



Identification and Classification of Technical Lignins by means of Principle Component Analysis and k-Nearest Neighbor Algorithm

Friedrich Fink,^[a, b] Franziska Emmerling,^[a, b] and Jana Falkenhagen*^[a]

The characterization of technical lignins is a key step for the efficient use and processing of this material into valuable chemicals and for quality control. In this study 31 lignin samples were prepared from different biomass sources (hardwood, softwood, straw, grass) and different pulping processes (sulfite, Kraft, organosolv). Each lignin was analyzed by attenuated total reflectance Fourier transform infrared (ATR-FT-IR) spectroscopy. Statistical analysis of the ATR-FT-IR spectra by means of principal component analysis (PCA) showed significant differences between the lignins. Hence, the samples can be separated by PCA according to the original biomass. The differences observed in the ATR-FT-IR spectra result primarily from the relative ratios of the p-hydroxyphenyl, guaiacyl and

syringyl units. Only limited influence of the pulping process is reflected by the spectral data. The spectra do not differ between samples processed by Kraft or organosolv processes. Lignosulfonates are clearly distinguishable by ATR-FT-IR from the other samples. For the classification a model was created using the k-nearest neighbor (k-NN) algorithm. Different data pretreatment steps were compared for $k=1 \dots 20$. For validation purposes, a 5-fold cross-validation was chosen and the different quality criteria Accuracy (Acc), Error Rate (Err), Sensitivity (TPR) and specificity (TNR) were introduced. The optimized model for $k=4$ gives values for $\text{Acc}=98.9\%$, $\text{Err}=1.1\%$, $\text{TPR}=99.2\%$ and $\text{TNR}=99.6\%$.

1. Introduction

After cellulose, lignins are the second most abundant biopolymers on our planet. Lignin is one of the main components in more highly developed vascular plants. The annual production of plant lignin is estimated at 20×10^9 tons per year.^[1] The research interest in lignins increased strongly since 2000, especially under the aspect of research on sustainable carbon sources and their use as an alternative to fossil carbon sources.^[2] The largest producer of technical lignins at present is the pulp and paper industry who produce approx. 50 to 70 million tons annually, but only a fraction is further processed into valuable chemicals.^[2] Lignin is also a by-product of biorefineries.^[2] The most common methods for separating lignin from other plant components are the Kraft process, the sulfite process, the soda-anthraquinone process and organosolv processes.^[3]


The chemical properties of technical lignins are determined by their source material and their respective extraction method, which leads to different potential applications. The smallest repetitive units in lignins are the phenolic propanoids: p-coumaryl alcohol (p-hydroxyphenyl, H) and coniferyl alcohol (guaiacyl, G) as well as sinapyl alcohol (syringyl, S). The ratio of these so-called monolignols to each other determines the structure and functional groups of the resulting lignin. The monolignol ratio varies between plant species. For example, lignins from softwoods are suitable for different applications as compared with lignins extracted from hardwoods, straw, or grasses. To produce lignin-modified phenol-formaldehyde resins, lignin with a relatively high proportion of free phenolic hydroxyl groups and free and reactive ortho positions is required.^[4] The use of lignins with a high proportion of H and G units and few S units, as is the case of softwood, is advantageous.


The pulping process also influences the further processing of the lignins. Lignosulfonates have some distinct properties in direct comparison to Kraft lignins. For example, due to the high content of sulfonate groups they are negatively charged and water soluble.^[5] Lignosulfonates are used for further processing into surfactants,^[6] animal feed,^[7] stabilizers in colloidal suspensions^[8] and plasticizers in concrete.^[9] Depending on the above mentioned factors lignins in general are further regarded as promising candidates for the production of high value materials and chemicals such as carbon fibres, synthesis gas, aromatic and functionalized hydrocarbons.^[10–11]

The major hurdle in the use and further processing of technical lignins is the complex structure inherent in the heterogenous material and the chemical change in structure

[a] F. Fink, Dr. F. Emmerling, Dr. J. Falkenhagen
Bundesanstalt für Materialforschung und -prüfung (BAM)
Richard-Willstätter-Strasse 11
12489 Berlin (Germany)
E-mail: jana.falkenhagen@bam.de

[b] F. Fink, Dr. F. Emmerling
Mathematische-Naturwissenschaftliche Fakultät
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin (Germany)

 Supporting information for this article is available on the WWW under <https://doi.org/10.1002/cmt.202100028>

 © 2021 The Authors. Published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution Non-Commercial NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

that can be caused by some pulping processes. It is therefore of crucial importance to develop suitable qualitative and quantitative analytical methods that meet industrial requirements in terms of time and cost. The technical possibilities to analyze lignin are vastly diverse. The analytical method which is most suitable depends on the respective problem and the kind of sample. For qualitative determination of native lignin in wood samples, for example, the staining and microscopy of microtomes are suitable.^[12–14] The elucidation of structural features is realized according to the current state of the art by 1D and 2D NMR (Nuclear Magnetic Resonance) techniques^[15–17] and various chromatography-based mass spectrometric methods^[18] as well as different optical spectroscopic methods. Among the latter, Fourier transform infrared spectroscopy (FT-IR) has become a fast and reliable method over the past decades. Many publications deal with the determination of the content of individual components in wood using near (NIR) or mid (MIR) infrared by computer-aided methods such as principal component analysis (PCA) and various multivariate regression methods like partial least squares regression (PLSR) or principle component regression (PCR).^[19–22] These approaches are also used in the investigation of technical lignins. For using this new raw material efficiently, it is of interest to determine which biomass and which extraction process is subject to lignin. A prominent classification tool in pattern recognition is the *k*-nearest neighbor algorithm (*k*-NN). The *k*-NN method was recently used to determine the heartwood and bark content of wood chips^[23] and for crop classification.^[24]

In the present work 31 samples of different technical lignins were examined obtained by different pulping methods and from different biomass. The analytical method of choice was Fourier Transform Infrared Spectroscopy (FT-IR). The aim was to evaluate whether and to what extent the fingerprint of the initial biomass is retained in the lignin during different pulping processes. Furthermore, it was examined whether it is possible to create a *k*-NN classification model with based on PCA from the available spectra to enable the identification of unknown technical lignins regarding their biomass.

