different range of spectra has no influence for the PCA and clustering applications. Based on the results, these two methods could not be used to filter out slat spectra from the dynamic data set. The best result for the evaluation in filtering out non-sugarcane (slat) spectra from the dynamic data set is from the statistical index application, 89.19% for overall accuracy. With this accuracy, eight sugarcane detections were identified as slat spectra whereas no slat detection was identified as the cane. This is what we expected is that a few cane spectral detections with low reflectance could be removed out as non-sugarcane spectra but keeping up slat detection as sugarcane spectra are not allowed. This statistical index was modified from the NDVI and GRVI, which are the normalized ratio relying on measuring the difference between two bands. The application of this index for identifying the difference between spectra of sugarcane and slat that characterizes their own pattern based on different standard deviation is satisfied. So, this index was applied to screen non-sugarcane (slat) spectra out from all the dynamic data set. The filtration result is shown in Figure 5, displaying the minimum signal reflection in the region around 750-900 nm of around 3%. This number is quite low compared to that (20%) of filtration proposed by Phetpan et al. (2018). However, this indicates the performance of this index to keep up only the sugarcane detecting signals. The result could also be confirmed by the pretreated spectra in Figure 5b with no presentation of the slat signal except for the sugarcane spectra. With this result, all of 176 spectral sets (100%) survived for this study whereas only 150 out of the 176 sets (85%) survived for the filtration process in Phetpan et al. (2018). Based on this, the statistical index proposed in this study works very well in classifying the sugarcane and non-sugarcane detection especially keeping up the sugarcane signal despite having a low signal reflection.

Table 2 The applications of PCA, cluster analysis and statistical index for filtering out the slat spectra from the dynamic data set

Methods for classification	Identified cane as slat	Identified slat as cane	% Overall accuracy
PCA *	38 (74)	0 (74)	48.65
PCA **	36 (74)	0 (74)	51.35
Cluster analysis *	31 (74)	1 (74)	56.76
Cluster analysis **	31 (74)	1 (74)	56.76
Statistical index	8 (74)	0 (74)	89.19

Note: * employed with the spectral range of 450-900 nm, whereas ** employed with that of 560-640 and 680-750 nm.

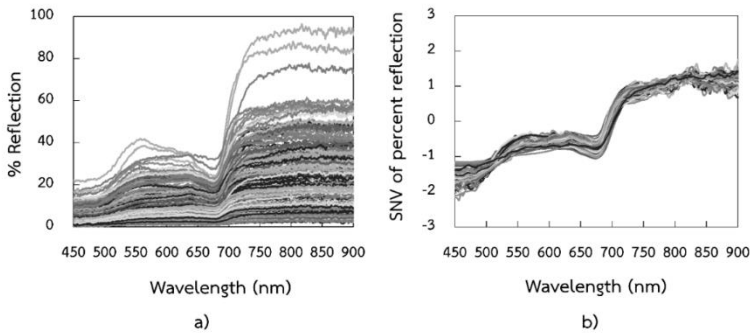

Figure 5 Spectral output remained from classifying by the statistical index, a) raw output and b) pretreated output.

A PLS model was developed using the spectra obtained from the filtration by statistical index to compare with the PLS model presented in Phetpan et al. (2018). So, the number of samples and pretreatment used for modelling were the same as the previous work. Figure 6 shows the result of PLS modelling employed using the spectra kept up by the statistical index. The model employed 4 latent variables (LVs) to account 92% and 77% in dynamic spectral and SSC variations, respectively. Two sugar related peaks at 755 nm (the 4th overtone of C-H stretching of sugar at 762 nm (Osborne et al. 1993) or the 3rd overtone of O-H stretching of sucrose in water at 740 nm) and 890 nm (the 3rd overtone of C-H stretching of sucrose in water at 910 nm (Golic et al. 2003)) were still found in the regression coefficient plot. The predictive performance of this model, validated by 50 samples in the prediction set, could be explained by root mean square error of prediction (RMSEP) of 0.44 °Brix.

