FEATURES' SELECTION BASED ON WEIGHTED DISTANCE MINIMIZATION, APPLICATION TO BIODEGRADATION PROCESS EVALUATION

A. Rammal[*], *H. Fenniri*[*], *A. Goupil*[*], *B. Chabbert*[†], *I. Bertrand*[†‡], *V. Vrabie*[*]

[*] URCA
CReSTIC
51687 Reims, France

[†] INRA
UMR 614 FARE
51100 Reims, France

[‡] INRA
UMR Eco&Sols
34060 Montpellier, France

ABSTRACT

Infrared spectroscopy can provide useful information of the biomass composition and has been extensively used in several domains such as biology, food science, pharmaceutical, petrochemical, agricultural applications, etc. However, not all spectral information are valuable for biomarkers construction or for applying regression or classification models and by identifying interesting wavenumbers a better processing and interpretation can be achieved. The selection of optimal subsets has been addressed through several variable or feature selection methods including genetic algorithms. Some of them are not adapted on large data, others require additional information such as concentrations or are difficult to tune.

This paper proposes an alternative approach by considering a weighted Euclidean distance. We show on real Mid-infrared spectra that this constrained nonlinear optimizer allows identifying the wavenumbers that best highlights the discrimination within the periods of the biodegradation process of the lignocellulosic biomass. These results are compared with previous ones obtained by a genetic algorithm.

Index Terms— Weighted Euclidean distance, feature selection method, genetic algorithm, infrared spectra, biodegradation process, lignocellulosic biomass.

1. INTRODUCTION

InfraRed (IR) spectroscopy provides useful information of the molecular composition of biological systems and has been widely used in several domains such as biology, food science, pharmaceutical, petrochemical, etc. For agricultural applications, Mid-InfraRed (MIR) spectroscopy, $400\,\mathrm{cm}^{-1}$ to $4000\,\mathrm{cm}^{-1}$, highlights the absorption of fundamental bands of molecular vibrations that are specific of chemical bonds. Being sensitive to both organic constituents (lignocellulose and soil organic matter) and mineral components (soil mineral phase), MIR is considered to be able to provide a performant tool in biomass analysis with relevant qualitative information in prediction models [1].

MIR has a growing interest for development of biomarkers related to intrinsic characteristics of plants and their mode of degradation. Development of rapid and robust MIR biomarkers is a crucial issue that applies to various industrial challenges including biorefineries, biotechnologies and environment, for example emission of greenhouse gases from soils [2, 3]. However, it is commonly assumed that not all spectral information is valuable for biomarkers construction and by identifying interesting wavenumbers, the degradation of the biomass can be better evaluated.

Selection of interesting wavenumbers can be done by variable or feature selection methods such as subset models, stepwise (multiple linear regression) methods, successive projections algorithm, competitive adaptive reweighted sampling, variable importance for projection, uninformative variable elimination, (backward/forward or moving window) interval partial least squares regression, simulated annealing, artificial neural networks-based methods, etc [4, 5, 6]. However, some of these methods require additional information such as concentrations or they do not scale correctly on large data.

The genetic algorithm (GA) is an interesting alternative heuristic optimization technique that has the advantage of exploring the space of all possible wavenumbers subsets fairly well in a large but reasonable amount of time. GA is revealed to be a highly effective method [4] and has been successfully applied to many frequency selection problems including the evaluation of the biodegradation of biomass through MIR and NIR spectra [7]. However, one of the main drawback is that GA requires numerous steps such as selection through a fitness function, crossover and mutation scheme, which can be addressed in different ways, as well as numerous parameters: initial population size, number and size of chromosomes, number of generations through which the process is allowed, etc. The sizes of chromosome and population were selected after evaluating the GA for different sizes and choosing those that give the minimum fitness function's value. The cross-over rate or mutation rate are parameters which should be set up carefully. Genetic algorithms are thus difficult to tune. Besides, the GA makes a huge *a priori* assumption about the shape of

The research for this paper was financially supported by the EMERGENCE SSELVES Grant of the Champagne-Ardenne Regional Council, France.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

the interesting features, that is a few sparse entries of vector.

