Nonparametric Functional Classification of Timber Species from DTG Curves. Comparison with Multivariate Classification Methods Using a New Approach for DTG Curves Discretization

Naya, Salvador (1st author)
University of A Coruña, Department of Mathematics
Higher Polytechnic School. Campus de Esteiro
Ferrol 15403, Spain
E-mail: salva@udc.es

Francisco-Fernández, Mario (2nd author)
University of A Coruña, Department of Mathematics
Faculty of Computer Science. Campus de Elviña
A Coruña (15071), Spain
E-mail: mariofr@udc.es

Tarrío-Saavedra, Javier (3rd author)
University of A Coruña, Department of Applied Economics
Faculty of Economics and Business. Campus de Elviña
A Coruña (15071), Spain
E-mail: jtarrio@udc.es

Lpez-Beceiro, Jorge (4th author)
University of A Coruña, Department of Industrial Engineering
Higher Polytechnic School. Campus de Esteiro
Ferrol 15403, Spain
E-mail: jlopezb@udc.es

Artiaga, Ramn (5th author)
University of A Coruña, Department of Industrial Engineering
Higher Polytechnic School. Campus de Esteiro
Ferrol 15403, Spain
E-mail: rartiga@udc.es

Introduction

Correct classification of wood samples is a complicated task, which requires qualified and largely trained personnel and a great amount of knowledge and observation. The cost corresponding to train these professionals is high, and the number of those increasingly rare. Thus, the application of statistical models for recognition of wood samples is useful and justified. There are some works related to wood identification from the data obtained by two different techniques: image-based and spectrum-based processing systems. Non-supervised classification was performed by making use of Fourier Transform Raman (FTR) spectroscopy as data, Neural Networks (NN) as classification method and spectral features extraction in Lewis et al. (1994). In Yang and Lewis (1999), NN were also applied to FTR spectra in order to distinguish temperate woods from tropical woods. More recently, K Nearest Neighbor classifiers (KNN), Linear, Quadratic, and Support Vectors Machines (SVM) classifiers were used to classify fluorescence spectra corresponding to wood of different species (Piuri and Scotti 2010). An alternative consisted in classifying by nonparametric functional analysis the curves obtained by

Thermogravimetric Analysis (TG) (Tarrío-Saavedra et al. 2010). In the same study, Differential Scanning Calorimetry (DSC) data conducted to worse results. However, the possibility of using Differential Thermal Analysis (DTG) curves (smoothed using the local linear regression estimator) as a source of data for statistical classification of wood species has not been studied yet. A DTG curve gives the mass loss rate of a particular material with respect to temperature. It provides important information related to the material thermal stability (Menczel et al. 2009). Specifically, the present work should give an answer to the following questions: is it possible to observe these differences among species in the mass loss rate of wood samples?, is it possible to classify woods from the differences present in the DTG curves?

Experimental conditions and data collecting

Seven different wood species of industrial interest were tested: 5 hardwoods (beech or *Fagus sylvatica*, chestnut or *Castanea sativa*, oak or *Quercus robur*, jatobá or *Hymenaea courbaril* and *Eucalyptus globulus*) and 2 softwoods (Scots pine or *Pinus silvestris* and insignis pine or *Pinus pinaster*). Seven samples per each specie obtained from wood of different trees are slected. The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the time of experimentation. The thermogravimetric tests were performed in a TA SDT 2960 TA Instruments simultaneous analyzer. Each sample consisted of a single chunk in the 6-8 mg range. The experimental conditions were the following: open alumina pan, a heating rate of 20°C min^{-1} from room temperature to 600°C, applying a nitrogen flow rate of 50 mL min^{-1}.

Classification methods

A nonparametric functional technique based on kernel methods (K-NPFDA) and a nonparametric functional method based in the boosting algorithm (B-NPFDA) are applied to construct a classification rule to discriminate between the different wood species (Ferraty and Vieu 2006), based on a sample of 49 DTG curves. A DTG curve is classified as belonging to the specie or the group to which the highest posterior probability is obtained. The functional Nadaraya-Watson kernel nonparametric method, shown in (1), is applied. Given a new DTG curve, $x = x(t)$, obtained from a material to classify, the estimator of the posterior probability of belonging to a class g, with $g \in \{0, 1, \ldots, G\}$, is given by:

$$\text{(1)} \qquad \hat{r}_h^{(g)}(x) = \frac{\sum_{i=1}^{n} I_{\{Y_i=g\}} K\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{h}\right)},$$

where the observed DTG curves, $X_i = X_i(t)$, are a sample of explanatory variables, while the response sample consists of the observations Y_i of a discrete random variable taking values in the set {0,1,. . . ,G}, the different classes.