Comparison the $SECV$ between the two PLS models found that the modelling using the data set obtained from the statistical index in this work tends to increase the standard error of prediction, $SECV$ of 0.43, whereas the model presented in previous work had 0.34 in term of $SECV$. However, there is no statistically significant difference between these two PLS models when F (0.05, 99, 99) < F limit (1.26 < 1.39).

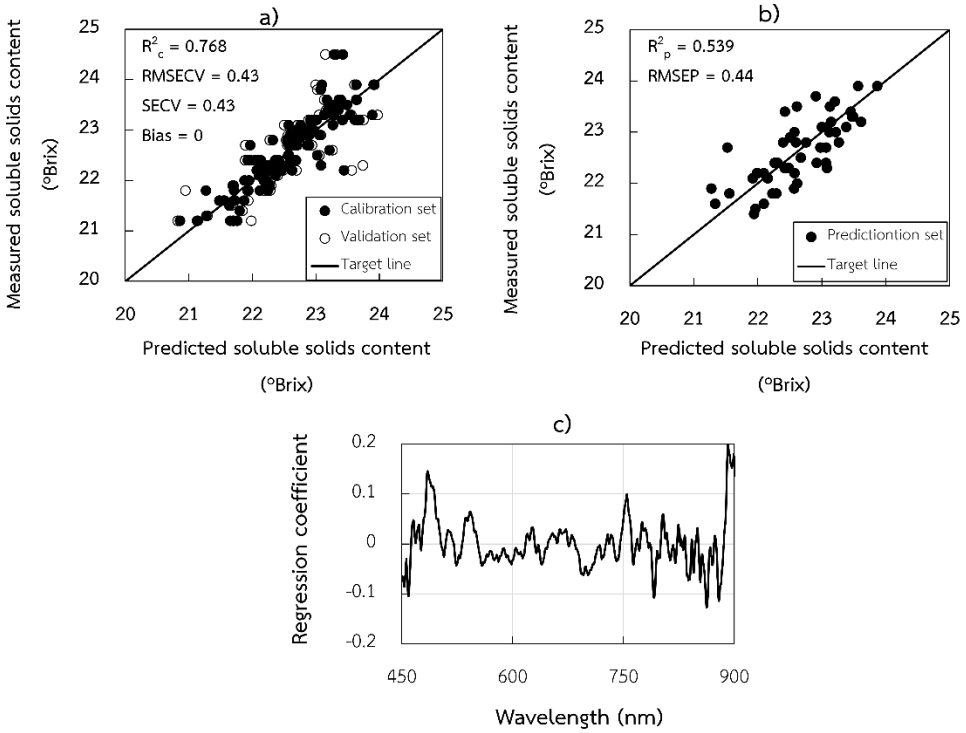

Figure 6 The result of PLS model constructed using spectra kept up by the statistical index; a) scatter plot the PLS modelling and its internal validation, b) scatter plot of external validation and c) the regression coefficient plot.

4 Conclusion

The application of principal component analysis (PCA), cluster analysis and statistical index showed a good tendency in screening out non-sugarcane spectra from the static data set (ideal case) with clear grouping between the different objects.

The statistical index was the best for removing the slat detection signals from the testing data set (randomly selected 74 spectra from the dynamic data set), showing 89.19% for overall accuracy, compared to the PCA and cluster analysis reflecting the accuracy between 48.65-56.76%.

The PLS modelling that employed the spectra obtained from the filtration using the statistical index against their SSC values displayed a coefficient of determination of calibration (R^2) of 0.768. It was internally validated and showed a root mean square error of cross-validation (RMSECV) of 0.43 °Brix,

whereas the RMSECV of 0.33 °Brix was shown in previous work. However, no statistically significant difference was observed ($p > 0.05$) from the error increased. So, this index would be applied as the filtration process for further work in developing the on-line soluble solid content measuring system of sugarcane on the conveyor.