This paper proposes an alternative approach based on a constrained nonlinear optimizer. Selecting a set of entries of the data vectors in order to better separate the clusters can be seen as modifying the distance between the input vectors. This distance can be constrained to be a weighted Euclidean distance and other constraints such as ℓ_∞ norm or ℓ_1 norm may be considered for the weights.

From an application's point of view, the objective of this study is to investigate the potential of such approach in order to identify the wavenumbers that best highlights the discrimination within the periods of the degradation process of the lignocellulosic biomass. Because biodegradation is a dynamic process, one of the challenges is to capture the changes in the spectra over time to then be able to predict the extent of biomass degradation. We show that the proposed alternative highlights the same principal vibrations of chemical functional groups of compounds that undergo degradation/conversion during the biodegradation of the lignocellulosic biomass, allowing however a better discrimination within the periods of the degradation process than considering the wavenumbers selected by GA.

2. DESCRIPTION OF THE METHOD

Our method is described in this section by, first, setting up the data model and by presenting the classical Davies-Bouldin index. The objectives is then presented with a solution proposed in a previous paper [7]. The new solution is then described.

2.1. Data model and DB criterion

The dataset is given as a collection of M points $X_1, \dots X_M$, which are real vectors of length N. Each data X_i is labeled by a class number k between 1 and K.

The labeling partitions naturally the data into clusters $C_1, \dots, C_K$. The location of each cluster is given by its centroid A_k. The scatter of the data points of the k-th cluster C_k may be evaluated with the mean distance to the centroid,

$$S_k = \frac{1}{|C_k|} \sum_{X \in C_k} d(X, A_k), \tag{1}$$

where $|C_k|$ is the size of the cluster.

The pairwise distance between cluster's centroids $M_{i,j} = d(A_i, A_j)$, provides a measurement of how far the clusters i and j are each other.

Using the previous measurements, the good separation of the clusters may be evaluated thanks to the Davies-Bouldin (DB) index [8] which computes the mean maximum ratio between the scatter of two clusters and their centroid's distances,

$$DB(X, d) = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \frac{S_k + S_{k'}}{M_{k,k'}}. \tag{2}$$

As can be read from its definition, lower Davies-Bouldin indices indicate better separations of the clusters. Other criteria may be used in order to evaluated the separation between clusters such as the Dunn index, the Fisher ratio or other validity indices [9].

2.2. Objective and previous approach

Given the dataset and the labeling, the Davies-Bouldin index indicate how well the clusters are apart. The objectives is to find the specific features of the dataset which gives the best separation of the clusters.

A genetic algorithm was chosen in [7] in order to select these features. In these case, the features are the entries of the data vectors X_i. The algorithm peeks less than a decade entries among the 520 coordinates and build a new dataset using only the sub-vectors that correspond to these entries. The feature selection is done in order to minimize the Davies-Bouldin index. This was imposed by the fact that we want to group the samples for each period of biodegradation and separate them according to the different periods of the biodegradation process.

GA are a type of evolutionary optimization computation based on the concept of natural selection of solutions. Each solution may be considered as a population where each element is represented in the form of a chromosome, with selected data vector entries as genes. The steps of the GA reproduce the various evolutionary operations such as crossover and mutation allowing to select for each generation the best chromosomes and to identify at the end and optimal chromosome with respect to an optimization criterion defined by a fitness function.

This solution gives very good results and the entries selected by the GA have a meaningful chemical interpretation, see [7] and below. However, this solution makes a huge *a priori* assumption about the shape of the interesting features, that is a few sparse entries of vector.

GA is also difficult to tune. Its population size or the crossover rate or mutation rate are parameters which should be set up carefully. The chromosome representation is also sensitive to the application: in the paper [7], either an entry is selected or not.

2.3. Our new approach

Selecting a set of entries of the data vectors in order to compute the Davies-Bouldin index can be seen as modifying the distance $d(\cdot, \cdot)$ between the input vectors. Indeed, let $\mathcal{E}$ be the set of selected entries of data vectors, the distance $d_{\mathcal{E}}(x, y)$ between vectors $x = (x_1, \dots, x_N)^T$ and $y = (y_1, \dots, y_N)^T$ is given by,

$$d_{\mathcal{E}}(x, y)^2 = \sum_{i \in \mathcal{E}} \big(x_i - y_i\big)^2. \tag{3}$$

Therefore, the genetic algorithm [7] seeks the optimum

$$\begin{aligned} &\mathcal{E}_{\text{opt}} = \arg\min_{\mathcal{E}} DB\big(X, d_{\mathcal{E}}(\cdot,\cdot)\big) \\ &\text{such that} \quad |\mathcal{E}| = e, \end{aligned} \tag{4}$$

where e is the number of entries to be selected. It is a parameter set by the operator.