In our research, the Gaussian kernel, K, is used. On the other hand, the smoothing parameter, h, is chosen as the value that minimizes the probability of missclassifying a future observation and it is selected according to the cross-validation method. This method consists in minimizing the cross-validation function:

$$CV(h) = n^{-1} \sum_{i=1}^{n} I_{\{Y_i \neq d_h^{-i}(X_i)\}},$$

where d_h^{-i} is the classification rule built up without the i-th observation:

$$d_h(x) = \text{argmax}_{0 \leq j \leq G} \left\{ \hat{r}_h^{(j)} \right\}.$$

Additionally, classical multivariate methods and machine learning approaches were used to classify the wood samples. To perform this analysis, a new technique for DTG curve discretization is considered. This discretization is done using the parameters resulting from the fit of a nonlinear parametric model to the DTG curves. The proposed model to be fit consists of a mixture of 4 generalized logistic derivative functions related to the wood main components, cellulose, hemicellulose, lignin and water:

$$x(t) = \sum_{i=1}^{n} \frac{c_i \cdot b_i \cdot \exp\left(-b_i \cdot (t - m_i)\right)}{\left[1 + \tau_i \cdot \exp\left(-b_i \cdot (t - m_i)\right)\right]^{(1+\tau_i)/\tau_i}},$$

where the c parameter represents the mass involved in the degradation process, b is related to the decomposition rate or rate of change, τ accounts for the asymmetry, m represents the temperature at the maximum rate of change and t is the temperature. The optimal fittings are obtained by minimizing the average squared error (ASE). Once the fit and the discretization process are carried out, several multivariate classification methods, such as Linear Discriminant Analysis (LDA), Logistic Regression, Naïve Bayes (NBC), K Nearest Neighbors (KNN), Support Vector Machines (SVM) and Neural Networks (NN) are applied. In all the cases, the samples have been successfully classified. The classification methods used in this study are validated through leave-one-out cross-validation, which is a widely used technique.

Results and discussion

Figure 1 shows all the smoothed (using the local linear estimator) DTG curves obtained from the experimental tests, which correspond to the 7 studied wood species. A direct plug-in bandwidth selection method (Ruppert et al. 1995) to select the smoothing parameter was used. As previously pointed out, first, FDA techniques were used to classify wood species. The application of linear local smoothing with an optimal bandwidth significantly improves the correct classification percentages obtained by K-NPFDA method: from 78% to 82%.

Figure 1. Smoothed DTG curves obtained for the different species.

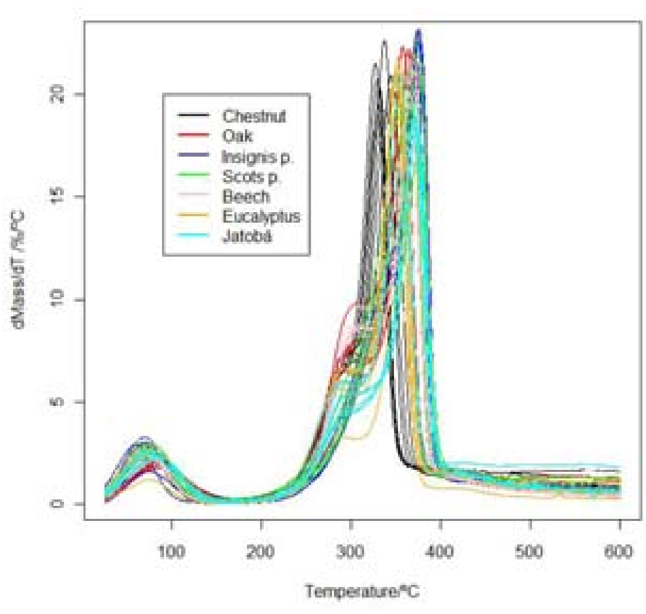

Table 1 shows the percentages of correct classification when the different approaches are used. Note that when we want to classify among the 7 wood species with the DTG curves using both functional and multivariate methods, the results are excellent, especially those corresponding to SVM (0.92) and logistic regression (0.92). Multivariate methods seem to perform better than functional ones. Thus, results in Table 1 show that it is possible to classify among the 7 studied species using these 16 parameters, obtained from the logistic fits as original multivariate data, jointly with the proposed classification methods. We can say that there are differences in wood species mass loss rate and that the 16 parameters corresponding to each DTG curve summarize correctly these differences.

Functional data classification can also be done from the successive derivatives of the DTG curves. The probability of correct classification obtained from the DTG second derivative is particularly high (0.90) and competitive with respect to the results from the multivariate methods. Furthermore, applying functional techniques presents the advantage of not having to perform any regression model fit to obtain the data.

Table 1. Prediction probabilities for 7 different classes obtained by each classification method, using extracted features from DTG curves, DTG curves, and first and second derivatives from DTG curves.