2. Results and Discussion

2.1 Spectral Data and Characteristic Vibrational Bands

Spectra of the samples in Table 2 were analyzed by ATR-FT-IR and are shown in Figure 1 as normalized second derivatives, the raw spectra are shown in Figure S1 in the Supporting Information. Three spectra were recorded per sample, which were averaged for clarity. The samples were grouped according to their biomass. Distinctive absorption bands were plotted with dashed lines. The vibration bands in the technical lignin samples have been identified with the help of literature (Table 1). In the ROI the spectra show several characteristic bands.

The absorption maxima are shown as minima since they are the second derivatives. The higher the derivative used, the sharper the maxima and the better hidden features are

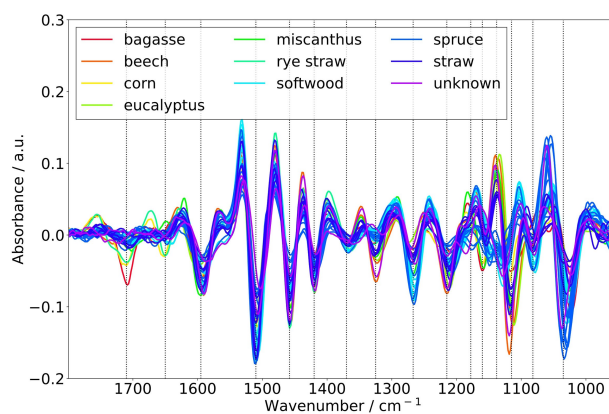


Figure 1. Second derivative of the ATR-FT-IR spectra of various technical lignins in the spectral range 1800–950 cm^{-1} .

Table 1. Assignment of bands of technical lignin samples in the IR spectrum.^[26–29]

Wave-number/ cm^{-1}	Vibration band assignment
1710	unconjugated C=O stretching
1650	conjugated C=O stretching
1610–1590	C=C aromatic stretching (aromatic skeleton) with C=O
1512	C=C aromatic stretching (aromatic skeleton)
1458	C–H deformation asymmetric in CH_2 and CH_3 , with C–H in-plane deformation
1420	C=C stretching aromatic skeleton, with C–H in-plane deformation aromatic skeleton
1370	O–H in-plane deformation (phenolic), C–H in CH_3
1325	S-ring ring breathing with C=O stretching, or G-ring substituted in C5
1267	G-ring ring breathing with C–O stretching
1215–1208	C–C with C–O and C=O stretching
1178	S=O symmetric stretching
1160	C=O stretching conjugated ester
1138	C–H in-plane aromatic deformation G-ring
1115	C–H in-plane aromatic deformation S-ring
1082	C–O deformation of secondary alcohols and aliphatic esters
1070–1030	S=O symmetric stretching of lignosulfonat salts
1040–1030	C–H in-plane aromatic deformation with C–O deformation (primary alcohols and ethers) and unconjugated C=O stretching and
915	C–H aromatic bending out-of-plane

elucidated. However, the spectrum becomes increasingly noisy and the more interpolation points must be selected for smoothing. A high degree of smoothing leads to suppression of small spectral features. So, it is important to find the sweet spot between increasing background noise and smoothing. A good compromise resulted in taking 21 smoothing points and a second derivative. A baseline correction is automatically performed by the derivation. Further a normalization is important and a crucial step for the later PCA since different features come in different units of measurement. In the case of IR spectra, normalization is advisable because the detector sensitivity depends on the wavelength. This offset was removed by means of unit vector normalization. The Samples were stored in the dark under dry conditions, moisture content was neglectable.

Softwood, hardwood and straw consist of different proportions of the three monolignols (H, G, S). For example, softwood consist of 90% G-units, whereas hardwoods are composed of a mixture of G- and S-units.^[25] Straw and grasses contain additional H-units.^[25] The typical vibration bands for guaiacyl are approximately at 1267 cm⁻¹ and 1138 cm⁻¹, those for syringyl appear at approximately 1325 cm⁻¹ and 1115 cm⁻¹. In general, the spectra are similar. However, there are also significant differences at some positions. Strong vibrational bands at 1710 cm⁻¹ was only found in samples from bagasse, corn and rye straw, indicating a larger quantity of free carbonyl stretching vibrations.

The syringyl vibration bands at 1325 cm⁻¹ and 1115 cm⁻¹ appear most strongly in lignins from beech, eucalyptus, straw, and a sample of unknown origin (OL-3). In softwood samples, S-units are generally not observed, but are weakly visible in some cases. The opposite is true for the vibrations of G-units. The softwood-like samples show increased absorption values, whereas hardwood, straw and grass samples lack strong vibrational bands.

The lignosulfonate samples show bands in a range around 1180 cm⁻¹ and 1050 cm⁻¹–1030 cm⁻¹, which indicates sulfonate stretching vibrations.^[27] This is consistent with the sulfite-process pulping procedure used for these samples.