This new point of view developed above about the method of [7] allows a simple generalization of the optimization. Indeed, our new method is given by the problem

$$\begin{aligned} &d_{\text{opt}}(\cdot,\cdot) = \arg\min_{d(\cdot,\cdot)} DB\big(X, d(\cdot,\cdot)\big) \\ &\text{such that} \quad d(\cdot,\cdot) \text{ is a suitable distance.} \end{aligned} \tag{5}$$

Of course, the "suitable" distance is application specific and the space of distance should be constrained enough in order to be meaningful. Otherwise, a distance function which assigns 0 for vector inside the same cluster and 1 for vectors belonging to distinct clusters is optimal for the Davies-Bouldin index while this distance is useless.

In order to be more specific for the application described below, the distance $d(\cdot,\cdot)$ is constrained to be a weighted Euclidean distance,

$$d_w(x,y)^2 = \sum_i w_i \, (x_i - y_i)^2. \tag{6}$$

If the weight w_i are 0/1 valued, the $d_w(\cdot,\cdot)$ distance is equivalent to a $d_{\mathcal{E}}(\cdot,\cdot)$ distance used by the previous method.

Using the distance $d_w(\cdot,\cdot)$, our method relies on the following optimization problem,

$$\begin{aligned} &w_{\text{opt}} = \arg\min_{w} DB\big(X, d_w(\cdot,\cdot)\big) \\ &\text{such that} \quad \begin{cases} 0 \le w_i \quad \text{for } 1 \le i \le N \\ \text{others constraints,} \end{cases} \end{aligned} \tag{7}$$

where the others constraints may be normalization constraints such as ℓ_2 or ℓ_1 norm: $\sum_i w_i^2 = 1$ or $\sum_i |w_i| = 1$ respectively. The weights may also be limited in magnitude, $w_i < 1$, also known as ℓ_∞ norm constraint. The choice of future constraints depends on the application.

Remark that the method may be looked at as a kind of kernel trick. The data points X_i are mapped by ϕ into another metric space, the feature space, and our method tries to find the best map ϕ according to the Davies-Bouldin index. For example, the distance $d_w(x,y)$ is the Euclidean distance between $\phi_w(x)$ and $\phi_w(y)$ where $\phi_w(x) = \big(\sqrt{w_1}\,x_1, \ldots, \sqrt{w_N}\,x_N\big)^T$.

3. APPLICATION

3.1. Mid-infrared spectra of lignocellulosic biomass

Maize (Zea mays L.) roots samples from two inbred parental lines (F2 and F292) and two mutants of these lines (F2bm1 and F292bm3) represent the lignocellulosic biomass that has been analyzed at $K = 5$ periods of biodegradation in soil: 0, 14, 36, 57 and 112 days [10]. Samples were dried at $40\,^\circ\mathrm{C}$ in a ventilated oven for 3 days and ground to $80\,\mu\mathrm{m}$ prior to Diffusion Reflectance Infrared Fourier Transformed (DRIFT) spectroscopy using an IRTF Nicolet 6700 Thermo electron spectrometer. All spectra (64 accumulations for each sample) were acquired with a spectral resolution of $4\,\mathrm{cm}^{-1}$. The MIR wavenumber region was chosen to be $800\,\mathrm{cm}^{-1}$ to $1800\,\mathrm{cm}^{-1}$, which corresponds to the principal vibrations of chemical functional groups associated to the lignocellulosic components [11]. The dataset is made up by 4×5 spectra.

All spectra were preprocessed with a first-order Savitzky-Golay filter with a fourth order polynomial and a smoothing of 17 points, followed by a Standard Normal Variate preprocessing, which have been used in the previous investigation [7].

Visualization of the clusters of this data set can be done thanks to the multidimensional scaling (MDS) [12] as shown in Figure 1(a). This representation can be linked with a scatter plot obtained by a principal component analysis. Samples at each period of the biodegradation process are represented with the same symbol and color.

