Departamento de Matemáticas
UNIVERSIDADE DA CORUÑA

# EVALUACIÓN Y CLASIFICACIÓN DE MATERIALES: UN ENFOQUE ESTADÍSTICO

**Javier Tarrío Saavedra**

**Septiembre 2012**

**Tesis Doctoral**

# Evaluación y clasificación de materiales: un enfoque estadístico

*Tesis doctoral*

**Javier Tarrío Saavedra**

Directores:
Salvador Naya Fernández
Ramón Artiaga Díaz
Mario Francisco Fernández

**Departamento de Matemáticas**
**Facultade de Informática**
**Universidade da Coruña**

**Septiembre 2012**

Los abajo firmantes certifican que son los directores de la Tesis Doctoral titulada "Evaluación y clasificación de materiales: un enfoque estadístico", llevada a cabo por Javier Tarrío Saavedra dentro del programa de doctorado "Estadística e Investigación Operativa", ofertado por el Departamento de Matemáticas de la Universidad de A Coruña, dando su consentimiento para que su autor, cuya firma también se incluye en las siguientes líneas, proceda a su presentación y defensa.

Fecha: 28 de septiembre, 2012

Directores:

Salvador Naya Ramón Artiaga Mario Francisco

Alumno:

Javier Tarrío Saavedra

*A Cuba.*
*A Monelos.*
*A los de la calle Sanmartín.*

*Take down this book*
*And slowly read...*
*W. B. YEATS*

# Agradecimientos

*La gratitud en silencio no sirve a nadie.*

Gladys Bronwyn Stern

*Es tan grande el placer que se*
*experimenta al encontrar un hombre*
*agradecido que vale la pena arriesgarse a*
*no ser un ingrato.*

Séneca

El autor de este trabajo agradece encarecidamente toda la ayuda prestada para la elaboración de este proyecto. Particularmente, éste no sería posible sin la colaboración, guía y apoyo de Salvador Naya, Ramón Artiaga, Mario Francisco, Iñaqui Ullibarri y Jorge López. También quisiera expresar mi agradecimiento a Manuel Febrero, coautor de la librería fda.usc en R, por sus orientaciones y comentarios acerca del ANOVA funcional y la clasificación funcional no paramétrica. Igualmente a José María Matías, por su atención en lo tocante a clasificación con SVM y a Ricardo Cao, por sus elegantes ideas. Hago extensivo este agradecimiento a Abhirup Mallik, Carlos Gracia, Aldana G. Montoro y a toda la gente con la que trabajo, tanto en el Laboratorio de Análisis Térmico en la EPS de Ferrol como en el Departamento de Matemáticas, en A Coruña.

Los diversos estudios de investigación de los que consta la presente monografía han sido parcialmente financiados por el Ministerio de Ciencia e Innovación, en particular mediante los fondos correspondientes a los proyectos MTM2008-00166 y MTM2011-22392.

# Prefacio

*Llega siempre un tiempo en que hay que*
*elegir entre la contemplación*
*y la acción.*

Albert Camus

La tesis doctoral que aquí se muestra se presenta como el compendio de cuatro de los artículos de investigación elaborados y publicados en revistas indexadas en el *Journal Citation Report* durante el período en el que el autor ha estado matriculado en el programa de doctorado de Estadística e Investigación Operativa de la Universidad de A Coruña (Universidade da Coruña). Seguidamente se muestran las referencias correspondientes a cada uno de los artículos:

1. Javier Tarrío-Saavedra, Salvador Naya, Mario Francisco-Fernández, Jorge López-Beceiro, Ramón Artiaga (2011). Application of functional ANOVA to the study of thermal stability of micronano silica epoxy composites. *Chemometrics and Intelligent Laboratory Systems*, **105**, 114-124. ISSN: 0169-7439.

2. Mario Francisco-Fernández, Javier Tarrío-Saavedra, Abhirup Mallik, Salvador Naya (2012). A comprehensive classification of wood from thermogravimetric curves. *Chemometrics and Intelligent Laboratory Systems*, In Press, DOI: http://dx.doi.org/10.1016/j.chemolab.2012.07.003. ISSN: 0169-7439.

3. Javier Tarrío-Saavedra, Salvador Naya, Mario Francisco-Fernández, Jorge López-Beceiro, Ramón Artiaga (2011). Functional nonparametric classification of wood species from thermal data. *Journal of Thermal Analysis and Calorimetry*, **104**, 87-100. ISSN: 1388-6150.

4. Abhirup Mallik, Javier Tarrío-Saavedra, Mario Francisco-Fernández, Salvador Naya (2011). Classification of wood micrographs by image

segmentation. *Chemometrics and Intelligent Laboratory Systems*, **107**, 351-362. ISSN: 0169-7439.

Aparte de la resolución de los problemas concretos planteados en cada artículo, esta serie de trabajos pretenden representar un nexo de unión o bisagra entre la Estadística, con todas las oportunidades que ofrece, y la Ciencia de Materiales, dotando al conjunto del texto de un carácter marcadamente aplicado. En esta tesis se abordan problemáticas de la Ciencia de los Materiales que son novedosas o que no han sido resueltas hasta el momento mediante métodos estadísticos clásicos. Para ello se recurre a técnicas recientemente desarrolladas, o incluso en desarrollo, relacionadas principalmente con el Análisis de Datos Funcionales (FDA) y el Aprendizaje Máquina en un contexto multivariante. Por lo tanto, aún sin ser su principal motivación, en la presente monografía se busca igualmente contribuir a la difusión de estos nuevos procedimientos en el campo de Ingeniería de Materiales.

Esta memoria está dividida en dos partes principales, una relacionada con la evaluación de micro-nanocompuestos de matriz orgánica epoxi y carga inorgánica humo de sílice, y otra que hace referencia a la clasificación de materiales (en particular, a la clasificación de especies de madera) a partir de datos térmicos y de las características obtenidas después de un proceso de segmentación de imágenes.

# Índice

# Índice de figuras

# Índice de Tablas

# Capítulo 1

# Introducción y objetivos generales

*Aquella teoría que no encuentre aplicación práctica en la vida, es una acrobacia del pensamiento.*

Swami Vivekananda

*Los que se enamoran de la práctica sin la teoría son como los pilotos sin timón ni brújula, que nunca podrán saber a dónde van.*

Leonardo da Vinci

Muchos problemas en la Ciencia de Materiales pueden ser tratados y resueltos mediante el uso del análisis estadístico de datos y, de hecho, ésta es una práctica habitual en los ámbitos de la industria y la investigación académica. En esta monografía se tratan problemáticas pertenecientes a la Ciencia de Materiales que son novedosas o que no han sido resueltas definitivamente hasta el momento empleando procedimientos estadísticos clásicos. La resolución de cada problema concreto representa el principal objetivo del presente texto, y para ello se recurre (aparte de a metodologías más comunes) a la aplicación de técnicas desarrolladas recientemente, o incluso en proceso de desarrollo, relacionadas, entre otros campos, con el Análisis de Datos Funcionales (FDA) y el Aprendizaje Máquina en un contexto de datos multivariantes. Por lo tanto, aún sin ser la principal motivación de la tesis aquí introducida, se pretende contribuir a la difusión de estos nuevos procedimientos en el campo de Ingeniería de Materiales, incluyendo también la generalización de metodologías como son los estudios de simulación estadística, la utilización del remuestreo o "bootstrap", el diseño de experimentos,

el tratamiento de imágenes y la aplicación de procedimientos fiables de validación. En concreto, en este trabajo se presentan los siguientes problemas: (a) la evaluación de nuevos micro y nanomateriales, en concreto micro y nanocompuestos epoxi-humo de sílice, y (b) la clasificación automática de materiales, centrada en el caso particular de las especies de madera.

La presentación y tratamiento estadístico de cada uno de estos problemas se corresponde, respectivamente, con cada una de las dos partes principales en las que se divide la tesis que aquí se introduce.

El estudio, desarrollo y aplicación de los micro y nanomateriales ha experimentado un gran auge en los últimos 20 años. Una de las razones es la posibilidad de obtener compuestos con propiedades mecánicas únicas, mejorando el comportamiento que presentan los materiales tradicionales. En lo que respecta a los compuestos de matriz polimérica epoxídica y refuerzo inorgánico de micro y nanopartículas de sílice, su estudio y uso está muy extendido. Esto es debido, por una parte, a la rigidez, estabilidad y ligereza de la resina epoxi y, por otra, a la mejora en la tenacidad, estabilidad térmica, resistencia a la llama, etc. que aporta la adición de micro y nanopartículas de sílice. La obtención de los valores óptimos de la cantidad de partículas, grado de dispersión, tamaño medio y forma, se vuelve crucial en el desarrollo de estos materiales. En la presente tesis doctoral se propone medir la influencia de la adición de nano partículas de sílice en la estabilidad térmica de una resina epoxi, aplicando para ello el denominado ANOVA funcional a las curvas obtenidas mediante Análisis Termogravimétrico (TG). Esta metodología, relacionada con el FDA, ha demostrado aportar una valiosa información, complementaria e incluso alternativa a la obtenida por otros procedimientos experimentales y estadísticos.

En la segunda parte de este estudio, se aborda el problema de la clasificación de materiales, concretamente, se evalúa el poder discriminante de las curvas TG y de las curvas obtenidas mediante Calorimetría Diferencial de Barrido (DSC) por varios métodos de clasificación, funcionales y multivariantes, paramétricos y no paramétricos, novedosos y tradicionales. Para ello, se elige uno de los materiales más difíciles de clasificar debido a su heterogeneidad estructural y química: la madera. La posibilidad de clasificación de especies de madera a partir de las curvas obtenidas mediante la aplicación del análisis térmico, había sido apuntada por Salvador Naya en su tesis doctoral (Naya, 2003). Junto al trabajo que aquí se describe, representan los únicos acercamientos, hasta la fecha, a la clasificación de maderas mediante curvas TG. Por otro lado, también se presenta el primer estudio de clasificación de maderas a partir de las curvas obtenidas mediante DSC para el mismo fin. Debido a la alta variabilidad que existe dentro de cada especie (se pueden

observar diferencias apreciables entre muestras pertenecientes a distintos árboles e incluso a un mismo árbol), se ha juzgado no sólo útil sino también necesaria la inclusión de un estudio de simulación de curvas TG. Este es un procedimiento todavía muy poco usual en el área del Análisis Térmico, a través del cual se obtienen nuevas muestras artificiales en escenarios más desfavorables que el que describen los datos experimentales, ahorrando en tiempo de experimentación y coste del material que conlleva. Mediante este procedimiento se evalúa de una forma exhaustiva el poder discriminante de las curvas TG.

Además, también se ha probado una nueva alternativa para resolver el mismo problema, pero en este caso partiendo de datos completamente diferentes; este enfoque está basado en la obtención y procesamiento de imágenes, metodologías mucho más utilizadas que el análisis TG y DSC para resolver problemas de clasificación y detección de defectos. Se han clasificado muestras pertenecientes a diferentes especies de madera, utilizando para ello las micrografías obtenidas a 1500× aumentos mediante Microscopía Electrónica de Barrido (SEM) y procesadas mediante el procedimiento de segmentación de imágenes. El uso de micrografías SEM como fuente de datos para la clasificación de maderas (previo procesado y extracción de características) representa una novedad en la Ciencia de Materiales y la Quimiometría. Hay que destacar que la discriminación de materiales en sus diferentes clases es un importante problema práctico con una aplicación directa en la industria. En concreto, la clasificación de las especies de madera es útil en las industrias del mueble, la producción de paneles de madera, o incluso en la Arqueología. Hoy en día, la identificación de la madera se realiza con una precisión que varía en gran medida dependiendo del operador. Además, la formación de trabajadores cualificados puede resultar muy costosa, tanto en tiempo como en dinero. Por tanto, la aplicación de métodos automáticos de reconocimiento de muestras de madera puede tener una aplicación inmediata dentro de la tecnología de este material. A partir de los resultados alcanzados, se han obtenido interesantes conclusiones relacionadas con la estructura físico-química de la madera.

Pese a la aparente heterogeneidad de las partes que constituyen esta monografía, una misma idea une las dos partes principales de este texto: la aplicación de diversas técnicas estadísticas al estudio de materiales, más concretamente, al estudio de la degradación de materiales. Dada la naturaleza de los datos a tratar, gran parte de las técnicas utilizadas en este trabajo tienen relación con el FDA, como son el ANOVA funcional o los diferentes métodos de clasificación supervisada utilizados. Pero, ya sea para aplicar las anteriores o de forma independiente, en las siguientes páginas se utilizan herramientas

como el bootstrap paramétrico, simple y suavizado, la simulación de Monte Carlo, la regresión no lineal según una suma de componentes logísticas, la regresión spline, el Análisis de Componentes Principales (PCA), la clasificación multivariante aplicando diversos clasificadores como son el Análisis Discriminante Lineal (LDA), Bayes Naive (NBC), Regresión Logística, Clasificación Cuadrática, $k$ vecinos más próximos ($k$-NN), Máquinas de Soporte Vectorial (SVM) o Redes Neuronales (NN). El principal objeto de aplicación de estas técnicas han sido las curvas resultantes de los experimentos obtenidos por análisis térmico, que, por sus características, pueden considerarse datos funcionales. Por un lado, se han obtenido curvas TG, que describen la masa que posee un espécimen en relación a la temperatura que soporta (o al tiempo) y, por otro, curvas DSC, que expresan el intercambio de energía en una muestra respecto a su temperatura o al tiempo. De forma global, el trabajo que aquí se presenta se ocupa del estudio de las diferencias existentes en los procesos de degradación entre materiales, ya sean provocadas (ANOVA funcional) o preexistentes (muestras de diferentes especies de madera) y de la simulación de dichos procesos de degradación (obteniendo nuevas muestras artificiales en diferentes escenarios). Adicionalmente, se presenta un estudio en el que se prueba la idoneidad del uso de micrografías SEM segmentadas para la clasificación supervisada de materiales.

Es muy importante tener en cuenta que la presente tesis doctoral se estructura en torno a un conjunto de cuatro artículos de investigación publicados durante el período en el que el autor estuvo matriculado en el programa de doctorado en Estadística e Investigación Operativa impartido en la Universidad de A Coruña. En el primer artículo se aborda el primer problema planteado anteriormente: evaluación de la influencia del contenido en micro-nanopartículas en la estabilidad térmica de una resina epoxi (Tarrío-Saavedra et al., 2011). Este estudio representa por sí solo la primera parte de esta tesis. Los tres artículos restantes constituyen la segunda parte: la clasificación de materiales; en particular, la clasificación de especies de madera a partir de datos térmicos (incluyendo un estudio de simulación) y micrografías segmentadas. Siguiendo este esquema, el Capítulo 3 se corresponde con Francisco-Fernández et al. (2012), el Capítulo 4 con Tarrío-Saavedra et al. (2011) y el Capítulo 5 con Mallik et al. (2011).

En las siguientes líneas se muestra una introducción a cada uno de los problemas planteados, describiendo previa y brevemente los antecedentes. Estos últimos representan los trabajos de investigación que, aunque no forman parte del cuerpo de esta tesis doctoral, han llevado al planteamiento y realización de los artículos de los que consta el presente documento.

## 1.1. Antecedentes

En esta sección se muestran aquellos artículos de investigación (en los que figura como firmante el autor de la presente memoria) que han servido como punto de partida y motivación de los cuatro artículos que componen el cuerpo de esta tesis. Todos ellos presentan el denominador común del empleo del Análisis Térmico, en particular la obtención de las curvas TG y DSC para el estudio y caracterización de materiales. El Análisis Térmico es un conjunto de técnicas englobadas dentro del grupo más amplio de la Química Analítica, y a él pertenecen técnicas como la termogravimetría, la calorimetría, el Análisis Termomecánico (TMA) y el Análisis Mecánico Dinámico (DMA) (Prime et al., 2009). Por otro lado, en mayor o menor medida, en cada uno de estos estudios se han aplicado diferentes técnicas estadísticas relacionadas con el diseño de experimentos, la regresión no lineal, la obtención de números índice o el remuestreo. Los resultados obtenidos en estos estudios, así como las necesidades crecientes de información a partir de los datos obtenidos en laboratorio, llevaron al planteamiento y elaboración de los cuatro artículos de los que consta el presente trabajo.

Primeramente, fue de particular relevancia la experiencia y resultados obtenidos en Tarrio-Saavedra et al. (2008). En dicho trabajo se prepararon y estudiaron diversos compuestos formados por una matriz polimérica epoxídica y una carga inorgánica llamada humo de sílice (un subproducto industrial de la industria del silicio). Se estudió su degradación térmica mediante análisis TG en atmósfera oxidante y se estimó su estabilidad térmica calculando una serie de números índice clásicos, obtenidos a partir de las curvas TG. Estos fueron la Temperatura de Descomposición Inicial (IDT), la temperatura correspondiente a la velocidad máxima de pérdida de masa (Tmax) y la Temperatura de Descomposición por Procedimiento Integral (IPDT). Se calcularon antes y después de sustraer matemáticamente, a cada curva TG, la masa del humo de sílice añadido (carga); dicha masa se correspondía con la masa final de los experimentos, pues el humo de sílice no se degrada a la temperatura programada. Por tanto, después de la sustracción de esta masa, se obtuvo el camino de degradación de, únicamente, la resina epoxi que forma parte del compuesto; si no se resta esta masa, la estabilidad térmica del compuesto aumenta aparentemente con el contenido en sílice. Sin embargo, después de la sustracción de la masa inorgánica, el camino de degradación térmica de la resina epoxi (curvas TG) se vio sólo muy ligeramente afectado por el contenido de sílice.

Este trabajo (Tarrio-Saavedra et al., 2008) planteó la necesidad de emplear técnicas estadísticas de diseño de experimentos para evaluar si estas

diferencias en la estabilidad térmica eran o no significativas. Adicionalmente, dada la naturaleza funcional de las curvas estudiadas (TG), surgió la necesidad de emplear técnicas FDA para su estudio, en lugar de un vector finito-dimensional formado por ciertas características de la curva, cuyo uso suponía un apreciable desaprovechamiento de información. En definitiva, este estudio sentó las bases para la elaboración del artículo que constituye la primera parte de esta tesis.

Posteriormente, se realizaron diversos estudios en los que se comenzaron a aplicar técnicas de diseño de experimentos sobre datos univariantes. Así, en Tarrío-Saavedra et al. (2010a) se midió el efecto de la adición de humo de sílice en el proceso de curado y en las propiedades termomecánicas de los compuestos de matriz epoxídica. Se tomó como dato la energía relacionada con dicha reacción de curado, calculada a partir de la integración numerica de las curvas DSC, obtenidas a partir de compuestos con un contenido de sílice entre el 5 wt % y el 50 wt % (tanto por ciento en peso de la muestra). Por otro lado, en Tarrío-Saavedra et al. (2010b), se estudiaron los factores influyentes en la estabilidad a la oxidación del biodiésel. En este trabajo se cuantificó la influencia que tiene el tipo de biodiésel, la presión de oxígeno y la concentración de un antioxidante sobre la estabilidad respecto a la oxidación del biodiésel. Para ello se realizó un diseño experimental considerando dos tipos de biodiésel, distintas presiones de oxígeno y adiciones de un antioxidante comercial. Como variable respuesta se eligió la llamada Temperatura de Inicio de la Oxidación o "Oxidation Onset Temperature" (OOT), evaluada mediante Calorimetría Diferencial de Barrido a Presión (PDSC). Se encontró que el diseño que mejor se ajusta a los datos es el modelo de efectos principales con tres factores tratamiento, replicado. La secuencia de factores, ordenados en orden decreciente según su influencia en el OOT resultó ser: Presión >Tipo de biodiésel >Antioxidante. Para llevar a cabo experimentos de DSC modulados a altas presiones (TMPDSC), también se investigó (empleando técnicas de diseño de experimentos) el uso conjunto de una celda de presión y la técnica de Calorimetría Diferencial de Barrido con Modulación de Temperatura (TMDSC) (Gracia-Fernández et al., 2011). No se han encontrado estudios anteriores acerca del uso de TMPDSC, por lo que este trabajo representó el punto de partida de esta metodología. El método propuesto se aplicó a la medición del efecto de la presión sobre la reacción de curado de un sistema epoxi. Se efectuaron experimentos de curado cuasi-isotermos modulados a diferentes presiones para evaluar el tiempo de vitrificación y se llevaron a cabo con éxito experimentos en los que se aplicó un calentamiento a velocidad constante con modulación de temperatura (a diferentes presiones) para separar el efecto "reversing", debido a la transición vítrea, de la reacción

de curado exotérmico residual; se evaluó la entalpía de curado, la conversión frente a la temperatura y la transición vítrea de los termoestables completamente curados. Todos los parámetros estudiados resultaron afectados por la presión en el intervalo comprendido entre la presión atmosférica y 35 bar. Resultó, a su vez, que la entalpía de curado, la velocidad de reacción y la conversión a un tiempo dado aumentan con cualquier incremento de la presión. Gracias a este trabajo, la utilidad de TMDSC para caracterizar el curado de los termoestables se extendió a las situaciones en que se aplica alta presión, como es el caso de la industria aeronáutica.

Aparte de todos estos trabajos relacionados con el diseño de experimentos, en López-Beceiro et al. (2011b) se introdujo un estudio simple de clasificación supervisada. El objetivo general de este trabajo era evaluar la estabilidad termooxidativa de diferentes tipos de aceites vegetales (los aceites de oliva Arbequina, Picual, Hojiblanca y Cornicabra, el aceite de maíz, el de soja y el de girasol) mediante el OOT obtenido partir de las curvas PDSC, por un procedimiento basado en la norma ASTM E2009 (ASTM International E2009-08, 2008). La mayoría de los problemas de seguridad alimentaria relacionados con los aceites tienen su origen en los procesos de oxidación producidos a las altas temperaturas empleadas cuando se fríe, por lo tanto, la estabilidad a la oxidación térmica, estimada por el OOT, es un parámetro importante en el estudio de aceites comestibles. Adicionalmente, se extrajeron una serie de características con significación física a partir de las curvas PDSC, empleándose conjuntamente con el OOT para poder identificar automáticamente cada tipo de aceite. A modo de conclusión, se investigó la relación entre la composición química y las características extraídas.

Finamelmente, en López-Beceiro et al. (2011a) se empleó un nuevo modelo de regresión compuesto por una suma de componentes logísticas generalizadas para ajustar curvas TG y sus derivadas, de forma que se pudieran estudiar separadamente los procesos de degradación solapados correspondientes a una muestra de alumbre. De este tipo de modelo se sirve el artículo Tarrío-Saavedra et al. (2011), incluído en esta tesis, para seleccionar las características representativas las curvas TG obtenidas a partir de muestras de madera.

El estudio Artiaga et al. (2011) también presenta una aplicación muy útil de los modelos de regresión no lineales al análisis DSC e, igualmente, cabe reseñar una serie de artículos, relacionados en mayor o menor grado con la presente tesis doctoral, en los que ha participado el autor (Tarrío-Saavedra et al., 2012; Sebio-Puñal et al., 2012; López-Beceiro et al., 2012; Gracia-Fernández et al., 2010a,b; Tarrío-Saavedra et al., 2011).

## 1.2. Aplicación del ANOVA funcional en el estudio de la estabilidad térmica de micro-nano compuestos epoxi-humo de sílice

En este capítulo, se ha evaluado el efecto de la adición de humo de sílice (compuesto de nano y micropartículas) en la estabilidad térmica de una resina epoxi mediante un test ANOVA de carácter funcional (FANOVA). Se han considerado tres niveles diferentes de adición de humo de sílice: 0, 10 y 20 wt %. La prueba ANOVA se ha aplicado tanto a las curvas TG reescaladas como a las curvas DTG y previamente a la aplicación del test, se ha realizado un preprocesado, con el objeto de preparar la base de datos convenientemente antes de ser analizada. Con tal motivo se han aplicado técnicas estadísticas como los $b$-splines penalizados y el concepto de profundidad de datos.

Principalmente, la bibliografía empleada en esta parte se compone de las dos monografías que mejor ilustran e introducen al lector en el campo del FDA: la escrita por Ramsay y Silverman (Ramsay y Silverman, 2005) y, en un contexto no paramétrico, la correspondiente a Ferraty y Vieu (Ferraty y Vieu, 2006). Por otro lado, los libros editados por Ferraty y Romain (Ferraty y Romain, 2010) y Ferraty (Ferraty, 2011) presentan algunos de los últimos avances en este campo. Concretamente, en el ámbito de los contrastes ANOVA con respuesta funcional, el test en el que se basa gran parte de este capítulo fue ideado y desarrollado por A. Cuevas, M. Febrero y R. Fraiman en Cuevas et al. (2004). Otros enfoques de este mismo problema se pueden encontrar en Brumback y Rice (1998), Ramsay y Silverman (2005); Ramsay et al. (2009), Fan et al. (1998), Shen y Faraway (2004) y Shen y Faraway (2004). Recientemente, también se han realizado estudios de gran interés para tratar problemas que presentan más de un de factor tratamiento (Cuesta-Albertos y Febrero-Bande, 2010).

## 1.3. Clasificación de materiales

La clasificación supervisada de materiales es un tema de importancia capital en Ingeniería, correspondiéndose con la segunda parte de esta tesis, en la que se estudia el caso particular de la madera. En concreto, se han considerado diferentes especies de interés industrial: roble, haya, castaño, pino insigne, pino rojo, eucalipto, nogal y jatobá. La elección de este tipo de material está relacionada, por un lado, con la importancia relativa que la industria de la madera tiene en Galicia y, por otro, con la particular dificultad

que la clasificación de la madera comporta, debido a su gran heterogeneidad (existente incluso en muestras de un mismo árbol).

La aportación fundamental de los capítulos 3, 4 y 5 a la tecnología de la madera consiste en la utilización y adaptación de nuevas técnicas no paramétricas pertenecientes al análisis de datos funcionales, basadas principalmente en el estimador de Nadaraya-Watson y el algoritmo Adaboost, así como su aplicación a curvas TG y DSC. Su comparación con métodos de clasificación supervisada multivariante representa también una aportación importante. Estos últimos se pueden emplear gracias al desarrollo y/o aplicación de diversas técnicas de extracción de características a partir de curvas TG y micrografías SEM previamente segmentadas: (a) utilización de los parámetros de ajuste correspondientes a un modelo de regresión compuesto por una suma de 4 componentes logísticas, cada una relacionada con el proceso de degradación de cada constituyente principal de la madera (agua, hemicelulosa, celulosa y lignina); (b) las componentes principales que explican el 99 % de la variabilidad de las curvas TG; (c) el área media, circularidad, rectangularidad, número y distancia entre la traqueidas mostradas en las micrografías SEM. El modelo de regresión propuesto ha servido igualmente para llevar a cabo un completo estudio de simulación de curvas TG correspondientes a las especies estudiadas. Mediante este estudio se pueden obtener muestras artificiales en escenarios con diferentes grados de dependencia entre parámetros y con una variabilidad diferente a la obtenida en los estudios experimentales. Ello permite evaluar de una forma más precisa y completa la capacidad discriminante de las curvas TG.

Este estudio abre un camino para el tratamiento de las curvas obtenidas mediante técnicas de análisis térmico como datos funcionales, mientras que las metodologías propuestas proporcionan nuevas vías de estudio y clasificación rápida y, relativamente, eficaz de materiales a partir de sus curvas TG, DSC y las micrografías SEM tomadas a 1500× aumentos.

## 1.4. Objetivos específicos y estructura

Con respecto a la primera parte de este texto, los objetivos principales son los siguientes:

- Evaluar el efecto de la adición de humo de sílice (compuesto de nano y micropartículas) en la estabilidad térmica de una resina epoxi mediante un test ANOVA de carácter funcional (FANOVA), considerando tres niveles diferentes de adición de humo de sílice: 0, 10 y 20 wt % para los que se han obtenido las curvas TG reescaladas y las derivadas de

las curvas termogravimétricas (DTG).

- Determinar el grado de interacción entre el humo de sílice y la resina epoxi, relacionado con la aparicion de una interfase orgánica-inorgánica.

Respecto a la segunda parte de este texto, los objetivos del capítulo 3 son, por un lado, la comparación de diferentes métodos para clasificar siete especies comerciales de madera (castaño europeo, haya europea, eucalipto, jatobá, roble europeo, pino rojo y pino insigne) a partir de sus curvas TG. Estos métodos incluyen técnicas no paramétricas funcionales, utilizando las curvas TG completas, y enfoques multivariados de clasificación supervisada, como LDA, $k$-NN, NBC, NN y SVM. Estos últimos se aplican a las puntuaciones o "scores" correspondientes a las componentes PCA que explican una mayor parte de la variabilidad de las curvas, o a los parámetros relativos al ajuste de un modelo de suma de componentes logísticas generalizadas. Por otro lado, estos métodos de clasificación, aplicados en un contexto funcional, se comparan de una forma más completa a través de un estudio de simulación. De hecho, la generalización de este tipo de estudios en el análisis térmico es uno de los objetivos del capítulo. Las curvas sintéticas obtenidas tratan de imitar las curvas TG en escenarios diferentes al de las curvas experimentales.

