Journal of Applied Science and Agriculture, 9(11) Special 2014, Pages: 382-388

AENSI Journals

Journal of Applied Science and Agriculture

ISSN 1816-9112

Journal home page: www.aensiweb.com/JASA

Using a Genetic Algorithm as an Optimal Band Selector in the Mid-Near Infrared: Evaluation of the Biodegradation of Maize Roots

[1]Abbas Rammal, [1]Eric Perrin, [1]Valeriu Vrabie, [2,3]Brigitte Chabbert, [2,3]Isabelle Bertrand, [2,3]Brieuc Lecart

[1]*Université de Reims Champagne-Ardenne, CReSTIC-Châlons EA 3804, Chaussée du Port, 51000 Châlons-en-Champagne, France*
[2]*INRA, UMR 614 Fractionnement des AgroRessources et Environnement, 51100 Reims, France*
[3]*Université de Reims Champagne Ardenne, UMR 614 Fractionnement des AgroRessources et Environnement, 51100 Reims, France.*

ARTICLE INFO

Article history:
Received 25 June 2014
Received in revised form
8 July 2014
Accepted 10 August 2014
Available online 30 August 2014

Keywords:
Genetic algorithm,
Davies Bouldin index,
Calinski-Harabasz index,
Xie Beni index,
Separation index,
Silhouette statistic index,
Fisher index,
Discriminant spectral bands,
Biodegradation process,
Lignocellulose.

ABSTRACT

Mid- and near- infrared spectroscopies can provide useful information of the biomass composition. These methods have been extensively used in several applications such as biorefineries, biotechnologies, environment, especially for predicting plant composition or classify plant samples. However, the numbers of molecular descriptors in mid- and near- infrared spectra are important and sometimes redundant. Generally, only few spectral bands (wavenumbers) are relevant for applying regression or classification models. The selection of optimal subsets has been addressed through several methods including genetic algorithms. These algorithms require an adapted fitness function in order to identify the most well suited spectral bands. This study intends to analyze the influence of several fitness functions both from a quantitative point of view (which subset better describes the biodegradation process) and from a qualitatively point of view (which subset highlights known chemical functional groups). Results obtained on spectra recorded on maize roots samples at several periods of the biodegradation process show that the Davies-Bouldin fitness function, which is a measure of the separability between clusters and within cluster scatter, gives the best results.

To Cite This Article: Abbas Rammal, Eric Perrin, Valeriu Vrabie, Brigitte Chabbert, Isabelle Bertrand, Brieuc 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.,*** *9(11): 382-388, 2014*

INTRODUCTION

Mid-infrared (MIR) and Near-infrared (NIR) spectroscopies are non-destructive methods that can provide rapid and simple analysis for the determination of biomass composition and its chemical evolution during biodegradation in natural or industrial processes. The MIR highlights fundamental molecular vibrations specific of chemical bonds, being sensitive to organic (lignocellulose and soil organic matter) and mineral (soil mineral phase) components. The NIR reflects the overtones and combinations of fundamental vibrations and bonds, being generally employed for quantitative and qualitative investigation of biomass composition (Xu, F., 2013).

Most of studies involving plant biomasses analysis by using MIR or NIR spectroscopies aim at predicting plant composition (as for example using a Partial Least Squares (PLS) analysis (Chadwick, D., 2014)) or at classifying plant samples to facilitate process (as for example by determining chemical variations across botanical fractions (Ye, X., 2008) or analyzing the stages of a biodegradation process (Rammal, A., 2014). However, due to the high dimensionality of recorded IR spectra, challenging problems such as redundancy or intensive computation need to be addressed. Indeed, the use of the entire MIR or NIR spectral regions implies long processing times and maybe unadapted data processing techniques, since all recorded spectral bands do not corresponds to independent molecular descriptors. Moreover, it has been proved that selecting spectral bands (wavenumbers) improves the prediction of calibration models (Bangalore, A.S, 1996). Furthermore, the biodegradation process is dynamic and spectra acquisition at several times is necessary to follow it, thus generating a high number of bands. Therefore, in such process, identifying the spectral regions which are changing with time is a challenging task.