3.2. Previous results

On this data set, the method based on the genetic algorithm allowed to identify the following wavenumbers [7]:

- $860\,\mathrm{cm}^{-1}$: aromatic skeletal vibrations combined with CH wag,
- $953\,\mathrm{cm}^{-1}$: C-O-C stretching of the polysaccharides,
- $1385\,\mathrm{cm}^{-1}$: cellulose with lignin (Aliphatic CH stretching in CH3),
- $1709\,\mathrm{cm}^{-1}$: hemicellulose (C=O stretching unconjugated ketones, carbonyls and in ester.

These wavenumbers correspond to principal vibrations of chemical functional groups of compounds that undergo degradation/conversion during the biodegradation of the lignocellulosic biomass. The discrimination of the samples according to the periods of the biodegradation process is highly improved, as emphasized by the MDS shown in Figure 1(b) as compared with the result obtained using the entire MIR wavenumber region (see Figure 1(a)). This is obvious when examining the DB index values. As was pointed out [7], not all spectral information is valuable; by identifying interesting wavenumbers, we can better describe the effect of the degradation time on biomass characteristics.

Beside the choice of adapted steps, the genetic algorithm required numerous parameters. The maximum number of generations, the fraction of crossover, the number of elites, and the stop parameters were empirically chosen, see [7]. The sizes of chromosome and population were selected after evaluating the GA for different sizes and choosing those that gave the minimum fitness function's value, as it usually done with GA.