Los objetivos del capítulo 4 de este estudio son:

1. Evaluar el potencial de los métodos no paramétricos funcionales de análisis discriminante para la clasificación entre maderas duras y blandas, y para clasificar entre las 7 especies de madera diferentes partiendo de las curvas TG y DSC.

2. Comparar el desempeño de los procedimientos de clasificación a partir de curvas TG y DSC, para distinguir entre las distintas especies y, además, entre maderas blandas (coníferas) y duras (frondosas). Se han utilizado los estimadores kernel (núcleo) y su versión $k$-nearest neighbors ($k$-vecinos más próximos, en lo relativo a la elección de la ventana óptima $h$) (Ferraty y Vieu, 2006). Para completar el estudio, se han aplicado dos métodos de clasificación basados en el algorimo Adaboost y que emplean el análisis PCA y la representación de los datos funcionales según una base $b$-spline (Ramsay y Silverman, 2005, 2002; Bühlmannand y Hothorn, 2007), respectivamente.

3. Encontrar el rango de temperatura en las curvas TG y DSC en el que se alcanza la mayor probabilidad de clasificación correcta.

4. Relacionar la proporción de clasificación correcta, obtenida en cada intervalo estudiado, con los intervalos de temperaturas a los que se

degradan la celulosa, la lignina y la hemicelulosa en una atmósfera inerte (nitrógeno).

Con respecto al capítulo 5, los objetivos se describen a continuación:

1. Verificar que es factible llevar a cabo una correcta clasificación de las especies de madera (también entre maderas blandas y duras), utilizando las características obtenidas a partir de la segmentación de las micrografías SEM, tomadas a 1500× aumentos, en las que se muestra la estructura traqueidas.

2. Mostrar las ventajas de la segmentación de imágenes para el caso particular del problema de clasificación de la madera a partir de sus micrografías.

3. Evaluar el potencial de los métodos de clasificación supervisada, como LDA, clasificación cuadrática, la regresión logística, $k$-NN, NBC, SVM y NN para distinguir entre castaño europeo, haya europea, eucalipto, jatobá, nogal, pino insigne y pino rojo partiendo de las características obtenidas a partir de segmentación de imágenes (también para distinguir entre madera duras y blandas).

4. Valorar, a su vez, el poder discriminante de dos estimadores diferentes de la dimensión fractal, tomados a partir de las micrografías.

La estructura de la presente tesis doctoral es la siguiente: En la Parte I se describe la aplicación del ANOVA funcional en el estudio de la estabilidad térmica de micro-nano compuestos epoxi-humo de sílice. En la Parte II, se aborda el problema de la clasificación de especies de madera: en el Capítulo 3 a partir de curvas termogravimétricas aplicando métodos de clasificación funcionales y multivariantes, en el Capítulo 4 buscando el intervalo de temperatura en el que se obtiene la mayor probabilidad de buena clasificación aplicando métodos de clasificación FDA y, en el Capítulo 5, mediante la aplicación de métodos multivariantes a las características obtenidas mediante segmentación de micrografías SEM. Por último, en el Capítulo 6 se recogen las conclusiones acompañadas de una breve discusión general de los resultados, enumerándose, finalmente, las líneas futuras de investigación.

Nótese que, dependiendo del capítulo, existen diferentes variantes para los acrónimos de algunos conceptos. Esto es debido a que se ha optado por respetar lo más estrictamente posible la versión original de cada uno de los artículos que componen esta tesis.

# Parte I

# Evaluación de micro-nanocompuestos epoxi-humo de sílice

*...Et tant de trucs encore*
*Qui dorment dans le crânes...*
*Tant de choses à voir*
*A voir et à z-entendre*
*Tant de temps à attendre*
*A chercher dans le noir.*

Boris Vian

En la primera parte de esta tesis se presenta el problema de la caracterización de materiales poliméricos; en este caso, una resina epoxi, a la que se le añade una cantidad variable de nano y micro partículas inorgánicas obtenidas como subproducto de la actividad industrial, humo de sílice. Su contenido se corresponde con el trabajo titulado *Application of functional ANOVA to the study of thermal stability of micro-nano silica epoxy composites*, publicado en el año 2010 en la revista *Chemometrics and Intelligent Laboratory Systems* e incluida en el *Journal Citation Report* dentro de la categoría *Statistics and Probability*. El texto aparece integramente en inglés, respetando la versión publicada por la revista. En el Apéndice A, se muestra la versión original de publicación en la mencionada revista.

La principal motivación del estudio que comprende esta primera parte es la evaluación de la estabilidad térmica de los nanocompuestos, variable respuesta funcional (curvas TG, DTG), y su estudio con respecto a la variación de un factor: la cantidad de carga inorgánica añadida. Se elabora un diseño de experimentos de una vía, a 3 niveles del factor cantidad de humo de sílice: 0, 10 y 20 % del peso total de la muestra. La principal novedad de esta parte, que comprende el Capítulo 2, es la utilización de un test ANOVA de carácter funcional, en contraposición con las opciones tradicionales univariantes y multivariantes. La metodología aquí propuesta permite un estudio más preciso e informativo de la variación de la estabilidad térmica en este tipo de compuestos, además de apoyar la hipótesis de la existencia de una interfase orgánica-inorgánica. El estudio de variaciones en las propiedades del material final según la variación de un factor es fundamental en el proceso de elaboración y caracterización de nuevos nanomateriales, así como también lo es la identificación y caracterización de la interfase. Ésta última condiciona en gran medida las propiedades de los compuestos. En el Capítulo 2 también se aplicarán conceptos como la profundidad de datos funcionales y el remuestreo bootstrap para la obtención de bandas de confianza.

# Capítulo 2

# Application of functional ANOVA to the study of thermal stability of micro-nano silica epoxy composites

**Resumen:** El principal objetivo de este capítulo es el uso de una nueva técnica que combina el análisis de datos funcionales (FDA) y el diseño de experimentos, ANOVA funcional de una vía, para medir la influencia que tiene la adición de nano y micro partículas de humo de sílice en la estabilidad térmica de una resina epoxi. Con tal objeto se ha realizado un diseño de experimentos consistente en un factor tratamiento (cantidad de humo de sílice) con tres niveles diferentes (0, 10 y 20 wt %). Los datos se obtienen mediante la aplicación del análisis termogravimétrico (TG), resultando cinco curvas de degradación por nivel. El ANOVA funcional utiliza toda la información correspondiente a cada curva o dato funcional. Los resultados obtenidos mediante esta metodología, a partir de las curvas TG reescaladas y sus derivadas (DTG), indican que la cantidad de humo de sílice empleada afecta significativamente a la estabilidad térmica del compuesto, y de la resina en particular. Este hecho puede ser indicativo de la interacción entre la fase orgánica y las partículas inorgánicas añadidas. Por último, con el objeto de discernir qué niveles del factor provocan diferencias en el camino de degradación de la resina, se ha propuesto el empleo de comparaciones dos a dos, conocidas como "pairwise", usando la metodología del ANOVA funcional, además de un test bootstrap basado en las distancias.

**ABSTRACT:** The main purpose of this work is to use a new technique that combines functional data analysis and design of experiments, functional ANOVA for a one way treatment, to measure the influence of adding fumed silica on the thermal degradation of an epoxy resin. To achieve this, a design of experiments with a treatment factor (the amount of fumed silica) at three different levels (0, 10 and 20 wt %) is performed. The data are obtained through the use of Thermogravimetric Analysis (TG), resulting in five degradation curves per level. The functional ANOVA uses all the information of each curve or functional data. The results obtained using this methodology with the TG rescaled data and their derivatives (DTG) indicate that the amount of fumed silica significantly affects the thermal stability of the compound. These facts may be indicative of the interaction between the organic phase and the inorganic particles. In addition, pairwise comparisons using the functional ANOVA method and a bootstrap distance based test are carried out to discern which factor levels provide different ways of degradation.

## 2.1. Introduction

Although epoxy resins are widely used thermostable polymers, their use is somewhat limited due to the high stiffness, caused by the dense cross-linking structure of these materials. However, their mechanical properties can be improved by the addition of inorganic particles (Harsch et al., 2007). The shape, volume, size, surface characteristics and dispersion of particles within the matrix determine the mechanical properties of resulting composites (Lee y Lichtenhan, 1999; Mehta et al., 2004; Shao-Yun et al., 2008). The nanocomposites with organic matrix and inorganic fillers have proven capacity of providing simultaneous increases in properties such as thermal stability, flame retardation, glass transition temperature and dimensional stability, as well as the decrease of the dielectric constant (Zhang et al., 2006; Preghenella et al., 2005; Yousefi et al., 1997).

In the present work, we perform an experimental design to evaluate the effect of the addition of fumed silica on the thermal degradation of the resulting material. The fumed silica epoxy-resin composites are prepared and characterized by Thermogravimetric Analysis (TG) and Differential Thermogravimetric Analysis (DTG), usual techniques in assessing the thermal stability of a material (Tarrio-Saavedra et al., 2008). A non-conventional epoxy resin based on trimethylolpropane (TMP), particularly suitable for the manufacture of composite materials, is chosen. Moreover, fumed silica used is a byproduct of the manufacture of silicon and ferrosilicon. It is produced at the top of the melting furnaces, thus its production method is different from the

conventional processes for synthetic $SiO_2$ (Mohammad y Simon, 2006). Due to the special characteristics of the production method, the fumed silica used has a variable purity, depending on the operating conditions in the furnaces. In any case, the silica weight ratio is never less than 95 %. It is also variable in particle size. In fact, fumed silica consists of nano and micro-particles, taking into account the Schadler approach (Schadler, 2003) (diameter$<$ 100 nm implies that it is a nano-particle). This particular size distribution suggests a possible combination of micro and nano effects.

To perform the analysis previously mentioned, a new statistical method, called one-way functional ANOVA, is applied. This procedure allows to test the possible differences in responses according to the treatments used, considering that the data are functions or curves. In our case, the silica content in each sample, with three levels (0, 10 and 20 wt %, weight percentage, of fumed silica) is chosen as the treatment factor or explanatory variable. Five experiments or replicates for each level are considered, which gives a balanced design. The number of replicas selected is set to reach an acceptable compromise between the adequate representation of the variability within each level and the total experimental time required. The response variables or dependent variables are functions, where each one is a curve representing the mass of material depending on the temperature at which it is subjected to. To obtain them, a constant increase in temperature of 10 °C/min is scheduled. All the curves decrease because the material degrades, or loses mass, by increasing the temperature to which it is subjected to. Therefore, each curve represents the particular way of degradation of each sample tested. Our analysis allows answer questions like, will the way of degradation be different for different levels of the factor *amount of silica*? or, can it be said that the thermal stability of the material increases or decreases with statistical evidence?

On the other hand, an important part of this study is to determine the degree of interaction between fumed silica and epoxy resin, related to organic-inorganic interphase. This region is defined as starting at the point where the filler differs from the rest of the load and it finishes at the point of the matrix in which its properties are the same as in the rest of the matrix (Schadler, 2003). The existence of this interphase affects some properties, such as thermal stability and glass transition temperature. In fact, the variation of these properties can be taken as an index of its existence (Tarrio-Saavedra et al., 2008).

To be able to compare all the data conveniently, the mass of each sample is expressed as a percentage of the initial amount (Prime, 1997). Thus, all curves start with a value of 100 % in the vertical axis.

The content of the paper is as follows. In Section 2.2, a comprehensive review on the issue of Functional Data Analysis (FDA) is presented, and the statistical method used in our research is described. In Section 2.3 the experimental process carried out to obtain the data is described. Section 2.4 presents a descriptive analysis of the data under consideration as well as the data preprocessing needed to apply the statistical methods. The results and discussion are included in Section 2.5, while Section 2.6 collects the main conclusions.

## 2.2. Statistical methods

In this Section, the main statistical technique used in our analysis is briefly described. This method takes advantage of the functional nature of the data under consideration, producing more reliable results. First, a comprehensive review about FDA is presented.

### 2.2.1. Functional data analysis (FDA)

FDA deals, in general, with experiments whose data and/or results are curves. According to Ferraty and Vieu (2006), we could say that a random variable $X$ is called a functional variable if it takes values in an infinite dimensional space (or functional space normed or semi-complete normed) and therefore, an observation $x$ of $X$ is called a functional data and, in addition, a functional dataset $x_1, \ldots, x_n$ is the observation of $n$ functional variables $X_1, \ldots, X_n$ identically distributed as $X$. The functional data, also called longitudinal data, turn out to be associated with processes continuously monitored in time. That is, when a variable is measured on a discrete and finite set of arranged values, considering that this variable follows a continuous functional relationship. A special case is when the functional variable $X$ belongs to a Hilbert space, as it is the case of continuous functions on an interval Ramsay and Silverman (2005). This is the case of the TG curves used in the present work, $X \in L_2([0, T])$ Naya (2003). In addition, FDA often makes use of the information in the slopes and curvatures of curves, reflected in their derivatives. This is the case of the TG derivatives curves (DTG) also used in our research.

This relatively new research field has received a lot of attention by the scientific community over the last two or three decades, although the study of probabilistic tools for infinite dimensional variables started in the beginning of the 20th century (González Manteiga and Vieu, 2007) . Nevertheless, lately, the interest in FDA methods has considerably increased, since the

technological progress allows collecting observations of infinite dimensional objects. The books by Ramsay and Silverman (2005) and that of Ferraty and Vieu (2006) are good introductory texts about this kind of data. On the other hand, the recent monograph of Ferraty and Romain (2010) presents the latest progress on this topic. Many databases belonging to very different branches of science are likely to be treated as functional data: econometrics, medicine, environmetrics, geophysics, biostatistics and, of course, chemometrics. Apart from the examples studied in Ramsay and Silverman (2005) or in Ferraty and Vieu (2006), there are many other datasets where functional analysis techniques are successfully applied. Many of them are treated in the special issues presented in González Manteiga and Vieu (2007), Valderrama (2007), Ferraty (2010) and Tsiatis and Davidian (2004). For example, climatologic and environmetrical curves are studied in López Pintado and Romo (2007) , economical curves in del Barrio et al. (2007), spectrometric curves in Antoniadis and Sapatinas (2007) , and Ferraty et al. (2007), geophysics curves are analysed in Nerini and Ghattas (2007), and biostatistics datasets are treated as functional data in Antoniadis and Sapatinas (2007) .

Given the wide range of databases that can be studied as functional data, and given the technical possibility of treating them, a considerable effort is being made for adapting the standard statistical methods to the functional context. For example, principal component analysis for functional data are studied in Locantore et al. (1999). Regression models with functional covariates (and scalar or functional response) are analysed in Cardot et al. (1999), Cuevas et al. (2002), Ferraty and Vieu (2002), Cardot et al. (2007) and Dabo-Niang and Rhomari (2009). Functional data classification is other important field in FDA (Ferraty and Vieu, 2003). This terminology includes supervised and unsupervised classification. Finally, in relation with the method used in the present paper, the procedure proposed in Ferraty et al. (2007) is particularly interesting.

Some of the references listed above use nonparametric methods (kernel estimation, wavelets, splines,...) to analyse functional data, for example, Antoniadis and Sapatinas (2007), Cardot et al. (2007), and Ferraty and Vieu (2003). In this framework, it is important to stress the monograph by Ferraty and Vieu (2006), where many of these two authors'contributions to the nonparametric estimation with functional data are summarized. Nonparametric techniques do not assume, in general, any parametric shape for the functions to be estimated or a specific distribution for the variables under consideration. In this sense, nonparametric functional data techniques are powerful and flexible tools.

When working with functional data, previously to the application of a spe-

cific statistical procedure, sometimes, a data preprocessing should be done to prepare the datasets to be analysed. For example, in practice, sample curves are usually observed in a finite set of points. These points can be not equally-spaced or can be different for the observed curves. In this case, the usual step is using a smoothing process, representing the functions in a proper functional basis, for example (using $b$-splines, for instance), to try to recover the original functional relationship. On the other hand, if the curves show a similar pattern, but present variations in their range or in their domain, transformations of the curves can be useful to align special features or to minimize variability. Different procedures are proposed for this task in the literature in many fields. *Marker registration* involves identifying the timing of specified features in the curves, and then transforming the argument of the curves so that these marker events occur at the same moment. A comprehesive reference is Bookstein (1991). *Time warping*, term mainly used in the engineering literature, is applied in the procedure proposed in Salkoe and Chiba (1978), while its statistical aspects are studied in Kneip and Gasser (1988). These methods, however, can be sensitive to errors in feature location, and these features may even be missing in some curves. Moreover, substantial variation in the argument of the curves may remain between widely separated markers. Under the name of *curve registration*, in Silverman (1995), a technique that does not require markers is developed, and in Ramsay and Li (1998) the Silverman's method is extended by using a flexible smooth monotone transformation family developed in Ramsay (1998). These approaches involve defining an entire warping curve for each observation, and they can be somewhat complex to program. In Kneip et al. (2000), a local nonlinear regression technique, computationally convenient, is described for identifying the smooth monotone transformations.

In the analysis of the real curves presented in the next sections, preliminary to the application of the functional ANOVA test, some approaches related with those previously described are used. On one hand, each curve is written as a linear combination of the elements of a functional basis. To select a proper basis, $b$-splines and penalized $b$-splines fits are used. On the other hand, in line with the results pursued with the registration of the curves, in our case, a simple rescaling of the curves can give a clue of possible existence of a chemical interaction between resin and inorganic. A detailed description about this is given in Section 2.4.

### 2.2.2. Functional ANOVA

When the data are functional, an alternative to the classical Analysis of Variance (ANOVA) is the named functional ANOVA (FANOVA) (Cuevas et al., 2004). The covariates are factors while the response is functional. This technique, compared with the classical one, has the advantage of using all the information in the curves, instead of some specific values on them.

Following the nomenclature of Cuevas et al. (2004), each functional data can be written as $X_{ij}(t)$, where $t$ usually represents the time, with $t \in [a, b]$, $i$ is the subscript that indicates the level of factor and $j$ the replication number ($j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$). Variables $X_{ij}(t)$ can be considered as $k$ independent samples of trajectories drawn from $L_2$-processes $X_i$, $i = 1, \ldots, k$.

In the present study, as the variable of interest, *the mass of the sample*, is evaluated originally at every second, the data can be considered functional. In this case, the temperature is directly proportional to the time (10 °C/min), and therefore, $t$ can be considered as the values of the temperature instead of the time.

The mean for each level or independent sample is given by $E(X_i(t)) = m_i(t)$, while the covariance between two specific values of a curve, $Cov(X_i(s), X_i(t))$, in the most restrictive case (existence of heteroscedasticity), can be estimated by $K_i(s, t)$:

$$K_i(s,t) = \sum_{j=1}^{n_i} \frac{\left(X_{ij}(s) - \overline{X}_{i\cdot}(s)\right)\left(X_{ij}(t) - \overline{X}_{i\cdot}(t)\right)}{n_i - 1} \tag{2.1}$$

We want to test:

$$H_0 : m_1 = m_2 = \ldots = m_k \tag{2.2}$$

The statistic implemented by Cuevas et al. (2004) to test (2.2) is as follows:

$$V_n = \sum_{i<j} n_i \left\| \overline{X}_{i\cdot} - \overline{X}_{j\cdot} \right\|^2 \tag{2.3}$$

The use of (2.3) avoids the requirement of the hypothesis of homoscedasticity in the usual ANOVA.

Assuming:

1. $n_i, n \to \infty$ in such a way that $n_i/n \to p_i > 0$ for $i = 1, \ldots, k$,
2. and the observations $X_{ij}(t)$ with $j = 1, \ldots, n_i$, corresponding to independent samples of size $n_i$ from $k$ $L_2$-processes with mean zero and

covariance $Cov(X_i(s), X_i(t))$,

it can be proved that the asymptotic distribution of $V_n$, under $H_0$, coincides with that of the statistic

$$V = \sum_{i<j}^{k} \|Z_i(t) - C_{ij} \cdot Z_j(t)\|^2 , \tag{2.4}$$

where $C_{ij} = (p_i/p_j)^{1/2}$ and $Z_1(t)$, ..., $Z_j(t)$ are independent Gaussian processes with mean zero and covariance $Cov(X_i(s), X_i(t))$.

To apply the test, if the $n_i$ are large enough, hypothesis $H_0$ is rejected, at a level $\alpha$, whenever $V_n > V_\alpha$ where $P_{H_0}(V > V_\alpha) = \alpha$.

In a practical situation, the distribution of $V$ under the null hypothesis can be approximated by applying a parametric bootstrap and the Monte Carlo method. This allows to estimate the value of the $\alpha$-quantile, $V_\alpha$.

In this case, the use of the parametric bootstrap is justified because the distribution of $V$ is a complicated function of $k$ Gaussian processes. This bootstrap procedure consists of resampling from these Gaussian distributions, but replacing their parameters with the maximum likelihood estimators calculated from the original sample. A detailed description of this procedure with the data analysed in this paper can be found in Section 2.5.

In the previous procedure the word *functional* in the term functional ANOVA refers to the functional nature of the data. A similar approach to the one used here has been already considered in Ramsay and Silverman (2005), and Ramsay et al. (2009), but there the problem is formulated from a regression point of view and a (pointwise) $F$ statistic at each $t$ is proposed. However, in the same functional context as here, there are several works developing an $F$ test defined by an unique value. In Fan et al. (1998), a HANOVA (high dimensional ANOVA) test, relying on wavelet thresholding techniques, which limits the size of each curve (discretized curves) to satisfy some asymptotic assumption, is proposed. A $F$ test for choosing among two nested functional linear models is developed in Shen and Faraway (2004). A discussion of the ways in which the functional ANOVA can be treated is shown in Brumback and Rice (1998). Other interesting reference is Ferraty et al. (2007), where standard approaches based on factorial analysis for comparing groups of multivariate data are extended to the infinite-dimensional framework. One of the differences between this and other tests and the one applied in the present article is in the discretization process. As stated in Cuevas et al. (2004), the test applied here is purely functional in the sense that our test statistic is a functional of the sample trajectories and its motivation is also given in functional terms. Since it is conceivable that the increasing interest on functional data could lead to measurement devices whose

outputs provide *true* functions, with *analytical* expressions (obtained maybe by nonparametric smoothing) instead of finite dimensional approximations, our procedure could give more reliable results.

It is very important to note that the term *functional* ANOVA is also used for different models to the one used here by some authors. Specifically, they employ these words when a multivariate function is represented by a decomposition in terms of functions of fewer variables, linked with non-linear models. In that context, there are also substantial works where functions of the predictors are estimated and tested. Several applications of this methodology can be seen in the following papers. In Stone et al. (1997), analysis of variance type models are considered for a regression, conditional probability, density and hazard functions using polynomial splines to model the main effects and interaction components. In Huang et al. (2000), the logarithm of the relative risk function in a proportional hazards model involving one or more possibly time-dependent covariates is treated as a sum of a constant, main effects and selected interaction terms. Moreover, in Kawaguchi et al. (2008), a methodology for modeling covariate effects on the time-to-event data is developed using polynomial splines. Therefore, it is important to point out that the statistical community is (curiously) using the same wording (FANOVA) for two very different models.

## 2.3. Materials and instrumental methods

An epoxy resin matrix consisting of two components is used. It is based on the diglycidyl ether of trimethylolpropane, Triepox GA, manufactured by Gairesa, SA. The curing agent used is an aromatic amine, 1,3-benzenedimethanamine, supplied by Aldrich. Triepox GA is a highly thixotropic resin that also possesses a low density and the ability to cure at room temperature in the absence of plasticizers or additives.

The fumed silica has been provided by Ferroatlántica I + D, Spain. It is obtained as a byproduct in the production of silicon in electrical melting furnaces. This process involves the reduction of high purity quartz, at temperatures above 1800 °C. Fumed silica is formed when SiO gas, resulting from the reduction of quartz, is mixed with oxygen at the top of the furnace, resulting in the production of spherical particles of silica. It is a fine powder varying in colour from nearly black to slightly off-white, according to their carbon content. Its average particle size is 0,15 microns and 41,9 % of the particles have a diameter less than 0,2 microns, as shown in Table 2.1. The surface area is about 20 $\mathrm{m^2\,g^{-1}}$.

As regards to the chemical composition, fumed silica consists of variable

| Diameter ( $\mu$m) | Mass ( %) |
|---|---|
| $50 - 100$ | 1,2 |
| $20 - 50$ | 2,0 |
| $10 - 20$ | 0,2 |
| $5 - 10$ | 0,5 |
| $2 - 5$ | 1,4 |
| $1 - 2$ | 1,5 |
| $0{,}5 - 1$ | 8,3 |
| $0{,}2 - 0{,}5$ | 43,0 |
| $< 0{,}2$ | 41,9 |

Tabla 2.1: Particle size distribution in the fumed silica.

purity amorphous $SiO_2$. Table 2.2 shows the main physical properties and composition.

| | |
|---|---|
| Moiseture 110 °C | 0,50 % |
| Loss on ignition at 1000 °C | 2,78 % |
| Real density | 2,26 g cm$^{-3}$ |
| Apparent density | 0,66 g cm$^{-3}$ |
| $SiO_2$ | +95 % |
| CaO | 0,68 % |
| MgO | 0,22 % |
| $Na_2O$ | 0,10 % |
| $K_2O$ | 0,22 % |
| Cl | 0,006 % |
| $SO_4$ | 0,076 % |

Tabla 2.2: Physical properties and chemical composition of fumed silica.

The samples are prepared for contents of 0, 10, and 20 wt % of fumed silica. Both resin and hardener are mixed in a stoichiometric ratio. To obtain the compounds corresponding with 10 and 20 wt %, the silica and resin mixtures are stirred for 15 minutes in order to obtain a distribution as uniform as possible. Then, an ultrasonic treatment is applied for 5 minutes at room temperature to disperse the silica agglomerates. The paste thus obtained is poured into a silicone mold with cavity dimensions of $0{,}8 \times 4 \times 30$ mm. In this area, the samples are cured at room temperature for 24 hours and then a post-curing is applied at 90 °C for 2h.

The TG experiments are carried out using a thermo-balance STA 1500, Rheometric Scientific. All samples are subjected to a heating ramp of 10 °C/min in a temperature range between 20 and 600 °C. All experiments are performed under nitrogen atmosphere, maintaining an air flow of 50 $\mathrm{mL\,min^{-1}}$.

## 2.4. Descriptive analysis and data preprocessing

In order to obtain further information from the experiments performed, three different datasets are considered and analysed: the TG curves directly obtained from the experimentation, the rescaled TG curves (used to find the true way of the epoxy matrix degradation within the composite, obtained by removing the mass at the end of each experiment, fumed silica, since the fumed silica is not degraded), and the derivatives with respect to the temperature of TG curves (DTG).

To investigate the amount of silica influencing on the thermal stability of the composites, 19 experiments are conducted: 7 with unloaded epoxy resin, 7 with 10 wt % silica content, and 5 with 20 wt % (see Figure 2.1). Each experiment corresponds to a functional datum which represents the mass of the sample in functional relation with the temperature at which it is carried out. As already indicated, each sample is heated to a rate of 10°C/min with temperatures ranging from 20 to 600 °C. At the end of each 600 °C trial, the organic phase (epoxy resin) is completely degraded, leaving only the added mass of fumed silica, which is much more heat resistant. It is important to note that the mass of the sample is represented in percentage, that is, the initial mass is assigned to the 100 % value.

Originally, each curve consists of a variable number of points around 3480, one per second, depending on the environment temperature at which the testing machine is. To facilitate further calculations, without losing information, 581 points are chosen within each experimental curve, one for each Celsius degree, in a range of temperatures from 20°C to 600°C.

The first step in our study is to find a proper functional basis to write each curve as a linear combination of the elements of it, and to achieve a smooth functional relationship. Therefore, each functional datum is represented, discretized, by a finite basis, so that an explicit form for the function is obtained (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006; Cuevas

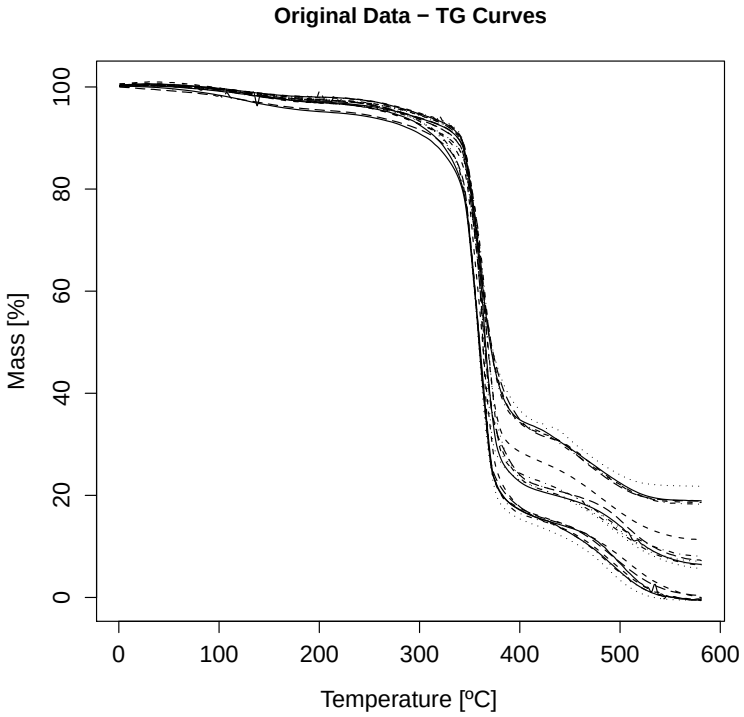


Figura 2.1: Experimental data: starting TG curves.

et al., 2006a),

$$y(t) = X(t) + \varepsilon(t) = \sum_{k=1}^{n} c_k \phi_k(t) + \varepsilon(t), \tag{2.5}$$

where $\phi_k$, with $k \in N$, is a set of known and independent functions, such that any function can be approximated by the linear combination of $K$ of these (elements of the basis). On the other hand, $\varepsilon(t)$ is the experimental error non explained by the adjusted model.

