Comparison of Machine Learning Approaches for Spectroscopy Applications

Ioanna Aslanidou [a,*], Jerol Soibam [a]

[a] *Mälardalen University, Box 883, 72123, Västerås, Sweden*
[*] *ioanna.aslanidou@mdu.se*

Abstract

In energy production the characterization of the fuel is a key aspect for modelling and optimizing the operation of a power plant. Near-infrared spectroscopy is a well-established method for characterization of different fuels and is widely used both in laboratory environments and in power plants for real-time results. It can provide a fast and accurate estimate of key parameters of the fuel, which for the case of biomass can include moisture content, heating value, and ash content. These instruments provide a chemical fingerprint of the samples and require a calibration model to relate that to the parameters of interest.
A near-infrared spectrometer can provide point data whereas a hyperspectral imaging camera allows the simultaneous acquisition of spatial and spectral information from an object. As a result, an installation above a conveyor belt can provide a distribution of the spectral data on a plane. This results in a large amount of data that is difficult to handle with traditional statistical analysis. Furthermore, storage of the data becomes a key issue, therefore a model to predict the parameters of interest should be able to be updated continuously in an automated way. This makes hyperspectral imaging data a prime candidate for the application of machine learning techniques. This paper discusses the modelling approach for hyperspectral imaging, focusing on data analysis and assessment of machine learning approaches for the development of calibration models.

Keywords: machine learning, near-infrared spectroscopy, hyperspectral imaging; calibration models

1. Introduction

In the energy and process industry the characterization of the feedstock material is a key aspect for modelling and optimizing the operation of a process or power plant. The operating parameters are continuously adjusted in order to provide output that meets certain specifications, which can be the quality of the end product or the power output. These are dependent on the quality of the feedstock, as a difference in its key parameters will result in different requirements for its processing. Detailed knowledge of the properties of the fuel can be used to improve the operation of the plant using feed-forward control approaches. A sensor that can provide this information is one of the foundations for a learning system that can support optimal operation and decision-making.

Near-infrared (NIR) spectroscopy is a well-established method for characterization of different fuels and is widely used both in laboratory environments and in power plants for real-time results. The method itself is based on the excitation and vibration of the molecules, which in turn provides the chemical information for the material. This needs to be correlated to the parameters of interest of the fuel, which is typically done using statistical analysis. Near-infrared spectroscopy is one of the technologies used for the development of smart sensors in the learning system described in Rahman et al. (2021). Near-infrared spectroscopy has been widely used in the literature (Tsuchikawa et al., 2003; Skvaril et al., 2017) as it can provide a fast and accurate estimate of key parameters of the fuel, which for the case of biomass can include moisture content, heating value, and ash content.

Near-infrared spectroscopy can only provide single point measurements. This information is well suited for homogeneous mixtures, where it can provide a good estimate for the parameters of interest of the entire batch. In recent years, hyperspectral imaging (HSI), which combines spectral information and conventional imaging, is also increasingly used for fuel characterisation. This allows the collection of spatial data for the near-infrared spectrum and can be applied in real environments (e.g. above a conveyor belt) to provide real-time information about the fuel. The simultaneous acquisition of spectral and visual information without the need for synchronization is another advantage, which also makes the use in real environments more realistic. In hyperspectral imaging, the instrument

can acquire images of the sample, as well as spectral information for each pixel, providing a hypercube of data. This can provide multiple opportunities, as discussed by Mäkelä and Geladi (2017), who used HSI to distinguish different materials (from different feedstocks or prepared under different temperatures) and evaluate their homogeneity. Another application of HSI in biomass characterization is for pelleting of biomass feedstocks where the spatial resolution allows the classification of images to assess the efficiency of the mixing of different biomass streams (Gillespie et al., 2016).

Regardless of the method or instrument used to acquire spectral information for the samples, a model is required to correlate that information to the parameters of interest and provide quantitative information about the parameters of interest. Linear regression techniques are considered the standard for quantitative characterization, with Partial Least Squares Regression (PLSR) being the most commonly used approach (Skvaril et al., 2017). Non-linear methods, such as Artificial Neural Networks (ANNs), have been shown to improve the results but are more demanding in terms of computational power. Advances in computing power have allowed machine learning techniques to be used to extract information from spectral data, and recent literature presents results from different applications.

