Biomass Characterization through NIR Spectra

1 Introduction

Reducing the moisture content of biomass material has several purposes in different energy applications and in the transportation and storage of materials. Water content in the fuel mass must be driven off before the first combustion stage can occur. This process requires a lot of energy, thus reducing overall system efficiency. This increases the risk of incomplete fuel combustion, leading to the emission of tars and creosote.

High moisture content also impacts the storage and transportation of biomass, as the net energy density of moist biomass is decreased, as water adds weight to the mass.

In this analysis, different pre-processing techniques are applied to a data set of NIR spectral data, which references moisture according to standardized laboratory methods, to predict the moisture value in each sample. The data set was obtained on solid biomass samples of pine and spruce wood chips, bark, and other formats collected at biomass facilities. Predicting moisture levels with NIR spectral data is beneficial, as it is more cost-efficient than weighing the biomass materials.

2 Data Analysis

Spectral data were obtained by using an FT-NIR spectrometer. The data set has the following parameters:

- Recorded Spectral Range: 834-2500 nm (12000-4000 cm-1)
- Spectral Rounds: 5-7
- Spectral Resolution: 16 cm-1
- Number of Data Points in each Spectra: 1037
- Number of tested samples: 125

The number of data points in each spectral round is high, thus feature reduction will be important in this analysis.

2.1 Data Set

Each spectral round of samples is used to train the models to predict the moisture value. Each sample ID is extracted as a SampleID_Base and the spectral round per sample as Spectral_Round.

The number of samples is low for a machine-learning task, with only 125 unique samples. In contrast, the data is very high-dimensional as the number of features is 1037. Before addressing these issues, a visual data analysis will be conducted.

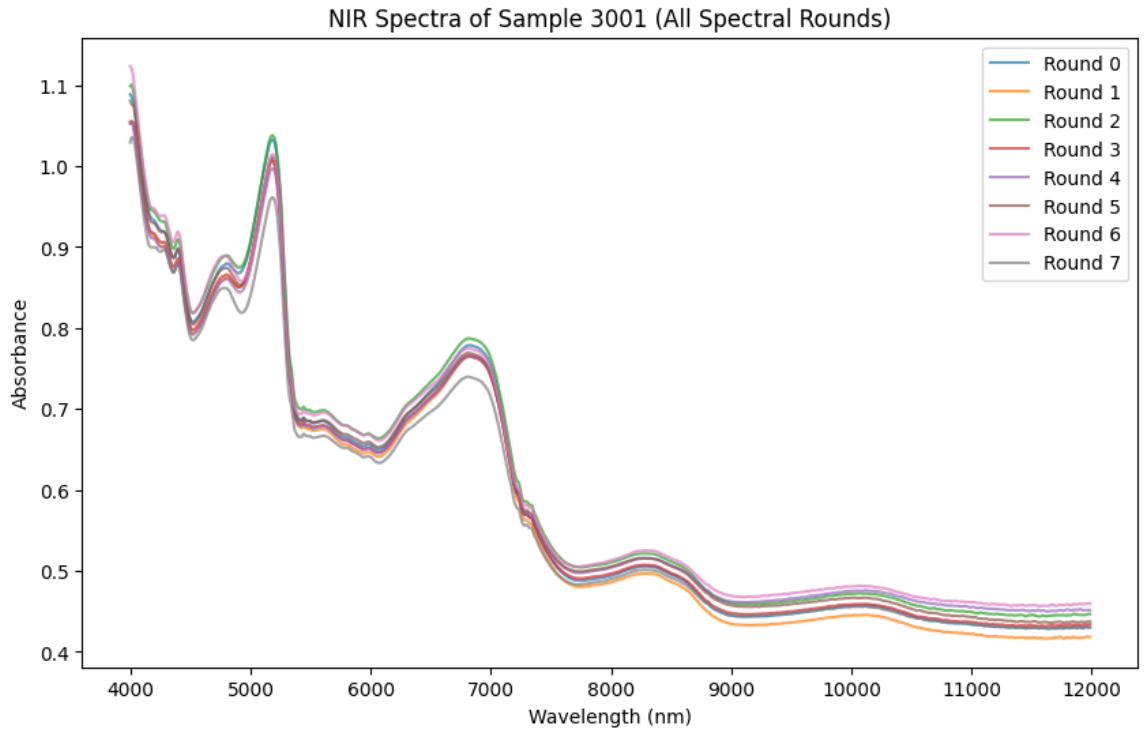

Figure 1 NIR Spectra of Sample 3001

2.2 Visual Data Analysis

A brief visual analysis of the data.

Spectral_Round	count
0	125
1	125
2	125
3	125
4	125
5	125
6	12
7	11

Figure 2 Spectral Rounds per Sample

Most samples have five spectral rounds each, 12 samples have five rounds and 11 have seven rounds, as shown in Figure 2.