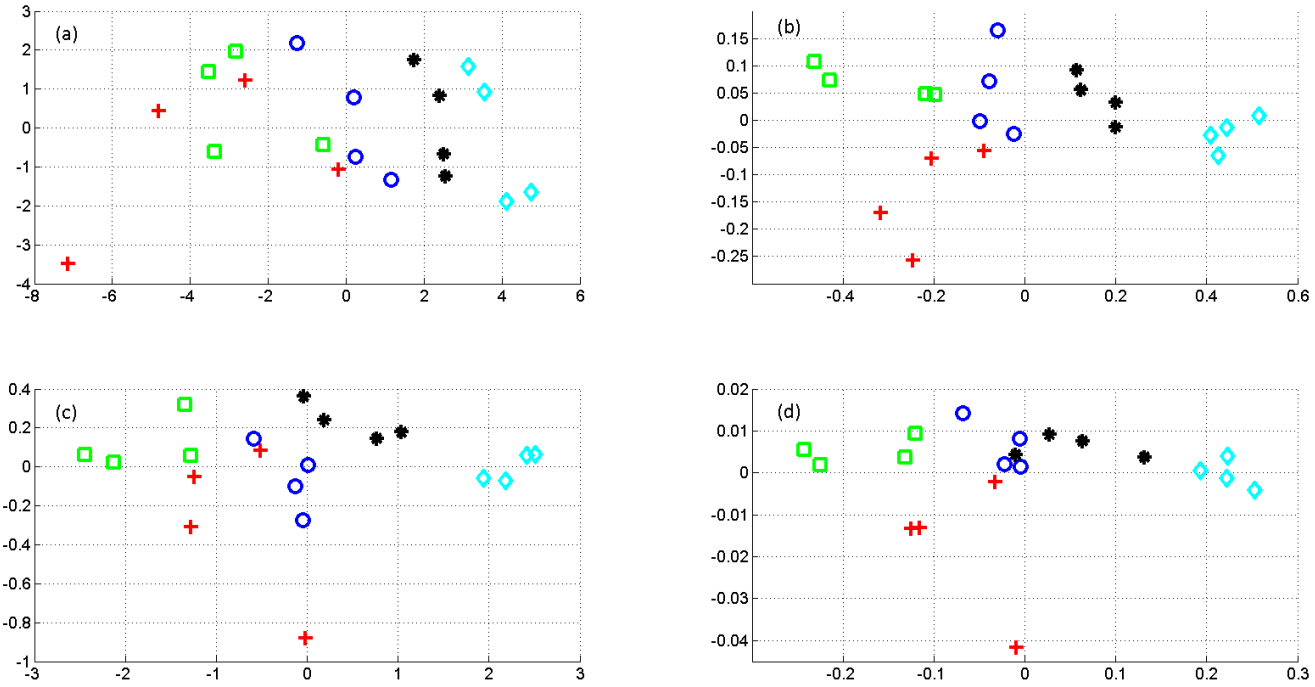

Fig. 1. Discrimination of samples according to the periods of the biodegradation process (+: t = 0; □: t = 14; o: t = 36; ∗: t = 57; and ◇: t = 112 days). MDS obtained on: (a) entire MIR wavenumber region, DB index= 1.5654; (b) wavenumbers identified by the GA on the same dataset [7], DB index= 0.6037; (c) wavenumbers identified by the proposed approach such that $0 \leq w_i$, DB index= 0.5861; (d) wavenumbers identified by the proposed approach such that $0 \leq w_i$ and $\sum_i |w_i| = 1$, DB index= 0.5578

3.3. Results with the proposed approach

The approach presented in Eq. (7) was applied on the same data set in a first instance without any normalization constraint. Optimization are performed using a toolbox implementing the sequential quadratic programming method (SQP). Contrary to the GA, which makes *a priori* assumption about the shape of the interesting features, *i.e.* few sparse entries of vector, this approach does not impose any particular shape on the weights w_i.

Figure 2(c), shows the distribution of the resulting weights w_i. Comparing with results obtained by the GA (shown in Figure 2(b)) we find out that the same spectral information are highlighted, with some slight differences. Firstly, the information at $1385\,\mathrm{cm}^{-1}$ (cellulose with lignin) is the most predominant, both in amplitude and width, as the neighboring bins at $1383\,\mathrm{cm}^{-1}$ is also identified. This is an interesting result since it is known that the cellulose changes with the biodegradation time. Secondly, a small "bump" around $912\,\mathrm{cm}^{-1}$: cellulose, hemicellulose, lignin (Anomere C-groups, C-H deformation with ring valence vibration) appears. This is also another interesting result since we put into evidence another principal vibrations of chemical functional groups of compounds that undergo degradation/conversion during the biodegradation of the lignocellulosic biomass.

With this new information, the discrimination of the samples (see figure Figure 1(c)) according to the periods of the biodegradation process is further enhanced, the DB index value decreasing from 0.6037 (with the GA) to 0.5861. However, without taking the DB index into consideration, a better discrimination can be found for GA since the 5 classes are not overlapped. Numerically, both algorithms try to minimize the DB index, which represents the mean maximum ratio between the scatter of the clusters and their centroid's distances. For the proposed approach, the DB index is lower since samples are better gathered at quite all periods of the biodegradation process, especially at t=14, 36 and 112 days (green, blue, and cyan). This is due to the fact that the approach allows to estimate relative weights associated to the wavenumbers. Besides, it is not easy to identify the predominant functional group that appeared around $912\,\mathrm{cm}^{-1}$.

For this reason, we have applied the proposed approach considering a complementary ℓ_1 normalization constraint. Result shown in Figure 1(d) indicates that the discrimination of the samples according to the periods of the biodegradation process has been slightly improved. Samples within a class are more grouped, the DB index value lowering to 0.5578. The small "bump" identified previously was transformed into a "pick" at $912\,\mathrm{cm}^{-1}$ (see Figure 2(d)) which, although small, highlights the cellulose, hemicellulose, and lignin. As previously, the information at $1385\,\mathrm{cm}^{-1}$ (cellulose with lignin)

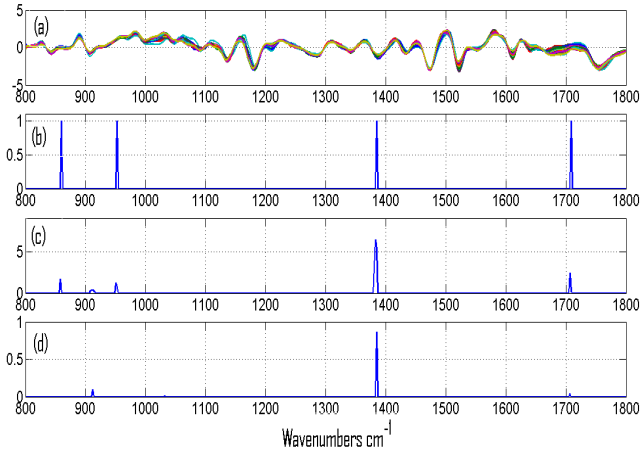

Fig. 2. (a): the 20 preprocessed spectra recorded on the 4 samples at the $K = 5$ periods of biodegradation in soil. (b): positions of wavenumbers identified by the GA (previous result). (c): weights w_i identified by the proposed approach with $0 \leq w_i$. (d) weights w_i identified by the proposed approach with $0 \leq w_i$ and $\sum_i |w_i| = 1$