Machine learning in combination with IR spectroscopy has been widely used for classification purposes. Mancini et al. (2020) used NIR to study the supply chain for biomass pellets and applied different classification algorithms to predict pellet quality. A similar approach was used by Tiitta et al. (2020) who employed electric impedance spectroscopy to classify wood chips of different origin, which can then allow the derivation of more accurate models for moisture content. Pitak et al. (2021) focused on the biomass pellet production process, using machine learning for wavelength selection and PLS regression for their calibration model. Tao et al. (2020) obtained IR spectra of biomass and waste with an attenuated total reflectance (ATR) and used ML for classification and characterization, employing regression techniques. Ahmed et al. (2018) applied different methods for the characterization of biomass wood chips using NIR, namely ANN, Gaussian Process Regression (GPR), Support Vector Regression (SVR) and traditional PLSR, with GPR showing the best results.

The use of hyperspectral imaging results in a larger amount of data that what is obtained with NIR and storage of the data becomes a key issue. The calibration model for hyperspectral imaging should be able to be updated continuously in an automated way, which makes hyperspectral imaging data a prime candidate for the application of machine learning techniques. Gewali et al. (n.d.) present a review of the literature in the use of ML for HSI, primarily for analysis of hyperspectral images captured from earth observing satellites and aircraft. They looked into techniques used for classification of images based on land cover, concluding that deep learning is a promising approach. For the estimation physical/chemical parameters related to agriculture, Bayesian methods were considered to be more suitable due to their flexibility, ability to handle uncertainty, and capacity to perform well with limited data.

This paper discusses the use of hyperspectral imaging and machine learning for biomass characterization. The focus is on data analysis and assessment of machine learning (ML) approaches for the development of calibration models. A comparison of different ML approaches for HSI for the prediction of biomass properties is not available in the literature, and neither is a comparison of ML with a conventional model to assess when the use of ML is beneficial in such applications. The contribution of the paper therefore lies in discussing the suitability of different methods depending on the purpose of the analysis and the type of data available, as a first step towards a thorough study of the use of ML for HSI data analysis in such applications.

2. Methodology

A set of biomass samples was analysed with a hyperspectral imaging camera and different machine learning techniques were used to create a calibration model. The following sections discuss the methods used to acquire the data, pre-processing techniques, and methods used to build calibration models, followed by validation and testing of their predictive capabilities.

2.1. Sample preparation and data acquisition

A set of 100 biomass fuel samples were used in this study. The spectral data of the samples was obtained with a push-broom line scanning hyperspectral imaging Specim FX17e camera (Specim Spectral Imaging Ltd, Finland). The camera is equipped with an InGaAs based NIR detector with spectral range of 900-1700 nm, 224 spectral bands, and 640 pixels over the cross-track field of view (FOV). The samples were illuminated with six halogen light sources of 150W and moved on a laboratory scanning table (20cm x 40cm) at a velocity of approximately 90mm/s. The acquisition was done under constant ambient conditions, at frame rate of 300fps and an exposure time of 5ms, in order to acquire images with the correct aspect ratio. The setup for data acquisition is shown in Figure 1.

Reflectance calibration was carried out to correct for background response of the instrument. The dark reference image (D) was acquired by closing the shutter of the camera lens and white reference image (W) was obtained from a 99% reflectance ceramic tile surface. The reflectance value R was calculated from the measured signal (S) on a pixel-by-pixel basis, as shown in equation 1, where i is the pixel index.

$$R_i = \frac{S_i - D_i}{W - D_i} \tag{1}$$

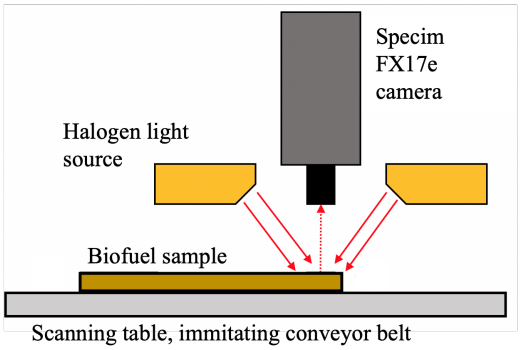


Figure 1: The experimental setup

A hyperspectral image of a biomass sample is shown in Figure 2. A set of spectral data is acquired for each pixel of the image. The spectra obtained from the camera are shown in Figure 3. It should be noted that the noisy parts of the spectra at the lower and higher wavelengths that contain no useful information have been removed in this figure and for all data before pre-treatment. The reference to no pre-treatment in the rest of this paper refers to data in which the noisy parts have been removed.