Two procedures to select an appropriated basis are tested: $b$-splines and penalized $b$-splines. Both techniques provide bases with the required flexible structure. On the other hand, Fourier bases would not be as appropriate here as the data do not show an apparent periodic trajectory. Moreover, given the already smooth appearance of the original data, it seems reasonable not to test Wavelets bases.

Generally speaking, a $b$-spline is a spline (a function defined piecewise by polynomials) that has minimal support with respect to a given degree,

smoothness, and domain partition. It can be written as a linear combination of normalized $b$-splines blending functions and some control points.

To avoid very variable fits, penalized $b$-splines can be used. The idea of this procedure is to fit a $b$-spline, but penalizing the variance of the fit represented by the second derivative of its density function. Specifically, the amount of residual fitting for a penalized $b$-spline basis respond to the expression

$$PRSS = \|Y - X\beta\|^2 + \lambda \int \left(f''(x)\right)^2 dx, \tag{2.6}$$

where $Y$ is the data to be fitted, $X\beta$ is the $b$-spline fitting, $f''(t)$ is the second derivative of the fit, and $\lambda$ the smoothing parameter (which penalizes the second derivative, that is, it restricts the internal variance the fit may have). More information can be found in Ferraty and Vieu's work (Ferraty and Vieu, 2006), or in Ramsay and Silverman (2005, 2002).

To choose the number of elements in the optimal basis, the Generalized Cross Validation (GCV) criterion is applied. Bases with number of elements ranging between 4 and 480 are tested. The number of elements corresponding to the minimum GCV for each functional datum is chosen. Finally, the number of elements in the basis is selected as the minimum of the minima computed for each one of the 19 GCV expressions (each one for each functional datum). This prevents against the risk of oversmoothing.

The result of minimizing the GCV criterion for one of the original curves can be seen in Figure 2.2.

Table 2.3 shows the number of elements in the optimal basis according to the GCV criterion. It shows that for $b$-splines, the optimal GCV is smaller than that obtained for penalized $b$-splines, but the number of bases is too large, with the risk of interpolating the data. However, as observed in that table, a acceptable GCV is obtained with a basis of 80 elements. In that case, a smooth fit without departing from the path of the original data is obtained. A similar GCV is obtained in the case of penalized $b$-splines with a basis of 80 elements. Figure 2.2 (right panel) corroborates these arguments. Additionally, in Figure 2.2 (left panel), it can be observed that the GCV decreases sharply for a given number of elements in the basis. This supports the decision to opt for a smaller basis, corresponding to a number of elements closer to the beginning of this drop. If a smaller number of elements in the base were selected, an unacceptable model error would be possibly obtained. In that case, the fits would be far from experimental data in the step slope changes, where it is very important that the data are faithfully reproduced by the fits. Therefore, a four order penalized $b$-spline basis with 80 elements is selected. The fit is perfect, very slightly smoother than that obtained with

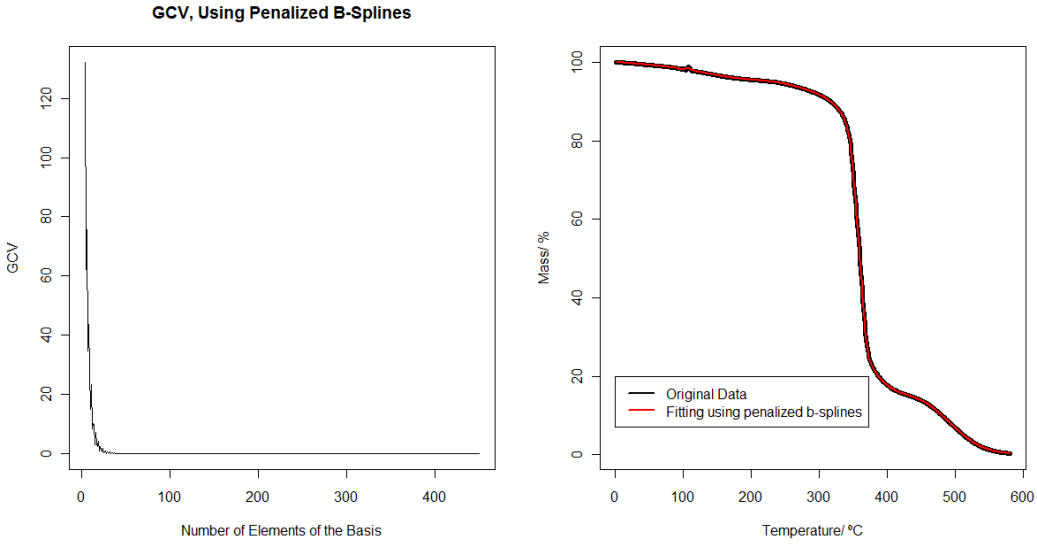


Figura 2.2: Left panel: GCV versus the number of elements in a penalized $b$-spline basis, for a given functional data. Right panel: Experimental datum (epoxy resin) and fitting with a penalized $b$-spline basis with 80 elements.

a $b$-splines basis. In addition, the small number of data does not cause a large computational cost.

| BASIS | Optimal GCV | N° of Elements in the Basis | GCV (with 80 elements) |
|---|---|---|---|
| B-Splines | $2{,}0 \cdot 10^{-7}$ | 375 | $3{,}1 \cdot 10^{-3}$ |
| Penalized B-Splines ($\lambda = 0{,}5$) | $3{,}1 \cdot 10^{-4}$ | 182 | $6{,}4 \cdot 10^{-3}$ |

Tabla 2.3: Number of elements in the optimal basis according to the GCV criterion.

The smoothed functional data using a penalized $b$-spline basis of 80 elements can be seen in Figure 2.3.

### 2.4.1. Descriptive analysis of rescaled TG data

Each level of the factor appears in Figure 2.3 relatively well differentiated from the others. The differences are especially found at high temperatures.

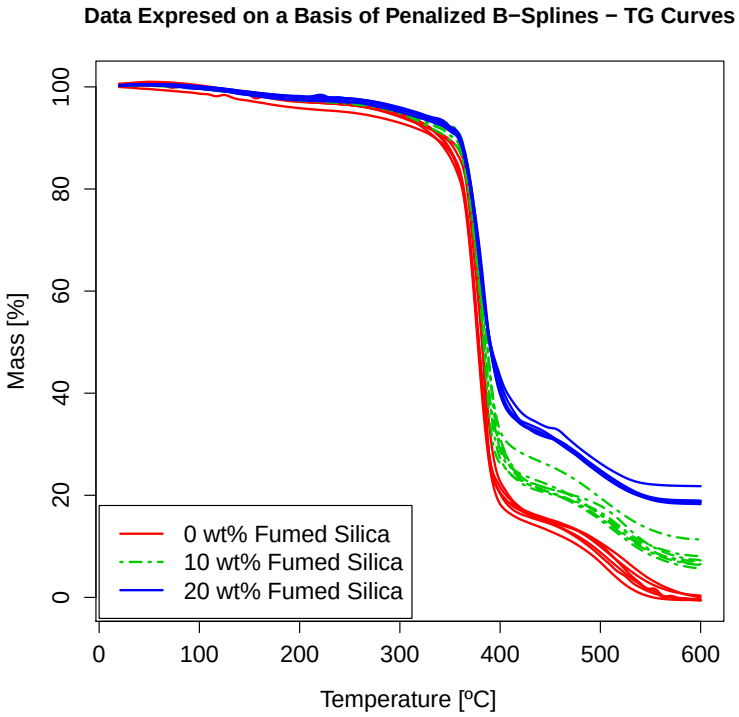


Figura 2.3: Functional data smoothed with a penalized $b$-spline basis of 80 elements.

The resin is degraded gradually. The resin without fumed silica is completely degraded while there is still around 10 and 20 wt % of mass for groups with 10 to 20 wt % of fumed silica.

At a first glance three steps in the curves can be observed. The first, not always perceptible, corresponds to the loss of volatiles and moisture (located at temperatures around 100 °C). The second step is singularly important, as it shows the main degradation process. When going from 0 to 10 wt % of fumed silica content, the mass remaining at the beginning of this step is slightly larger at a given temperature (e.g. around 320 °C). This means that, apparently, the thermal stability of the resulting composite material increases. However, when moving from 10 to 20 wt % of fumed silica, this increment is much lighter. The third step corresponds to the disappearance of the carbon residue resulting from the above reaction. Here we can see that the differences are very evident among the different factor levels, mainly because the fumed silica is not degraded at these temperatures.

If we look at the original TG curves, it is clear that the addition of fumed silica to an epoxy resin leads to differences in the path of degradation. In turn, since the curves are riding on each other by increasing the amount of silica added, it is also evident that its thermal stability increases (a less quantity of mass is lost at the same temperature). This is mainly because a material that does not degrade in the temperature range tested is being added: fumed silica. Are there other reasons?

One might then ask oneself these questions: does fumed silica really interact with epoxy resin forming an interphase?, does the addition of inorganic phase influence in the way of degradation of the epoxy resin put into each sample?, and what is the best way to study the degradation of the neat resin (without taking into account the added inorganic matter)?

The answer lies in the rescaling of the data (Tarrio-Saavedra et al., 2008). First, the mass at the end of the experiment is subtracted for calculating the TG curve only for the epoxy resin. This gives its way of degradation. This mass corresponds to the real mass of fumed silica added, that does not degrade itself. Then, each curve is rescaled, so that the initial value corresponded to 100 % of the sample and the end to 0 %, according to the degradation curve with epoxy resin. Once all the data obtained are rescaled, if significant differences in TG curves between factor levels are observed, a clue of the possible existence of a chemical interaction between resin and inorganic fillers will have been discovered.

The curves of Figure 2.4 represent the way of degradation of the epoxy resin (neat epoxy resin) in each sample of composite material. Eliminating mathematically the inorganic mass proportion (inert mass), it is noted that the curves corresponding to 10 and 20 wt % levels are more similar to the way of degradation of the epoxy resin without silica. If it were possible to prove that at least one of the means of a group is different from the others, it would be possible to prove the existence of an organic-inorganic interphase. In fact, in Tarrío-Saavedra et al. (2010a), dynamic mechanical tests (DMA) performed on the same material already suggested the existence of this interphase. Therefore, the results of the present study could support the DMA ones.

For simplicity, a balanced design is intended to be performed. Therefore, two functional data for the 0 wt % level and two others for the 10 wt % level should be rejected. In fact, there are some slight differences in the experimental conditions for some experimental data. Therefore, it would be interesting to explore their depth.

Depth is a concept that explains how central a datum is with respect to a set of points belonging to a population. Following this criterion, different

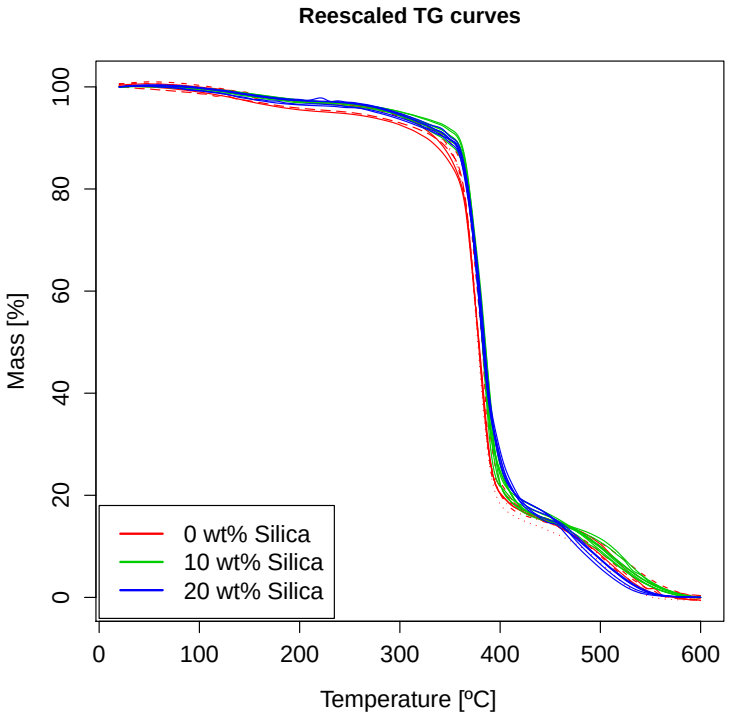


Figura 2.4: Functional data: rescaled TG curves.

functional data, from a sample of a given population, can be sorted: deeper curves are identified as closer to the center (the deepest datum is defined as the median) (Cuevas et al., 2006a).

The depths given by Fraiman and Muniz (2001) (FM), called the Median depth, the Mode depth (the deepest point would be equivalent to the mode of the data) (Cuevas et al., 2006a), and the depth of Random Projections (RP) (Cuevas et al., 2006b) are calculated for the groups of 0 and 10 wt %, separately. We are interested in reducing the number of data just for these groups. The results appear in Tables 2.4 and 2.5.

The curves chosen as less deep are the sample 7 (for the 0 wt % in silica group) and the sample 8 (for 10 wt %). In the case of sample 7, the three criteria coincide, while for the sample 8, there is an overlap in the criteria RP and Mode, being the second less deep datum according to FM criterion. Note the numbers in bold in the tables.

It is still necessary to eliminate another curve in these two groups. Sample 3 is chosen in the 0 wt % group. This curve is the less deep datum according

| Depth | Sam.1 | Sam.2 | Sam.3 | Sam.4 | Sam.5 | Sam.6 | Sam.7 |
|---|---|---|---|---|---|---|---|
| FM | 0,766 | 0,757 | 0,746 | 0,766 | **0,718** | 0,799 | **0,661** |
| Mode | 0,645 | **0,345** | 0,480 | 1,000 | 0,725 | 0,910 | **0,000** |
| RP | 0,247 | 0,233 | **0,215** | 0,296 | 0,251 | 0,262 | **0,210** |

Tabla 2.4: Depths for the 7 samples of epoxy resin without fumed silica according to the 3 criteria. Rescaled data.

| Depth | Sam.8 | Sam.9 | Sam.10 | Sam.11 | Sam.12 | Sam.13 | Sam.14 |
|---|---|---|---|---|---|---|---|
| FM | **0,701** | **0,665** | 0,760 | 0,704 | 0,747 | **0,761** | 0,873 |
| Mode | **0,000** | 0,249 | 0,301 | 0,562 | 0,687 | **0,070** | 1,000 |
| RP | **0,202** | 0,242 | 0,239 | 0,258 | 0,277 | **0,202** | 0,294 |

Tabla 2.5: Depths for the 7 samples of epoxy resin and 10 wt % silica according to the 3 criteria. Rescaled data.

to the RP criterion and the third less deep according to the Mode and FM criteria. In the case of 10 wt %, RP and Mode criteria select sample 13. So, this time, sample 13 is selected. It is worth noting that the curves identified as less deep correspond to that obtained with slightly different experimental conditions (a smaller quantity of experimental mass, not homogeneous dispersion of the load in the matrix and, therefore, slightly different values in 10 wt % curves). Then, those data selected as less deep, that is, samples 3 and 7 for 0 wt %, and samples 8 and 13 for 10 wt % of fumed silica, are finally removed.

With 15 functional data, 5 per level, mean and median from the functional data are calculated. Figure 2.5 shows the mean of each group (0, 10 and 20 wt %) jointly with the confidence bands obtained using the bootstrap naive approach. In Figure 2.6, the medians and the corresponding confidence bands, calculated using a smoothed bootstrap method with a smoothing parameter $h = 0{,}07$ are presented. Smoothed bootstrap is mainly used in cases where the simple bootstrap does not provide a clear picture of what may be the confidence interval, requiring the addition of an additional random component (to fill gaps, obtaining confidence bands). The mean is given by:

$$\overline{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t), \tag{2.7}$$

and the medians are computed using the depth concept of Fraiman and Muniz (2001).

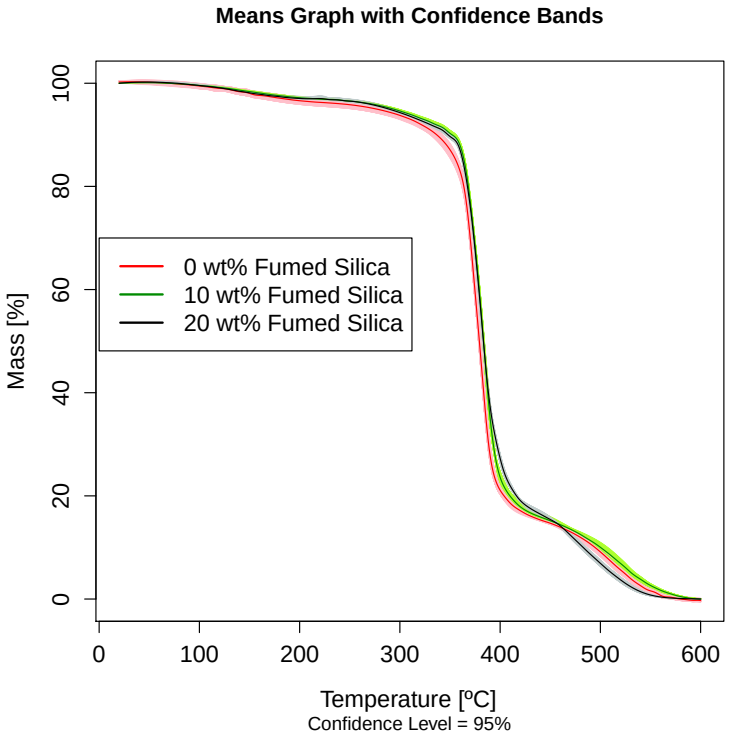


Figura 2.5: Means of each group of data, with confidence bands developed using bootstrap.

Figures 2.5 and 2.6 show that there is a greater overall variability in the data with the epoxy resin alone. The variability decreases slightly for the 10 wt % group and, finally, is much smaller for 20 wt % (this can be seen in the confidence bands of mean and median). This trend may be due to the heterogeneity of the samples or to the learning effect of the operator. In fact, the first samples tested correspond to 0 wt %, and the last ones to 20 wt % (those with less variability). In the case of 0 wt %, another possible reason to take into account is the testing of samples with different moisture contents.

Comparing the statistics (mean and median) of the rescaled curves with those of the original data, one can see that the differences between the curves are substantially lower in the rescaled case. Still, in Figure 2.4 and 2.5, different ways are observed according to the group, with very little dispersion. The differences are especially conspicuous between the groups 0 and 10 wt %, or the 20 wt % class. On the other hand, the differences are slighter between 10 wt % and 20 wt % levels; they seem mainly differ in the last step of degradation. In the region of the second step, the most important, the mean of the

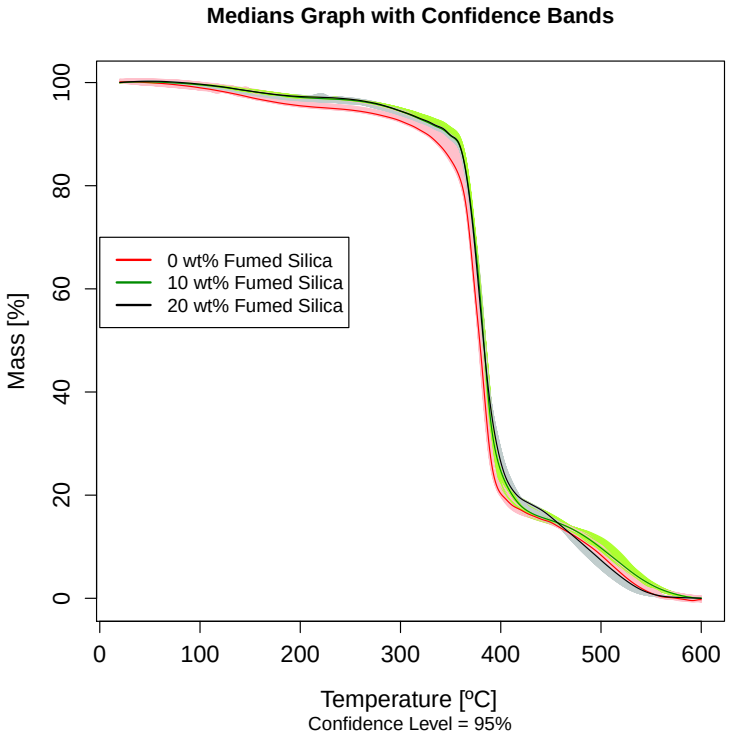


Figura 2.6: Medians of each group of data, with confidence bands developed using smoothed bootstrap.

10 and 20 wt % groups are significantly above the mean of the 0 wt % group. As a result, the thermal stability increases when going from a level of 0 wt % fumed silica to any other. The difference is especially noticeable at the top of the second step, which is the area where the thermal stability of a material is commonly evaluated. However, smaller differences are observed between the degradation pathways for 10 and 20 wt % groups. On the other hand, some differences are observed in the third step: the 20 wt % curves start located at a higher level to then fall much more sharply than those of other groups. Definitely, the addition of 20 wt % of fumed silica causes a decrease in thermal stability in the last stage of the third step of degradation. In this last stage only a char residue exists. The limits of the confidence bands (especially for the means, which are those to be used to build the statistic of functional ANOVA) could suggest that these findings, done from a descriptive analysis of the data, have statistical significance. This is shown in Section 2.5.

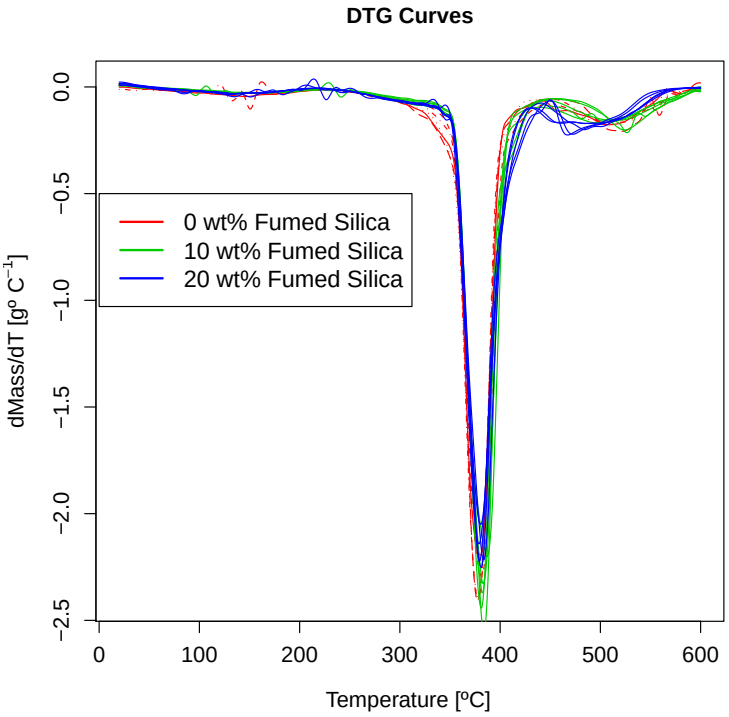


Figura 2.7: DTG curves for the different groups.

### 2.4.2. Descriptive analysis of DTG curves

The study of the Differential Thermogravimetric Analysis (DTG) curves is very common in thermal analysis. Through the study of the derivative, apart from supporting the above results, we intend to complete the study by tackling an important concept: the degradation rate of the samples. Does the amount of fumed silica significantly affect the degradation rate of the epoxy? To answer this question, the derivatives of the rescaled TG data using the statistical package R (R Development Core Team, 2008) are calculated (Figure 2.7). Then, each data is fitted using a penalized splines basis consisting of 60 elements, obtaining a GCV= 0,006 and a spar coefficient equal to 0,35. Spar is the integrated squared derivative of order 2 which controls the amount of smoothing. It is related with lambda parameter. The optimum is obtained using a basis of 124 elements, giving GCV= $9{,}32 \cdot 10^{-6}$, but due to the same reason as in the rescaled data, a smaller basis is chosen. The parameters are selected in order to smooth the data conveniently in the derivative, with much more noise than that in the original data.

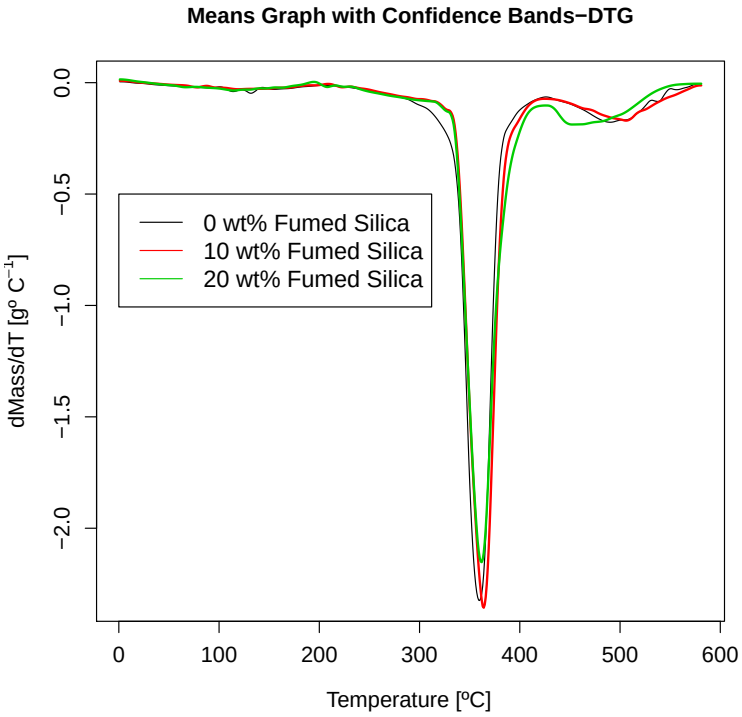


Figura 2.8: DTG mean curves for each group.

In the case of the DTG curves, the semi-metric given by:

$$d(X_i, X_j) = \sqrt{\int_a^b \left(X_i'(t) - X_j'(t)\right)^2 dt}$$

is used

The mean of the DTG curves are plotted in Figure 2.8.

As in the case of not derived rescaled data, differences in mean between groups seem to be small.

The main differences between groups are observed at the beginning and the end of the main degradation process (Figures 2.7 and 2.8). At the beginning of the main degradation process, the degradation rate in 0 wt % group is significantly greater than in groups of 20 and 10 wt %. The DTG mean curves of these two groups are over the first one. Nevertheless, the differences between the degradation rate in 10 and in 20 wt % groups are smaller. On the other hand, at the end of the main degradation process, the degradation rate in 0 wt % group is slightly lower than in 20 and 10 wt % levels. However, the

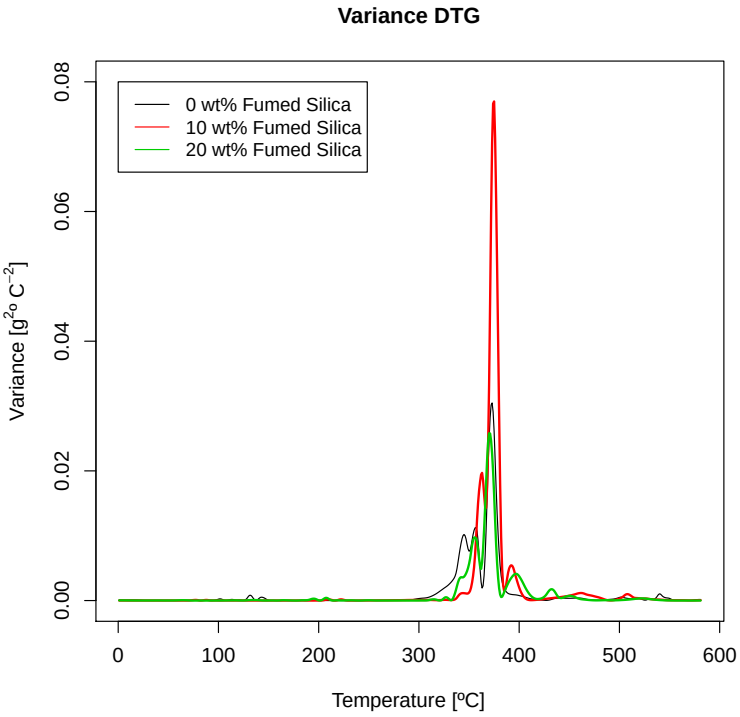


Figura 2.9: Functional variance curves.

differences are slighter between 10 wt % and 20 wt % groups. Additionally, at the beginning of the last degradation process, the degradation rate in the 20 wt % group is significantly greater than in other groups. The addition of 20 wt % of fumed silica accelerates the degradation of the epoxy resin at very high temperatures. This is consistent with the conclusions obtained using the rescaled data without deriving. It may be due to the confinement effect of silica agglomerates on the resin, or even to the fact of better heat transmission within the sample through thinner layers of resin. Finally, the temperature at which the maximum mass loss rate occurs, hardly varies. Although it decreases slightly in module for the group of 20 wt %.