2.2 Principle Component Analysis

The samples in Table 2 were analyzed by PCA. Since the sample mixtures are complex, PCA was performed for 10 PCs to ensure that nearly the entire variance of the data set was recorded. Furthermore, a systematic cross-validation was performed internally over all samples according to Wold et. al.,^[30] the explained variance is given for PC_i to PC_k over:

$$R_{CV}^2 = 1 - \frac{\sum_i^k \sum_j^n (x - \hat{x})^2}{\sum_i^k \sum_j^n x^2}$$

Validation is not always necessary for purely exploratory data analysis and for an initial review of the data. If the PCs are subsequently used for any regressions or classifications, validation is necessary. Low validation values indicate a poorly calibrated model that describes only noise in the data structure and has no relation to the actual data. The cumulative explained variance of the PCs for calibration and validation is given in the supporting information (Figure S2 in the Supporting Information).

PC 1 and PC 2 describe 45.2% and 19.2% of the spectral variance. PC 3 is responsible for 8.8% of the variance. The first three PCs together describe 73.3% of the spectral variance. A score plot for PC 1 and PC 2 is shown in Figure 2. Four main clusters formed and a confidence ellipse with the size of two standard deviations was drawn for each cluster. In the first quadrant and partly in the fourth quadrant are all samples originally obtained from hardwoods, grasses or straw, as well as the sample OL-3 of unknown origin. The second quadrant contains all lignosulfonates produced from spruce and the

Table 2. Overview of lignin samples with information on the respective extraction process and underlying biomass.

sample	pulping process	biomass
190	kraft	softwood
L3	organosolv	rye straw
L2	organosolv	corn
L1	organosolv	bagasse
K1	Bergius-Hägglund ^{[a][32]}	spruce
AL	kraft	unknown
OL-3	organosolv	unknown
A1	sulfite	spruce
A2	sulfite	spruce
A3	sulfite	spruce
A5	sulfite	spruce
A6	sulfite	spruce
A7	sulfite	spruce
B1	organosolv	straw
B2	organosolv	straw
B3	organosolv	straw
B4	organosolv	straw
B5	organosolv	straw
C1	organosolv	beech
C2	organosolv	beech
D1	organosolv	spruce
D2	organosolv	spruce
E1	kraft	softwood
E2	kraft	softwood
E3	kraft	softwood
F1	kraft	eucalyptus
G1	organosolv	miscanthus
O1	sulfite	softwood
O2	kraft	pine
O3	kraft	pine
ZPR	kraft	softwood

[a]This is no pulping process, but a method for hydrolyzing carbohydrates in biomass with concentrated hydrochloric acid.

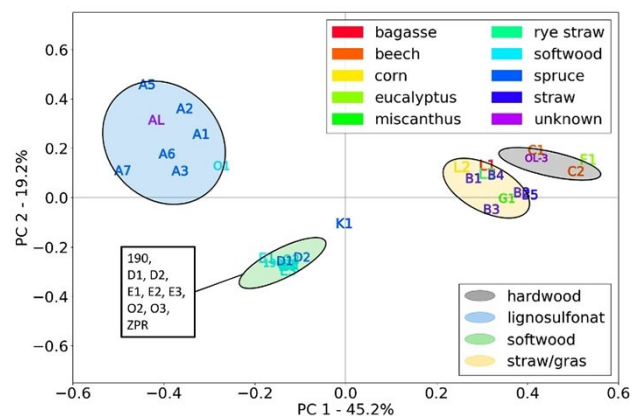


Figure 2. Scores plot for PC 1 and PC 2 of the different technical lignin samples.

sample O1 from softwood. The remaining samples are grouped in the third quadrant, which is also made of softwood.

The reason for the subdivision into the different groups results from the loadings for PC1 and PC2 (Figure 3). The loadings of the respective PC plotted against the wave numbers indicate which wave number is responsible for the displacement in positive or negative direction. The exact position on a PC is given by the scalar product of the loading weights of that PC with the mean centered data.

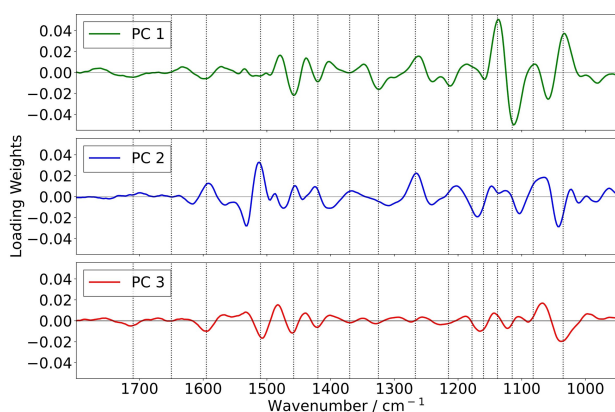


Figure 3. Loadings for PC 1 (45.2%), PC 2 (19.2%) and PC 3 (8.8%), important vibrations are shown as dashed lines.

With regard to the loadings for PC 1, it becomes clear that the wave numbers around 1512 cm^{-1} , 1460 cm^{-1} , 1325 cm^{-1} , 1267 cm^{-1} , 1138 cm^{-1} , 1115 cm^{-1} and $1070\text{--}1030\text{ cm}^{-1}$ are mainly responsible for the position of a sample on PC 1. Depending on the strength of the absorption, the samples are positioned in the scores plot according to their loading weight. For example, the C–H deformation vibration of the G-ring in softwood at 1138 cm^{-1} has a strong positive value in the loadings of PC 1. Samples that have a strong absorption at this wavenumber are shifted strongly in negative direction on PC 1 and vice versa. Similar results are obtained for samples containing S-units. These are shifted strongly in the positive direction along PC 1 due to the absorption bands at 1115 cm^{-1} and 1325 cm^{-1} . It can be concluded that samples of softwoods are grouped in the negative area of PC 1 and samples of hardwoods and straw or grass are grouped in the positive area, due to their respective loading values. Furthermore, it can be assumed that the liginosulfonates were also obtained from softwood, since they are also positioned in the negative region of the scores plot. The liginosulfonate cluster is additionally shifted more strongly in the negative direction by absorptions of 1178 cm^{-1} and 1035 cm^{-1} .