The selection of optimal subsets of wavenumbers from the MIR or NIR spectral band can be addressed through several methods (Skurichina, M., 2006) such as: the *all subset models* method, which identifies the best optimal subset by testing all the possible combinations, but cannot be practically applied on large data such as IR spectra; the *sequential search* method (Miller, A., 2002) that is a meta-heuristic method not adapted for IR spectra; *StepWise* regression methods (Hastie, T., 2009), which are based on F-tests or on the minimization of

Corresponding Author: Abbas Rammal, CReSTIC-Châlons, Université de Reims Champagne-Ardenne, Chaussée du Port 51000 Châlons-en-Champagne, France
E-mail: abbas.rammal@etudiant.univ-reims.

some score functions, however requiring additional information; genetic algorithms, etc. The latest ones have been successfully used for several applications, such as dimension reduction in the analysis of hyperspectral data, and more recently on agricultural related applications. For example, they have been used to discriminate vegetation species by selecting individual wavelengths and the number of latent variables employed in a PLSR calibration model (Ullah, S., 2012).

A genetic algorithm is a bottom-up approach that selects an optimal spectral subset according to the concept of the "survival of the fittest", each chromosome competing with the others. That is, chromosomes are evaluated for their quality according to a predefined fitness function. However, each fitness function will identify a specific subset that might be more or less representative of the biodegradation process. Besides, a particular fitness function might leads to one or more genes (wavenumbers) in the survival chromosome that not correspond to a known principal chemical functional group.

This study analyzes the influence of classical fitness functions based on Davies Bouldin, Calinski-Harabasz, Xie Beni, Separation, Silhouette statistic, and Fisher indexes. For that, we use the Principal Component Analysis score plots as an indicator of the separability of the analyzed samples according to the periods of the biodegradation process. The aim of this analysis is to identify which fitness function is better adapted while attempts to classify samples related to biodegradation of maize root in soil. From an application point of view, we analyze at five different dates, maize root residue of different genotypes following soil biodegradation (Machinet, G., 2009) Results show that the Davies-Bouldin fitness function allows selecting wavenumbers that correspond to known principal chemical functional groups of lignocellulosic biomass and wavenumbers that best highlight the five periods of the biodegradation process.

2. Methodology:

Let X be a matrix of dimension $R^{r \times J}$, where each column corresponds to an infrared spectra $\underline{x}_j^t(\underline{y}) = [x_{j1}, \dots x_{jr}]^T \in R^r$ recorded over the wavenumbers $\underline{y} = [y_1, \dots y_r]^T \in R^r$. We suppose that J spectra have been recorded on several types of samples at K periods in their biodegradation process t = {t_1, t_2….. t_K}.

The idea is to find with a genetic algorithm (tested with different fitness functions) a sub-vector of wavenumbers $\underline{z}_o$ (extracted from $\underline{y}$) where the amplitudes of the spectra associated to these wavenumbers, i.e. the sub matrix X ($\underline{z}_o$) gives the best classification according to the K periods of the biodegradation process.

2.1. Wavenumber selection by Genetic Algorithm (GA):

Genetic algorithms are a type of evolutionary optimization computation that became popular through the work of Holland (Holland, J., 1975). These algorithms are based on the concept of natural selection of solutions by copying its main principles. Each solution may be considered as a population where each element is represented in the form of a chromosome with selected wavenumbers positioned as genes. The GA steps for each generation reproduce the various evolutionary operations such as crossover and mutation. The different steps of the algorithm are not presented here, but can be found in different publications such as (Vishwakarma, P., 2012; Mitchell, M., 1995; Yang, M., 2014).

The initialized parameters are: the chromosome size L (the number of genes corresponding to the wavenumbers to be selected); the population size N (the number of chromosomes per generation); the number of elites N_e representing the chromosomes with the best fitness values in the current generation that are guaranteed to survive to the next generation; the fraction F_c that limits the number of chromosomes selected to perform crossover and mutation. The stop parameters are: the maximal number of iterations, T, and the tolerance for the chosen fitness function, ε.

Chromosomes are randomly generated to form an initial population: $P(0) = \{\underline{z}_i = [z_{i1} \dots z_{il} \dots z_{iL}]^T \in R^L\}_{i=1}^N$, where each gene z_{il} is a wavenumber randomly chosen in the vector $\underline{y}$.