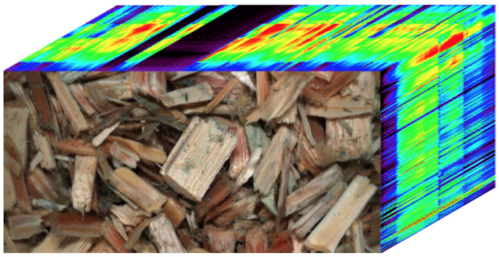


Figure 2: Hyperspectral image of a biomass sample

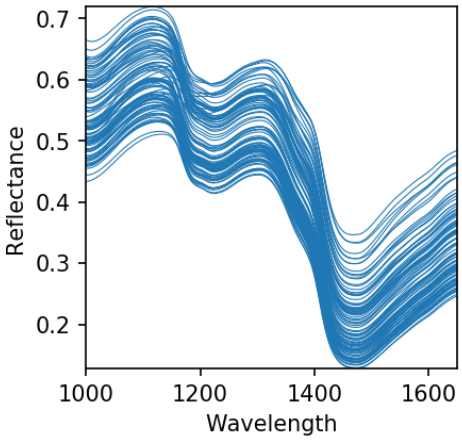


Figure 3: Spectra with no pre-treatment

The moisture content in the samples was determined according to the European standard EN ISO 18134. The samples were oven dried for 20 hours in 105°C and weighed before and after the process. The resulting moisture content range was from 31.0% to 55.8%.

2.2. Data Preprocessing

The spectral data was pretreated to enhance the differences among the samples in order to provide a better calibration model. Noisy parts of the spectrum were removed as they contain no useful information and can instead confound the model. Two different pre-processing techniques were applied: Savitzky-Golay first derivative (SG1) and Standard Normal Variate (SNV), which were shown to perform best in similar samples analysed with NIR spectroscopy Ahmed et al. (2018). The results were also compared to those obtained without any pre-treatment of the data. The pre-treated data was also scaled in accordance with the requirements of the data analysis method, using either the mean and standard deviation or the range of the dataset to obtain a range from 0 to 1.

2.3. Methods for data analysis

In this paper SVR and ANN were compared with PLSR. The different techniques were implemented in Python using the Scikit-learn module (Pedregosa et al., 2011).

SVR is an extension of the Support Vector Machines (SVM) method for classification problems to solve regression problems. It can allow the user to determine the maximum error that is acceptable in the model and find an appropriate hyperplane to fit the data. Hyperparameters C, gamma, and epsilon were adjusted in order to obtain a model that can provide the best prediction. Three different kernel functions (linear, polynomial, radial basis function - RBF) were tested to allow the separation of the data and allow for a better model to be obtained. The hyperparameters were tuned using a grid search with K-fold cross-validation for all kernel functions.

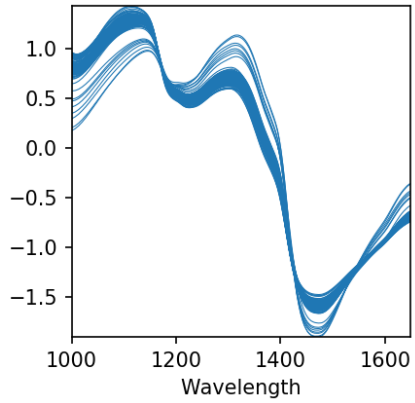
ANNs are widely used for regression problems and can provide good results when the underlying relationship between the different parameters is non-linear. A network with two hidden layers was used, with 128 nodes in the first layer and 32 nodes in the second layer. A third hidden layer was not found to improve the results, which were also similar for 32, 64, and 128 neurons in the hidden layers. The optimal learning rate was selected based on the cross-validation results and the epoch with the lowest error was selected using a callback function.

PLSR is considered the standard approach for spectroscopy applications and performs well when the underlying relationship is linear. Principal component analysis was performed to select the number of components that gives the highest prediction accuracy.