The separation of hardwood and straw/grass is achieved by S-units, hardwoods have a higher proportion of S-units,^[31] resulting in extremely positive values on PC 1.

Along PC 2, the cluster of hardwoods and the cluster of grass- and straw-like samples differ minimally. The group of liginosulfonates, on the other hand, is widely dispersed, whereas the cluster of softwoods is densely packed. The scattering is influenced by the sulfonate vibrational bands at $1070\text{--}1030\text{ cm}^{-1}$ and 1180 cm^{-1} . Medium to strong absorptions shift the liginosulfonates on PC 2 strongly into the positive range. The stronger the absorption, the more positive the value for PC 2. Hence, the value of PC2 allows conclusions to be drawn regarding the degree of sulfonation, since different levels of absorption indicate higher or lower functionalization. PC 3 describes predominantly conjugated and unconjugated C=O and C–O stretching and deformation vibrations in secondary alcohols or aliphatic ester groups.

An interesting question is to what extent the pulping process is reflected in the IR spectra. If the samples in Figure 3 are labelled according to the pulping methods (Figure S3 in the Supporting Information), it appears the clusters can also be explained by the different pulping processes rather than the original biomass. The primary reason for the apparent formation of clusters according to the pulping process is because softwoods have been treated mainly by Kraft and sulfite procedures for several decades. Hardwoods are usually also processed by the established sulfate procedure. More recent methods, such as organosolv processes, are typically used for annual plants, such as various types of straw or grasses, and for agricultural waste.

In direct comparison with sulfate and sulfite processes, organosolv processes offer the possibility to produce lignins of relatively high quality.^[2] Furthermore, the solvents used can be recovered by simple distillation, yielding less water pollution and thus preserving the environment.^[2] Samples D1 and D2 in the coniferous wood cluster and OL-3, C1 and C2 in the hardwood cluster are probably test samples to evaluate to what extent organosolv processes might be used for plant material other than annual plants.

Separation is primarily achieved according to biomass signature. The organosolv lignins D1 and D2 arrange themselves optimally in the softwood cluster, which consists mainly of Kraft lignin. Sample K1 lies slightly outside the softwood cluster. The reason for this offset could be the process used. K1 is a residual product from a hydrochloric acid process to hydrolyse carbohydrates.^[32] Due to condensation processes these lignins have a higher proportion of carbohydrates and lignin-carbohydrate complexes.^[33]

The only Kraft lignin of hardwood between the hardwood organosolv lignins is sample F1, but it is not noticeable by any significant shift within the hardwood cluster. Both, the kraft and organosolv methods modify lignin to an extent that still allows detailed conclusions to be drawn about the original biomass.

In contrast to organosolv lignins, kraft lignins contain sulphur components. The reason why PCA does not show any separation between the two processes is probably due to several factors.

Firstly, it depends on the form in which the sulfur is present. Svensson et al.^[34] proposed that the sulfur is mainly bound as organic sulfur ($\sim 70\%$), as inorganic sulfur ($\sim 29\%$) and elemental sulfur ($\sim 1\%$). Half of the organic sulfur is present in the form of disulfides R_2S_2 and the other half as thiiranes ($R\text{--}S\text{--}R$) and thiols ($\text{--}SH$). The inorganic sulfur is mainly present as sulfate ion (SO_4^{2-}).^[34] Evdokimov et al.^[35] stated that there are many more species, but in less extent and studies about that topic are scarce and inconclusive. Secondly, the infrared vibrations of the sulfur containing groups in kraft lignin are outside of the ROI that was used for the data analysis. The PCA only recognizes the vibrations in the range of $1850\text{--}950\text{ cm}^{-1}$. Thiols absorb around $2600\text{--}2400\text{ cm}^{-1}$ and $800\text{--}600\text{ cm}^{-1}$, thiiranes absorb at $800\text{--}600\text{ cm}^{-1}$ as well.^[27] Furthermore, sulfate ions are present in less extent and may not be recognized by IR spectroscopy, due to the general broad and overlapping vibrational bands. It could also be possible, that the technical lignin underwent

some washing processes after extraction to remove inorganic salts, which is probably the case. Thirdly, the total sulphur content in kraft lignins is only 2–3%.^[35]

However, lignin is highly functionalized with sulfonate groups in the sulfite process.^[36] These sulfonate groups are not present in the other samples, resulting in a significant separation of the lignosulfonate samples from both the hardwood and other softwood lignins in the PCA. The sample AL, although classified as a Kraft lignin, is counted amongst the lignosulfonates. This is commercially acquired lignin from Sigma-Aldrich; in the CoA there is a clear reference to sulfonate components,^[37] which causes the shift to the lignosulfonate cluster.

The results of the PCA are generally consistent with the study on technical lignins by Lancefield et al.,^[38] Cotrim et al.^[39] and Beoriu et al.,^[40] who were also able to distinguish lignins using PCA based on the original biomass. What distinguishes the present work from the others in the next section is the automatic classification using k nearest neighbor algorithm based on the extracted PCs.