In Figure 2.9, the functional variance in each one of the three groups of both the TG and DTG curves are shown. It can be observed that they are very small throughout the range, compared with the values of the functional means (Figures 2.5 and 2.8). Therefore, the variability per group is properly represented with just 5 curves per level, optimizing the total experimental time required.

## 2.5. Results and discussion

Once the data to be analysed were described in the previous Section, now, the statistical methods presented in Section 2.2 are applied to those data in order to measure statistically the influence of adding fumed silica on the thermal degradation of an epoxy resin. The results obtained also give statistical significance to some of the conclusions derived in the previous descriptive analysis.

The FANOVA test described in Subsection 2.2.2 is applied to the rescaled TG curves and also to the DTG curves. The null hypothesis to be tested is

$$H_0 : m_1 = m_2 = m_3,$$

where $m_i$ is the mean of the functional data within each of the three levels studied belonging to the factor *amount of fumed silica*.

First, the statistic $V_n$ given in Eq. (2.3) is calculated. The number of replicates for each one of the three levels (0, 10 and 20 wt % of fumed silica), $n_i$, is equal to 5 (balanced design). As explained in Subsection 2.2.2, under the null hypothesis, the asymptotic distribution of $V_n$ coincides with that of the statistic $V$, given by Eq. (2.4). The distribution of $V$, under $H_0$, can be approximated by using a parametric bootstrap and Monte Carlo with the following steps:

1. The variance-covariance matrix (assuming heteroscedasticity) for each level is estimated, computing $K_i(s,t)$ in Eq. (2.1), where $s$ and $t$ are two given moments within each functional data.

2. Starting from the original sample, matrices $K_i\left(t_p,t_q\right)_{1\leq p,q\leq m}$ are computed. Next, $B = 2000$ bootstrap resamples by level, following a normal distribution with zero mean and covariance matrix given by $K_i\left(t_p,t_q\right)_{1\leq p,q\leq m}$, are generated.

3. Therefore, 2000 values, $Z_{il}^* = (Z_{il}^*(t_1),\ldots,Z_{il}^*(t_m))$, by level are obtained, with $l = 1,\ldots,2000$ and $i = 1,2,3$. These values approximate the $Z_i(t)$ continuous paths or trajectories with discrete versions, evaluated in a grid $a \leq t_1,\ldots,t_m \leq b$.

4. Finally, 2000 values of the expression

$$\hat{V}_l = \sum_{i<j} \left\| Z_{il}^*(t) - C_{ij}\cdot Z_{jl}^*(t) \right\|^2 .$$

   are calculated. These values approximate the distribution of $V$ when $H_0$ is true.

5. Then, it is possible to estimate the $\alpha$-quantile (denoted by $V_\alpha$) such that, $P(V > V_\alpha) = \alpha$, under $H_0$, for any $\alpha$.

6. If $V_n > V_\alpha$, the test is significant, and at least one of the functional means is different from the others.

The result of the application of this procedure in the case of the rescaled TG curves is the following: $V_n = 34470{,}8$ and $V_\alpha = 1235{,}671$. Therefore, $V_n >> V_\alpha$. The test is highly significant. At least one of the functional means is different from the others. This agrees with the arguments explained in the descriptive analysis of the data (Subsection 2.4.1). The addition of silica causes changes in the functional means of the rescaled TG curves. The thermal stability of epoxy resin, which forms part of the composite material, undergoes a highly significant statistical increment with the addition of an increased amount of fumed silica (it supports higher temperatures before degrading). It could be said that its particular path of degradation has a different shape with the addition of fumed silica, at least for one level. This was the main aim of our research. This is the indicator of an interaction between the epoxy resin and fumed silica, which could result in the creation of a matrix-filler interphase. This result is supported by the DMA tests performed with the same material (Tarrío-Saavedra et al., 2010a). At least one level is different to the others, but what levels are really different?

However, once it is statistically proved that these difference exist, the next step of our research is trying to find out what groups are really different. This kind of analysis would correspond to the named *post-hoc analysis* in the context of design of experiments. To our knowledge, there are no specific procedures for this task in the setting of functional data. Nevertheless, since only three groups are considered here, three pairwise comparisons, using the same functional ANOVA method, can be used as a first attempt to tackle this problem. To correct the problem of multiple testing, a Bonferroni correction is used (Miller, 1991). The idea behind this approach is to consider a new significance level, $\alpha_{\text{Bonf}} = \frac{\alpha}{J}$, $J$ being the number of groups to be compared ($J = 3$, in our case), and compute individual tests using this new level. Table 2.6 shows the results of all pairwise comparisons with the functional ANOVA test using $\alpha = 0{,}05$ and $\alpha_{\text{Bonf}} = \frac{0{,}05}{3} \approx 0{,}015$. As it can be observed, according to these tests, the three groups are significant different.

It is possible that the previous approach can suffer of lack of power, due to the small sample size in each one of the applications of the test (just 10 curves). To solve this drawback, additionally to the previous proposal, we perform a Tukey-like test specially designed for this situation of functional data. Following analogous ideas as those used in the Tukey test, the method

| Groups compared | $V_{0,05}$ | $V_{0,015}$ | $V_n$ | Result |
|---|---|---|---|---|
| 0 wt %–10 wt % | 1506,59 | 2252,66 | 13654,93 | Significant |
| 0 wt %–20 wt % | 1439,34 | 2135,89 | 16199,65 | Significant |
| 10 wt %–20 wt % | 774,42 | 1121,88 | 4616,23 | Significant |

Tabla 2.6: Pairwise comparisons using the functional ANOVA test with TG curves.

| Groups compared | Mean distance | Quantile | Result |
|---|---|---|---|
| 0 wt %–10 wt % | 0,058 | 0,038 | Significant |
| 0 wt %–20 wt % | 0,103 | 0,038 | Significant |
| 10 wt %–20 wt % | 0,045 | 0,038 | Significant |

Tabla 2.7: Pairwise comparisons using a Tukey-like bootstrap test with TG curves.

proposed consists in computing the differences of pairwise curve means, using the $L_2$ distance, and comparing these values with the corresponding quantile of the distribution of the maximum distance between the means of the groups, under the null hypothesis of no difference between groups. The distribution under the null hypothesis is approximated using a bootstrap procedure. Assuming that there are no differences between groups, resamples (with replacement) from the original 15 curves are obtained. These are divided in three subsets and the maximum of the distance between the pairwise curve means of these groups is obtained. This process is repeated a large number $B$ ($B = 1000$ in our case) of times and the corresponding quantile is computed. The result of the application of this procedure to our data with $\alpha = 0{,}05$ is shown in Table 2.7.

The previous naive bootstrap test could be affected by possible outliers due to the small sample size, so that, we also perform a parametric bootstrap. The only difference with respect to the naive method is in the resampling process. Under the null hypothesis, the resamples are now obtained as the sum of the mean curve (computed with 15 original curves) and a error curve generated from a Gaussian distribution with zero mean and variance-covariance matrix estimated form the original 15 curves. In this case, the curves as treated as a of 581-dimensional vector (the number of values where the curves are evaluated) in order to generate new curves. The quantiles obtained with this parametric bootstrap approach are slightly different to those shown in Table 2.7, but the conclusion is the same: the three groups are statistically different.

| Groups compared | $V_{0,05}$ | $V_{0,015}$ | $V_n$ | Result |
|---|---|---|---|---|
| 0 wt %–10 wt % | 4,135 | 7,051 | 25,348 | Significant |
| 0 wt %–20 wt % | 2,887 | 4,769 | 21,78 | Significant |
| 10 wt %–20 wt % | 3,752 | 5,763 | 10,24 | Significant |

Tabla 2.8: Pairwise comparisons using the functional ANOVA test with DTG curves.

In addition, other competitive models like unfunctional ANOVA models based on discretized curves (MANOVA test) and the poinwise $F$ test presented in Ramsay and Silverman (2005) were applied to the TG and DTG curves. In the first case, a MANOVA test (Wilks) was performed. For this, the curves were discretized just using 12 points per curve, because the sample size here is 15. Depending of the election of these points, the results of the test were different. Moreover, the relatively small sample size can make that important information could get lost in the process of discretization. In this sense, our test takes the full functional nature of the curves into account and more reliable results are obtained. On other hand, a pointwise $F$ test was applied, obtaining similar results to the functional test performed in the present paper.

In the case of the DTG curves, the application of the FANOVA test gives as results: $V_n = 57{,}37$ and $V_\alpha = 3{,}58$. Therefore $V_n >> V_\alpha$. The resulting test is highly significant. At least one of the functional means is different from the others. Therefore, the arguments explained in the descriptive analysis have a statistically significance. The addition of silica causes changes in the functional means of the derivatives of the rescaled TG curves. The thermal stability of epoxy resin experienced a statistically significant change with an increasing amount of fumed silica. It can be said that the rate of degradation really experiences the changes indicated in the descriptive study, at least for a level. This result is also an indicator of some interaction between the epoxy resin and fumed silica, which could result in the creation of a matrix-filler interphase, like in the rescaled data.

Following the same steps as before, the functional ANOVA pairwise comparisons are applied to distinguish which groups (0, 10 and 20 wt %) are really different, that is, what levels of silica provide different degradation rates (Table 2.8). Additionally, the above proposed Tukey based test is performed to complete the study (Table 2.9).

Observing the results, we could say that the degradation rate for each group is significantly different from the others.

| Groups compared | Mean distance | Quantile | Result |
|---|---|---|---|
| 0 wt %–10 wt % | 0,0017 | 0,0016 | Significant |
| 0 wt %–20 wt % | 0,0044 | 0,0016 | Significant |
| 10 wt %–20 wt % | 0,0029 | 0,0016 | Significant |

Tabla 2.9: Pairwise comparisons using a Tukey-like bootstrap test with DTG curves.

## 2.6. Conclusions

In this paper, a functional ANOVA (FANOVA) test is used to evaluate the effect of adding fumed silica on thermal degradation of an epoxy resin. Three different levels (0, 10 and 20 wt %) of amount of fumed silica are considered. The procedure is applied to rescaled TG curves and the derivatives of the rescaled TG. Previously to the application of the test, some data preprocessing is done to prepare the datasets to be analysed. Statistical techniques like penalized $b$-splines or the concept of depth are used for this purpose.

All tests performed have resulted highly significant. Therefore it can be said that the addition of fumed silica affects the way of degradation of epoxy resin involved in the sample, at least in one group. Moreover, by performing pairwise comparisons using the functional ANOVA method and a bootstrap distance based test, it can be said that the way of degradation for each group is significantly different from the others.

As one moves from 0 to 10 wt % or 20 wt % of fumed silica, the thermal stability of pure epoxy resin increases at temperatures corresponding to the second degradation step. This is an indication of the interaction between the organic and inorganic phases, with statistical significance.

Regarding the thermal degradation rate, it can be said that this is different with the addition of fumed silica, by applying functional ANOVA and pairwise comparisons. The mass loss rate decreases at the beginning of the main degradation process. The temperature at which the maximum mass loss rate occurs hardly changes, independently on the filler content, but the maximum mass loss rate module decreases with the addition of 20 wt % of fumed silica.

In the groups of 10 and 20 wt % of fumed silica, the rescaled curves and their derivatives are very similar. They mainly differ in the last degradation process, but significant differences in the second degradation step are also appreciated. The addition of 20 wt % in silica can cause a decrease in thermal stability at very high temperatures, because of the morphology of the material

itself.

It would be interesting, in a future research, studying some extensions of this work, analyzing the possible influence of more factors (for example, the particle size distribution) in the response variable.

## 2.7. Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 (ERDF included) and by Xunta de Galicia PGIDIT07PXIB105259PR. The authors wish to express special thanks to Manuel Febrero Bande and Aldana González Montoro, for their valuable comments. In addition, the authors thank two referees for constructive comments that improved the presentation of this article.

# Parte II

# Clasificación supervisada de materiales

*Es preciso suponer que en todo lo que se combina hay muchas cosas de todas clases, y semillas de todas las cosas, que tienen formas diversas y colores y sabores diferentes.*

Anaxágoras

La segunda parte de la presente tesis aborda el problema general de la clasificación supervisada de materiales. Para ello se ha escogido el caso particular de la madera, debido a la especial dificultad que presenta su clasificación en diferentes especies (dada su alta heterogeneidad estructural y química). Además, la clasificación de una muestra de madera en diferentes especies predeterminadas es una tarea necesaria y habitual en la industria, llevada a cabo con una total dependencia del factor humano. Por lo tanto, la aplicación de metodologías de clasificación supervisada estadística tiene una utilidad inmediata y viene a resolver el problema de la escasez del personal cualificado y su alto coste de formación.

La Parte II se divide en 3 capítulos diferentes. El Capítulo 3 se corresponde con el artículo titulado *A comprehensive classification of wood from thermogravimetric curves*, publicado *on-line* en julio de 2012 (en su versión *In Press*) por la revista *Chemometrics and Intelligent Laboratory Systems*, incluida en el *Journal Citation Report*, dentro de, entre otras, la categoría *Statistics and Probability* (situada en el primer tercil). El Capítulo 4 reproduce íntegramente el texto correspondiente el artículo titulado *Functional nonparametric classification of wood species from thermal data*, publicado en el año 2011 en la revista *Journal of Thermal Analysis and Calorimetry*, incluida en el *Journal Citation Report*, dentro de la categoría *Chemistry, Analitycal* (situada en el segundo tercil). Finalmente, el Capítulo 5 comprende el artículo de título *Classification of wood micrographs by image segmentation*, publicado en el año 2011 por la revista *Chemometrics and Intelligent Laboratory Systems*, incluida en el *Journal Citation Report*, y situada en el primer cuartil de la categoría *Statistics and Probability*.

En el Capítulo 3 se aborda la clasificación supervisada de materiales, en este caso particular la madera, a partir de las curvas termogravimétricas, aplicando y comparando técnicas de análisis de datos funcionales, por un lado, y multivariantes, por otro. En el caso del análisis multivariante, para llevarlo a cabo se ha desarrollado un nuevo procedimiento de extracción de características representativas: los parámetros de ajuste de un modelo de regresión no lineal compuesto por una suma de funciones logísticas generalizadas. Se ha desarrollado un completo estudio de simulación de curvas TG, en muy diversos escenarios definidos por el grado de dependencia de los parámetros que definen cada curva y la variabilidad existente. Los estudios de simulación no son muy habituales en anális térmico, por lo que el presente ofrece una alternativa para completar los resultados experimentales.

En el Capítulo 4 se particulariza el estudio de la clasificación de especies de madera dentro de la perspectiva del análisis de datos funcionales. Se utilizan dos bases de datos diferentes: curvas TG y curvas DSC, a las que se aplican 4 tipos de clasificadores diferentes, 2 basados en el estimador de Nadaraya-Watson y otros 2 basados en el algoritmo Adaboost. Estos se aplican a los segmentos de curva correspondientes a más de 1000 intervalos diferentes de temperatura, permitiendo establecer el rango en el cual se obtiene una clasificación óptima. Este procedimiento permite establecer interesantes conclusiones relacionadas con la estructura química de las maderas.

El Capítulo 5 aborda el problema expuesto en los capítulos anteriores pero utilizando para ello una base de datos completamente diferente. Si en los Capítulos 3 y 4, los datos empleados eran los obtenidos a partir de técnicas

de análisis térmico (como también lo eran aquéllos utilizados en el Capítulo 2), en el Capítulo 5 se utilizan las imagenes, micrografías, obtenidas mediante Microscopía Electrónica de Barrido a 1500× aumentos. Después de un proceso de mejora y segmentación de las imágenes, se efectúa la extracción de 5 características representativas de la estructura de las traqueidas de la madera, además del cálculo de diferentes estimadores de la dimensión fractal de cada imagen, aplicándose a continuación los métodos de clasificación multivariante.

Los textos aparecen integramenten en inglés, respetando las versiones publicadas por cada revista. Cada uno de los artículos que se exponen en la Parte II se muestra en su versión original de publicación en los Apéndices B, C y D, respectivamente.

# Capítulo 3

# A comprehensive classification of wood from thermogravimetric curves

**Resumen:** La madera es uno de los materiales más difíciles de clasificar en sus diferentes grupos o especies. En este capítulo, se han empleado las curvas termogravimétricas (TG) correspondientes a 49 muestras de madera, con el objeto de clasificarlas dentro de 7 especies predeterminadas. Para ello se han aplicado tanto métodos estadísticos de clasificación supervisada pertenecientes al análisis de datos funcionales (FDA) como métodos de clasificación multivariantes: un estimador funcional no paramétrico tipo kernel basado en el estimador de Nadaraya-Watson (K-NFDA), aplicado directamente a las curvas TG, y métodos de clasificación supervisada a través de un enfoque multivariante como son el Análisis Discriminante Lineal (LDA), $k$-Nearest Neighbors ($k$-NN) o $k$-vecinos más próximos, Bayes Naive (NBC), Redes Neuronales (NN) y Support Vector Machines (SVM) o máquinas de soporte vectorial. Previamente a la aplicación de estas técnicas multivariantes, teniendo como objetivo discretizar las curvas TG (resumir en unas pocas variables la información existente en las mismas), se ha empleado el análisis de componentes principales (PCA) de las curvas TG y, por otro lado, el ajuste de un modelo de regresión compuesto por cuatro componentes logísticas generalizadas. Los resultados muestran que el método clásico de LDA, aplicado a los parámetros logísticos, proporciona el mejor rendimiento, aunque también se obtuvieron altos porcentajes de clasificación correcta mediante los demás enfoques alternativos. El trabajo se completa con un estudio de simulación exhaustivo, comparando las técnicas de clasificación en diferentes escenarios. Las nuevas curvas TG artificiales se generan mediante el modelo logís-

tico antes mencionado, estableciéndose conclusiones adicionales acerca de la clasificación de la madera. Dada la alta heterogeneidad de la madera y la dificultad de obtener una muestra realmente representativa, este estudio de simulación representa una opción muy útil para describir los peores escenarios y para evaluar con mayor precisión las metodologías de clasificación propuestas.

**ABSTRACT:** Wood is one of the most complicated materials to be classified in different classes or species. In this paper, the thermogravimetric (TG) curves of 49 wood samples are used to classify them in 7 predetermined species. Different functional and multivariate statistical supervised classification methods are used for this task: a nonparametric Nadaraya-Watson kernel functional estimator (K-NFDA), using the complete TG curves, and multivariate supervised classification approaches, such as linear discriminant analysis (LDA), $k$ Nearest Neighbors ($k$-NN), Naïve Bayes (NBC), Neural Networks (NN), and Support Vector Machines (SVM). Before applying the multivariate techniques, the TG curves are discretized using principal component analysis (PCA) or fitting a four-component generalized logistic model. The results show that the classical method of LDA using the logistic parameters had the best performance, although high correct classification percentages were also obtained with the rest of the approaches. The work is completed with a comprehensive simulation study, comparing the classification techniques in different scenarios. Artificial TG curves are generated using the logistic model and additional conclusions on wood classification are established. Due to the heterogeneity of wood, this simulation study is very useful to describe worst-case scenarios and to assess more accurately the proposed classification methodologies.

## 3.1. Introduction

The discrimination of a specific material in different classes represents an important practical problem with direct industrial applications. When the material under consideration is wood, the correct classification into different species becomes an extremely difficult task. The reasons are complex compositions of this material, the wide variety of existing species and the anatomical heterogeneity of its elements. Depending on wood species, a timber presents different physical-chemical properties determining its industrial applications and price. Wood identification is often made on the basis of readily visible characteristics such as color, odor, density, presence of pitch, or grain pattern which may result error arising from human-bias. To get a

more accurate classification, in case of particular difficulties, it is essential to use microscopy techniques, physical hardness tests and chemical analyses (Guindeo Casasús et al., 1997; Lewis et al., 1994; Miller, 1999). These kind of analyses are useful in the furniture industry, the wood panel production, or even in archeology, where it is crucial to know the kind of wood used to combat fraud (Khalid et al., 2008; Hayek et al., 1990a; Arno et al., 1993; Hoadley, 1990). Therefore, the implementation of quantitative models and automatic wood sample recognition methods are justified and can be of immediate application in these fields.

In this paper, the thermogravimetric (TG) curves of different wood samples are used to classify them in different species. The TG curves explain the mass loss when the temperature is increased. They are the result of applying a common technique, called thermogravimetry, belonging to the thermal analysis of materials (Prime et al., 2009). Different multivariate techniques have been applied to thermograms (Lukasiak et al., 2006; Pomerantsev y Rodionova, 2005; Miltyk et al., 2010). However, to our knowledge, the present research is the first of its kind, where different classification methods are evaluated and compared using experimental and simulated TG curves.

Figure 3.1 shows the 49 TG curves of the wood samples (7 samples per class of wood) used in our analysis (see Section 3.3.1 for a description on the wood species and their corresponding TG curves employed in this research). A particular wood species is highlighted in each panel. The last panel (in row 4, column 2) shows all the TG curves.

The functional nature of these curves suggests the use of functional supervised classification methods for this task. On the other hand, if the curves are properly discretized, multivariate classification techniques can also be applied here. Different classification approaches in these two settings are compared in this work. Specifically, classical multivariate supervised classification methods, such as linear discriminant analysis (LDA), some machine learning techniques (including Naïve Bayes (NBC), $k$ Nearest Neighbors ($k$-NN), Neural Networks (NN) or Support Vector Machines (SVM)), and the nonparametric Nadaraya-Watson kernel functional (K-NFDA) method are used and compared. Regarding the multivariate methods, an important problem arising is how to discretize the curves. In other words, what features are actually representative of each obtained TG curve? In the present study, two approaches are followed. On the one hand, using principal component analysis (PCA), some components of the TG curves which are together explaining most of the variation of the data are selected. We set a cut-off for the percentage of variation which we wish to achieve. The rest of the components of the curves are then neglected. Note that, in practice, many traditional methods including

discriminant analysis cannot be applied directly when the number of components exceeds the number of observations and, therefore, this fact must be taken into account in the process of selecting the principal components. On the other hand, a new nonlinear regression model is proposed to fit the TG curves and to extract some representative features from them. This model can be written as the sum of four generalized logistic components, one per principal constituent of wood (hemicellulose, cellulose and lignin) and one component corresponding to the water involved. The parameters obtained from the fit of each TG curve (4 for each component, 16 in total) are used as a vector of features, ensuring the representativity of them.

In our research, tests for seven different wood species and seven samples per class are performed. This can be considered a moderate sample size and, while important conclusions can be deduced from this analysis, a more comprehensive study to compare the different classification methods would require of samples collected in different scenarios. Nevertheless, this is a problem from a practical point of view. The cost of each thermogravimetric test is not negligible, both in time and money. The time spent on sample preparation must be added to the duration of each test. In addition, while performing the tests, the device is occupied, being impossible its use in other applications. Moreover, due to the extreme heterogeneity of wood (there are differences in wood according to the tree from which it is extracted, or even between different parts of the same tree), it is very difficult to get a fully representative sample. For this reason, apart from using the real wood samples, a complete simulation study to compare the different approaches is carried out. The TG curves are mimicked using the previously mentioned generalized logistic model. Artificial TG curves are simulated using the parameters obtained from fitting the generalized logistic mixture model to the real curves, considering different covariance matrices and sample sizes. This represents a general proposal that can be applied not only in the context of the present paper. Representative samples in less time and in a variety of scenarios can be then produced. An alternative way to simulate TG curves, based on the Arrhenius model, was presented in Ferriol et al. (2003). They focussed on the thermal behavior of poly(methl methacrylate) (PMMA) and described an algorithm to simulate the overall weight loss of PMMA in any given experimental condition.

Some statistical classification methods have been previously compared in different papers. For example, in Ferraty and Vieu (2003), the K-NFDA method (also used in the present work) was defined and compared with several existing curve discrimination techniques. This comparison was performed using two real data sets and through simulations. The same real data sets

were used in Bin and Qingzhao (2008) to compare their functional segment discriminant analysis proposal with standard classification methods. Some simulation experiments complete that paper. Working with microarray data, in Lee et al. (2005), 21 classification methods were compared on 7 data sets. With these (and other) studies in mind, an important aim of the present paper lies in comparing the performance of different classification methods applied to functional data, but focussing on the practical problem of wood species discrimination, using their TG curves. The comparison of nonparametric functional methods and classical and machine learning multivariate techniques in this framework represents a relevant challenge in the wood industry. It is also important to stress that the proposed logistic model to extract representative features from the TG curves is also a novel approach with physical-chemical interpretations in this context. Moreover, it can be used to generate artificial TG curves, making it possible to extend the comparison in a simple way to different scenarios. Note that statistical simulation studies are unusual in thermal analysis, being an exception the work of Artiaga et al. (2011). Therefore, the simulations experiments carried out in this research are a novelty in themselves in this framework, giving the opportunity not only of establishing a comparison among the different classification techniques, but also to extract some conclusions related with wood discrimination from a chemometric point of view.

Accordingly, the objectives of the present study are, on the one hand, comparing different methods to classify seven commercial wood species (European chestnut, European beech, eucalyptus, jatobá, European oak, Scots and insignis pine) on the basis of their TG curves. These methods include nonparametric functional techniques, using the complete TG curves, and multivariate supervised classification approaches, such as LDA, $k$-NN, NBC, NN, and SVM, using some components of the curves selected by PCA or fitting a generalized logistic mixture model. Additionally, these classification methods applied in a functional context are compared through a comprehensive simulation study. The synthetic curves used to test the different methods try to imitate the TG curves in different scenarios.

The structure of the paper is as follows. In Section 3.2, the functional and multivariate data classification methods used in the present research are explained. In Section 3.3, these methods are applied to the real wood samples using their TG curves, analyzing the results obtained. Previously, in this section, the materials and experimental techniques used to obtain the TG curves are also described. In Section 3.4, the performance of the different classification techniques are compared in different scenarios through a simulation study. Finally, Section 3.5 collects the main conclusions.

## 3.2. Classification techniques

The process of assigning an observation to one of several predetermined groups is called supervised classification. The principal aim is to obtain a discriminant function summarizing the information contained in the observations (curves or multidimensional vectors, in the framework of the present paper), and using this function to classify a new observation in one of the given groups. In other words, given a learning sample consisting of observed curves from known groups, our issue is to predict the group membership of a new incoming curve. In the statistical literature, several multivariate and functional methods have been developed to address supervised classification problems. In the next sections, the approaches used in our experiments are briefly described. The free statistical software R (R Development Core Team, 2008) was used to implement these methods. The specific R packages and libraries employed for this task, as well as the range of values used to select the optimal parameters in each case, are cited in the corresponding description of each method. This can demonstrate to practitioners how to apply the different approaches.

### 3.2.1. Linear Discriminant Analysis

The classical Linear Discriminant Analysis (LDA), proposed by Fisher (Fisher, 1936), is one of the most widely used technique for data classification. It works by maximizing the ratio of between-class variance to the within-class variance, being optimal under the assumptions of Gaussian likelihood and equal covariance matrices between groups. There has been many modifications proposed to this method, such as, Orthonormal LDA (Okada and Tomita, 1985), Non parametric LDA (Fukunaga, 1990), etc. But in our case, we have found the performance of classical LDA satisfactory. LDA has been successfully applied for classification and pattern recognition in fields as diverse as engineering, economics, computing science, biology, etc. Related to the topic of the present paper, there are some interesting works where LDA is applied to the features extracted from wood micrographs (Mallik et al., 2011) and fluorescence spectra (Piuri and Scotti, 2010; Camorani et al., 2008; Labati el al., 2009). The R `MASS` library (Venables and Ripley, 2002)nnet was used to apply the LDA in our experiments. Specifically, the function named `lda` is designed to perform this classification method.

### 3.2.2. Naïve Bayes

Naïve Bayes classifier (NBC) is a supervised multivariate classification technique based on the Bayes theorem. Using this rule, we intend to calculate the posterior probability that a sample belongs to a particular class (from a group of possible classes), given the feature values that define the sample. Then, the class of a test sample is estimated using the largest posterior probability obtained. NBC assumes that the conditional probabilities of the independent variables are statistically independent and, therefore, the posterior probabilities can be rewritten in a simpler way. The function `naiveBayes` included in the R `e1071` library (Dimitriadou et al., 2011) was used in our analysis to apply the NBC method. Many papers showing the application of this technique in a variety of classification problems can be found in the scientific literature. For instance, in computer science, addressing the problem of classifying E-mails in spam and non-spam; in medicine, to solve medical diagnosis; in acoustics, performing automatic classification of sound and voice, or in image classification. Some references are, for example, Kononenko (2001) and Tóth et al. (2005). For the particular case of wood classification, we can cite the studies of Mallik et al. (2011) and Gasson et al. (2010), where NBC technique is applied to features extracted from wood micrographs at different magnification.