Then each chromosome $\underline{z}_i$ may be evaluated by a fitness function F, which assigns a value F_i:

$$F_i = F(\underline{z}_i) \ \forall i = 1 \dots N.$$

Fitness function is arguably the most important part of a GA. The role of a fitness function is to measure the quality of the chromosome in the population according to the given optimization objective. Since we want to classify spectra within the K periods of the biodegradation process, we propose to test six well-known and widely used validity indexes (Yang, M., 2014; Xu, R., 2012) as fitness function to evaluate the qualities of the obtained clusters. All of these have an important feature as they differently quantify the separability of classes:

- Xie and Beni (XB) index aims to quantify the ratio of the total variation within clusters and the separation of clusters (Maulik, U., S. Bandyopadhyay, 2003),
- Separation Index (SI) measures the ratio of the sum of compactness and separation of the clusters (Yang, M., 2014),

Journal of Applied Science and Agriculture, 9(11) Special 2014, Pages: 382-388

- Calinski and Harabasz (CH) index measures the dispersion ratio of the spectra within a cluster and between the clusters respectively (Xu, R., 2012),
- Fisher Index (FI) represents the ratio of within-class scatter to between class scatter (Chiang, L., 2004),
- Silhouette (SIL) value measures the similarity of each spectrum to the other spectra in its own cluster compared to spectra in other clusters (Xu, R., 2012),
- Davies-Bouldin (DB) index calculates the ratio of the sum of dispersion within a class to the separation between the classes:

$$DB\left(\underline{z}_i\right) = \frac{1}{K}\sum_{k=1}^{K} max_{k',k\neq k'}\left\{\frac{S_k(\underline{z}_i)+S_{k'}(\underline{z}_i)}{d_{kk'}}\right\},$$

where $d_{kk'}$ is the Euclidian distance between the centers of classes k and k' and the dispersion $S_k\left(\underline{z}_i\right)$ within the class t_k is given by:

$$S_k(\underline{z}_i) = \sqrt{\frac{1}{card\,(t_k)}\sum_{j=1,-j} d\left(\underline{x}_j^{t=t_k}(\underline{z}_i) - m_k(\underline{z}_i)\right)},$$

with $card\,(t_k)$ the number of spectra in the kth class and $d(.)$ the Euclidian distance between the spectrum $\underline{x}_j^{t=t_k}(\underline{z}_i)$ and the center $m_k(\underline{z}_i)$ of that same kth class (Sushmita, M., 2004).

For each fitness function, the values F_i are ordered and the number of the surviving chromosomes, i.e. the number of elites, represents the best N_e chromosomes.

The selection of chromosomes for crossing may be implemented in different ways. We have chosen here the stochastic universal sampling selection (Ranjini, A, B. Zoraida, 2013), the ordered F_i values allows selecting $N_p = (F_c + 1)N - 2N_e$ chromosomes for crossover. For the crossover step, we have used the uniform crossover method since it gives good results in a majority of the cases (Picek, S., M. Golub, 2010). After crossing, one gene is randomly selected either from the first or from the second parent. The crossover operation gives $N_c = F_c$N $-N_e$ children. To maintain genetic diversity from one generation of a population to the next one, we have chosen the Gaussian mutation operator, since it produces the best results for most of the fitness functions (Hinterding, R., 1995). This operator adds a unit Gaussian distributed random value to $N_p - 2\,N_c$ chromosomes.

These different steps are repeated until the maximal number of iterations T is reached or when GA has converged, i.e. the average relative change in the fitness function value is less than the tolerance. This procedure gives an optimal chromosome $\underline{z}_o = [z_{o1} \ldots z_{ol} \ldots z_{oL}]^T \in R^L$ allowing to extract a new sub matrix X($\underline{z}_o$) of under-dimensioned spectra on which we can apply methods of data analysis.

2.2. Analysis and quantification of the biodegradation process

To analyze the influence of these fitness functions, the Principal Component Analysis (PCA) score plots can be used as an indicator of the separability of the analyzed samples according to the biodegradation process. The PCA is a classical data analysis method that has been widely used in many applications including environmental and agricultural context (Dong, Y.-W., 2011). PCA provides two types of information: the matrix of loadings L, that indicates the strength level of correlation between components and the wave numbers and the matrix of scores S that indicates the location of the sample along a component. For agricultural applications, the S scores plots are used to evaluate the separability of the analyzed samples.

In our application, the separability of the entire spectral information $X\,(\underline{y})$ recorded on the J analyzed samples, or of sub matrices X($\underline{z}_o$) selected by the GA with one of the fitness functions above, into the K periods of their biodegradation process {t_1, t_2….. t_K}, can be assessed qualitatively by analyzing the PCA scores plots.