2.3 k Nearest Neighbor

With the k -NN method several classification models for $k = 1 \dots 20$ were created because different values for k generate different error rates and accuracies. As data sets the mean-centered raw data, the derived, and the derived normalized data were used. From each data set a PCA was performed and the first two PCs were used for the model building. Here it will be shown how data pretreatment can influence on the performance of the model. The robustness of each model was evaluated with a 5-fold cross validation. Here, 20% of the samples were randomly selected to be retained as a test set. The model was trained with the remaining 80%. This selection-training procedure was repeated 5 times. Since every sample was measured three times only complete subsets were used for testing and training. For each cross-validation step the figures of merit Acc, Err, TPR and TNR were calculated and averaged at the end. Figure 4 shows the mean error rates of the training and test data sets over all values for k and Figure 5 shows the mean values for sensitivity and specificity over all values for k for the test data sets.

As expected, the error rate of the training sets for $k=1$ is $\text{Err}=0.0$ which is equivalent to an accuracy of 100%. This is because the nearest neighbor of a training data point will almost always belong to the same class as itself. The model is therefore overfitted at the boundaries for $k=1$. It would be wrong to conclude that this is the optimal value for k . Furthermore, the differences of the models with respect to the selected data pretreatment becomes visible. Looking at the error rate of the raw data for the training data set, it increases for $k=2$ abruptly. After rising to 25.5%, the error rate increases continuously for increasing values of k .

The mean error rate of the test data set fluctuates around an approximate value of 67%. This is equivalent to an accuracy of 33%. Regardless of the k -values selected, a reliable classi-

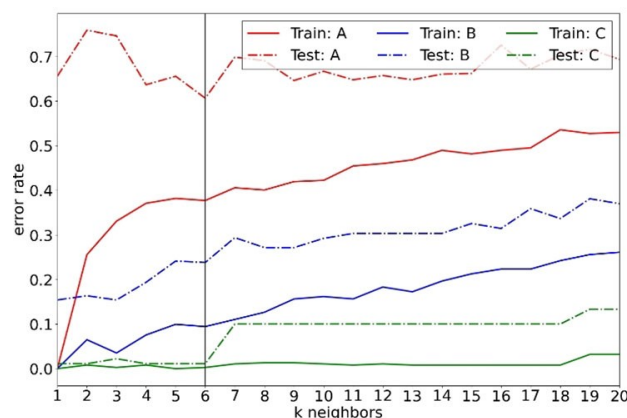


Figure 4. Error rate of the training and test data set during different data pretreatment steps. A: mean centered raw data, B: second derivative, C: second derivative + UVN.

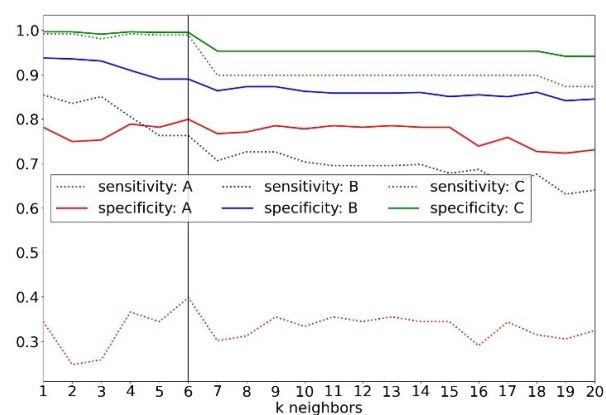


Figure 5. Sensitivity and specificity of the training and test data set during different data pre-treatment steps. A: mean centered raw data, B: second derivative, C: second derivative + UVN.

fication of unknown samples by k -NN models based on the mean centered raw spectra is not possible without significant error rates. As data preprocessing progresses, the performance of the model also improves. With the preceding PCA it is possible to reduce the feature space, which was previously spread out from more than 400 wave numbers in the ROI, to two meaningful features. The models created from the derived data are on average more than twice as accurate compared to the models from the mean centered raw data. This is a significant change in comparison to the mean centered data. The best performance is achieved with the fully optimized model. At this point it is necessary to discuss the optimal value for k . For $k=7 \dots 18$ the values Acc, Err, TNR and TPR remain constant and drop for $k=19$ and $k=20$. Due to these facts there is no reason to choose values for $k > 7$. Another reason is the fact that the four biomass classes used have different sizes. With increasing values for k , the risk of samples being incorrectly assigned increases. The probability that the k closest neighbors of a sample, a class with more specimens, will be assigned increases. This means, for example, that the nearest sample of a sample to be classified belong to a smaller class,

but the nearest $k-1$ neighbors belong to a larger class, and thus the sample is incorrectly classified. The optimum value for k with respect to the figures of merit Err (1.1%), Acc (98.9%), TPR (99.2%) and TNR (99.6%) for 5-fold cross validation is $k=4$. According to Figure 4 and 5, values for k of $k=5$ and $k=6$ are also conceivable since they provide nearly identical results. In view of the different sizes of the classes and the relatively small sample size of 31 individual samples, a low k value should be chosen for the present problem.

To show which samples were assigned to which class and where the decision boundaries of the k -NN classifier run, they were entered into the scatter plot of PC 1 vs. PC 2 (Figure 6). As the clusters were sharply separated from the beginning, there are nearly no samples that were wrongly assigned by the optimized model. As the only incorrectly classified sample, one of the triple measured L1 samples from bagasse is in the hardwood sector. This may be a measurement error, or the bulk sample was poorly homogenized.