### 3.2.3. $k$ Nearest Neighbors

$k$ Nearest Neighbors ($k$-NN) is a multivariate nonparametric supervised classification method introduced by Fix and Hodges (1951). It can also be applied in populations where the assumption of normality is not required. In principle, this represents an advantage over the LDA method, which is not optimal for non Gaussian populations. The basic idea of $k$-NN is the following: the class of a given sample will be the most repeated class corresponding to the surrounding neighbors. The $k$-NN procedure starts choosing an appropriate distance (mainly the Euclidean or Mahalanobis distances) between samples, represented by vectors of features. Then, the distances between the test sample, $x_0$, and the other samples are calculated. The $k$ nearest samples to those we want to classify are selected. Next, the proportion of these $k$ samples belonging to each of the studied populations is calculated. Finally, the sample $x_0$ is classified within the population corresponding to the highest existing frequency. Among the different methods available for choosing the value of $k$, the minimization of the cross-validation error is one of the most used. A weighted version of $k$-NN was used in the present study. Each $k$ nearest observation is weighted according to its distance to the new ob-

servation that we want to classify (Hechenbichler and Schliep, 2004). The distances are transformed by the application of some type of kernel (Gaussian, triangular, rectangular, Epanechnikov, etc). Then, the winning class is chosen as the class estimation which shows a weighted majority of the $k$ nearest neighbors. The model parameters, number of nearest neighbors and type of kernel, were obtained by the inner loop corresponding to a double cross-validation process (see Section 3.3.2). For this purpose, `kknn` library (Schliep and Hechenbichler, 2008) was used. Specifically, the function `kknn` of that package is designed to apply this method. We fix the range of possible values for the number of nearest neighbors between 1 and 25, while rectangular, triangular, Epanechnikov, Gaussian and rank kernels were tested. The $k$-NN technique have been successfully applied in diverse fields, such as chemistry, biology, medicine, computer science, genetics and material science, identifying wood species (Tarrío-Saavedra et al., 2011; Mallik et al., 2011; Piuri and Scotti, 2010; Camorani et al., 2008; Labati et al., 2009).

### 3.2.4. Support Vector Machines

Support Vector Machines (SVM) is a comparatively new machine learning technique developed by Vapnik and co-workers (Vapnik, 1998). SVM is a non-probabilistic classifier. It works by constructing a hyperplane or a set of hyperplanes maximizing its distance from the nearest data point on each side, therefore achieving largest separation. This hyperplane is called the maximum margin hyperplane. For non linear cases the hyperplane is constructed by a nonlinear kernel function in place of dot products. Homogeneous polynomial and Gaussian radial basis functions are mostly used kernels. Note that in the original form, SVM are binary classifiers, i.e. they discriminate between two classes. In a situation like the one studied in this research, where there are 7 possible classes, a common procedure to overcome this drawback is to turn one multi-class problem into several two-class problems. In the present paper, a C-SVM model and a Gaussian kernel were tested. The C (cost of constraints violation) and gamma parameters were obtained by the inner loop of a double cross-validation procedure (see Section 3.3.2). Values between 1 and 21 for the C parameter, and between 0.001 and 0.051 for the gamma parameter were used in the present research. The `svm` function of the R package `e1071` (related to the `libsvm` library) (Dimitriadou et al., 2011) was employed to apply SVM. It is important to point out that the `libsvm` library uses one-against-one classifications for multiclass problems with $k$ levels, $k > 2$, in which $n$ choose $k$ binary classifiers are trained. The class which has won most number of times is chosen as winner class. During

recent years SVM is being successfully used in classification (Karatzoglou et al., 2006), text and data mining (Tong and Koller, 2001) and image retrieval (Chang et al., 2005) problems. Regarding wood classification problems, SVM was applied in Mallik et al. (2011), and Piuri and Scotti (2010).

### 3.2.5. Neural Networks

Neural Networks (NN) is a machine learning technique that is motivated to imitate the structural and functional aspects of biological neural networks (Ripley, 1994). In most cases, NN consists of a network of artificial neurons, which changes its structure based on external or internal information flowing through the network at learning phase. NN are becoming widely popular because of their wide area of applicability such as functional approximation, data modeling, classification (Michie et al., 1994; Lippmann, 1989), robotics, etc. NN are computationally demanding, but they are very useful because they provide flexibility of choosing various learning algorithms and cost functions, thus offering superior control over learning rate of the model and its accuracy. For our purpose, we have used a single hidden layer feed forward network to train and test our dataset. The model parameters, in this case, number of units in the hidden layer and weight decay, were obtained by the inner loop corresponding to a double cross-validation process (see Section 3.3.2). We used the R `nnet` library (Venables and Ripley, 2002) to implement this classification method, selecting a range between 1 and 3 for the number of units in the hidden layer, and between 0.00001 and 0.001 for the weight decay. In relation with the topic of this paper, there are several interesting studies where NN was applied to classify wood species (Mallik et al., 2011; Labati et al., 2009; Jordan et al., 1998).

### 3.2.6. Nonparametric Functional Data Analysis

Statistics for functional data is a recent field of research popularized by the monographs of Ferraty and Vieu (2006), Ramsay and Silverman (2002) or Ferraty (2010). Assuming that a set of $n$ curves $X_i = X_i(t)$, $i = 1, \ldots, n$, have been observed, each one belonging to a known class $g$, with $g \in \{0, 1, \ldots, G\}$, and given a new curve, $x = x(t)$, we are interested in classifying $x$ in one of the classes $\{0, 1, \ldots, G\}$. This problem is solved estimating the posterior probability that $x$ belongs to a class $g$, for each $g \in \{0, 1, \ldots, G\}$, and selecting the class with largest estimated posterior probability. It is clear that this problem can be tackled from a regression point of view and, therefore, the issue of estimating those posterior probabilities is equivalent to that of

estimating the corresponding regression functions. Taking this account and given the functional nature of the data, the posterior probability estimator that $x$ belongs to a class $g$, with $g \in \{0, 1, \ldots, G\}$, used in the present work is the functional Nadaraya-Watson nonparametric kernel method (K-NFDA) shown in (4.1), and defined in Ferraty and Vieu (2003):

$$\hat{r}_h^{(g)}(x) = \frac{\sum_{i=1}^{n} I_{\{Y_i=g\}} K\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{h}\right)}, \tag{3.1}$$

where $Y_i$ are the corresponding classes of the observed curves $X_i$, with $i = 1, \ldots, n$. In equation (4.1), $I_{\{.\}}$ is the indicator function, the parameter $h$ is the bandwidth or smoothing parameter, $K$ is a kernel function, and $\|\cdot\|$ is a semi-norm. The kernel function is typically a density function chosen by the user (in our experiments, we used the Gaussian kernel) and $h$ regulates the amount of smoothing to be used. Although the choice of the kernel function is of secondary importance, the smoothing parameter plays a crucial role in kernel estimation. We chose $h$ as the value that minimizes the probability of misclassifying a future observation (it is selected according to the cross-validation method, searching a bandwidth in a grid of values ranging between 0.001 and 2). Finally, we used as semi-norm, the $L_2$ norm, measuring the differences between curves by:

$$d(X_i, X_j) = \int_a^b \left(X_i(t) - X_j(t)\right)^2 dt, \tag{3.2}$$

where $[a, b]$ is the interval studied. The R Packages `fda` (Ramsay et al., 2011) and `fda.usc` (Febrero-Bande et al., 2011) were mainly used to perform the classification applying nonparametric functional analysis.

## 3.3. Classification of wood samples

In this Section, we describe the wood data used in our analysis and the experimental process to obtain their TG curves. Then, the statistical methods previously presented are applied to these TG curves to classify the corresponding timber samples in different species. Some programs coded in R using the specific packages and functions cited in the previous section were employed to compare the approaches, using the real data and through simulations.

### 3.3.1. The data

Tests for five different hardwoods (European beech or *Fagus sylvatica*, European oak or *Quercus robur*, chestnut or *Castanea sativa*, *Eucalyptus globulus* and jatobá or *Hymenaea courbaril*) and two softwoods (Scots pine -*Pinus silvestris*- and insignis pine -*Pinus radiata*) were carried out. Seven samples per each one of the above mentioned species, obtained from wood of different trees were tested. The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the experimentation time. The samples were not dried in order to avoid disturbance to their structure and composition as much as possible, and to test the automatic classification methods with a minimal sample preparation, under the worse conditions.

The test was performed on a SDT 2960 TA Instruments thermo balance. This apparatus provides TG curves used in the classification analysis. A heating ramp of $20\,^{\circ}\mathrm{C} \cdot \mathrm{min}^{-1}$ was applied in the range from 20 to 600 °C, at a rate of 50 $\mathrm{mL} \cdot \mathrm{min}^{-1}$ of $N_2$ (Prime et al., 2009). The nitrogen was purged for 10 min, before starting the heating program for establishing an inert environment. The used heating rate was chosen to obtain a proper balance between time test and resolution (Prime et al., 2009). It aims to assess the discriminatory power of the resulting curves, using the minimum experimental time. The sample mass chosen was between 6 and 8 mg. Alumina crucibles were used.

### 3.3.2. Results and discussion

Figure 3.1 in Section 3.1 shows the 49 TG curves obtained (7 per class). A descriptive analysis of these data was shown in Tarrío-Saavedra et al. (2011). Different trends are observed for almost all types of wood. Note also that the TG curves of some species tend to overlap with other species. This problem arises because of their similar densities, hardness and mechanical properties. More specifically, the shape of the TG curves obtained in a pyrolysis test (as it is the case of the present work) is directly related to wood composition (Yang et al., 1999; Alén et al., 1996; Gašparovič et al., 2009; Roberts, 1970; Grønli et al., 2002; Müller-Hagedorn et al., 2003; Wang et al., 2009). In fact, wood degradation in an inert atmosphere is dominated by the degradation behavior of its three main components (cellulose, lignin, and hemicellulose) as was reported in Yang et al. (1999) and Gašparovič et al. (2009). Cellulose represents about 40 and 60 % in the overall weight of dry wood (23-33 % of the mass of softwoods), 23-33 % of lignin in softwoods (16-25 % in hardwoods), and 25-35 % of hemicellulose (Miller, 1999; Grønli et al., 2002). The propor-

tion of each wood component varies depending on the species, to a greater or lesser extent (Grønli et al., 2002; Müller-Hagedorn et al., 2003; Wang et al., 2009; Roberts, 1970). Therefore, it is expected that the TG curves are different depending on the type of wood to which they belong. Furthermore, effects on the shape of the TG curves, such as the existence of other components as ethanol extractives (Sebio-Puñal et al., 2012) and even the origin of lignin and hemicellulose are not absolutely negligible. Differences in the pyrolysis of lignin and hemicellulose depending on whether these come from softwoods or hardwoods, or even from different species, were also observed (Sebio-Puñal et al., 2012; Müller-Hagedorn et al., 2003; Wang et al., 2009; Mohan et al., 2006). Considering all these studies, it seems reasonable to suggest that TG analysis has the potential to be useful to discriminate between classes of wood species. Moreover, it is important to note that the hemicellulose, cellulose and lignin decompose in temperatures ranging between and 200-260 °C, 240-350 °C, 280-500 °C, respectively (Yang et al., 1999; Alén et al., 1996; Gašparovič et al., 2009; Roberts, 1970; Grønli et al., 2002; Müller-Hagedorn et al., 2003; Wang et al., 2009).

As observed in Figure 3.1, each curve represents a functional data and, therefore, it is quite natural to use functional classification methods (see Section 3.2.6) to carry out the classification process. On the other hand, after discretizing the TG curves, the multivariate classification methods described in the previous section can also be applied. As pointed out in the Introduction Section, two methods were employed for this process. First, using PCA to select the curve components explaining most of the variation of the data (99 % in our case). Second, fitting a generalized logistic model to the TG curves and using the parameters of this model as representative curve features. The proposed model to be fitted consists of a mixture of 4 generalized logistic functions related to the wood main components, i.e. cellulose, hemicellulose, lignin and water:

$$Y(t) = \sum_{i=1}^{4} c_i \left( 1 - \frac{1}{\left(1 + \tau_i \cdot e^{-b_i \cdot (t - m_i)}\right)^{1/\tau_i}} \right) \tag{3.3}$$

where the $c_i$ parameter is related to the mass involved in the degradation process, $b_i$ is related to the decomposition rate or rate of change, $\tau_i$ accounts for the asymmetry, $m_i$ represents the temperature at the maximum rate of change, $t$ is the temperature, and $Y$ the fitted TG curve. The optimal fittings were obtained by minimizing the average squared error (ASE), using the Nelder-Mead algorithm (Nelder and Mead, 1965). All the resulting fits presented coefficient of determination higher than 0,999. An example of the goodness-of-fit of this model can be observed in Figure 3.2, where an original

TG curve of beech and the corresponding fit using Eq. (3.3) is presented. The four logistic components of this fit are also included in this plot. Using these 16-dimensional vectors (or the corresponding vectors obtained with the PCA approximation), it is possible to apply the multivariate supervised classification methods previously described to perform the curve classification.

In the case of the nonparametric functional approach, after a careful examination of the TG curves, it was found that the curves can be better discriminated if they were standardized, using a linear transformation on the original curves based on their mean and variances (López-Granados et al., 2008; Tarrío-Saavedra et al., 2011).

Two different procedures were used to validate the different approaches, cross-validation and external validation. Regarding the first proposal, a double cross-validation process was performed (Wehrens, 2011). It consists of an inner loop where the optimal parameters corresponding to the classification methods (SVM, NN, $k$-NN, etc.) were obtained, and an external loop to assess the prediction error in each case. This methodology was used to avoid the bias resulting from the use of the same sample to estimate the parameters of each classification method, and to obtain the misclassification errors (Wehrens, 2011). In the present study, the inner and external loops were consisted of two leave-one-out cross-validation procedures. This process works by leaving out one TG curve; then a model is trained with the remaining thermograms and, finally, the developed model is used for the classification of the left out TG curve. This is repeated until all the curves have been left out once. As the dataset available contain 49 samples, 48 samples were used for training and 1 sample for testing. It is important to note that the model is trained by another leave-one-out procedure using the 48 remaining curves (inner loop). This process was repeated 49 times, and the percentages (measured as per one) of correct classification were calculated. Figure 3.3 shows a kind of flowchart describing the validation process previously described for the multivariate methods. Similar steps are followed when the K-NFDA method is validated through leave-one-out cross-validation, but obviating the feature extraction step.

The error estimates obtained by leave-one-out cross-validation can present a relatively high variance, although without bias. On other hand, multiple cross-validation (e.g. 10-fold) is about ten times faster and presents less variability in the error estimations (Efron and Tibshirani, 1993; Wehrens, 2011), although a bias can occur because the model is based on a data set that is smaller than the real data set, giving slightly pessimistic misclassification error estimations. For avoiding these disadvantages of leave one out cross validation as far as possible, and for completing the results obtained by this

technique, an external validation procedure was performed. Considering the relatively small number of samples per class, a first set of samples (one per class) is randomly extracted as external validation set. The remaining samples are used as training set to obtain the classification model parameters, according to a 10-fold cross-validation in this case. This global procedure is repeated 100 times to be sure that, with high probability, each sample is included in the test set at least once. Figure 3.4 shows a similar flowchart to that presented in Figure 3.3, but now for the external validation process.

Table 3.1 and 3.2 shows the misclassification errors obtained by the different methods using leave-one-out cross-validation and the external validation procedure, respectively. In the second framework, the standard deviations are also included between brackets in Table 3.2.

| Classification | Misclassification errors | |
|---|---|---|
| Methods | PCA | Logistic |
| LDA | 0,18 | 0,14 |
| NBC | 0,33 | 0,27 |
| $k$-NN | 0,27 | 0,39 |
| SVM | 0,16 | 0,24 |
| NN | 0,24 | 0,24 |
| K-NFDA | 0,22 | |

Tabla 3.1: Misclassification errors obtained by each classification method and leave-one-out cross-validation. The multivariate classification methods were tested using PCA and the generalized logistic fits.

The results obtained by cross-validation and external validation are generally very similar. This fact support the validity of the obtained error estimations. In general, Tables 3.1 and 3.2 show a relatively high probability of correct classification, especially taking into account wood heterogeneity and the results obtained in other studies (Brandtberg, 2002; Tou et al., 2009). Therefore, classification using the logistic parameters or PCA has been shown feasible. It is interesting to note that SVM, LDA and K-NFDA methods gave the highest probabilities of good classification, but LDA and K-NFDA required less computing times. Since the probability of misclassification is larger than zero, an interesting question is: which are the species tending to get confused using these features and methods? The answer is in the confusion matrices shown in Tables 3.3 and 3.4.

Table 3.3 shows the confusion matrices corresponding to LDA, $k$-NN, SVM and K-NFDA, applied to classify between 7 different types of wood,

| Classification Methods | Misclassification errors | |
|---|---|---|
| | PCA | Logistic |
| LDA | 0,20 (0,12) | 0,18 (0,14) |
| NBC | 0,29 (0,15) | 0,28 (0,17) |
| $k$-NN | 0,27 (0,15) | 0,39 (0,16) |
| SVM | 0,17 (0,12) | 0,23 (0,15) |
| NN | 0,34 (0,17) | 0,35 (0,17) |
| K-NFDA | 0,26 (0,15) | |

Tabla 3.2: Misclassification errors obtained by each classification method and the external validation process. The multivariate classification methods were tested using PCA and the generalized logistic fits. Standard deviations are included between brackets.

when using PCA to discretize the TG curves in the multivariate classification approaches. Table 3.4 shows the same information, but when the features of the TG curves are selected by fitting the generalized logistic model given in Eq. (3.3) (K-NFDA is not included in Table 3.4. Instead, the results for NBC are presented).

Tables 3.3 and 3.4 show that, among the 7 species studied, Scots pine is the most difficult species to be classified. When we classify using the logistic parameters or the raw TG curves, there are some confusion between Scots pine and insignis pine, and between oak and beech. This may be due to their similar chemical, physical and mechanical properties. It is also interesting to observe that jatobá samples can be classified as Scots pine samples and vice versa when the features obtained by PCA are used. Misclassification errors in species like oak vary considerably depending on the method and the features used as dataset. This should be taken into account for further implementation in the industry.

## 3.4. Simulation study

This section shows a simulation study comparing the classification methods previously applied to the real TG curves. Using the parameters obtained from fitting the generalized logistic model (Eq. (3.3)) to the real TG curves, artificial curves imitating the real ones are generated. This allows to compare the classification approaches and, additionally, to establish conclusions on wood classification in different scenarios, saving time and money. Note

that the extreme heterogeneity of wood makes it very difficult to perform a comprehensive comparison of the different techniques with real samples. So, the generation of artificial TG curves, mimicking the real ones, in different scenarios represents a useful tool in this setting. The analysis of the simulation results, jointly with the conclusions derived from the experiments with the real samples, could give clues of a general good approach to be recommended to practitioners in this field.

### 3.4.1. Data-generating process

The generation of synthetic TG curves is an important point of this simulation study. The process is the following:

1. The generalized logistic model, given in Eq. (3.3), is fitted to each real TG curve, obtaining a 16-dimensional vector of parameters for each TG curve:

   $$((c_1, b_1, \tau_1, m_1), (c_2, b_2, \tau_2, m_2), (c_3, b_3, \tau_3, m_3), (c_4, b_4, \tau_4, m_4)).$$

   After these fits, 49 16-dimensional vectors (7 for each kind of wood) are obtained.

2. For the $r$th class of wood ($r = 1, 2, \ldots, 7$), $n_r$ 16-dimensional vectors, $(x_1, y_1, z_1, t_1, \ldots, x_4, y_4, z_4, t_4)$, are generated from a multivariate normal distribution, $N_{16}(\mu_{(r)}, \Sigma_{(r)})$, where $\mu_{(r)}$ is the sample mean and $\Sigma_{(r)}$ represents a variance-covariance matrix, both computed from the 7 vectors of parameters of the $r$th class of wood, obtained from the logistic fits. We establish the condition of simulating a new vector if any of the components of the generated vector is negative.

3. Finally, we define,

   $$c_j = \frac{x_j}{\sum_{l=1}^{4} x_l} 100, b_j = y_j, \tau_j = z_j, m_j = t_j, \quad j = 1, 2, \ldots, 4,$$

   and using the logistic model (Eq. (3.3)), we obtain the artificial TG curves ($n_r$ for each class of wood, with $r = 1, 2, \ldots, 7$).

The variance-covariance matrix $\Sigma_{(r)}$ for the $r$th kind of wood, $r = 1, 2, \ldots, 7$ is defined by:

$$\Sigma_{(r)} = \begin{pmatrix} \Sigma_{(r),1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Sigma_{(r),2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Sigma_{(r),3} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \Sigma_{(r),4} \end{pmatrix},$$

with $\Sigma_{(r),j}$ a variance-covariance matrix of the $j$th component ($j = 1, 2, \ldots, 4$) of the logistic model. In our studies, we use two parameters, $\alpha$ and $\beta$, to control the amount of variance and the degree of the dependence, respectively, in the matrix $\Sigma_{(r),j}$. Specifically, we consider

$$\Sigma_{(r),j} = \alpha\Sigma^{(1)}_{(r),j} + \beta\Sigma^{(2)}_{(r),j},$$

with

$$\Sigma^{(1)}_{(r),j} = \begin{pmatrix} \sigma^2_{(r),1,j} & 0 & 0 & 0 \\ 0 & \sigma^2_{(r),2,j} & 0 & 0 \\ 0 & 0 & \sigma^2_{(r),3,j} & 0 \\ 0 & 0 & 0 & \sigma^2_{(r),4,j} \end{pmatrix}$$

and

$$\Sigma^{(2)}_{(r),j} = \begin{pmatrix} 0 & \sigma_{(r),12,j} & \sigma_{(r),13,j} & \sigma_{(r),14,j} \\ \sigma_{(r),21,j} & 0 & \sigma_{(r),23,j} & \sigma_{(r),24,j} \\ \sigma_{(r),31,j} & \sigma_{(r),32,j} & 0 & \sigma_{(r),34,j} \\ \sigma_{(r),41,j} & \sigma_{(r),42,j} & \sigma_{(r),43,j} & 0 \end{pmatrix}.$$

Therefore,

$$\Sigma_{(r),j} = \begin{pmatrix} \alpha\sigma^2_{(r),1,j} & \beta\sigma_{(r),12,j} & \beta\sigma_{(r),13,j} & \beta\sigma_{(r),14,j} \\ \beta\sigma_{(r),21,j} & \alpha\sigma^2_{(r),2,j} & \beta\sigma_{(r),23,j} & \beta\sigma_{(r),24,j} \\ \beta\sigma_{(r),31,j} & \beta\sigma_{(r),32,j} & \alpha\sigma^2_{(r),3,j} & \beta\sigma_{(r),34,j} \\ \beta\sigma_{(r),41,j} & \beta\sigma_{(r),42,j} & \beta\sigma_{(r),43,j} & \alpha\sigma^2_{(r),4,j} \end{pmatrix},$$

where the matrix

$$\begin{pmatrix} \sigma^2_{(r),1,j} & \sigma_{(r),12,j} & \sigma_{(r),13,j} & \sigma_{(r),14,j} \\ \sigma_{(r),21,j} & \sigma^2_{(r),2,j} & \sigma_{(r),23,j} & \sigma_{(r),24,j} \\ \sigma_{(r),31,j} & \sigma_{(r),32,j} & \sigma^2_{(r),3,j} & \sigma_{(r),34,j} \\ \sigma_{(r),41,j} & \sigma_{(r),42,j} & \sigma_{(r),43,j} & \sigma^2_{(r),4,j} \end{pmatrix}$$

is the sample variance-covariance matrix of the $j$th component of the logistic model, $j = 1, \ldots, 4$. Note that a value of $\beta = 0$ indicates that the logistic model parameters are independent.

Different values of $\alpha$ and $\beta$ were used in the simulation study. Each simulation setting was repeated $B = 1000$ times and results were obtained by averaging over the $B$ replicates. As an example, Figures 3.5 and 3.6 show the artificial TG curves in two scenarios, $\alpha = 0{,}05$, $\beta = 0$, and $\alpha = 2$, $\beta = 0{,}5$, respectively. Note the similarity with the original curves in Figure 3.1, but with different variability between the curves in both scenarios.

### 3.4.2. Results

In the first part of the study, 140 artificial TG curves (20 of each class) were simulated. Table 3.5 shows the probabilities of misclassification (averages over the 1000 replicates) obtained using leave-one-out cross-validation in a variety of scenarios. Three datasets were considered: the raw TG curves, the regression parameters resulting from the application of the logistic mixture model, given in Eq. (3.3), and the values obtained using PCA, explaining of the 99 % of the total variance. Nine scenarios were chosen, corresponding to different values of $\alpha$ and $\beta$.

Table 3.5 shows that the best results were obtained using the logistic parameters as dataset and LDA as the classification method. Successful classifications were performed applying LDA on the above mentioned dataset in all the studied scenarios: the worst result was obtained when $\alpha = 4$ and $\beta = 0$ (probability of misclassification equal to 0,26), and the best for $\alpha = 0{,}05$ and $\beta = 0$ (probability of correct classification equal to 1). As expected, the smaller value of $\beta$ (less variance), the higher probabilities of correct classification. The application of SVM, NBC and NN produced competitive results. In fact, SVM gave the lower probability of misclassification in the case of scenarios with a high variance and covariance, such as $\alpha = 2, \beta = 2$ (applied to the logistic parameters). The results of the K-NFDA method were generally worse than those obtained by LDA, SVM, NN and NBC, and similar to those produced by $k$-NN. Nevertheless, the probabilities of correct classification with this method were very high for the cases of independence ($\beta = 0$) and small variances, needing a shorter computing time than the other methods.

Figure 3.7 shows boxplots of the misclassification errors over the 1000 replicas for the different classification methods, using leave-one-out cross-validation. Four scenarios are included in this plot, $\alpha = 0{,}25$ and $\beta = 0$, $\alpha = 1$ and $\beta = 0$, $\alpha = 4$ and $\beta = 0$, and $\alpha = 2$ and $\beta = 1$. For the multivariate approaches, boxplots using the parameters selected by the logistic fits (denoted by 'P') and by PCA (denoted by 'PC') are shown in this figure.

An external validation procedure, based on a random selection of the test and training samples, was proposed to supplement the results obtained by cross-validation. 1000 curves were simulated. The number of samples of each species was assigned by simulating a multinomial distribution (with equal probabilities to each one of the 7 classes or species). Then, a test and a training samples were randomly selected, and the misclassification probability was computed. This process was repeated 50 times. A similar procedure was carried out, for example, in Ferraty and Vieu (2003) or Bin and Qingzhao (2008). In this part of the study, we used the raw artificial curves to validate the K-NFDA method, and the corresponding parameters

obtained from the logistic fits, in the case of the multivariate approaches. A 10 % of the total sample was selected as the training sample, while the remaining 90 % corresponded to the test sample. The use of relatively small training samples would be in line with the frameworks usually found in real situations. The results shown in Table 3.6, obtained by external validation, are quite similar to the probabilities of misclassification when using the cross-validation procedure.

Additionally, samples using a multinomial distribution with non-equal probabilities for the different species were generated. We used probabilities of 0,3 for Scots pine and beech, and 0,08 for the remaining ones. The results are presented in Table 3.7.

Table 3.7 shows the robustness of the classification methods studied in this paper when the number of samples corresponding to each species and used in the training sample are changed.

Figure 3.8 shows boxplots of the misclassification errors over the 50 replicas for the different classification methods in the external validation experiments. Four scenarios are included in this plot, $\alpha = 0{,}25$ and $\beta = 0$, $\alpha = 1$ and $\beta = 0$, $\alpha = 4$ and $\beta = 0$, and $\alpha = 2$ and $\beta = 1$. The frameworks of equal probabilities of generating a curve of each group (denoted by 'I') and non-equal probabilities (denoted by 'NI') are shown in this figure.

In general, the conclusions obtained from the simulations are similar to those in the initial application in scenarios of small variance and covariance. When these parameters increase, the best methods are LDA, SVM and NBC. In general, the multivariate methods work better when the parameters obtained with the generalized logistic model are used. Therefore, our general recommendation with this kind of data and in this framework could be to use LDA or SVM, using the parameters selected with the logistic model. They provide general good results no matter the degree of variance or covariance of the data (in other words, no matter the heterogeneity of the real samples). On the other hand, although the K-NFDA did not provide very good results in the simulations, this method could be useful if the interest is to find the temperature ranges where the highest probability of correct classification is reached. This problem is also interesting in this setting from a practical point of view (Tarrío-Saavedra et al., 2011).

## 3.5. Conclusions

In the present paper, the performance of different functional and multivariate classification methods to classify wood species, using their TG curves,

has been tested. The different approaches have been compared using real data and also validated through a comprehensive simulation experiment. Similar conclusions can be derived from both studies. They have shown that the classification of wood species is possible by applying these statistical techniques to their corresponding TG curves.