To quantify the separability of the samples within the biodegradation process and thus numerically assess the representation of the PCA, the Dunn Index (DI) was used to estimate the ratio between the minimal intracluster distance to maximal intercluster distance (Xu, R, 2012) in the PCA score plot plan S:

$$DI = \frac{d_{min}}{d_{max}}$$

where d_{min} denote the smallest distance between two spectra from different clusters and d_{max} the largest distance of two spectra from the same cluster.

The DI is a positive number and a higher value indicates a better separation of the clusters.

3. Analysis of the biodegradation of maize roots

The goal of our experimental study is to analyze the influence of function fitness by comparing the performance of the six fitness functions in mid- and near-infrared spectra from a quantitative point of view.

3.1. Samples

Maize roots from two distinct parental lines (F2 and F292) and two mutants of these lines (F2bm1 and F292bm3) represent the lignocellulosic biomass that we analyzed at K=5 biodegradation periods: $t_1 = 0$, $t_2 = 14$, $t_3 = 36$, $t_4 = 57$ and $t_5 = 112$ days (Machinet, G., 2009). These residues were incubated in an agricultural soil and at days 14, 36, 57 and 112 after the beginning of incubation, maize roots residues were separated from the soil and dried pending analyses and IR spectra acquisition.

3.2. Spectra

Spectra were acquired on all samples using MIR and NIR spectroscopies. For MIR spectra, we have chosen the 800 - 1800 cm^{-1} spectral range, which corresponds to principal vibrations of chemical functional groups of compounds that are useful in the lignocellulosic biomass analyses. For the NIR spectra, we have chosen the 4000 - 6000 cm^{-1} spectral range that represents the noise-free spectral range (Maurel, V., 2011).

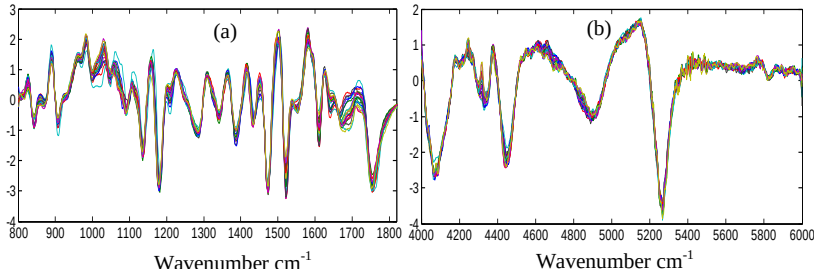

Fig. 1: Pre-processed spectra recorded at 5 biodegradation periods on 4 samples of maize roots: (a) Mid-Infrared spectra in the 800 - 1800 cm^{-1} spectral region; (b) Near-Infrared in the 4000 - 6000 cm^{-1} spectral region.

All spectra were pre-processed with a *first-order* Savitzky-Golay (SG) *filter* with a 4th order polynomial and a smoothing of 17 points, followed by a Standard Normal Variate (SNV) preprocessing. Figure 1 shows the pre-processed MIR and NIR spectra. This pre-processing was found as being efficient for discriminating maize roots samples by classifying MIR and NIR spectra (Rammal, A., 2014).

3.3. Optimization of the parameters of Genetic Algorithm (GA)

The initial parameters have been fixed as follows: the maximum number of generations T= 500, the fraction of crossover $F_c = 0.8$, the elites number $N_e = 2$, the tolerance $\varepsilon = 10^{-6}$. These values have been used for several implementation of GA since they give good results (Rammal, A., 2014; Ullah, S., 2012).

Then, we have evaluated the GA for different sizes of chromosomes, L = 3, 4, 5 …8, and different values of the population size, N = 50, 100,…, 500 for both MIR and NIR spectra. The values that give minima fitness function are chosen as the optimal ones. Figure 2 shows the case of the Davies Bouldin fitness values, N = 300, L = 4 give the smallest fitness value for the MIR spectra and N = 400, L = 4 for the NIR spectra.