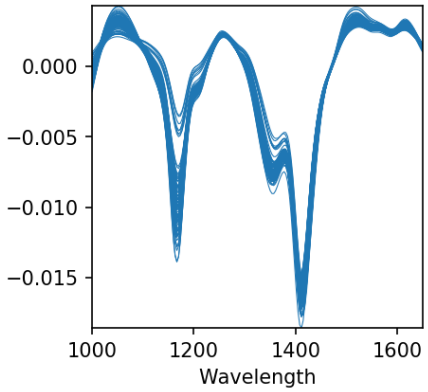
In this work the collected data was split into a training set of 80 samples and a testing set of 20 samples. K-fold cross-validation was also performed during the training to improve the prediction capability of the models and avoid over-fitting. The optimal setting for the cross-validation was found to be 15 folds using 10% of the training dataset.

3. Results and Discussion

The results of the pre-treatment are shown in Figure 4(a) for the Standard Normal Variate and in Figure 4(b) for the Savitzky-Golay first derivative. The setup for SG1 aimed to ensure enough information was retained in the spectra to provide a good model and was evaluated based on the cross-validation results for PLSR. As seen in figures 3 and 4, there are two clear dips in the spectra: one at 1180nm and one at 1430nm. These areas contain much useful information about the chemical composition of the samples. The wavelength of 1180nm is the fingerprint of the C-H stretching overtone, whereas the wavelength of 1430nm corresponds to the O-H overtone. It is this differentiation in the spectra of the different samples that can be coupled to the moisture content and be used to create a robust model.



(a) Spectra with SNV pre-treatment



(b) Spectra with SG1 pre-treatment

Figure 4: Spectra with pretreatment

The pre-treated spectra of the training set were used to build the calibration models. K-fold cross validation was employed to increase the predictive capability of the models. The models were then evaluated on an unseen test set. The results for the PLSR, SVR, and ANN regression for the different pre-treatment approaches are summarized in Table 1. The evaluation metrics used are the goodness of fit measure for linear regression R^2 and the root mean square error, $RMSE$, both for the prediction of the unseen test set. As seen in the results, both the the Savitzky–Golay 1st derivative (SG1) and the SNV pre-treatment methods provide an improvement in terms of fit and error for all modelling approaches.

Table 1: Results of cross-validated models for different pre-treatment methods evaluated on the test set

PLSR		
Pre-treatment	R^2	$RMSE$
None	0.977	1.087
SG1	0.975	1.196
SNV	0.984	0.772

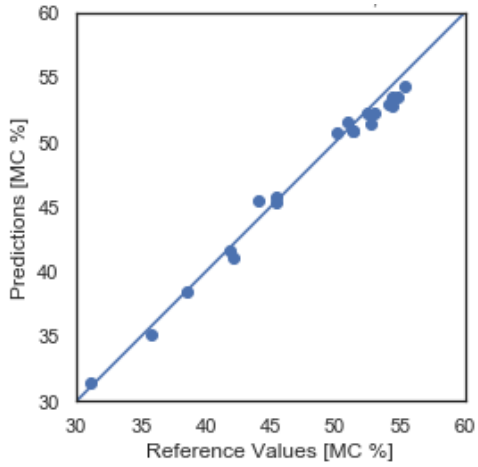
SVR		
Pre-treatment	R^2	$RMSE$
None	0.919	3.893
SG1	0.980	0.952
SNV	0.968	1.530

ANN		
Pre-treatment	R^2	$RMSE$
None	0.949	2.433
SG1	0.969	1.49
SNV	0.973	1.312

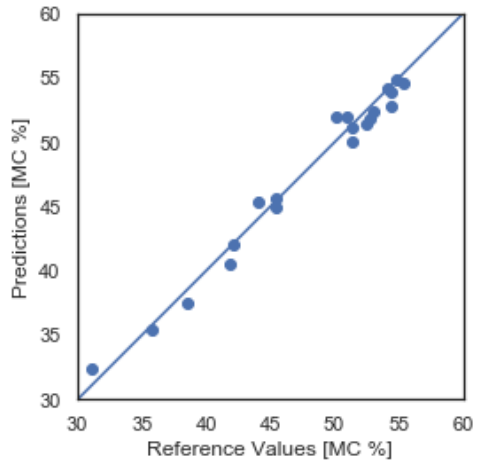
In the case of PLSR the difference between the untreated data and the pre-treated data is very small, which can be attributed to the cleaning of the data and the removal of the noisy parts of the spectra before the pre-treatment. In fact, SG1 pre-treatment performs worse than the untreated data. This is most likely due to the fact that there is little noise in the spectra, rendering the treated and untreated data very similar. It is possible that the window selected for the derivative was slightly larger than the optimal. The SG1 setup was evaluated based on the cross-validation set, and it appears that the results on the unseen test set point out that a different setup would be optimal for SG1. However, this is not possible to know before the models are tested.