The main contribution of the simulation study in the present research is the possibility to design scenarios with different values of the variances and covariances of the artificial TG curves. This allows to study the behavior of the methods proposed in more unfavorable situations than those obtained experimentally. These situations are indeed very likely, given the high heterogeneity of a material like wood (even within the same species). Three different databases have been considered: the raw TG curves, the parameters obtained from fitting a generalized logistic mixture regression model to each TG curve, and the principal components of the TG curves (accounting for the 99 % of the variability). Two different validation procedures were used: external validation based on the random selection of the training and test samples and leave-one-out cross-validation. In general, the higher probabilities of correct classification were obtained using the logistic parameters as dataset and LDA as classification method. In this case, the application of NBC, SVM and NN produced competitive results. Higher misclassification probabilities were obtained when applying the multivariate classification methods to the features selected with PCA. Regarding the K-NFDA method, the results were generally worse, although good probabilities of correct classification were obtained when the artificial TG curves were generated in scenarios of small variance and under independence. Note also that the K-NFDA method needed of a shorter computing time than the other methods.

In the case of the external validation, high probabilities of correct classification were obtained in all the scenarios considered. Even in the most unfavorable conditions, $\alpha = 2, \beta = 0{,}5$ or $\alpha = 4, \beta = 0$, and external validation with different sample size for each simulated species, LDA, NBC and SVM provided results over 80 % of correct classification. In addition, the methodology proposed have resulted robust when the number of samples of each species in the training sample is varied.

Given all these results, the authors recommend the application of LDA to the parameters obtained by the logistic fit to perform this type of classification problems. This method has provided a general good performance in the initial application and in the simulations no matter the degree of variance or covariance of the data (in other words, no matter the heterogeneity of the real samples). Moreover, it needs of a lower computing time for its application than other approaches.

Finally, it is important to stress that this research was carried out with some specific wood species, but the methodologies used here could be applied with different species, or even with different materials.

## 3.6. Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 (ERDF included) and Grant MTM2011-22392.

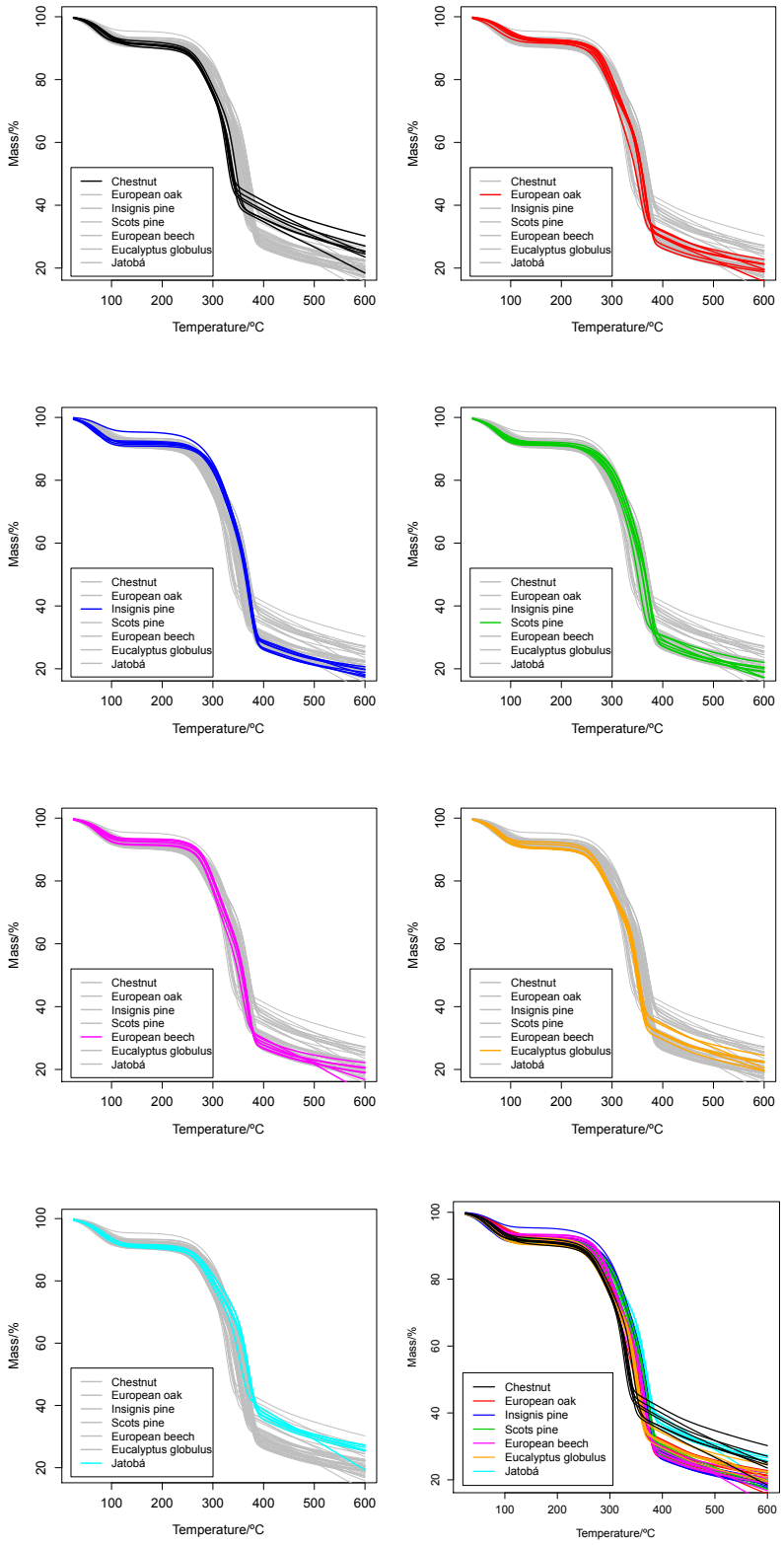


Figura 3.1: TG curves of the wood samples (7 per class) used in the analysis. Each particular wood species is highlighted in the corresponding panel. In the last panel (row 4, column 2) all the curves are presented.

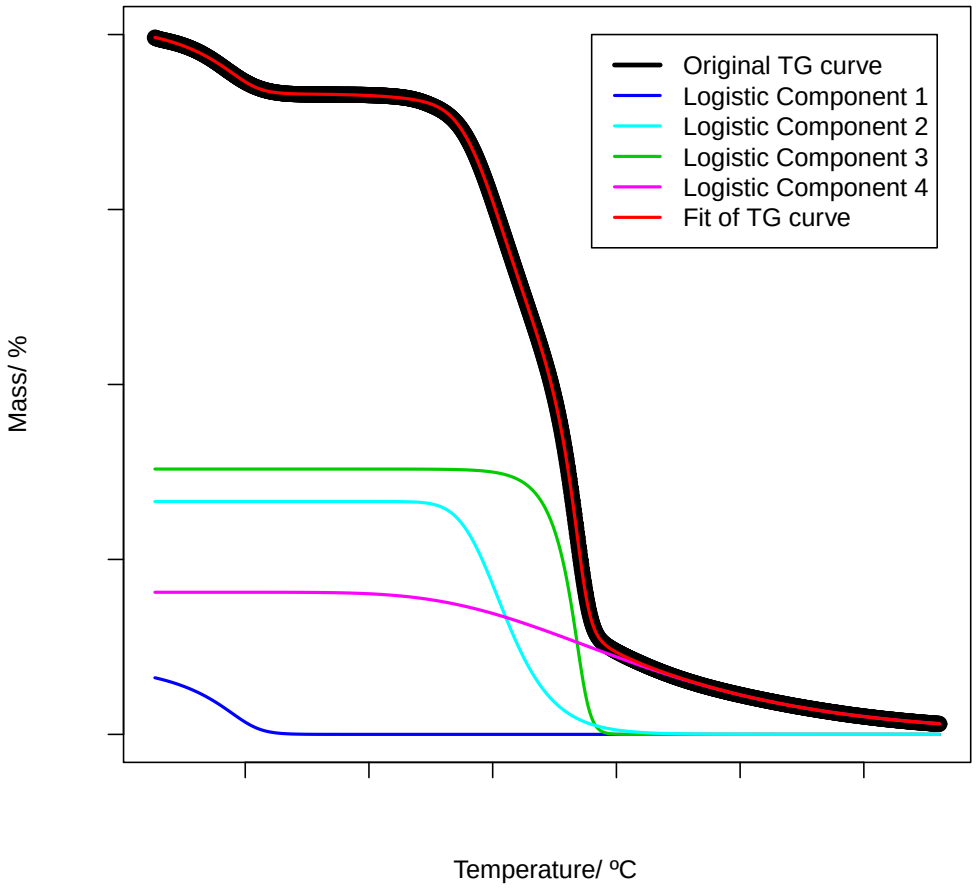


Figura 3.2: Original beech TG curve and the corresponding logistic fit.

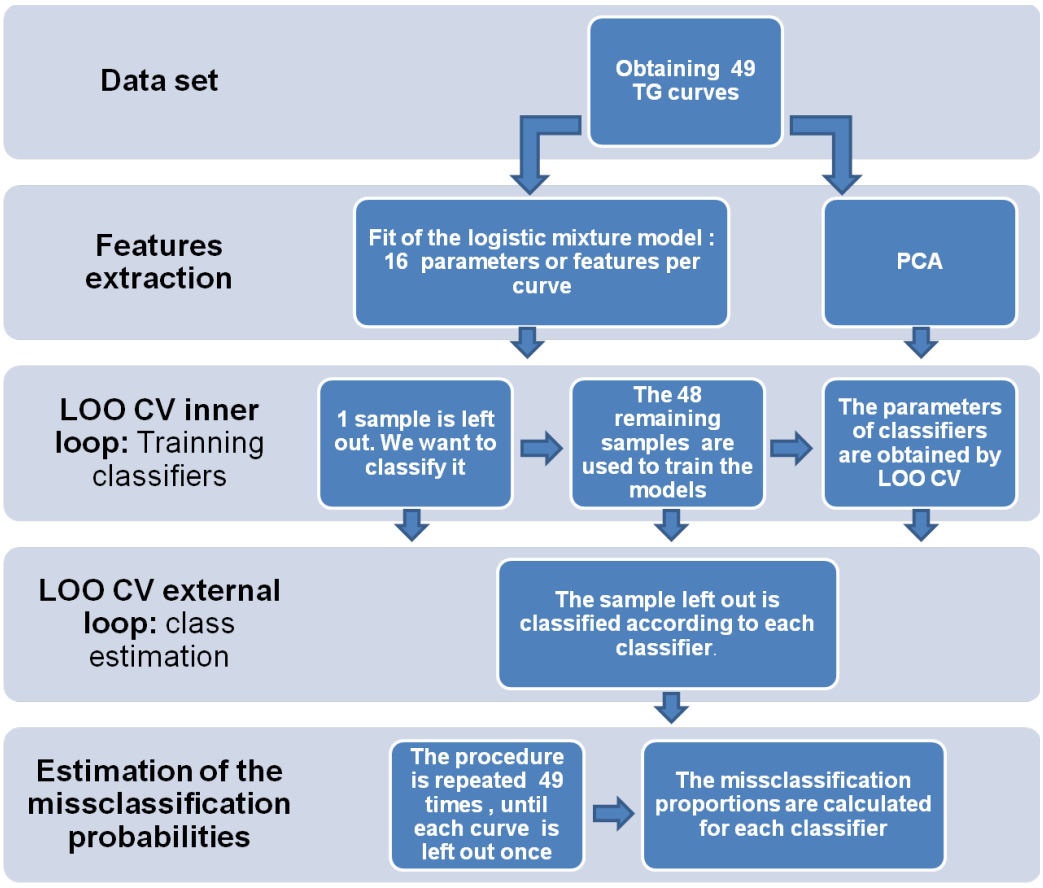


Figura 3.3: Flowchart of the leave-one-out cross-validation process for the multivariate approaches.

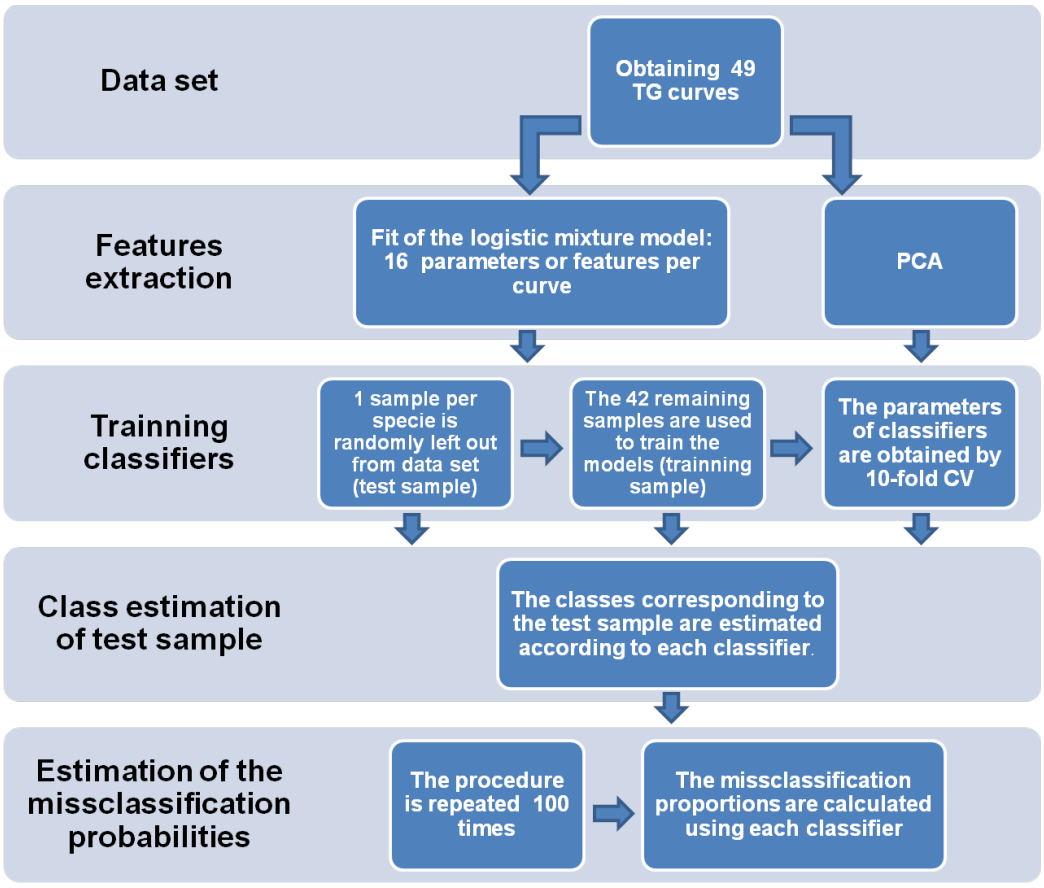


Figura 3.4: Flowchart of the external validation process for the multivariate approaches.

| | | Estimated | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | Actual | Oak | Beech | Ches. | Eucal. | Jat. | Insig. P. | Scots P. |
| LDA | Oak | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,00 | 0,00 | 0,00 | 0,00 | 0,71 | 0,00 | 0,29 |
| | Insig. P. | 0,14 | 0,29 | 0,00 | 0,00 | 0,00 | 0,57 | 0,00 |
| | Scots P. | 0,14 | 0,00 | 0,00 | 0,00 | 0,43 | 0,00 | 0,43 |
| $k$-NN | Oak | 0,71 | 0,29 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,14 | 0,00 | 0,86 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,00 | 0,00 | 0,00 | 0,00 | 0,57 | 0,00 | 0,43 |
| | Insig. P. | 0,00 | 0,00 | 0,29 | 0,00 | 0,00 | 0,71 | 0,00 |
| | Scots P. | 0,00 | 0,14 | 0,00 | 0,14 | 0,43 | 0,00 | 0,29 |
| SVM | Oak | 0,86 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,14 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,00 | 0,00 | 0,00 | 0,00 | 0,71 | 0,00 | 0,29 |
| | Insig. P. | 0,00 | 0,14 | 0,14 | 0,00 | 0,00 | 0,71 | 0,00 |
| | Scots P. | 0,14 | 0,00 | 0,00 | 0,00 | 0,29 | 0,00 | 0,57 |
| K-NFDA | Oak | 0,57 | 0,14 | 0,00 | 0,29 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,29 | 0,71 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,00 | 0,00 | 0,86 | 0,14 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,00 | 0,14 | 0,00 | 0,00 | 0,86 | 0,00 | 0,00 |
| | Insig. P. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,71 | 0,29 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 0,14 | 0,00 | 0,14 | 0,71 |

Tabla 3.3: Probabilities of correct classification, using leave-one-out cross-validation, in 7 different classes (chestnut, eucalyptus, beech, jatobá, insignis pine, oak and Scots pine) obtained by LDA, $k$-NN, SVM and K-NFDA, when using PCA to discretize the TG curves in the multivariate classification approaches. The probabilities are rounded using two significant figures.

| | | Estimated | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | Actual | Oak | Beech | Ches. | Eucal. | Jat. | Insig. P. | Scots P. |
| LDA | Oak | 0,57 | 0,29 | 0,00 | 0,14 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,14 | 0,00 | 0,00 | 0,00 | 0,86 | 0,00 | 0,00 |
| | Insig. P. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,86 | 0,14 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,29 | 0,71 |
| $k$-NN | Oak | 0,71 | 0,29 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,14 | 0,29 | 0,29 | 0,29 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,14 | 0,00 | 0,00 | 0,86 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,14 | 0,00 | 0,00 | 0,00 | 0,71 | 0,14 | 0,00 |
| | Insig. P. | 0,14 | 0,14 | 0,00 | 0,00 | 0,00 | 0,71 | 0,00 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 0,14 | 0,00 | 0,86 | 0,00 |
| SVM | Oak | 0,71 | 0,14 | 0,00 | 0,14 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,14 | 0,00 | 0,00 | 0,86 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,00 | 0,00 | 0,00 | 0,00 | 0,71 | 0,00 | 0,29 |
| | Insig. P. | 0,00 | 0,14 | 0,00 | 0,00 | 0,00 | 0,57 | 0,29 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,57 | 0,43 |
| NBC | Oak | 0,71 | 0,14 | 0,00 | 0,14 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,43 | 0,57 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Ches | 0,14 | 0,00 | 0,86 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 0,86 | 0,14 | 0,00 | 0,00 |
| | Jat. | 0,29 | 0,00 | 0,00 | 0,00 | 0,71 | 0,00 | 0,00 |
| | Insig. P. | 0,14 | 0,00 | 0,00 | 0,00 | 0,00 | 0,71 | 0,14 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,29 | 0,71 |

Tabla 3.4: Probabilities of correct classification, using leave-one-out cross-validation, in 7 different classes (chestnut, eucalyptus, beech, jatobá, insignis pine, oak and Scots pine) obtained by LDA, $k$-NN, NN and NBC, when the features of the TG curves are selected by fitting the generalized logistic model. The probabilities are rounded using two significant figures.

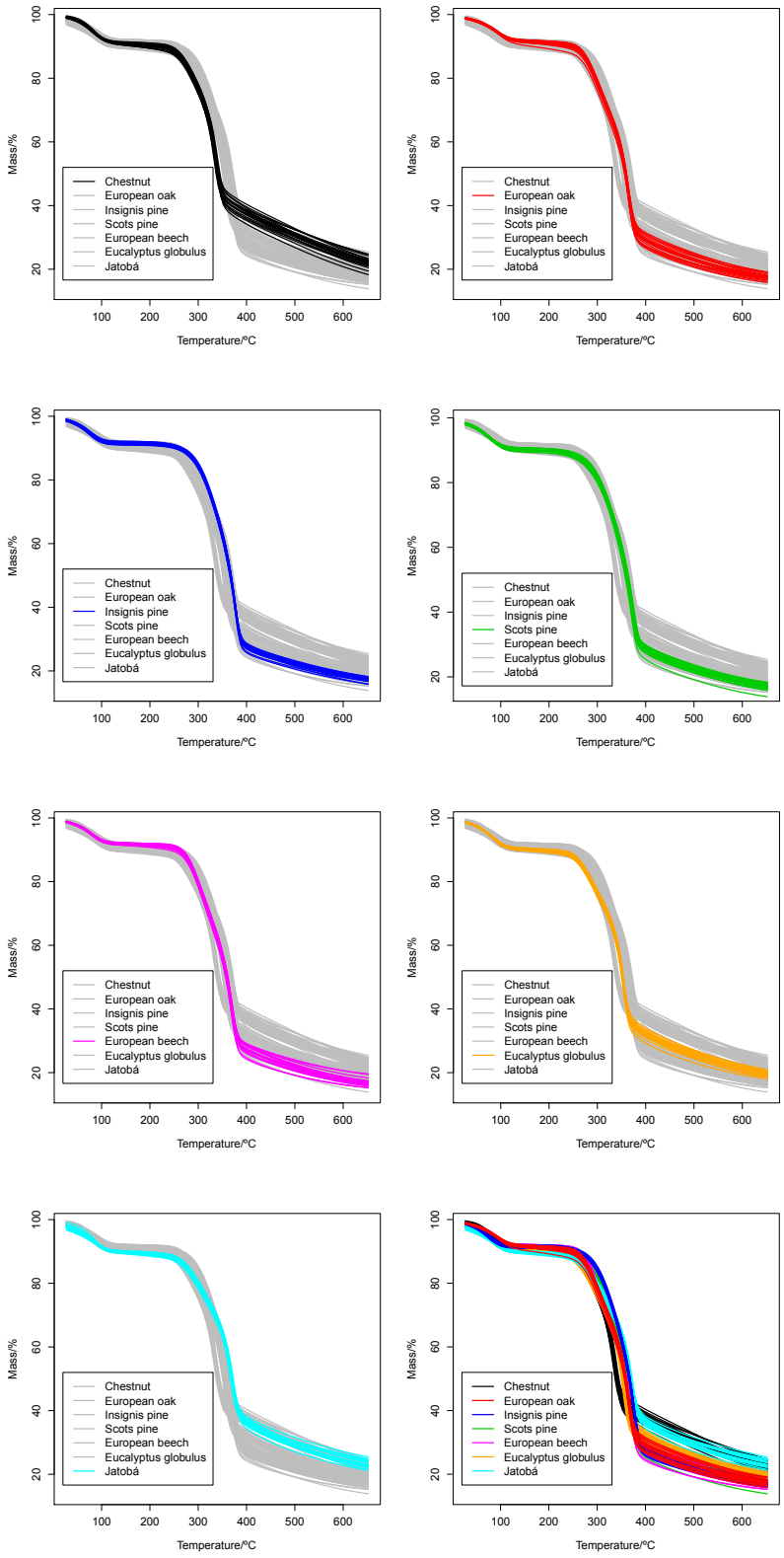


Figura 3.5: Artificial TG curves for $\alpha = 0{,}05$, $\beta = 0$. Each particular wood species is highlighted in the corresponding panel. In the last panel (row 4, column 2) all the curves are presented.

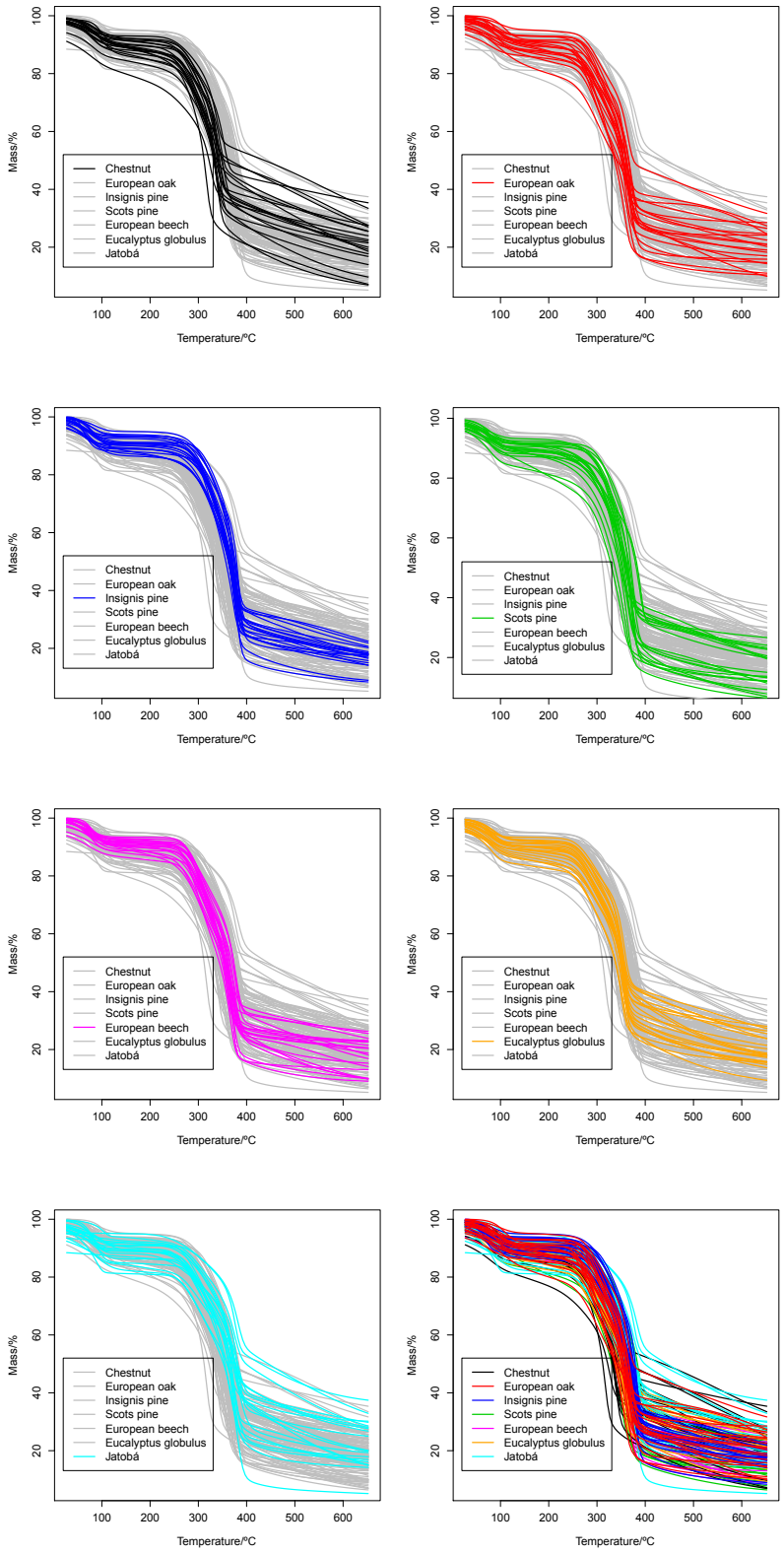


Figura 3.6: Artificial TG curves for $\alpha = 2$, $\beta = 0{,}5$. Each particular wood species is highlighted in the corresponding panel. In the last panel (row 4, column 2) all the curves are presented.

| Scenarios | | PCA of TG | | | | | Logistic parameters | | | | | Raw TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | LDA | NBC | $k$-NN | SVM | NN | LDA | NBC | $k$-NN | SVM | NN | K-NFDA |
| 0,05 | 0 | 0,00 | 0,03 | 0,01 | 0,01 | 0,01 | 0,00 | 0,00 | 0,01 | 0,00 | 0,00 | 0,01 |
| 0,25 | 0 | 0,03 | 0,09 | 0,09 | 0,06 | 0,07 | 0,01 | 0,01 | 0,08 | 0,01 | 0,01 | 0,12 |
| 0,5 | 0 | 0,10 | 0,15 | 0,17 | 0,14 | 0,15 | 0,02 | 0,03 | 0,16 | 0,05 | 0,04 | 0,22 |
| 1 | 0 | 0,20 | 0,24 | 0,27 | 0,24 | 0,26 | 0,08 | 0,10 | 0,27 | 0,11 | 0,11 | 0,32 |
| 2 | 0 | 0,32 | 0,35 | 0,39 | 0,37 | 0,38 | 0,17 | 0,19 | 0,37 | 0,19 | 0,21 | 0,41 |
| 4 | 0 | 0,45 | 0,49 | 0,53 | 0,51 | 0,50 | 0,26 | 0,27 | 0,45 | 0,27 | 0,31 | 0,51 |
| 2 | 0,5 | 0,32 | 0,35 | 0,39 | 0,37 | 0,38 | 0,17 | 0,19 | 0,36 | 0,18 | 0,21 | 0,42 |
| 2 | 1 | 0,31 | 0,33 | 0,37 | 0,35 | 0,37 | 0,18 | 0,19 | 0,34 | 0,17 | 0,21 | 0,42 |
| 2 | 2 | 0,21 | 0,24 | 0,25 | 0,22 | 0,26 | 0,12 | 0,14 | 0,19 | 0,07 | 0,11 | 0,37 |

Tabla 3.5: Misclassification probabilities corresponding to the simulation study, applying leave-one-out cross-validation to the raw TG curves, the logistic regression parameters and the PCA parameters (99 % of total variance) of the TG curves. The probabilities are rounded using two significant figures.