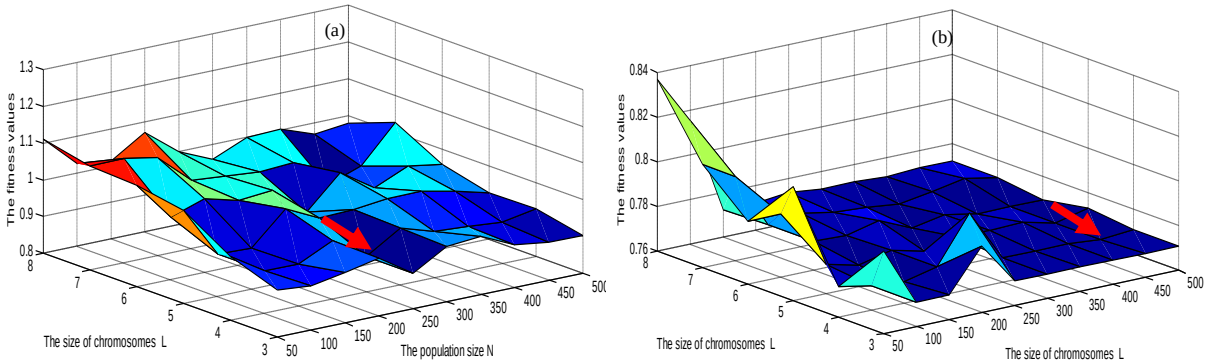

Fig. 2: Values of the Davies Bouldin fitness functions for different sizes of chromosomes, L, and population sizes, N, for: (a) MIR spectra and (b) NIR spectra. The optimal value is indicated by the red arrows.

3.4. Performances of different fitness functions

Figure 3 shows the score plots S of the PCA applied on MIR spectra recorded over the entire 800 - 1800 cm^{-1} spectral range (a) and on spectral MIR amplitudes selected by the GA with the following fitness functions: Davis Bouldin (b), Xie Beni (c), Calinski-Harabasz (d), Silhouette index (e), Separation index (f), and Fisher function (g). The high relative singular values indicated for the principal components show that the first two principal components efficiently describe each considered dataset. These figures indicate that selecting wavenumbers generally provides better discrimination within the periods of the degradation process

that using the entire MIR spectral band. Indeed, if we compare for example Figs. 3(a) and 3(b), K=5 classes corresponding to the {t_1, t_2….. t_K} biodegradation periods are more easily distinguished in the second figure. Identical results, which are not shown here, have been obtained on NIR spectra. Figure 3 also shows that the GA with the Davies Bouldin fitness function provides the best qualitative results, all 5 biodegradation periods being well separated as compared with the others fitness functions. The Silhouette index gives also a good result, however one sample from day 0 can be confused with one from 14 days.