5 Acknowledgments

The authors thank the Near Infrared Spectroscopy Research Center for Agricultural product and Food (www.nirsresearch.com) and Precision Agriculture Laboratory at King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, for laboratory space and instruments; the authors also thank Mr. Ampol Jongsomboonpokha for providing the sugarcane samples from his field. We acknowledge the financial support from the Royal Golden Jubilee Ph.D. Scholarship (PHD/0102/2558) of the Thailand Research Fund (TRF).

6 References

Bramley, R.G.V., Quabba, R.P. 2001. Opportunities for improving the management of sugarcane production through the adoption of precision agriculture - An Australian perspective. Proceedings of the 24th Congress of the International Society of Sugar Cane Technologists 38-46.

Cookson, C., 2012. A tank of sugar: how Brazil runs on biofuel. FT Magazine, United Kingdom.

Golic, M., Walsh, K., Lawson, P., 2003. Short-wavelength near infrared spectra of sucrose, glucose, and fructose with respect to sugar concentration and temperature. Applied Spectroscopy 57, 139-145.

Kingston, G., Hyde, R.E. 1995. Intra-field variation of commercial cane sugar (CCS) values. Proceedings of the Australian Society of Sugar Cane Technologists 17, 30-38.

Klute, U. 2007. Microwave measuring technology for the sugar industry. International Sugar Journal 109(1308), 1-6.

Mccarthy, S., Billingsley, J. 2002. A sensor for the sugar cane harvester topper. Sensor Review 22(3), 242-246.

Molla, N., Bakardzhiyski, I., Manolova, Y., Bambalov, V., Cozzolino, D., Antonov, L. 2016. The Effect of Path Length on the Measurement Accuracies of Wine Chemical Parameters by UV, Visible, and Near-Infrared Spectroscopy. Food Analytical Methods 10(5), 1156–1163.

Motohka T., Nasahara K.N., Oguma H., Tsuchida S. 2010. Applicability of Green-Red Vegetation Index for remote sensing of vegetation phenology. Remote Sensing 2, 2369-2387.

Nawi, N.M., Chen, G., Jensen, T. 2013b. Visible and shortwave near infrared spectroscopy for predicting sugar content of sugarcane based on a cross-sectional scanning method. Journal of Near Infrared Spectroscopy 21, 289–297.

Nawi, N.M., Chen, G., Jensen, T., Mehdizadeh, S.A. 2013a. Prediction and classification of sugar content of sugarcane based on skin scanning using visible and shortwave near infrared. Biosystems Engineering 115(2) 154-161.

Nelson, S.O. 1987. Potential agricultural applications for RF and microwave energy. Transactions of the ASAE 30(3), 818-831.

Otto, M. 2017. Chemometrics: Statistics and Computer Application in Analytical Chemistry. New York, USA: Wiley-VCH.

Osborne, B., Fearn, T., Hindle, P., 1993. Practical NIR spectroscopy with applications in food and beverage analysis. Harlow, Essex, United Kingdom: Longman Scientific & Technical.

Phetpan, K., Udompetaikul, V., Sirisomboon, P. 2018. An online visible and near-infrared spectroscopic technique for the real-time evaluation of the soluble solids content of sugarcane billets on an elevator conveyor. Computers and Electronics in Agriculture 154, 460-466.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Shah, S., Joshi, M. 2010. Modeling microwave drying kinetics of sugarcane bagasse. International Journal of Electronics Engineering 2 (1), 159–163.

United Nations. 2017. News: World population projected to reach 9.8 billion in 2050, and 11.2

billion in 2100. Available at: https://www.un.org/development/desa/en/news/population/world-population-prospects-2017.html. Accessed on 4 January 2019.

Wickham, H., Francois, R., Henry, L., Müller, K. 2017. dplyr: A Grammar of Data Manipulation. R package version 0.7.4. https://CRAN.R-project.org/package=dplyr.