3. Conclusions

In the present work, technical lignins obtained from different biomass and with different industrial pulping methods were analyzed. By means of FT-IR and subsequent principal component analysis, chemical differences in the spectra of the individual lignin samples could be identified clearly and comprehensibly. The different samples were successfully separated according to their biomass in the scoresplot PC 1 vs. PC 2 and assigned to the four superordinate groups: coniferous wood, hardwood, straw/grasses and lignosulfonates. However, a clear fingerprint of the pulping method used could not be obtained from the IR spectra alone for all methods. Only the lignosulfonates stood out clearly from the other samples due to their high degree of sulfonate functionalization. Under the aspect of automatic classification, a model for the classification of FT-IR spectra of samples of unknown biomass was created using the simple yet powerful k -NN algorithm. The spectral range was manually limited to 1800–950 cm^{-1} . However,

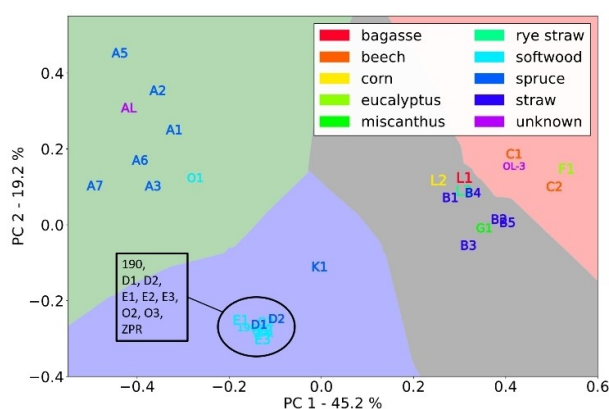


Figure 6. The k -NN decision boundaries plotted on top of PCA results. Color code: hardwood – red, straw/gras – grey, softwood – blue, lignosulfonates – green.

classification based on the raw spectra was not possible. Only after specific data pretreatment it was possible to create a reliable and optimized classification model. A second derivative according to Savitzky-Golay with 21 interpolation points and a second-order polynomial with subsequent unit vector normalization was found to be a suitable data pretreatment. The hyperparameter k for the optimized model resulted in $k=4$ with the processed data. The model was validated by 5-fold cross-validation and resulted in Err = 1.1%, Acc = 98.9%, TPR = 99.2% and TNR = 99.6% for the figures of merit. The k -NN classifier proves to be a reliable method for classifying unknown samples. To the best of our knowledge, this is the first-time k -NN algorithm was combined with FT-IR spectra of technical lignins for the purpose of automatic classification. This represents a step towards automated quality control or handheld devices, which are capable of fast classification. In future work it will be of interest to compare different classification methods, as the k -NN algorithm requires long computing time for large sample quantities.

Experimental Section

This section provides information on the samples and the equipment used, as well as the applied software and mathematical methods used.

Lignins

An overview of the 31 lignin samples kindly provided by industrial project partners is shown in Table 2.

Additional information on individual samples can be read from the columns pulping process and biomass. For some samples detailed information is available, such as the tree species or the exact pulping process, for others only the parent plant group is known. In some cases, information on biomass or pulping process is missing.

ATR-FT-IR

The IR spectra were recorded with a Nicolet Nexus 670 FT-IR (ThermoFisherScientific GmbH) in attenuated total reflection (ATR) mode and a “Golden Gate” sample holder. Before each measurement, the sample holder was cleaned with ethanol and acetone. The background spectrum was measured in air, then three different aliquants of one sample were measured. Per measurement 32 scans were taken with a resolution of 4 cm^{-1} . The spectral range was set to 4000–600 cm^{-1} and the spectra were recorded in absorption mode. The software OMNIC was used and the spectra were exported in CSV file format.

Python

The freely available Anaconda distribution (4.8.3) with the integrated development environment Spyder (4.0.1) and Python (3.7.6) was used. The import of the data was done with the additional package Pandas (1.0.3), numerical processing of the data was based on NumPy (1.18.1). PCA and k -NN algorithms were taken from the scikit-learn (0.22.1) library. Derivation and smoothing methods according to Savitzky-Golay (SG) are included in SciPy (1.4.1). The graphics were created with matplotlib (3.1.3).

Spectral Preprocessing

For the precise extraction of information from the spectrum, prior data pre-treatment is essential, while at the same time improving the predictive power of the model. First, the spectral range was narrowed down. The selected range of 1800–950 cm^{-1} is in the following referred to as region of interest (ROI). To work out any absorption maxima more clearly, a second derivative of the spectra is formed with the SG algorithm. The derivation of asymmetrical vibration bands may lead to slight displacements. The derived spectra were automatically smoothed with the SG algorithm, a second-degree polynomial with 21 interpolation points was used here. The exact specification of the function in Python is:

```
scipy.savgol_filter(window_length=21,polyorder=2,deriv=2).
```

The final step involves unit vector normalization (UVN), which converts each row of the data matrix, which can be understood as a vector in multidimensional space, into unit length:

$$x_i(\text{UVN}) = \frac{x_i}{\sqrt{\sum_{j=1}^n x_{ij}^2}}$$

Note that after smoothing the spectrum, $\frac{n-1}{2}$ positions at the edges of the ROI are set to zero. Before applying the UVN, the edges should be shortened by just those zero points, otherwise they will be included in the calculation of the normalization. All IR spectra were mathematically pre-treated in the same way.