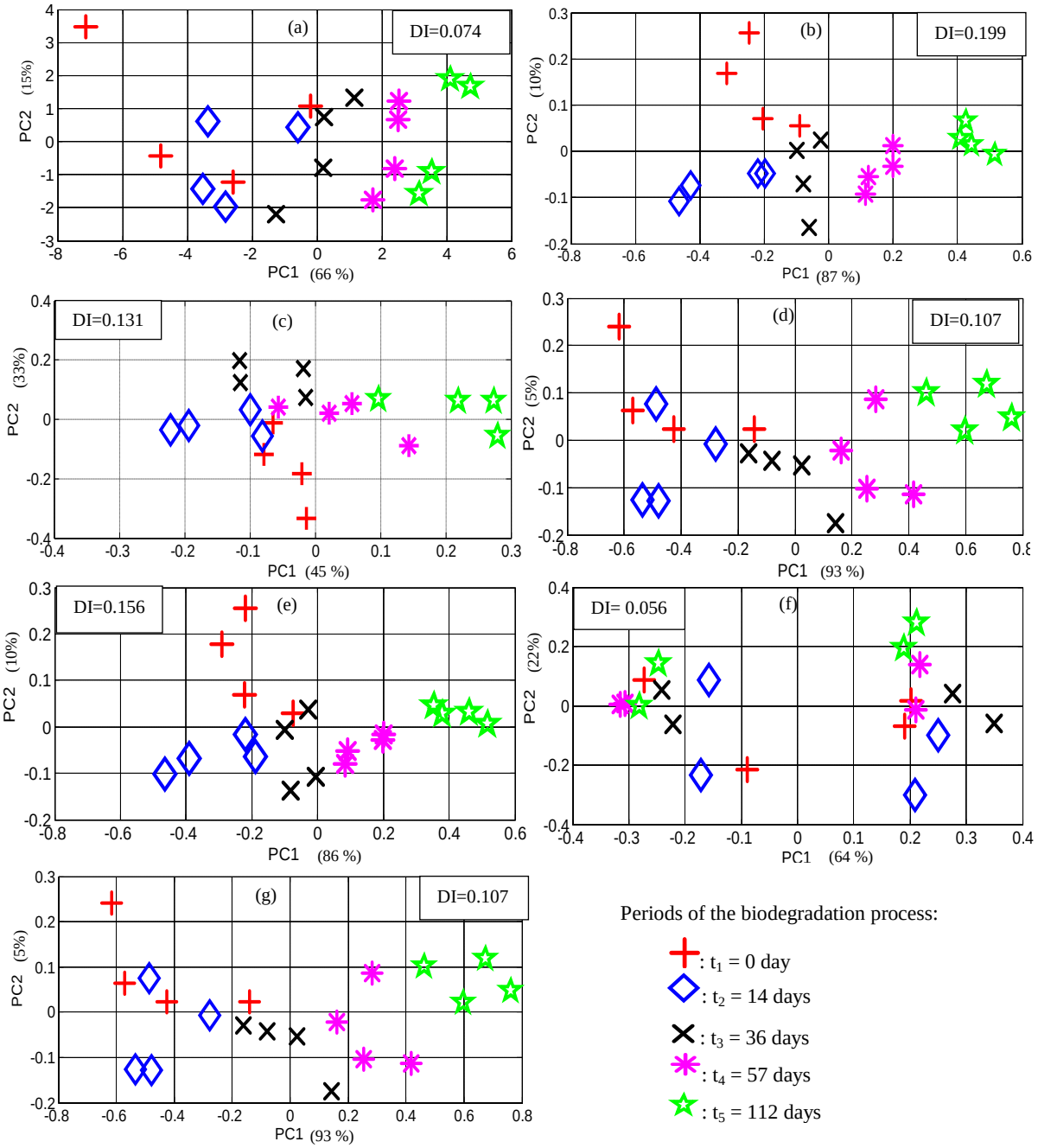

Fig. 3: Scores plots showing the discrimination according to the periods {t_1, t_2….. t_K} of the biodegradation process in terms of PC1 vs PC2. The PCA was applied on: (a) information recorded over the entire 800 - 1800 cm^{-1} MIR range, (b-g) MIR information selected at wavenumbers identified by the GA with the following fitness functions: (b) Davis Bouldin (DB), (c) Xie Beni (XB), (d) Calinski-Harabasz (CH), (e) Silhouette (SIL), (f) Separation (SI), and (g) Fisher (FI). The values of the Dunn Index are also reported.

To quantify the separability of samples according to the periods of the biodegradation process, we have computed the Dunn Index (DI) index for each PCA score plot shown in Fig. 3 as well as on the PCA score plots obtained on NIR spectra. These values, presented in Table 1, confirm that the Davies Bouldin (DB) function fitness best highlights the biodegradation process on both MIR and NIR spectra.

Table 1: Values of the Dunn Index (DI) computed on the score plots of PCA applied on spectral information

	on the entire spectral range	at wavenumbers selected by the genetic algorithm with the fitness functions:					
		Davis Bouldin (DB)	Xie Beni (XB)	Calinski-Harabasz (CH)	Silhouette (SIL)	Separation (SI)	Fisher (FI)
MIR	0.074	**0.1991**	0.131	0.107	0.156	0.056	0.107
NIR	0.039	**0.112**	0.055	0.063	0.101	0.064	0.063

3.5. Analysis of the wavenumbers selected by GA

For MIR spectra, we find that all wavenumbers selected by GA based on Davies Bouldin (DB) function corresponds to principal vibrations of chemical functional groups of compounds that are related to the lignocellulosic biomass: 860 cm^{-1} = aromatic skeletal vibrations combined with CH wag, 953 cm^{-1} = C-O-C stretching of the polysaccharides, 1385 cm^{-1} = cellulose with lignin (Aliphatic CH stretching in CH3), and 1709 cm^{-1} = hemicellulose (C=O stretching unconjugated ketones, carbonyls and in ester.

For the Calinski-Harabasz (CH) and Fisher (FI) functions, we observe that the 942, 1382, and 1706 cm^{-1} bands have the same chemical signification as the 953, 1385, and 1709 cm^{-1} bands selected by the Davies Bouldin function (DB), but one selected band, 1529 cm^{-1}, has no chemical signification.