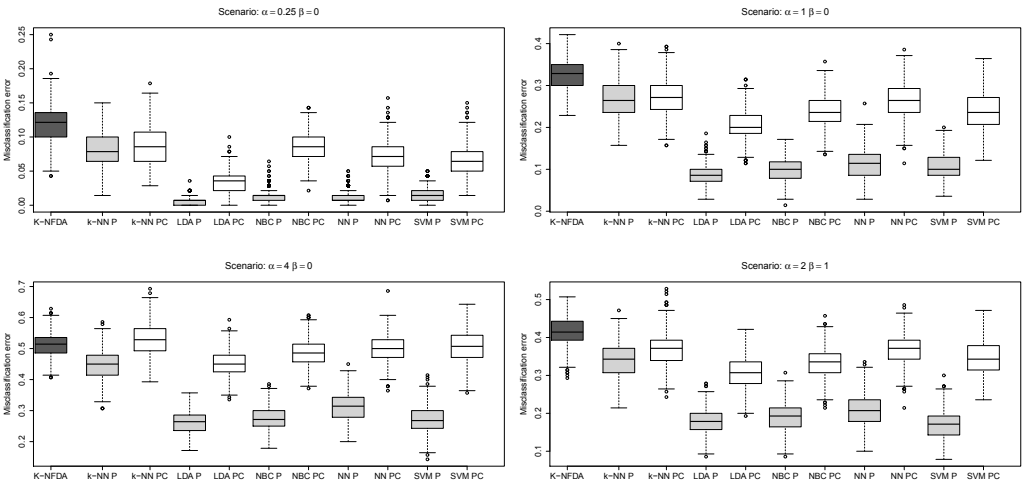


Figura 3.7: Misclassification errors over 1000 replicas using leave-one-out cross-validation. For the multivariate approaches, boxplots using the parameters selected by the logistic fits (denoted by 'P') and by PCA (denoted by 'PC') are shown.

| Scenarios | | Logistic parameters | | | | | Raw TG |
|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | LDA | NBC | $k$-NN | SVM | NN | K-NFDA |
| 0,05 | 0 | 0,00 | 0,00 | 0,01 | 0,00 | 0,00 | 0,01 |
| 0,25 | 0 | 0,00 | 0,01 | 0,06 | 0,02 | 0,01 | 0,12 |
| 0,5 | 0 | 0,03 | 0,03 | 0,13 | 0,05 | 0,04 | 0,24 |
| 1 | 0 | 0,09 | 0,08 | 0,23 | 0,09 | 0,10 | 0,33 |
| 2 | 0 | 0,16 | 0,15 | 0,31 | 0,15 | 0,20 | 0,42 |
| 4 | 0 | 0,25 | 0,21 | 0,38 | 0,22 | 0,29 | 0,51 |
| 2 | 0,5 | 0,17 | 0,15 | 0,30 | 0,15 | 0,19 | 0,43 |
| 2 | 1 | 0,17 | 0,15 | 0,28 | 0,13 | 0,19 | 0,41 |
| 2 | 2 | 0,11 | 0,11 | 0,14 | 0,05 | 0,09 | 0,35 |

Tabla 3.6: Misclassification probabilities corresponding to the simulation study, applying an external validation to the raw TG curves and to the logistic regression parameters. Scenario of equal probabilities of generating a curve of each group. The probabilities are rounded using two significant figures.

| Scenarios | | Logistic parameters | | | | | Raw TG |
|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | LDA | NBC | $k$-NN | SVM | NN | K-NFDA |
| 0,05 | 0 | 0,00 | 0,00 | 0,01 | 0,00 | 0,00 | 0,01 |
| 0,25 | 0 | 0,01 | 0,02 | 0,05 | 0,01 | 0,01 | 0,11 |
| 0,5 | 0 | 0,02 | 0,03 | 0,09 | 0,03 | 0,03 | 0,21 |
| 1 | 0 | 0,07 | 0,06 | 0,19 | 0,08 | 0,09 | 0,34 |
| 2 | 0 | 0,13 | 0,12 | 0,24 | 0,12 | 0,17 | 0,42 |
| 4 | 0 | 0,21 | 0,20 | 0,33 | 0,20 | 0,27 | 0,53 |
| 2 | 0,5 | 0,13 | 0,14 | 0,23 | 0,12 | 0,18 | 0,42 |
| 2 | 1 | 0,14 | 0,14 | 0,22 | 0,13 | 0,19 | 0,41 |
| 2 | 2 | 0,09 | 0,11 | 0,14 | 0,07 | 0,10 | 0,37 |

Tabla 3.7: Misclassification probabilities corresponding to the simulation study, applying an external validation to the raw TG curves and to the logistic regression parameters. Scenario of non-equal probabilities of generating a curve of each group. The probabilities are rounded using two significant figures.

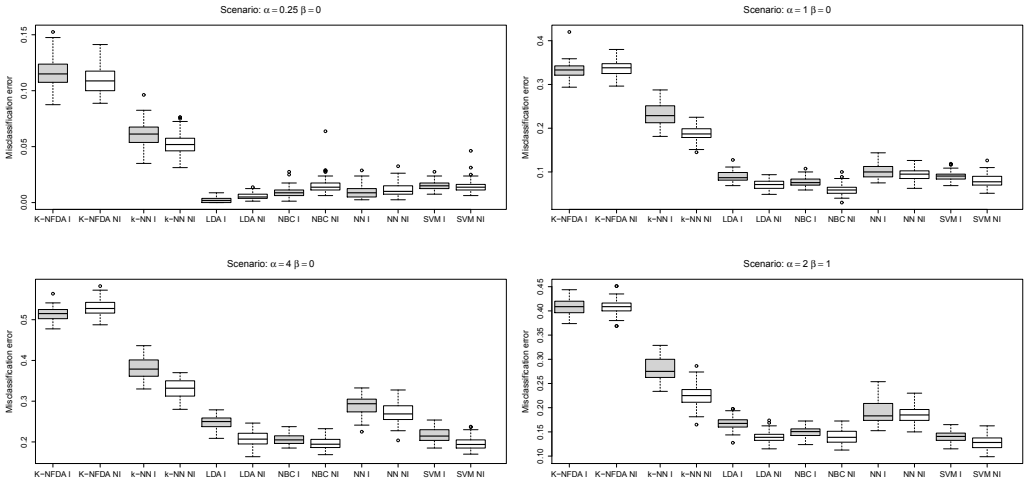


Figura 3.8: Misclassification errors over 50 replicas for the external validation process. Boxplots in the case of equal probabilities of generating a curve of each group (denoted by 'I') and non-equal probabilities (denoted by 'NI') are shown.

# Capítulo 4

# Functional nonparametric classification of wood species from thermal data

**Resumen:** En este trabajo, se han empleado las curvas termogravimetricas (TG) y calorimétricas (DSC), obtenidas por medio de un analizador simultáneo TG / DSC, además de métodos estadísticos de clasificación funcional no paramétrica, para clasificar materiales, en particular diferentes especies de madera. Además, se proporcionan los intervalos de temperatura para los que se obtiene una mayor probabilidad de clasificación correcta de las especies. Para esta labor, como en este trabajo cada observación es una curva, se ha empleado una técnica discriminante no paramétrica de caracter funcional basada en la regla de Bayes: el estimador de Nadaraya-Watson funcional, por medio del cual se consigue asignar una observación futura -en este caso una curva- a la clase predefinida para la que se estima una mayor probabilidad de pertenencia (clasificación supervisada). El parámetro de suavización necesario para la aplicación de este método no paramétrico se selecciona de acuerdo con la técnica de validación cruzada. El método propuesto se aplica a un conjunto de 49 muestras de madera (7 por clase) para distinguir entre las distintas especies y, además, para clasificar entre maderas duras (frondosas) y blandas (coníferas). La metodología propuesta ha sido aplicada con éxito, obteniendose altas proporciones de clasificación correcta, especialmente cuando el objeto de estudio son las curvas TG. Los resultados obtenidos han sido comparados con aquéllos resultado de la aplicación de otros métodos no paramétricos funcionales basados en algoritmos "boosting". Finalmente, se presenta una discusión acerca de la relación existente entre los resultados obtenidos y los rangos de temperatura de degradación

referenciados correspondientes a los 3 principales constituyentes de la madera.

**ABSTRACT:** In this study, thermogravimetric (TG) and differential scanning calorimetry (DSC) curves, obtained by means of a simultaneous TG / DSC analyzer, and statistical functional nonparametric methods are used to classify different wood species. The temperature ranges where the highest probability of correct classification is reached are also computed. As each observation is a curve, a nonparametric functional discriminant technique based on the Bayes rule and the Nadaraya-Watson regression estimator is used. It assigns a future observation to the highest probability predefined class (supervised classification). The smoothing parameter needed in this nonparametric method is selected according to the cross-validation technique. The method proposed is applied to a sample of 49 wood items (7 per wood class) and also to classify between hardwoods and softwoods. In all the cases, the samples have been successfully classified, obtaining better results with the TG curves. The results are compared with those obtained with other nonparametric methods based on boosting algorithm. A discussion about the relation of the obtained results with the referenced wood component degradation temperature ranks is presented.

## 4.1. Introduction

The identification of wood is one of the most difficult tasks to perform related with the technology of this material, due to the wide variety of species and anatomical heterogeneity of its elements. Wood identification can often be made on the basis of readily visible characteristics such as color, odor, density, presence of pitch, or grain pattern. This analysis is typical in the furniture industries and the wood panel production. Often, the performed analysis has a non uniform accuracy due to the operator. To achieve a correct classification is essential to use microscopy techniques, physical hardness tests and chemical analysis (Guindeo Casasús et al., 1997; Lewis et al., 1994; Miller, 1999). Therefore, the implementation of quantitative models and automatic recognition methods of wood samples are justified and can be immediately useful. While there are various computational procedures to evaluate and rate the quality of a timber by inspecting its defects by image processing techniques and spectral analysis (Cavalin et al., 2006; Fuentealba et al., 2004; Guindeo Casasús et al., 1997; Gu et al., 2009; Lampinen et al., 1995; Watanabe et al., 2010), these are not so generally used for species

identification, despite several studies addressing this problem exist (Khalid et al., 2008; Lavine et al., 2001; Lewis et al., 1994; Nuopponen et al., 2006; Piuri and Scotti, 2010; Yang et al., 1999).

A first step in a classification problem is to choose a discriminant feature from which it will be possible to classify. In the case of wood species classification, this discriminant feature could be the output of an experimental technique that really differentiates between them. In the literature, wood samples are mainly classified based on the results of two techniques: image- and spectrum-based processing systems. A method of classification of 20 types of tropical timber from image processing, using extracting textural wood features has been successfully tested in Khalid et al. (2008). Those authors obtained a good classification proportion of 95 %. On the other hand, in (Lewis et al., 1994), the Fourier Transform Raman (FTR) spectroscopy and neural network technology have been coupled for spectral feature extraction and non supervised classification. This represents the first time that both methodologies are combined. Later, neural networks and the FTR spectra for hardwoods and softwoods to differentiate temperate woods from tropical woods were also used (Yang et al., 1999). Genetic algorithms and principal components analysis were used to classify 98 Raman spectra of temperate softwoods, hardwoods and Brazilian and Honduran tropical woods (Lavine et al., 2001). Recently, in Piuri and Scotti (2010), an automatic wood type classification system based on the analysis of the fluorescence spectra, using nearest neighbor classifiers, linear and quadratic classifiers, and support vectors machines (SVMs) has been designed. However, it seems that the possibility of using thermal analysis as a source of data for statistical classification of wood species has not been enough studied yet. In the present paper the thermograms obtained by thermogravimetric analysis (TG) and differential scanning calorimetry (DSC) are used as a discriminant characteristic. These curves can be processed in a relatively simple way with functional analysis (Ferraty and Vieu, 2006; Ramsay and Silverman, 2005, 2002) and, as shown below, the shape of the TG curves is directly related to the wood composition. Therefore, TG analysis becomes an interesting option to discriminate between classes of timber.

In general, wood is defined as the set of xylem tissues forming the trunk, roots and branches of woody plants, excluding the bark. The tubular cells size, shape and distribution, along with other anatomical elements such as wood radios, the presence of resin canals or vessels, pores, etc., in addition to the variable proportion of its chemical components, define the different wood species and their properties (Alén et al., 1996; Gašparovič et al., 2009; Grønli et al., 2002; Guindeo Casasús et al., 1997; Miller, 1999; Müller-Hagedorn et

al., 2003; Raveendran et al., 1996; Roberts, 1970; Wang et al., 2009; Yang et al., 1999). Also, the different wood types can be generally divided in two broad categories: softwoods or conifers (gymnosperms) and hardwoods (dicot angiosperms), which can be subdivided into boreal, austral and tropical hardwood types (Guindeo Casasús et al., 1997; Miller, 1999). Is it possible to observe these differences among species in the shape of the TG curves in a pyrolysis test? According to existing studies, the answer is yes (Alén et al., 1996; Gašparovič et al., 2009; Grønli et al., 2002; Müller-Hagedorn et al., 2003; Raveendran et al., 1996; Roberts, 1970; Wang et al., 2009; Yang et al., 1999).

The wood degradation in an inert atmosphere is dominated by the degradation behavior of its three main components (Alén et al., 1996). These are cellulose, lignin and hemicellulose (Alén et al., 1996; Gašparovič et al., 2009; Grønli et al., 2002; Müller-Hagedorn et al., 2003; Raveendran et al., 1996; Roberts, 1970; Wang et al., 2009; Yang et al., 1999). Cellulose represents about 40 and 60 % in the overall weight of dry wood (it accounts for 23 %-33 % of the mass of softwoods), 23-33 % of lignin in softwoods (16-25 % in hardwoods) and 25-35 % of hemicellulose (more in hardwoods than in softwoods) (Miller, 1999; Grønli et al., 2002; Mohan et al., 2006). The three components decompose in temperatures ranging between 240-350 °C, 280-500 °C and 200-260 °C, respectively (Alén et al., 1996; Grønli et al., 2002; Mohan et al., 2006; Müller-Hagedorn et al., 2003; Roberts, 1970; Wang et al., 2009). As it was reported in (Gašparovič et al., 2009; Yang et al., 1999), the TG curve describing the pyrolysis of wood nearly coincides with the sum of the degradation of its constituents. In many cases, no significant interaction between them has been concluded (Raveendran et al., 1996). The proportion of each wood component varies depending on the species, to a greater or lesser extent (Grønli et al., 2002; Müller-Hagedorn et al., 2003; Roberts, 1970; Wang et al., 2009). Therefore, it is expected that the TG curves are different depending on the type of wood to which they belong. While the effect of the wood structure appears to exist (Roberts, 1970), this is much lower than that of the components (Yang et al., 1999; Gašparovič et al., 2009; Raveendran et al., 1996). Furthermore, differences in the pyrolysis of lignin and hemicellulose depending on whether these come from softwood or hardwood, or even of different species, were observed (Müller-Hagedorn et al., 2003; Wang et al., 2009). These results suggest the use of discriminant characteristic TG curves.

Accordingly, the objectives of this study are:

1. Evaluating the potential of functional nonparametric methods of discriminant analysis for the classification of hardwoods and softwoods

and then for the classification of European oak, European chestnut, European beech, jatobá, eucalyptus, scots and insignis pine on the basis of TG and DSC data.

2. Comparing the accuracy performance of TG or DSC curve classification to discriminate between wood species or between major groups. The supervised kernel nonparametric classification and kernel nonparametric classification using the $k$-nearest method to select the bandwidth $h$ are used (Ferraty and Vieu, 2006). In addition, two methods based on the boosting algorithm are also used to complete the study: using principal components analysis (PCA) and by the representation of functional data on a $b$-spline basis (Ramsay and Silverman, 2005, 2002; Bühlmannand and Hothorn, 2007).

3. Finding the temperature range in TG and DSC curves where the highest probability of correct classification is reached.

4. Relating the results of classification analysis in each interval with the referenced cellulose, lignin, and hemicellulose degradation temperature ranks in a nitrogen atmosphere.

## 4.2. Experimental

### 4.2.1. Materials

Tests for five different hardwoods (european beech or *Fagus sylvatica*, european oak or *Quercus robur*, chestnut or *Castanea sativa*, *Eucalyptus globulus* and jatobá or *Hymenaea courbaril*) and two softwoods (scots pine $-$*Pinus silvestris* and insignis pine $-$*Pinus radiata*) are carried out. Seven samples per each one of the above mentioned species, obtained from wood of different trees are tested. The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the time of experimentation. The samples are not dried in order to avoid disturb as much as possible their structure and composition, and to test the automatic classification method with a minimal sample preparation, under the worse conditions.

### 4.2.2. Measurement methods

The test is performed on a SDT 2960 TA Instruments thermo balance. This apparatus provides both TG and DSC curves used in the classification analysis. A heating ramp of $20\,^{\circ}\mathrm{C\ min^{-1}}$ is applied in the range from 20 to

600 °C, at a rate of 50 mL $\text{min}^{-1}$ of $N_2$ (Prime et al., 2009). The nitrogen is purged for 10 min, before starting the heating program for establishing an inert environment. The used heating rate is chosen to obtain a proper balance between time test and resolution (Prime et al., 2009). It aims to assess the discriminatory power of the resulting curves, using the minimum experimental time. The sample mass chosen is between 6 and 8 mg. Alumina crucibles are used. In particular, TG and DSC measurements are affected by some experimental parameters such as heating rate, amount of mass, type of atmosphere or sample geometry (Prime et al., 2009). Therefore, all these parameters are remained constant to obtain a better classification.

## 4.3. Classification techniques

Nonparametric functional techniques based on kernel methods (Ferraty and Vieu, 2006) and two nonparametric methods based in the boosting algorithm are applied to construct a classification rule to discriminate between hardwoods and softwoods, and between the different species: European beech, European oak, European chestnut, eucalyptus, jatobá, scot, and insignis pine, based on a sample of 49 TG and DSC curves. A DSC or TG 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 (K-NPFDA), shown in (4.1), is applied. Given a new TG or DSC 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:

$$\hat{r}_h^{(g)}(x) = \frac{\sum_{i=1}^{n} I_{\{Y_i=g\}} K\left(\frac{\|x-X_i\|}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{\|x-X_i\|}{h}\right)}, \tag{4.1}$$

where the observed TG/DSC 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. The parameter $h$ is the bandwidth or smoothing parameter and $\|\cdot\|$ denotes the following distance between curves:

$$d(X_i, X_j) = \int_a^b \left(X_i(t) - X_j(t)\right)^2 dt, \tag{4.2}$$

where $[a, b]$ is one of the 1280 temperature intervals studied.

After a careful examination of the TG/DSC curves, further processing of the data has been found useful for standardizing the curves (López-Granados et al., 2008). Denoting by $f(x)$ a curve, a linear transformation, $\tilde{f}(x) = \alpha f(x) + \beta$ with

$$\alpha = \frac{\sqrt{b-a}}{\sqrt{\int_a^b \left(f(t) - \frac{1}{b-a}\int_a^b f(s)ds\right)^2 dt}}$$

and

$$\beta = 1 - \frac{\int_a^b f(t)dt}{\sqrt{b-a}\sqrt{\int_a^b \left(f(t) - \frac{1}{b-a}\int_a^b f(s)ds\right)^2 dt}},$$

is done to achieve

$$\frac{1}{b-a}\int_a^b \tilde{f}(t)dt = 0$$

and

$$\frac{1}{b-a}\int_a^b \left(\tilde{f}(t) - \frac{1}{b-a}\int_a^b \tilde{f}(s)ds\right)^2 dt = 1$$

This transformation should act on the mean and variance to improve the discriminant power of the curves.

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 (Naya et al., 2006). This method consists in minimizing the cross-validation function:

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

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

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

It can be useful and efficient to replace the $h$ parameter, a real number, by an integer parameter $k$ from a finite subset. A way to achieve this is to consider a $k$ Nearest Neighbors (kNN) version of the kernel estimator (Ferraty and Vieu, 2006). In the present work, it is named KNN-NPFDA method. The number of neighbors and the bandwidth is selected using the cross-validation method.

When the kNN estimator is used, the parameter $h$ is replaced by $h_k$, which is the bandwidth allowing us to take into account $k$ terms in the weighted average (Ferraty and Vieu, 2006):

$$\hat{p}_k^{(g)}(x) = \frac{\sum\limits_{\{i:Y_i=g\}}^{n} K\left(h_k^{-1} \cdot d(x, X_i)\right)}{\sum\limits_{i=1}^{n} K\left(h_k^{-1} \cdot d(x, X_i)\right)},$$

where $h_k$ is a bandwidth such that

$$\#\{i : d(x, X_i) < h_k\} = k,$$

with # the cardinal of the set.

Two additional nonparametric methods based on the boosting algorithm, the B method and the B-PCA method, are implemented to be compared with the kernel methods. They are specially designed to perform a nonparametric supervised classification for functional data. The boosting algorithm used is the Adaboost algorithm for classification (Buhlmannand and Hothorn, 2007).

In the B method, the boosting algorithm estimates the optimal number of basis and the optimum depth of the tree partition using Functional Data Object for obtaining the best possible estimation.

In the B-PCA method, the Adaboost algorithm is applied to a set of data using Principal Component Analysis. The optimal number of Principal Components and the optimum depth for one or more classifiers are estimated.

The free statistical software R (R Development Core Team, 2008) is employed to implement the nonparametric functional methods used in this article. Mainly, the R packages `fda` and `fda.usc` are used to perform the classification applying nonparametric functional analysis. Adaboost algorithm is a modified algorithm of the function `adabag.M1` of `adabag` R package that fits adaboost algorithm with classification trees as individual classifiers.

## 4.4. Results and discussion

In this Section the methods previously presented are applied to the TG and DSC curves to classify between different species and main groups. First, a descriptive analysis of the data is shown. It is important to note that each method is validated through cross-validation, which is the technique widely used for the validation of an empirical model. It works by leaving out one TG/DSC curve; then a model is trained with the remaining thermograms and, finally, the developed model is used for the classification of the left

out TG/DSC curve. This is repeated until all the curves have been left out once. As the data set available contain 49 samples, 48 samples are used for training and 1 sample for testing. This process is repeated 49 times, and the percentage (measured as per one) of correct classification are calculated.

### 4.4.1. Descriptive analysis of the TG curves

Fig. 4.1 shows the 49 TG curves obtained (7 per class). As it can be observed, each curve represents a functional data. A particular wood specie is highlighted in each panel. The last panel (in row 4, colum 2) shows all the TG curves. Different trends are observed for almost all types of wood. Even so, the variability in each class prevents from discerning clearly in some intervals. Apart from this, some species such as oak and beech tend to overlap. This is correlated to their similar densities, hardness and mechanical properties. In addition, they belong to the broader group of boreal hardwood (Guindeo Casasús et al., 1997).

In Fig. 4.2a the means of the TG curves for each class of wood, as defined by Fraiman and Muniz (2001), are plotted. They can report on the possible degree of overlapping among the degradation trends for the different species and in what intervals this happens. In fact, it appears that differences between species are starting to take place from 200 °C onwards, coinciding with the beginning of the hemicellulose degradation (Roberts, 1970; Mohan et al., 2006). These differences become maxima in the range of temperatures where the cited maximum decomposition rate of the cellulose and lignin occurs (Yang et al., 1999).

As the storage period is long enough (over a year) and the storage conditions of all wood samples are the same, through the TG curves is possible to measure the water absorption capacity of each timber. In fact, it is observed that the height of the first step is slightly different for some species, being able to build two groups: oak, beech, insignis pine and, on the other hand, chestnut, eucalyptus and jatobá. It can be observed that the existing residue in the range of 400-600 °C is different depending on the species (chestnut and jatobá > eucalyptus > oak, scots pine, beech and insignis pine).

Fig. 4.3a shows the variability in each class. By working with functional data, the variance of the TG curves for a class is not a value, but a curve. The greatest variability occurs in the range of temperatures where the maximum decomposition rate of the cellulose and lignin is produced, according to Yang et al. (1999). Therefore, the use of classification methods should work worse at these temperatures (320-370 °C, see Fig. 4.3a), but this depends on the magnitude of the variability among species in this interval, which is also

higher (see mean differences in Fig. 4.2a). This variability is expected given the heterogeneity of wood (Guindeo Casasús et al., 1997; Miller, 1999).

Fig. 4.4 shows the DSC curves obtained for every sample tested, in the same form as in Fig. 4.1. There is a wide variability in each class, higher than in the TG curves. Therefore, it is expected that the DSC curves discriminate worse than TG between classes of wood. According to the literature, the endo and exo DSC events of hemicellulose, cellulose and lignin overlap in the 220-520 °C range (Yang et al., 1999; Roberts, 1970; Mohan et al., 2006). The differences among species can be set according to the displacement and magnitude of these three previous peaks (see Fig. 4.5), which may be due to different weight percentage of hemicellulose, cellulose and lignin, interactions or differences in the structure (Grønli et al., 2002; Müller-Hagedorn et al., 2003; Roberts, 1970; Wang et al., 2009). There are also differences in the peak corresponding to the water loss (see Fig. 4.5). The classification methods implemented may determine whether these differences are reliable.

By implementing a location-scale transformation, a reduction of the variance in each class is sought. Fig. 4.6 and 4.7 show the TG and DSC data after this processing step. A reduction in the dispersion of the curves in each class is observed (Fig. 4.3). There is also an increasing distance between curves from different classes in certain temperature ranges.

### 4.4.2. Result of the data transformation

We apply nonparametric kernel methods for functional data (K-NPFDA and KNN-NPFDA) and methods based on boosting algorithm (B and B-PCA). It is observed that the transformation of the data (López-Granados et al., 2008) significantly helps to distinguish among the different groups (hardwoods and softwoods) and also among species. For example, if the TG curves are analyzed by the K-NPFDA method, classifying samples from seven kinds of wood, a probability of correct classification equal to 0,79 is obtained in the best of possible intervals (180-330 °C) while that if we use the transformed data a probability of 0,88 is obtained in the range 192,5-292,5 °C. This is repeated for all models and data analysis, therefore, hereafter the results using the transformed data are shown. See Table 4.1.

### 4.4.3. TG curve classification

In Table 4.2 the probabilities of correct classification and the temperature ranges for which they are maxima are shown. They are computed in two settings, classifying among the seven different species and in the more general

| | Original Data | | Transformed Data | |
|---|---|---|---|---|
| Data | Number of Classes | Corr. Class. Prob. | Number of Classes | Corr. Class. Prob. |
| TG | 3 | 0,94 | 3 | 0,94 |
| TG | 7 | 0,79 | 7 | 0,88 |
| DSC | 3 | 0,42 | 3 | 0,79 |
| DSC | 7 | 0,24 | 7 | 0,57 |

Tabla 4.1: Correct classification probabilities using original and transformed data, with 3 or 7 classes.

case of classifying into three different groups. This is the result of evaluating the probabilities at 1280 intervals of eight different sizes, from 50 to 400 °C. The four methods of classification described above are calculated.

| | 3 Groups Classification | | 7 Groups Classification | |
|---|---|---|---|---|
| Methods | Optimal Interval/ °C | Corr. Class. Prob. | Optimal Interval/°C | Corr. Class. Prob. |
| K-NPFDA | 192,5-292,5 | 0,94 | 192,5-292,5 | 0,88 |
| KNN-NPFDA | 192,5-292,5 | 0,94 | 182,5-282,5 | 0,88 |
| B | 210,0-310,0 | 0,90 | 217,5-417,5 | 0,90 |
| B-PCA | 210,0-310,0 | 0,90 | 195,0-345,0 | 0,83 |

Tabla 4.2: Correct classification probabilities and optimal intervals obtained by each classification method. The TG data were tested with 3 (boreal hardwoods, softwoods, other hardwoods) and 7 classes.

The results in Table 4.2 can be grouped in two blocks, those techniques based on a kernel and the Nadaraya-Watson estimator and those based on boosting algorithm. In the case of classification in three groups or general classes of wood, the K-NPFDA (with $0{,}0005 < h < 0{,}001$) and KNN-NPFDA techniques, with an optimal number of neighbors equal to 1, are the best methods with an estimated probability of success of almost equal to 1 (0,94). It is interesting to note that the four methods practically coincide in the optimal range (192,5-292,5 °C and 210-310 °C), which in turn coincides with the region of hemicellulose degradation reported by several authors (Roberts, 1970; Mohan et al., 2006). The hardwoods tend to have a higher content of hemicellulose (Miller, 1999; Grønli et al., 2002; Mohan et al., 2006) and their weight proportion influences the total degradation of

wood. Classification between hardwoods and conifers has been successfully achieved. Also, the correct classification among austral, boreal and tropical hardwoods is obtained. In fact, there is very small confusion between the groups. In Table 4.3, the classification matrices in the optimal intervals presented in Table 4.2 using 3 different classes (boreal hardwoods, softwoods and other hard woods) are shown. It can be observed that using the K-NPFDA method the probability of classifying a boreal as a tropical or austral hardwood is only 0,05 and the probabilities of correct classification in each group are very high in this case (0,95 for boreal hardwoods, 0,86 for softwoods and 1 for other hardwoods). This result is similar to that obtained by other techniques using image- and spectrum-based processing systems (Khalid et al., 2008; Lavine et al., 2001; Lewis et al., 1994; Nuopponen et al., 2006; Piuri and Scotti, 2010; Yang et al., 1999).

| | | Theoretical | | |
|---|---|---|---|---|
| Methods | Estimated | B. Hardwoods | Softwoods | O. Hardwoods |
| K-NPFDA | B. Hardwoods | 0,95 | 0,07 | 0,00 |
| | Softwoods | 0,00 | 0,86 | 0,00 |
| | O. Hardwoods | 0,05 | 0,07 | 1,00 |
| B | B. Hardwoods | 0,97 | 0,00 | 0,23 |
| | Softwoods | 0,00 | 1,00 | 0,06 |
| | O. Hardwoods | 0,03 | 0,00 | 0,71 |

Tabla 4.3: Classification matrices using 3 different classes (boreal hardwoods, softwoods and other hard woods) obtained by different nonparametric classification methods, using TG data.