Principle Component Analysis

PCA is amongst the most used methods for dimensional reduction and exploratory data analysis. The data matrix $X_{m \times n}$ consisting of m samples and n features is approximately decomposed into a matrix with lower rank h , for which $h \ll n$ applies. In the case of IR spectra, the properties are the absorption values at different wavelengths. The new latent variables h are uncorrelated and are called principal components (PCs). The PCs are linear combinations of the original data and run towards the greatest variance. PCA is one of the unsupervised methods and is strictly speaking not a method for classifying and distinguishing different samples. However, it is possible to clearly show the differences contained in the data. The data speak for themselves in this sense. In this article a complete singular value decomposition (fullSVD) according to the method of Halko et al. included in the scikit-learn package was used.^[41] The following parameters were used:

```
sklearn.decomposition.PCA(n_components=10,svd_solver='auto').
```

k-Nearest Neighbor

The k -NN algorithm is one of the simplest yet most powerful classification and pattern recognition methods available.^[42] Starting from a data set used to train the model, which contains samples of a known category, samples of an unknown category of the test data set can be classified. The principle of the method is to find the k -nearest neighbors of a sample of the test data set in the feature space of the training data set. The choice for k depends strongly on the type of data. In general, very low values for k (e.g. $k=1$, $k=2$) are more prone to outliers in the data and the result generally appears noisier. Too large values for k , on the other hand, may outperform categories with few samples. In general, the k -NN algorithm is very intuitive, easy to install and to interpret. Furthermore, it has few hyperparameters, only k and the distance metric, which must be optimized. Despite the advantages, there are

also some limitations. k -NN is extremely memory intensive for large data sets because it is an instance-based method. Instance-based methods are also called lazy methods.^[43] This is because the entire training data set is loaded into memory and used for classification or prediction. Furthermore, the runtime for all predictions is $O(n)$, by using time-saving techniques like KD-tree or Ball-Tree, k -NN can be optimized.^[44–45] In the present work, however, less than 100 data points are available, so this problem can be neglected. It is worth mentioning that for very complex classification problems it might be that the k -NN method is surpassed by other “exotic” techniques, such as Support Vector Machines (SVM) or Neural Networks. Another point to consider is the “curse of dimensionality”.

The larger the feature space gets, the less effective the k -NN method is, since Euclidean distance is used, which is the real distance between two points that are connected by a straight line. The Euclidean distance for $x = \{X_1, \dots, X_n\}$ and $y = \{Y_1, \dots, Y_n\}$ is given as:

$$EUD(x, y) = \sqrt{\sum_{j=1}^n |X_j - Y_j|^2}$$

For this reason, a PCA was performed before applying the k -NN algorithm in order to limit the feature space to a few meaningful variables. In the present work no distance weight functions were used.

To test the performance of the model, various parameters suitable for classification were used. These include accuracy (Acc), error rate (Err), sensitivity (TPR) and specificity (TNR). All parameters are based on the true positive (TP), true negative (TN), false positive (FP) and false negative (FN) assignments of the model according to a confusion matrix. Accuracy is the ratio of correctly classified samples to all samples in the data set.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

The counterpart of accuracy is the error rate.

$$Err = 1 - Acc = \frac{FP + FN}{TP + TN + FP + FN}$$

The sensitivity of the model also known as true positive rate (TPR) or hit rate is the ratio of all positive and correctly classified samples to all positive classified samples.

$$TPR = \frac{TP}{TP + FN}$$

Specificity also known as true negative rate (TNR) or inverse recall is expressed as the ratio of correctly classified negative samples to the total number of negative classified samples.

$$TNR = \frac{TN}{FP + TN}$$

Conflict of Interest

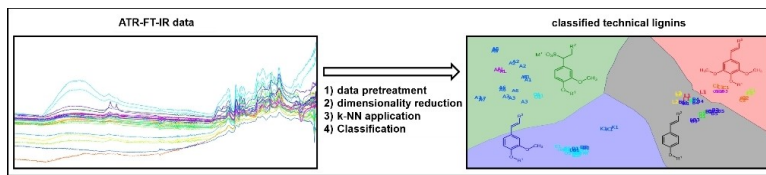
The authors declare no conflict of interest.