For the Xie Beni (XB) function, the 1365 and 1380 cm^{-1} bands have the same chemical signification as the 1385 cm^{-1} band. The 1050 cm^{-1} concerns the cellulose (C–O–C vibration), same chemical bond as the 953 cm^{-1} band, but the 1546 cm^{-1} band corresponding to lignin (the aromatic skeletal vibration) have a different chemical signification that the 1709 cm^{-1} band chosen by the DB who gives the better class separability.

The separation index (SI) function selected the 1098 and 1321 cm^{-1} bands that have no chemical signification, and the 825, 845 and 896 cm^{-1} bands have the same chemical signification as the 860 cm^{-1} band (also chosen by DB).

The Silhouette Index (SIL) function extracts bands that have the same chemical signification, but are less relevant than the bands selected by the Davies Bouldin function (DB).

For NIR spectra, wavenumbers selected by GA based on Davies Bouldin (DB) function have interesting signification: 4850 cm^{-1} = CO with CH3 and OH stretching + O-H deformation, 5195 cm^{-1} = water (O-H antisymmetric, stretching vibration + O-H deformation vibration of H2O), 5540 cm^{-1} = cellulose with lignin (CH functional group: 1st overtane CH3 and -CH = CH-), 5705 cm^{-1} = first overtone of the CH stretching.

The Calinski-Harabasz (CH) and Fisher functions select the 4308 and 4651 cm^{-1} bands that are not identical but they have almost the same chemical signification as the 4850 cm^{-1} band. The 4727 cm^{-1} band has not a chemical signification, while the 4850 cm^{-1} band was also found by the DB function.

The Xie Beni (XB) function selects the 4518 cm^{-1} band that has no chemical significations. The 4268, 4277, 4279 and 4804 cm^{-1} bands represent the cellulose (CH stretching + CH deformation, respectively OH stretching + CH deformation), but the classification is worse than by the DB.

The separation index (SI) function finds the 4092 cm^{-1} band that has no chemical signification and the 4669 and 4696 cm^{-1} bands have the same chemical significations as the 4850 cm^{-1} band chosen by DB.

The Silhouette Index (SIL) function, choose bands that have the same chemical signification, but are less relevant than the bands selected by the Davies Bouldin function (DB).

Overall, the Davies Bouldin (DB) fitness function allows identifying principal vibrations of chemical functional groups of compounds that are related to the lignocellulosic biomass. The others fitness functions partially identify the same bands or other spectral bands, but with same chemical signification, or spectral band without chemical signification.

4. Conclusion

We have shown that the implementation of a genetic algorithm with adapted steps leads to some first interesting results in the evaluation of the biodegradation process of maize roots. The Principal Component Analysis applied on wavenumbers selected by the genetic algorithm shows a better discrimination according to the periods of the biodegradation process of maize roots than the classical application over the mid-infrared or near-infrared entire spectral ranges. The Davies-Bouldin as fitness function gives the best results as compared with fitness function based on well-known validity indexes which are the: Calinski-Harabasz, Xie Beni, Separation, Silhouette, and Fisher indexes. Applied on spectra collected at different stages of degradation, the genetic algorithm with a Davies-Bouldin fitness function identifies interesting wavenumbers corresponding to principal chemical functional groups related to the lignocellulosic biomass in both mid- and near- infrared spectral ranges.

ACKNOWLEDGEMENT

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

REFERENCES

Xu, F., J. Yu, T. Tesso, F. Dowell, D. Wang, 2013. "Qualitative and quantitative analysis of lignocellulosic biomass using infrared techniques: A mini-review", *Applied Energy*, 104: 801–809.

Chadwick, D., K. McDonnell, L. Brennan, C. Fagan, C. Everard, 2014. "Evaluation of infrared techniques for the assessment of biomass and biofuel quality parameters and conversion technology processes: A review", *Renewable and Sustainable Energy Reviews*, 30: 672–681.

Ye, X., L. Liu, D. Hayes, A. Womac, K. Hong, S. Sokhansanj, 2008. "Fast classification and compositional analysis of cornstover fractions using Fourier transform near-infrared techniques", *Bioresource Technology*, 99: 7323-7332.