Moisture values range from 17.98 to 73.07, with a mean of 45.92, and a standard deviation of 14.44.

Figure 1 shows the NIR spectra of sample 3001 and all its spectral rounds. The absorbance values at each wavelength are plotted as graphs.

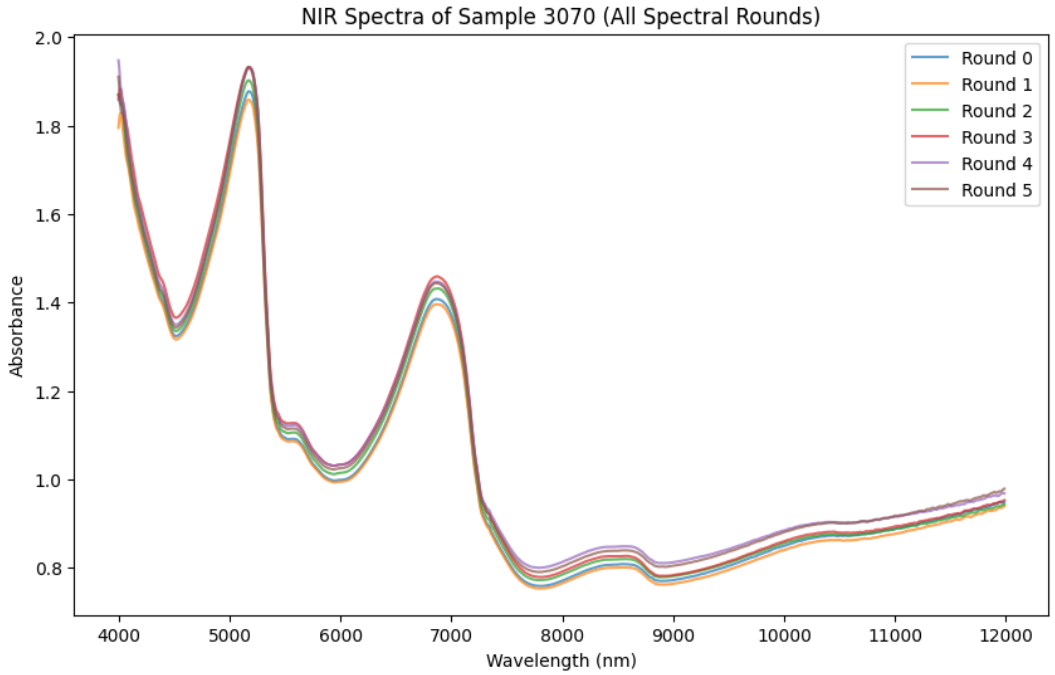

Figure 3 NIR Spectra of Outlier 3070

2.3 Outlier Detection

Outliers were detected by taking the mean for each sample's absorbance and standard deviation values at different wavelengths. The outliers were flagged based on having a Z-score higher than 2. 21 examples were flagged in total. Figure 3 shows sample 3070, which has higher absorbance values than 3001 at correlating wavelengths.

Figure 4 shows spectral round means for each outlier at each wavelength. Most outliers have high moisture values. For this analysis, outliers were dropped at this point to reduce noise, as we are working with a limited sample size.

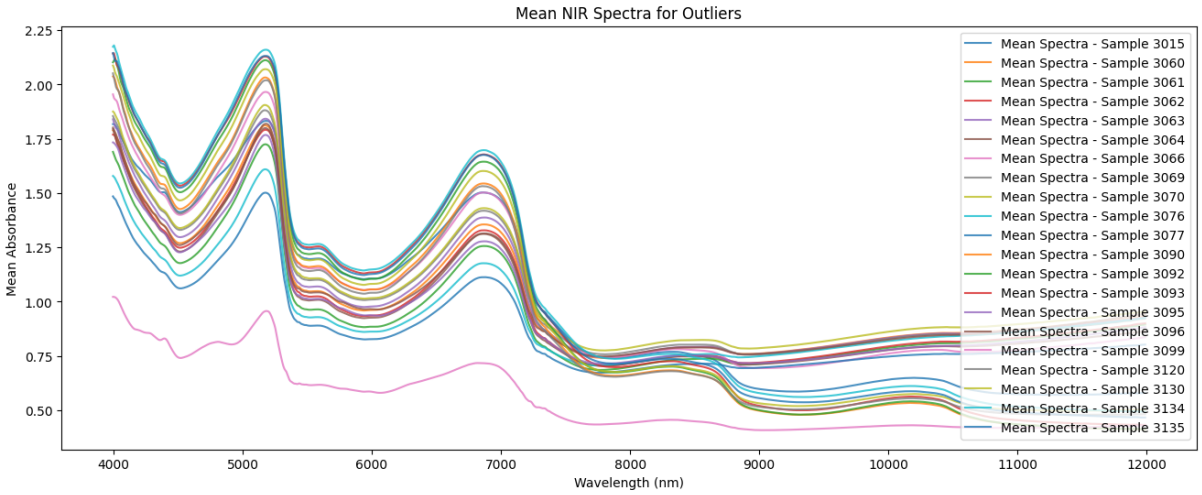

Figure 4 Mean NIR Spectra for Outliers

	PLS	SVR	RNN
Savitsky-Golay filtering	✓	✓	✓
Standard scaler	✓	✓	
PCA		✓	✓
MinMaxScaler			✓