Keywords: PCA · k-nearest neighbor · FT-IR · technical lignin · classification

- [1] M. Abhilash, D. Thomas, in *Biopolym. Compos. Electron.*, Elsevier, **2017**, pp. 405–435.
- [2] A. Tribot, G. Amer, M. A. Alio, H. de Baynast, C. Delattre, A. Pons, J.-D. Mathias, J.-M. Callois, C. Vial, P. Michaud, *Eur. Polym. J.* **2019**, *112*, 228–240.
- [3] A. G. Vishtal, A. Kraslawski, *BioResources* **2011**, *6*, 3547–3568.
- [4] A. Tejado, C. Pena, J. Labidi, J. Echeverria, I. Mondragon, *Bioresour. Technol.* **2007**, *98*, 1655–1663.
- [5] T. Aro, P. Fatehi, *ChemSusChem* **2017**, *10*, 1861–1877.
- [6] V. Hornof, R. Hombek, *J. Appl. Polym. Sci.* **1990**, *41*, 2391–2398.
- [7] A. Corey, K. Wamsley, T. Winowski, J. Moritz, *J. Appl. Poult. Res.* **2014**, *23*, 418–428.
- [8] J. D. Megiatto Jr, B. M. Cerrutti, E. Frollini, *Int. J. Biol. Macromol.* **2016**, *82*, 927–932.
- [9] G. Yu, B. Li, H. Wang, C. Liu, X. Mu, *BioResources* **2013**, *8*, 1055–1063.
- [10] A. M. Puziy, O. I. Poddubnaya, O. Sevastyanova, in *Lignin Chemistry*, Springer, **2020**, pp. 95–128.
- [11] R. T. Hilares, L. Ramos, M. A. Ahmed, A. P. Ingle, A. K. Chandel, S. S. da Silva, J. W. Choi, J. C. dos Santos, *Lignocellul. Biorefin. Technol.* **2020**, *247*–263.
- [12] M. A. Hubbe, R. P. Chandra, D. Dogu, S. van Velzen, *BioResources* **2019**, *14*, 7387–7464.
- [13] A.-Q. Duan, K. Feng, G.-I. Wang, J.-X. Liu, Z.-S. Xu, A.-S. Xiong, *Protoplasma* **2019**, *256*, 777–788.
- [14] C. Simon, C. Lion, C. Biot, N. Gierlinger, S. Hawkins, *Annu. Plant Rev.* **2018**, *1*, 1–32.
- [15] M. Y. Balakshin, E. A. Capanema, R. B. Santos, H.-m. Chang, H. Jameel, *Holzforchung* **2016**, *70*, 95–108.
- [16] A. V. Faleva, A. Y. Kozhevnikov, S. A. Pokryshkin, D. I. Falev, S. L. Shestakov, J. A. Popova, *J. Wood Chem. Technol.* **2020**, *40*, 178–189.
- [17] C. Zhao, J. Huang, L. Yang, F. Yue, F. Lu, *Ind. Eng. Chem. Res.* **2019**, *58*, 5707–5714.
- [18] J. Banoub, G. H. Delmas Jr, N. Joly, G. Mackenzie, N. Cachet, B. Benjelloun-Mlayah, M. Delmas, *J. Mass Spectrom.* **2015**, *50*, 5–48.
- [19] A. Alves, M. Schwanninger, H. Pereira, J. Rodrigues, *Holzforchung* **2006**, *60*, 29–31.
- [20] F. S. Poke, C. A. Raymond, *J. Wood Chem. Technol.* **2006**, *26*, 187–199.
- [21] L. M. Fahey, M. K. Nieuwoudt, P. J. Harris, *Cellulose* **2019**, *26*, 7695–7716.
- [22] D. P. Garcia, J. C. Caraschi, G. Ventorim, F. H. A. Vieira, T. de Paula Protásio, *Renewable Energy* **2019**, *139*, 796–805.
- [23] M. Tiitta, V. Tiitta, J. Heikkinen, R. Lappalainen, L. Tomppo, *Sensors* **2020**, *20*, 1076.
- [24] J. Zhang, Y. He, L. Yuan, P. Liu, X. Zhou, Y. Huang, *Agronomy* **2019**, *9*, 496.
- [25] S. Wang, G. Dai, H. Yang, Z. Luo, *Prog. Energy Combust. Sci.* **2017**, *62*, 33–86.
- [26] F. Xu, J. Yu, T. Tesso, F. Dowell, D. Wang, *Appl. Energy* **2013**, *104*, 801–809.
- [27] W. Gottwald, G. Wachter, *IR-Spektroskopie für Anwender*, Wiley-VCH-Verlag, **1997**.
- [28] O. Faix, *Holzforchung* **1991**, *45*, 21–28.
- [29] Z. Shi, G. Xu, J. Deng, M. Dong, V. Murugadoss, C. Liu, Q. Shao, S. Wu, Z. Guo, *Green Chem. Lett. Rev.* **2019**, *12*, 235–243.
- [30] S. Wold, *Technomet* **1978**, *20*, 397–405.
- [31] M. Bergs, G. Völkerling, T. Kraska, R. Pude, X. T. Do, P. Kusch, Y. Monakhova, C. Konow, M. Schulze, *Int. J. Mol. Sci.* **2019**, *20*, 1200.
- [32] F. Bergius, *J. Soc. Chem. Ind.* **1933**, *52*, 1045–1052.
- [33] A. E. Kazzaz, P. Fatehi, *Ind. Crops Prod.* **2020**, *154*, 112732.
- [34] S. Svensson, Degree Project, Mälardalens Högskola, **2008**.
- [35] A. N. Evdokimov, A. V. Kurzin, O. V. Fedorova, P. V. Lukanin, V. G. Kazakov, A. D. Trifonova, *Wood Sci. Technol.* **2018**, *52*, 1165–1174.
- [36] M. C. Iglesias, D. Gomez-Maldonado, B. K. Via, Z. Jiang, M. S. Peresin, *For. Prod. J.* **2020**, *70*, 10–21.
- [37] Sigma-Aldrich.
- [38] C. S. Lancefield, S. Constant, P. de Peinder, P. C. Bruijninx, *ChemSusChem* **2019**, *12*, 1139–1146.
- [39] A. Cotrim, A. Ferraz, A. Gonçalves, F. Silva, R. Bruns, *Bioresour. Technol.* **1999**, *68*, 29–34.
- [40] C. G. Boeriu, D. Bravo, R. J. Gosselink, J. E. van Dam, *Ind. Crops Prod.* **2004**, *20*, 205–218.
- [41] N. Halko, P.-G. Martinsson, J. A. Tropp, *SIAM Rev.* **2011**, *53*, 217–288.
- [42] A. Kataria, M. Singh, *Int. J. Adv. Res. Technol.* **2013**, *3*, 354–360.
- [43] D. Wettschereck, D. W. Aha, T. Mohri, *Artif. Intell. Rev.* **1997**, *11*, 273–314.
- [44] J. L. Bentley, *Commun. ACM* **1975**, *18*, 509–517.
- [45] S. M. Omohundro, ICSI Technical Report TR-89-063, **1989**.

Manuscript received: March 29, 2021

FULL PAPERS



FT-IR spectra of technical lignins of different biomass have been analyzed. Using common data pretreatment methods and subsequent dimensional reduction, a reliable model for the prediction of technical lignins of

unknown biomass based on the k nearest neighbor algorithm was established. For an optimized model with $k=4$ an accuracy of 98.9% could be achieved for the prediction.

*F. Fink, Dr. F. Emmerling, Dr. J. Falckenhausen**

1 – 9

Identification and Classification of Technical Lignins by means of Principle Component Analysis and k-Nearest Neighbor Algorithm