Rammal, A., E. Perrin, V. Vrabie, I. Bertrand, A. Habrant, B. Chabbert, 2014. "Selection of Discriminant Mid-Near infrared Wavenumbers Using a Genetic Algorithm: Evaluation of the Biodegradation of Maize Roots", Int. Conf. on Chemical, *Biological and Environmental Engineering (ICBEE 2014), Paris, France – to be published in APCBEE Procedia*.

Bangalore, A.S., R.E. Shaffer, G.W. Small, 1996. "Genetic Algorithm-Based Method for Selecting Wavelengths and Model Size for Use with Partial Least-Squares Regression: Application to Near-Infrared Spectroscopy", *Analytical Chemistry*, 68: 4200-4212.

Skurichina, M., S. Verzakov, P. Paclik, R.P.W. Duin, 2006. "Effectiveness of Spectral Band Selection Extraction Techniques for Spectral Data", *Structural, Syntactic, and Statistical Pattern Recognition, Lecture Notes in Computer Science*, 4109: 541-550.

Miller, A., 2002. "Subset selection in regression", *Chapman & Hall/CRC*.

Hastie, T., R. Tibshirani, J. Friedman, 2009. "The elements of statistical learning", *Springer*.

Ullah, S., T.A. Groen, M. Schlerf, A.K. Skidmore, W. Nieuwenhuis, C. Vaiphasa, 2012. "Using a Genetic Algorithm as an Optimal Band Selector in the Mid and Thermal Infrared (2.5–14 μm) to Discriminate Vegetation Species", *Sensors*, 12: 8755-8769.

Machinet, G, I. Bertrand, B. Chabbert, F. Watteau, G. Villemin, S. Recous, 2009. "Soil biodegradation of maize root residues: Interaction between chemical characteristics and the presence of colonizing micro-organisms". *Soil Biology & Biochemistry*, 41: 1253-1261.

Holland, J., 1975. "Adaptation in Natural and Artificial Systems", *MIT Press*.

Vishwakarma, P., Y. Kumar, R. Kumar, 2012. "Data Mining Using Genetic Algorithm (DMUGA)", *International Journal of Engineering Research and Development*, 5(2): 88-94.

Mitchell, M.,1995. "Genetic Algorithms: An Overview", *MIT Press*, 1: 31-39.

Yang, M., Y. Yang, T. Su, 2014. "An Efficient Fitness Function in Genetic Algorithm Classifier for Landuse Recognition on Satellite Images", *The Scientific World Journal*: doi: 10.1155/2014/264512.

Xu, R., J. Xu, D. Wunsch, 2012. "A Comparison Study of Validity Indices on Swarm-Intelligence-Based Clustering", *IEEE Transactions on systems, man, and cybernetic*, 42: 1243-1256.

Maulik, U., S. Bandyopadhyay, 2003. "Fuzzy Partitioning Using a Real-Coded Variable-Length Genetic Algorithm for Pixel Classification", *IEEE Transactions on geoscience and remote sensing*, 41(5): 1075 – 1081.

Chiang, L., R. Pell, 2004. "Genetic algorithms combined with discriminant analysis for key variable identification", *Journal of Process Control*, 14: 143–155.

Sushmita, M., 2004. "An evolutionary rough partitive clustering", *Pattern Recognition Letters*, 25: 1439-1449.

Ranjini, A, B. Zoraida, 2013. "Analysis of Selection Schemes for Solving Job Shop Scheduling Problem Using Genetic Algorithm", *IJRET: International Journal of Research in Engineering and Technology*, 2 (11): 2319-1163.

Picek, S., M. Golub, 2010. "Comparison of a Crossover Operator in Binary-coded Genetic Algorithms", *WSEAS Trans.on Computers*, 9(9): 1064-1073.

Hinterding, R., 1995. "Gaussian Mutation and Self-adaptation for Numeric Genetic Algorithms", *2nd IEEE Conf on Evolutionary Computation*, 384-389.

Dong, Y.-W, S.-Q Yang, C.-Y. Xu, Y.-Z. Li, W. Bai, Z.-N. Fan, Y.-N. Wang, Q.-Z. Li, 2011. "Determination of Soil Parameters in Apple-Growing Regions by Near- and Mid-Infrared Spectroscopy", *Pedosphere*, 21(5): 591-602.

Maurel, V., A. McBratney, 2011. "Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils - Critical review and research perspectives", *Soil Biology & Biochemistry*. 43: 1398-1410.