Figure 5 Pre-processing Techniques for Each Chosen Model

3 Pre-processing Techniques

This chapter briefly explores different techniques used in chemometric approaches. Different machine learning models have different needs and thus the combinations of these combinations vary. This chapter briefly explores different techniques used in chemometric approaches.

3.1 Savitsky-Golay Filter

Savitsky-Golay filters are often applied to spectral data to reduce noise. For each data point in the spectrum, the Savitsky-Golay algorithm selects a window of points neighboring a point in the data set. It fits a polynomial to the points in the selected window and replaces the data in the selected window with the value of the polynomial. The window length in this analysis is 7 and the polynomial is of the second order. The filter was applied for the training sets for all machine learning models in this analysis.

3.2 StandardScaler

Standardization of the data is required for many machine-learning models. Objective functions often assume that all features are centered around 0 and have variance in the same order. The data was scaled with this scaler for the partial least-square regression (PLSR) and support vector regression (SVR) models. The recurrent neural network has its own MinMaxScaler algorithm.

3.3 Principal Component Analysis

As the data set is highly dimensional with 1037 features, a dimensionality reduction technique is used to reduce the complexity. Principal Component Analysis (PCA) reduces the dimensionality without losing original information, by transforming potentially correlating variables into a smaller variable set.

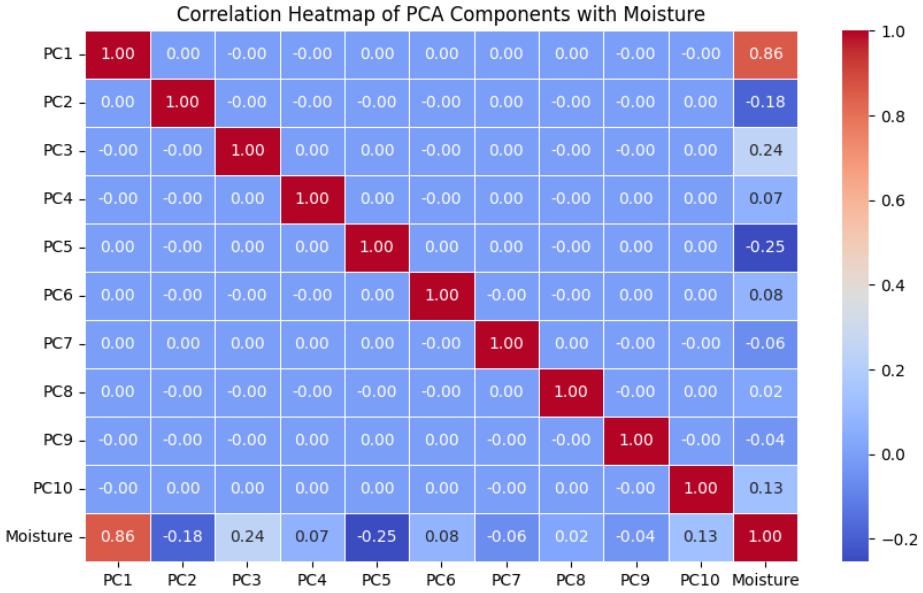

Figure 6 The PCA Components and Correlations

The dimensionality is reduced by combining the features into ten separate components. Figure 6 shows the correlations between the components and their correlation to the target moisture variable. PCA component PC1 has the highest correlation of 0.86, while other components have a low correlation with the target variable. In this analysis, PCA is used for SVM and RNN, as PLSR can handle high-dimensional data within itself.

3.4 MinMaxScaler

MinMaxScaler transforms the features into a given range for the RNN model, ensuring that features contribute equally to the prediction. The target variable is also scaled. Scaling improves model generalization and prediction accuracy.

4 Modeling Approaches

This section briefly discusses the chosen models' characteristics, usage scenarios, and benefits.

4.1 Partial Least Squares Regression

Partial Least Squares Regression (PLS) is a machine learning model with similar properties to PCA. PLS is efficient when the data set has more features than examples, as it reduces the complexity without losing important information.

PLS is used with 15 components after exploring different component values.

4.2 Support Vector Machine Regression

SVM regression differs from classification tasks by reversing the objective of fitting the largest possible street between two classes, by fitting as many instances as possible on the street while limiting margin violations. The hyperparameter epsilon controls the width of the street. Parameter C is the regularization parameter. We use the standard rbf kernel, epsilon 0.05, and C 1700.