In the case of SVR and ANN, the difference between the untreated and pre-treated data is much larger. This is due to the fact that the untreated data was not normalized for these cases, resulting in notably worse models than when the data was pre-treated. For SVR, the best models for each of the different pre-treatments were selected. The models with a linear or polynomial kernel were the best in all cases, whereas the models with the RBF kernel were sometimes slightly overfitted, despite the cross-validation. For ANN, the pre-treatment did not affect model selection as much, and the best models were not as good as those built with SVR. Nonetheless, the difference was not very large, despite the relatively small dataset.

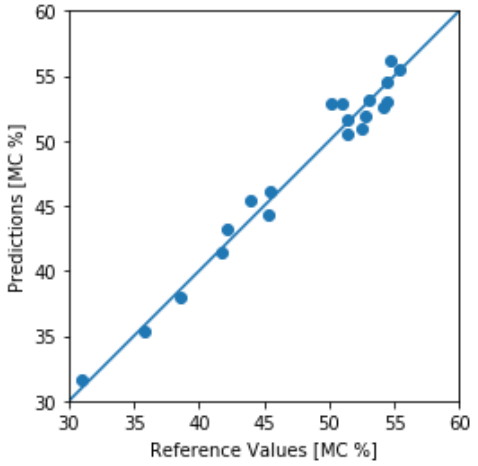
The predicted values of the unseen test set are plotted against the reference values in Figure 5 for the best models, selected based on the R^2 and $RMSE$ values.



(a) PLSR



(b) SVR



(c) ANN

Figure 5: Predicted vs reference moisture content (MC) of the unseen test set

These are the SNV pretreatment for PLSR (Figure 5(a)), SG1 pretreatment with a polynomial kernel for SVR (Figure 5(b)), and SNV pre-treatment for ANN (Figure 5(c)). It can be seen that the predictions of both the SVR and the ANN models are worse than those of the PLSR especially for the higher moisture content, and this could be addressed in future work.

4. Conclusion

This paper compared different methods to develop calibration models for hyperspectral imaging applied to the characterization of biomass samples. The HSI technique can provide good predictions of the moisture content in woody biomass fuel. The two different preprocessing techniques improved the results for SVR and ANN, with standard normal variate performing better than Savitzky-Golay first derivative in the best models developed with two out of the three methods. Overall, PLSR provided the best results, particularly with SNV pre-treatment of the data. Nonetheless, both SVR with SG1 pre-treatment and ANN with SNV pre-treatment were able to deliver accurate and robust models. It is also worth noting that the differences between SNV and SG1 pre-treatment were not very large for many of the models, and the simplicity in the setup of SNV compared to SG1 should be taken into consideration. When taking into account the time requirements to train and tune a model, PLSR and SVR are the best options for this application.

It was noted that in a number of models the results from the evaluation of the models based on their performance in cross-validation did not agree with those of the evaluation based on the test set, with the two pointing at different model setup. Model selection has to be based on the performance on the test set, but since the models will be tuned only with the cross-validation, it is important to be aware that the best model might still be somewhat overfitted with this process. This also points out the usefulness of a more extensive dataset. Nonetheless, the differences are very small and all models are providing acceptable predictions for a real application where this information can be used for the control of a power plant.

Further analysis is required to assess whether different machine learning methods, such as GPR, can be employed to develop a model with an even better predictive capability. GPR is based on a Bayesian approach and as such can also provide the uncertainty range of the predictions. Taking into account the uncertainty of the measurements can provide more information on the suitability of the different techniques for the development of more realistic, probabilistic models. An expansion of the dataset, taking advantage of the range of measurements obtained with the hyperspectral imaging camera, can help increase the accuracy of the models and understand whether the difference in performance is due to the size of the dataset.