	DTG	**Derivative DTG**	**2nd Derivative DTG**
Classification methods	**Prediction**	**Prediction**	**Prediction**
K-NPFDA	0.82	0.84	0.90
B-NPFDA	0.84	–	–
LDA	0.90	–	–
Logistic regression	0.92	–	–
NBC	0.78	–	–
KNN	0.78	–	–
SVM	0.92	–	–
NN	0.90	–	–

Table 2 shows the confusion matrix corresponding to 7 species, using the best resulting methods for multivariate and functional methods. Looking at the diagonal, all the particular class probabilities are very high. There is only a little of confusion between oak and beech, and between Scots and Insignis pine.

Table 2. Confusion matrix or probabilities of correct classification in 7 groups, obtained by second derivative DTG curves and NPFDA method, and by SVM using parameters extracted from the nonlinear fitting model of DTG curves.

		Actual						
Methods	**Estimated**	**Chesn.**	**Oak**	**Insig. P.**	**Scots P.**	**Beech**	**Eucal.**	**Jat.**
NPFDA	**Chestnut**	0.86	0.00	0.00	0.00	0.00	0.00	0.00
(from 2nd	**Oak**	0.00	0.57	0.00	0.00	0.00	0.00	0.00
derivative	**Insignis P.**	0.00	0.00	1.00	0.14	0.00	0.00	0.00
DTG)	**Scots P.**	0.00	0.00	0.00	0.86	0.00	0.00	0.00
	Beech	0.00	0.29	0.00	0.00	1.00	0.00	0.00
	Eucalyptus	0.00	0.14	0.00	0.00	0.00	1.00	0.00
	Jatobá	0.14	0.00	0.00	0.00	0.00	0.00	1.00
SVM	**Chestnut**	1.00	0.00	0.00	0.00	0.00	0.00	0.14
(from	**Oak**	0.00	0.86	0.00	0.00	0.00	0.00	0.00
DTG	**Insignis P.**	0.00	0.00	1.00	0.29	0.00	0.00	0.00
curves)	**Scots P.**	0.00	0.00	0.00	0.71	0.00	0.00	0.00
	Beech	0.00	0.14	0.00	0.00	1.00	0.00	0.00
	Eucalyptus	0.00	0.00	0.00	0.00	0.00	1.00	0.00
	Jatobá	0.00	0.00	0.00	0.00	0.00	0.00	0.86

Conclusions

The obtained probabilities of correct classification show that it is possible to observe differences among wood species studying thermal stability tested by DTG. Thus, DTG is an appropriate technique for classifying wood species.

Two functional nonparametric methods of discriminant analysis for the classification of wood species are successfully applied. The best results are obtained using DTG second derivative curves and K-NPFDA method (0.90).

A regression model consisting of 4 components is proposed to fit the DTG curves, obtaining very good performance. The model is justified on the basis that there are 4 different degradation processes. Multivariate classifiers as SVM with Gaussian kernel (0.92), NN (0.90) and Logistic regression (0.92) produce slightly better probabilities of correct classification than functional methods when the DTG curves are used, but the K-NPFDA is able to give similar results when the DTG second derivatives are used.

Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 (ERDF included).

REFERENCES

Ferraty, F., Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice, Springer-Verlag,New York.

Lewis, I.R., Daniel, N.W., Chaffin, N.C., Griffiths, P.R. (1994), Raman spectrometry and neural networks for the classification of wood types-1, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 50, 1943-1958.

Menczel J.D., Judovits L., Prime R.B., Bair H.E., Reading M., Swier S. (2009), Differential scanning calorimetry (DSC), in Thermal analysis of polymers. Fundamentals and applications, San Jose: John Wiley & Sons.

Piuri, V., Scotti, F. (2010), Design of an automatic wood types classi?cation system by using ?uorescence spectra, IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, 40, 358-366.

Ruppert, D., Sheather, S.J., Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association, 90, 1257-1270.

Tarrío-Saavedra, J., Naya, S., Francisco-Fernández, M., López-Beceiro, J., Artiaga, R. (2010), Functional nonparametric classification of wood species from thermal data, Journal of Thermal Analysis and Calorimetry,104, 87-100.

Yang, H., Lewis, I. R. (1999), Raman spectrometry and neural networks for the classification of wood types. 2, Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 55, 2783- 2791.

ABSTRACT

The principal aim of this work is to classify among seven commercial wood species through their thermal stability, evaluated by differential thermal analysis (DTG). New nonparametric functional data analysis (FDA) techniques, as well as classical and machine learning methods of multivariate supervised classification, like Linear Discriminant Analysis, Quadratic classification, Logistic Regression, Naïve Bayes, K Nearest Neighbors, Support Vector Machines and Neural Networks are used for this task. A DTG curve gives the mass loss rate of a particular material with respect to temperature. Each DTG curve is smoothed using the local linear regression estimator. Then, a nonparametric functional technique based on the Bayes rule and the Nadaraya-Watson regression estimator is used to discriminate between the different wood classes. The results are compared with those obtained using classical and machine learning multivariate methods. To obtain a multivariate dataset summarizing the information in the DTG curves, a logistic mixture regression model consisting of four components is proposed to fit the curves, using the resulting parameters as multivariate datasets.