4.3 Recurrent Neural Network

Usually, Recurrent Neural Networks (RNN) are used to predict time series, but it was chosen as a model as the spectral range can also be seen as a series of events. In this analysis, RNN is used with a few layers:

- SimpleRNN with 128 neurons
- Dropout 0.3
- SimpleRNN with 32 neurons
- Dense output layer with a linear activation function

The first layer with 128 neurons captures complex properties of the data set and a dropout layer reduces overfitting risk. The second recurrent layer has 32 neurons, which reduces the dimensionality. Finally, a single moisture value prediction is outputted with a linear activation function.

5 Results

For comparison of the results, cross-validation was used for PLS and SVM, while the R-squared and RMSE were calculated separately for RNN.

5.1 PLS

Using cross-validation, the R-squared score is 0.89 and the RMSE is 3.46. The model performs well even with high dimensionality and minimal samples.

The scores for each of the five folds: [0.96417074 0.82243432 0.86630828 0.90927406 0.89328315].

5.2 SVM

The R-squared score for cross-validation is 0.77 and RMSE 5.01.

The SVR model performed well on the validation and test data but has inconsistent R-squared performance on the folds in cross-validation: [0.83579723 0.56820434 0.83605354 0.86479835 0.73473191].

5.3 RNN

The RNN score was calculated best on the original training and testing data sets instead of cross-validation. As expected, the RNN model performed well on the sequential with an R-squared score of 0.93 and RMSE 2.89.

5.4 Comparison

A brief visual comparison of the model's essential scores:

	Model	R^2 Score	RMSE Score
0	PLSRegression	0.891094	3.458960
1	SVR	0.731563	5.433323
2	RNN	0.912509	3.218497

Figure 7 R-squared and RMSE results

As seen in Figure 7, RNN had the lowest RMSE and the highest R-squared score of all three models, closely followed by the PLS model. The SVR model performed decently on most cross-validation folds but had accuracy issues in two out of five folds, as seen in section 5.2.

6 Conclusions

This paper explored spectral biomass data to predict the moisture target value. Moisture estimation based on spectral data is essential for saving costs compared to weighing the materials separately. With additional information, such as biomass sources and harvesting weather conditions, there are also possibilities to gain valuable information on what can lead to higher moisture values in biomass.

The provided data set is small, with only 125 examples. To evaluate the suitability of the different models, all spectral rounds were kept providing more examples for the models. However, this did not increase the variance noticeably, as the spectral rounds of each sample were quite similar. To handle the possible noise of outliers, they were excluded from the analysis in this paper. It is noteworthy that many of the outliers had increased moisture levels compared to the non-outlier samples. With more examples later, the models can train with more variation in the data sets.

Different sets of pre-processing techniques were used for each model, as seen in Figure 5. The PLS and RNN models performed well with the chosen pre-processing techniques, even with a small data set. PLS handled dimensionality reduction within itself, while a PCA approach was used for the RNN model. The RNN model might have performed well without dimensionality reduction, but PCA was used to speed up the training.

SVMs are often used in NIR spectroscopy, but this model showed the lowest performance in this paper when cross-validating. On regular training, validation, and testing sets it had an average of 0.88 R-squared, but the cross-validation had folds with as low an R-squared as 0.56. The different preprocessing techniques could still be fine-tuned, and the sample size could be increased for further examination.

All three models are promising for predicting moisture levels in biomass based on spectral measurements. However, the combinations of pre-processing techniques can still be further explored and, the sample size increased. Machine learning models can have a big environmental impact, as high moisture levels in biomass can lead to emissions. Predicting moisture levels efficiently can also reduce costs while contributing to a more sustainable biofuel industry.

7 References

Course Materials

Effect of Moisture Content – Forest Research
url: https://www.forestresearch.gov.uk/tools-and-resources/fthr/biomass-energy-resources/fuel/woodfuel-production-and-supply/woodfuel-processing/effect-of-moisture-content/ retrieved 20.03.2025

Savitsky-Golay Smoothing Method – NIIRPY Research
url: https://nirpyresearch.com/savitzky-golay-smoothing-method/ retrieved 20.3.2025

StandardScaler – scikit learn documentation
url: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html retrieved 20.3.2025

What is Principal Component Analysis? – IBM
url: https://www.ibm.com/think/topics/principal-component-analysis retrieved 20.03.2025 retrieved 20.3.2025

MinMaxScaler – scikit learn documentation
url: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html retrieved 20.3.2025

Partial Least Squares Regression (PLSRegression) using Sklearn – Geeks for geeks
url: https://www.geeksforgeeks.org/partial-least-squares-regression-plsregression-using-sklearn/ retrieved 21.3.2025