Acknowledgment

The authors would like to thank Mikael Karlsson, Konstantinos Kyprianidis, and Robert Tryzell for their technical input and Lotta Lejdberg and Per Örvind for their support.

This work is part of the IFAISTOS and DYNOP projects. IFAISTOS has received funding in the framework of the joint programming initiative ERA-Net Smart Energy Systems' focus initiatives Smart Grids Plus and Integrated, Regional Energy Systems, with support from the European Union's Horizon 2020 research and innovation programme under grant agreements No 646039 and 775970. DYNOP was funded by the Swedish Knowledge Foundation (KKS).

Nomenclature

ANN	Artificial Neural Network
GPR	Gaussian Process Regression
HSI	HyperSpectral Imaging
MC	Moisture Content
ML	Machine Learning
NIR	Near InfraRed
PLSR	Partial Least Squares Regression
RBF	Radial Basis Function
SNV	Standard Normal Variate
SVM	Support Vector Machines
SVR	Support Vector Regression
SG1	Savitzky-Golay first derivative

References

Ahmed, M., Andersson, P., Andersson, T., Tomas Aparicio, E., Baaz, H., Barua, S., Bergström, A., Bengtsson, D., Skvaril, J. and Zambrano, J. (2018), Real-time biomass characterization in energy conversion processes using near infrared spectroscopy - a machine learning approach.

Gewali, U. B., Monteiro, S. T. and Saber, E. (n.d.), 'Machine learning based hyperspectral image analysis: A survey'. doi: 10.48550/ARXIV.1802.08701.

Gillespie, G. D., Farrelly, D. J., Everard, C. D. and McDonnell, K. P. (2016), 'The use of near infrared hyperspectral imaging for the prediction of processing parameters associated with the pelleting of biomass feedstocks', *Fuel Processing Technology* **152**, 343–349. doi: 10.1016/j.fuproc.2016.06.026.

Mancini, M., Mircoli, A., Potena, D., Diamantini, C., Duca, D. and Toscano, G. (2020), 'Prediction of pellet quality through machine learning techniques and near-infrared spectroscopy', *Computers Industrial Engineering* **147**, 106566. doi: 10.1016/j.cie.2020.106566.

Mäkelä, M. and Geladi, P. (2017), 'Hyperspectral imaging to determine the properties and homogeneity of renewable carbon materials', *ChemSusChem* **10**. doi: 10.1002/cssc.201700777.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), 'Scikit-learn: Machine learning in Python', *Journal of Machine Learning Research* **12**, 2825–2830.

Pitak, L., Laloon, K., Wongpichet, S., Sirisomboon, P. and Posom, J. (2021), 'Machine learning-based prediction of selected parameters of commercial biomass pellets using line scan near infrared-hyperspectral image', *Processes* **9**(2). doi: 10.3390/pr9020316.

Rahman, M., Fentaye, A. D., Zaccaria, V., Aslanidou, I., Dahlquist, E. and Kyprianidis, K. (2021), A framework for learning system for complex industrial processes, *in* K. Kyprianidis and E. Dahlquist, eds, 'AI and Learning Systems', IntechOpen, Rijeka, chapter 2. doi: 10.5772/intechopen.92899.

Skvaril, J., Kyprianidis, K. G. and Dahlquist, E. (2017), 'Applications of near-infrared spectroscopy (nirs) in biomass energy conversion processes: A review', *Applied Spectroscopy Reviews* **52**(8), 675–728. doi: 10.1080/05704928.2017.1289471.

Tao, J., Liang, R., Li, J., Yan, B., Chen, G., Cheng, Z., Li, W., Lin, F. and Hou, L. (2020), 'Fast characterization of biomass and waste by infrared spectra and machine learning models', *Journal of Hazardous Materials* **387**, 121723. doi: 10.1016/j.jhazmat.2019.121723.

Tiitta, M., Tiitta, V., Heikkinen, J., Lappalainen, R. and Tomppo, L. (2020), 'Classification of wood chips using electrical impedance spectroscopy and machine learning', *Sensors* **20**. doi: 10.3390/s20041076.

Tsuchikawa, S., Inoue, K., Noma, J. and Hayashi, K. (2003), 'Application of near-infrared spectroscopy to wood discrimination', *Journal of Wood Science* **49**, 0029–0035. doi: 10.1007/s100860300005.