is the most predominant, however is constrained here due to the ℓ_1 norm to only one bin. However, without taking the DB index into consideration, a better discrimination can always be found for GA. This can be explained by the fact that the kinetic of the degradation is not taken into consideration by the criterion itself. While the proposed approach allows to estimate relative weights, the DB criterion that both algorithms try to minimize does not take into account the mineralization process, which presents a maximum around $t = [14, 21]$ days and similar values (depending on the specimen) at t=0 and t around 40 days, which can a posteriori justify that the overlapped classes in Figure 1(c) might be a positive result, but further works should be carried out.

4. CONCLUSION

We consider the problem of finding an appropriate alternative distance metric that takes into consideration the distribution of our real mid-infrared spectra.

We presented a study experiment in order to examine the impact of different term weighting schemes on the clustering quality. The underlined values of Davies-Bouldin index indicate that a complementary ℓ_1 normalization constraint achieves a good numerical performance in those experiments in comparison with the results provided by the genetic algorithm.

The new method strengthens the choice of a set of wavenumbers as the best features to be selected. Moreover, new information is added by this method compared to the GA approach: the relative weights associated with the wavenumbers permits to sort their relative impacts.

REFERENCES

[1] J. Lupoi, S. Singh, B. Simmons, and R. Henry, "Assessment of lignocellulosic biomass using analytical spectroscopy: an evolution to high-throughput techniques," *BioEnergy Research*, vol. 7, pp. 1–23, 2014.

[2] J. Reeves, "Potential of near and mid-infrared spectroscopy in biofuel production," *Commun. Soil Sci. Plan.*, vol. 43, pp. 478–495, 2012.

[3] V. Bellon-Maurel and A. McBratney, "Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils — critical review and research perspectives," *Soil Biol. Biochem.*, vol. 43, pp. 1398–1410, 2011.

[4] R. M. Balabin and S. V. Smirnov, "Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data," *Anal. Chim. Acta*, vol. 692, pp. 63–72, 2011.

[5] T. Mehmood, K. Liland, L. Snipen, and S. Sæbø, "A review of variable selection methods in partial least squares regression," *Chemom. Intell. Lab. Syst.*, vol. 118, pp. 62–69, 2012.

[6] M. Vohland, M. Ludwig, S. Thiele-Bruhn, and B. Ludwig, "Determination of soil properties with visible to near- and mid-infrared spectroscopy: Effects of spectral variable selection," *Geoderma*, vol. 223–225, pp. 88–96, 2014.

[7] A. Rammal, E. Perrin, V. Vrabie, B. Chabbert, I. Bertrand, and B. Lecart, "Using a genetic algorithm as an optimal band selector in the mid-near infrared: Evaluation of the biodegradation of maize roots," *J. Appl. Sci. & Agric.*, vol. 9, pp. 392–388, 2014.

[8] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 1, no. 2, pp. 224–227, Apr. 1979.

[9] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Prez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recogn.*, vol. 46, pp. 243–256, 2013.

[10] G. E. Machinet, I. Bertrand, B. Chabbert, and S. Recous, "Decomposition in soil and chemical changes of maize roots with genetic variations affecting cell wall quality," *Eur. J. Soil Sci.*, vol. 60, no. 2, pp. 176–185, 2009.

[11] H. Chen, C. Ferrari, M. Angiuli, J. Yao, C. Raspi, and E. Bramanti, "Qualitative and quantitative analysis of wood samples by Fourier transform infrared spectroscopy and multivariate analysis," *Carbohydr. Polym.*, vol. 82, pp. 772–778, 2010.

[12] I. Borg, P. Groenen, *Modern Multidimensional Scaling: theory and applications*, Springer-Verlag (New York, 2nd ed.), 2005.