The overall probabilities of correct classification when one wants to discriminate among the seven existing species of wood is also very high. Especially interesting are the results obtained using the methods K-NPFDA and B (7 elements in the basis and depth of tree equal to 3). In the first case, a probability of correct classification of 0,88 for the interval 192,5-292,5 °C is obtained (see Table 4.2). This could be due to the different hemicellulose content and differences in hemicellulose degradation depending on the species, but we must do more experiments to prove it. The optimal interval (217,5-417,5 °C) obtained by the method B includes the degradation processes of the hemicellulose, cellulose and lignin getting a slightly higher probability of correct classification. In fact, according to Müller-Hagedorn et al. (2003), the differences in wood species are mainly due to the different thermochemical behavior of lignin degradation and that of the first step of the hemicellulose

degradation.

Table 4.4 shows the classification matrices in the optimal intervals presented in Table 4.2 using seven different classes. It can be observed that the results obtained by Methods B and K-NPFDA are complementary. The probabilities of correct classification of each kind of wood are relatively high in both cases. The K-NPFDA method only fails to discriminate the scots pine class (P = 0,58). Instead, a probability of correct classification equal to 1 of the scots pine class is obtained by the B method. Among other things, it can be due to the much larger optimal temperature range, encompassing much of the lignin degradation which provides more information to differentiate the species. On the other hand, the K-NPFDA method classifies slightly better oak and jatobá woods than the B method (see Table 4.4).

| | | Actual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | Estimated | Chesn. | Oak | Insig. P. | Scots P. | Beech | Eucal. | Jat. |
| K-NPFDA | Ches. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Oak | 0,00 | 0,72 | 0,00 | 0,14 | 0,00 | 0,00 | 0,00 |
| | Insig. P. | 0,00 | 0,00 | 1,00 | 0,14 | 0,00 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 0,58 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 0,14 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,14 | 0,00 | 0,00 | 0,00 | 1,00 | 0,14 |
| | Jat. | 0,00 | 0,00 | 0,00 | 0,14 | 0,00 | 0,00 | 0,86 |
| B | Ches. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Oak | 0,00 | 0,67 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Insig. P. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,17 |
| | Jat. | 0,00 | 0,33 | 0,00 | 0,00 | 0,00 | 0,00 | 0,83 |

Tabla 4.4: Classification matrices using 7 different classes (chestnut, oak, insignis pine, scots pine, beech, eucalyptus and jatobá) obtained by two different nonparametric classification methods, using TG data.

In general, it appears that the cited range of decomposition of the hemicellulose and, to a lesser extent, of lignin and cellulose is the temperature range where more differences between species exist.

Apart from using leave-one-out cross-validation, the prediction power of the K-NPFDA method is measured. For this, the whole set of 49 curves is divide in two groups, a training sample of 42 curves and a test sample of

7 curves (one for each class of timber). Our aim is try to classify correctly the test sample using the training sample. This problem is more common in industry. Table 4.5 shows the classification matrix when we want to classify among the seven considered species. In this table the results in the interval 207-307 °C are shown. This is the optimal interval using cross-validation with the training sample. It can be observed that the 100 % of the test sample is successfully classified using the K-NPFDA method. Same results for the case of classifying among the 3 main groups are shown in Table 4.6 (in the interval 192,5-292,5 °C), obtaining the same success.

| 207-307 °C | | Actual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| K-NPFDA | Estimated | Ches. | Oak | Insig. P. | Scots P. | Beech | Eucal. | Jat. |
| New | Ches. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| Sample | Oak | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| P=1 | Insig. P. | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 |
| | Jat. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 |

Tabla 4.5: Classification matrix using 7 different classes (chestnut, oak, insignis pine, scots pine, beech, eucalyptus and jatobá) obtained by K-NPFDA using a training sample with 42 TG curves. Probabilities of correct classification of a new sample consisting of 7 curves (one per class).

| 192.5-292,5 °C | | Actual | | |
|---|---|---|---|---|
| K-NPFDA | Estimated | B. Hardwoods | Softwoods | O. Hardwoods |
| New | B. Hardwoods | 1,00 | 0,00 | 0,00 |
| Sample | Softwoods | 0,00 | 1,00 | 0,00 |
| P=1 | O. Hardwoods | 0,00 | 0,00 | 1,00 |

Tabla 4.6: Classification matrix using 3 different classes (boreal hardwoods, softwoods y tropical and austral hardwoods) obtained by K-NPFDA using a training sample with 42 TG curves. Probabilities of correct classification of a new sample consisting of 3 curves.

In conclusion, it is proved that using TG curves as discriminant characteristic is possible to classify different species of wood.

### 4.4.4. DSC curves classification

Tables 4.7, 4.8 and 4.9 show similar results to those presented in Tables 4.2, 4.3 and 4.4, respectively, but using the DSC curves obtained by a simultaneous SDT. It can be observed that DSC curves have less discriminating power than TG curves. Nevertheless, very good results are obtained when we try to distinguish among boreal hardwoods, softwoods and tropical or austral hardwoods (Tables 4.7 and 4.8). The kernel nonparametric functional methods (K-NPFDA and KNN-NPFDA) work better than the others based on the boosting algorithm. A good classification probability equal to 0,80 is obtained in the temperature range of 322,5-485 °C. This interval corresponds to the region of maximum degradation rate of cellulose and lignin, reported by several authors (Yang et al., 1999; Roberts, 1970; Mohan et al., 2006). Therefore, the DSC curves have a higher classification power in this area; there are more differences between species. These differences may be due to the nature and weight percentage of lignin (Nuopponen et al., 2006).

| | 3 Groups Classification | | 7 Groups Classification | |
|---|---|---|---|---|
| Methods | Optimal Interval/°C | Corr. Class. Prob. | Optimal Interval/ °C | Corr. Class. Prob. |
| K-NPFDA | 322,5-485,0 | 0,80 | 322,5-472,5 | 0,57 |
| KNN-NPFDA | 322,5-485,0 | 0,80 | 322,5-347,5 | 0,60 |
| B | 330,0-575,5 | 0,60 | 325,0-375,0 | 0,46 |
| B-PCA | 330,0-575,5 | 0,67 | 325,0-375,0 | 0,38 |

Tabla 4.7: Correct classification probabilities and optimal intervals obtained by each classification method. The DSC data were tested using 3 (boreal hardwoods, softwoods, other hardwoods) and 7 classes.

| | | Theoretical | | |
|---|---|---|---|---|
| Methods | Estimated | B. Hardwoods | Softwoods | Ot. Hardwoods |
| K-NPFDA | B. Hardwoods | 0,90 | 0,29 | 0,21 |
| | Softwoods | 0,05 | 0,71 | 0,00 |
| | O. Hardwoods | 0,05 | 0,07 | 0,79 |

Tabla 4.8: Classification matrix using 3 different classes (boreal hardwoods, softwoods and other hard woods) obtained by different nonparametric classification methods, using DSC curves.

When we want to classify among the seven species, the DSC results are worse than the TG ones. The best methods are again K-NPFDA and KNN-NPFDA with a maximum probability of correct classification equal to 0,60 for the interval 322,5-347,5 °C. Thus, using the K-NPFDA method, good classification results are only obtained in the case of eucalyptus, chestnut and jatobá (Table 4.9) but the two kinds of pine are often misclassified and the beech curves are often classified as oak curves (P = 0,29). In fact, both pairs of species have very similar mechanical properties, hardness and density (Guindeo Casasús et al., 1997).

| | | Actual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | Estimated | Ches. | Oak | Insig. P. | Scots P. | Beech | Eucal. | Jat. |
| K-NPFDA | Ches. | 0,86 | 0,14 | 0,00 | 0,13 | 0,00 | 0,00 | 0,14 |
| | Oak | 0,00 | 0,57 | 0,14 | 0,00 | 0,14 | 0,14 | 0,14 |
| | Insig. P. | 0,00 | 0,00 | 0,29 | 0,29 | 0,14 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,00 | 0,43 | 0,29 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 0,29 | 0,14 | 0,29 | 0,43 | 0,00 | 0,00 |
| | Eucal. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,86 | 0,00 |
| | Jat. | 0,14 | 0,00 | 0,00 | 0,00 | 0,29 | 0,00 | 0,72 |

Tabla 4.9: Classification matrix using 7 different classes (chestnut, oak, insignis pine, scots pine, beech, eucalyptus and jatobá) obtained by K-NPFDA method, using DSC data.

In Table 4.10 we measure the prediction power of a new sample taking a training sample of 42 items. The new sample consists of seven curves, six of which have been classified successfully (P = 0,86). A higher misclassification is obtained precisely in the most heterogeneous group (other hardwoods).

| 322,5-485,5 °C | | Actual | | |
|---|---|---|---|---|
| K-NPFDA | Estimated | B. Hardwoods | Softwoods | O. Hardwoods |
| New | B. Hardwoods | 1,00 | 0,00 | 0,50 |
| Sample | Softwoods | 0,00 | 1,00 | 0,00 |
| P=0.86 | O. Hardwoods | 0,00 | 0,00 | 0,50 |

Tabla 4.10: Classification matrix using 3 different classes (boreal hardwoods, softwoods y tropical and austral hardwoods) obtained by K-NPFDA using a training sample with 42 DSC curves. Probabilities of correct classification of a new sample consisting of 3 curves.

## 4.5. Conclusions

Classifying different species of wood using the TG curves as discriminant characteristic has been proved possible (percentage of correct classification = 90 %). Also, the classification between hardwoods, softwoods and tropical or austral hardwoods have been successfully carried out using these curves (percentage of correct classification = 94 %). The results are comparable to those obtained from image- and spectrum-based processing systems. It was observed that the temperature ranges corresponding to the higher probabilities of correct classification basically match with those reported for the single components decomposition (mainly hemicellulose).

The DSC curves obtained by a simultaneous SDT have less discriminant power than that of the TG curves. Nevertheless, using these curves, very good results are obtained when we try to distinguish among boreal hardwoods, softwoods and tropical or austral hardwoods (percentage of correct classification = 80 %) and among certain types of hardwoods (chestnut, jatobá and eucalyptus). Moreover, the referenced temperature range corresponding to the maximum rate of decomposition of lignin and cellulose is the range where more differences among species were found, using DSC curves.

In general, K-NPFDA and KNN-NPFDA methods, based on the non-parametric Nadaraya-Watson functional regression estimator, have provided probabilities of correct classification superior to the others based on the boosting algorithm, and with a shorter computing time.

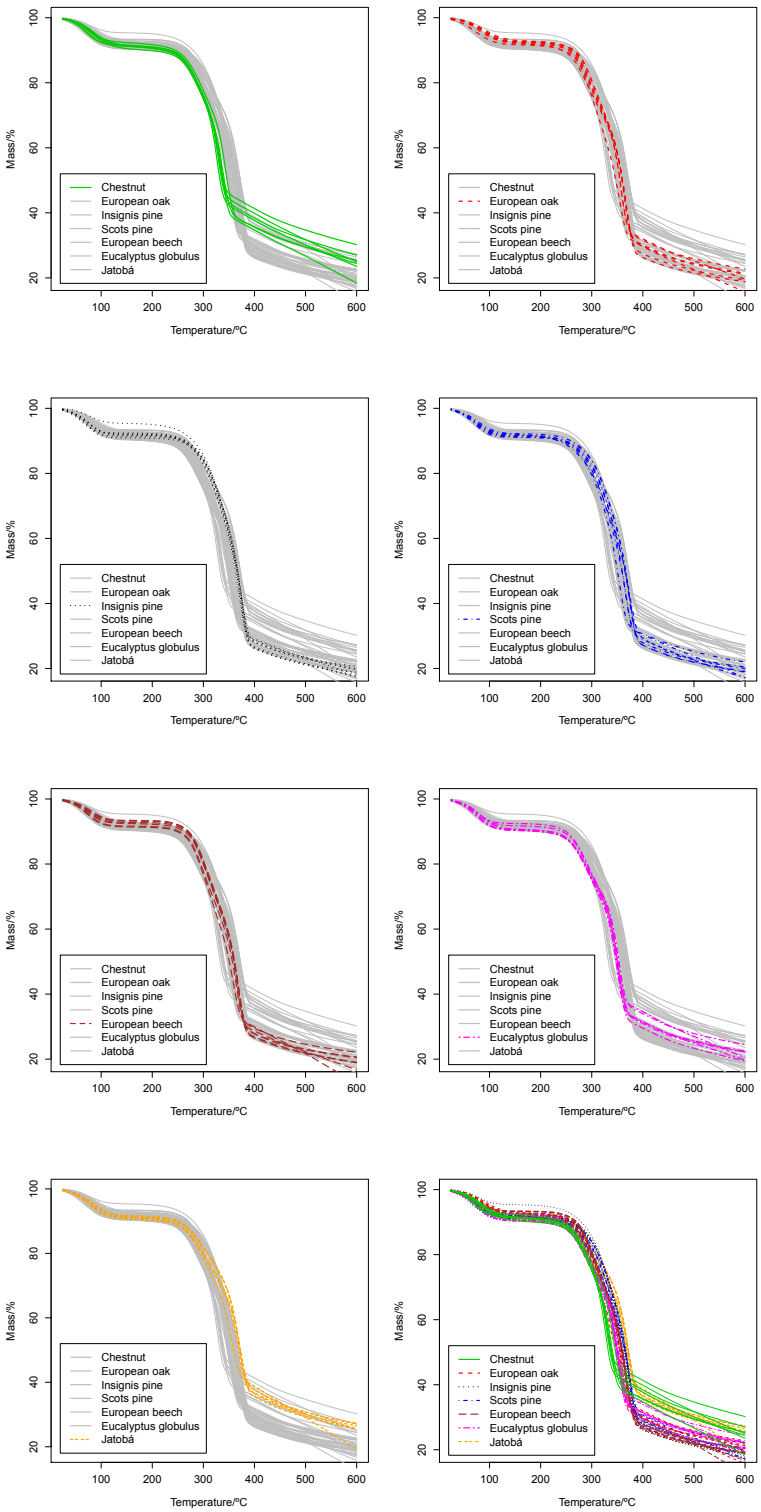


Figura 4.1: Original TG curves (7 per class), where each particular wood specie is highlighted in the corresponding panel. In the last panel (row 4, colum 2) all the curves are presented

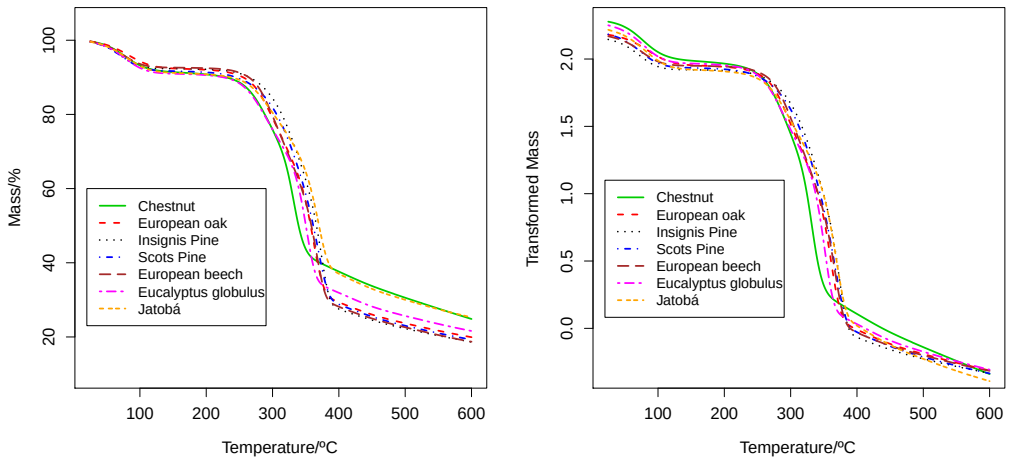


Figura 4.2: **a** Functional means of the original TG curves for each specie (left panel). **b** Functional means of the location-scale transformed TG curves (right panel).

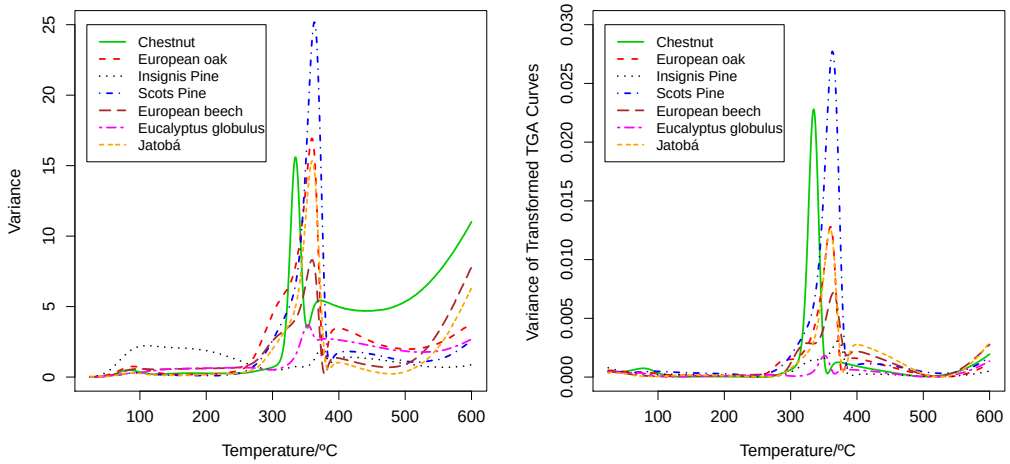


Figura 4.3: **a** Functional variances of the original TG curves for each specie (left panel). **b** Functional variances of the location-scale transformed TG curves (right panel).

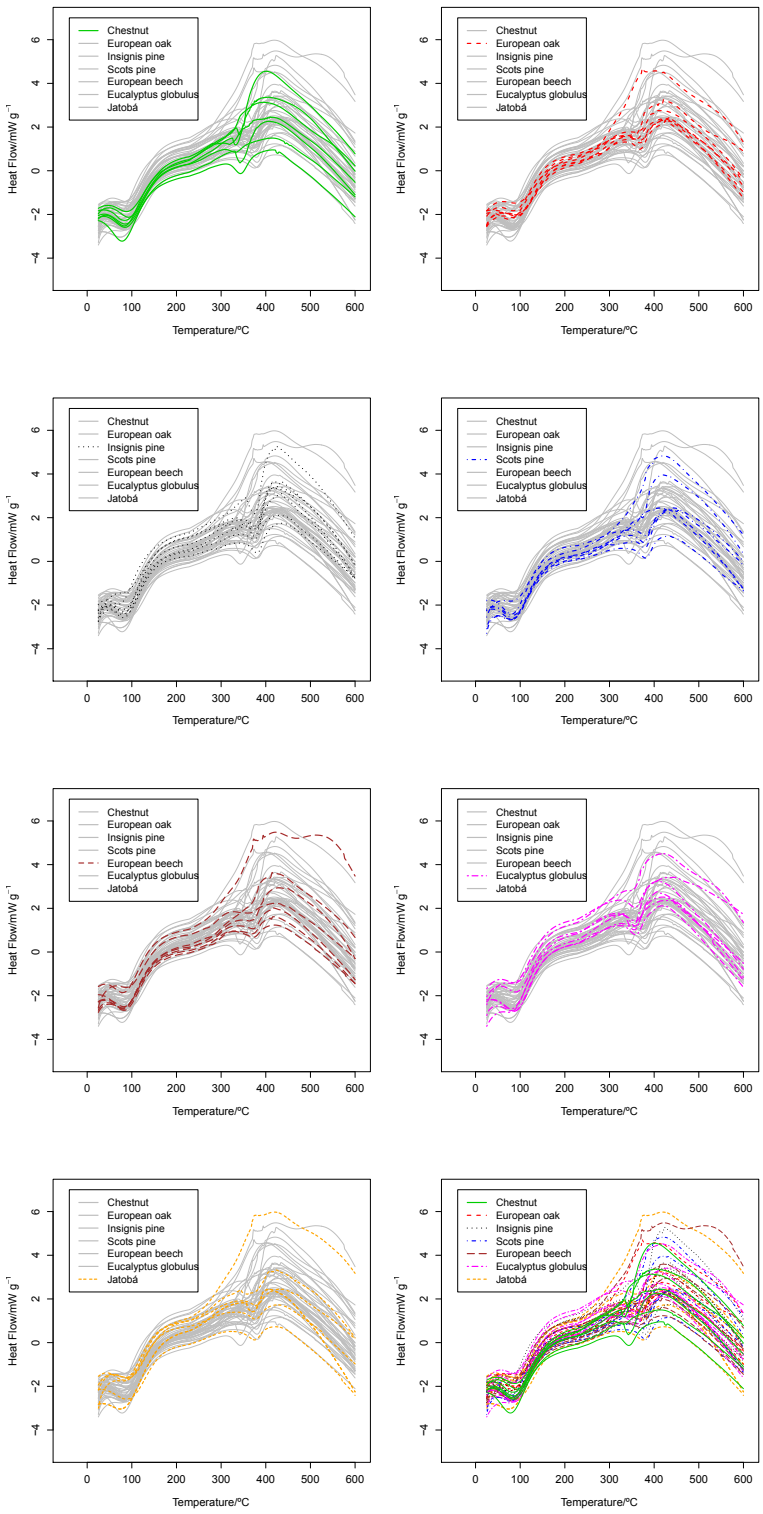


Figura 4.4: Original DSC curves (7 per class), where each particular wood specie is highlighted in the corresponding panel. In the last panel (row 4, colum 2) all the curves are presented.

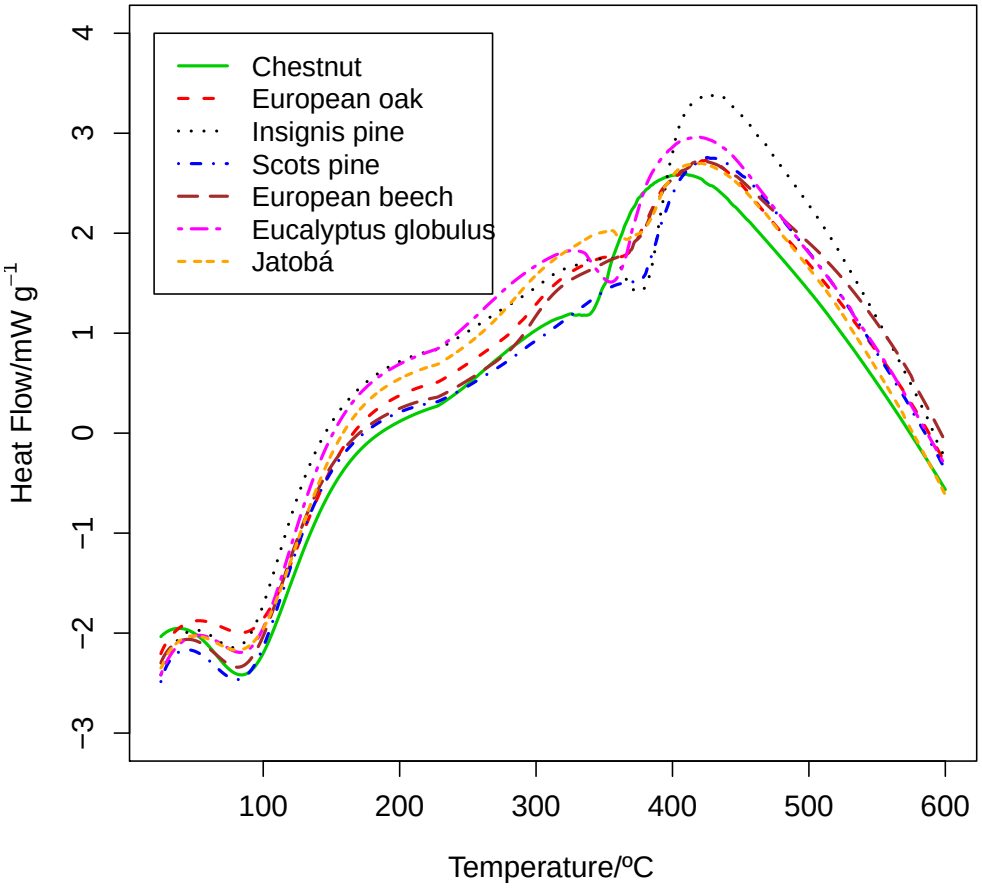


Figura 4.5: Functional means of DSC curves.

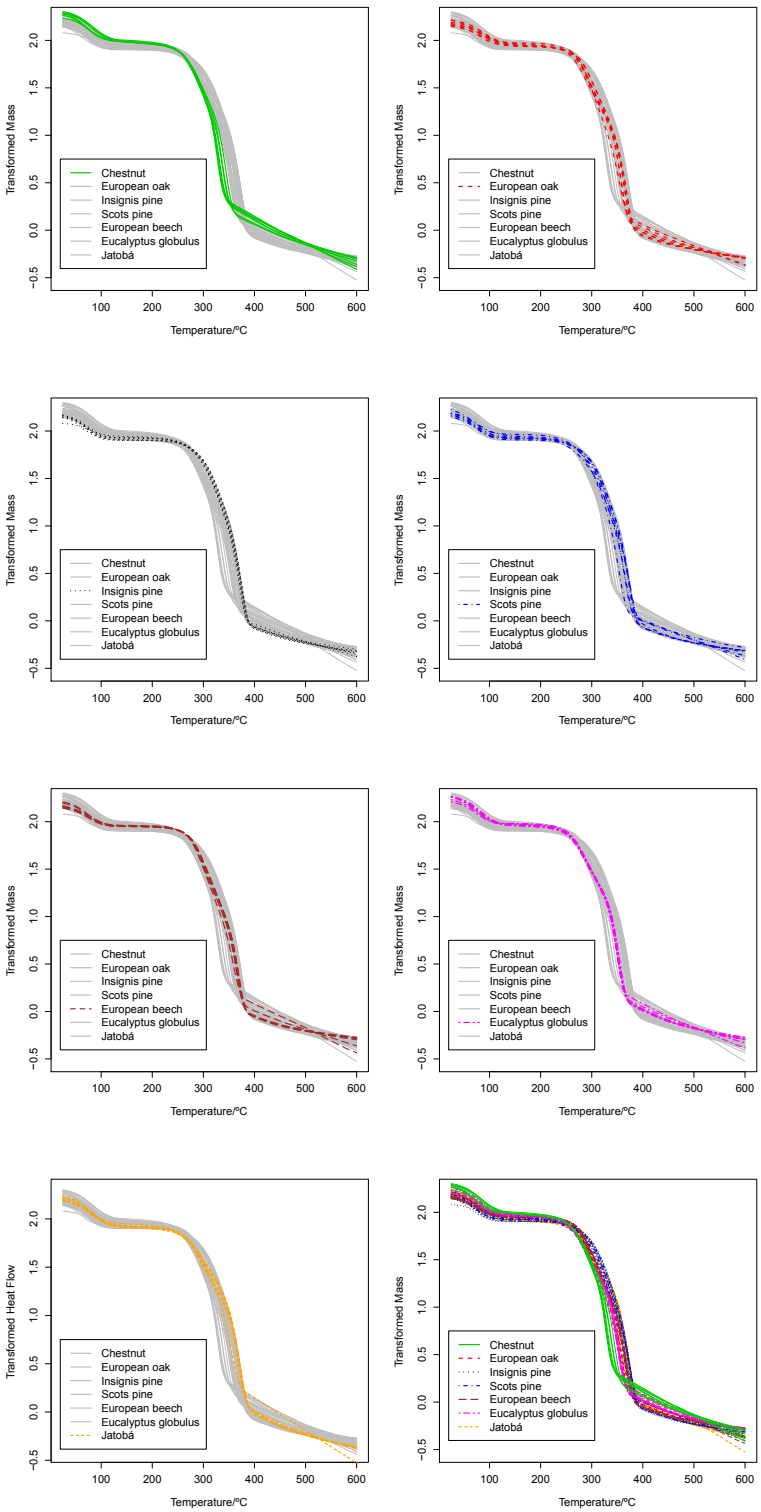


Figura 4.6: Location-scale transformed TG curves.

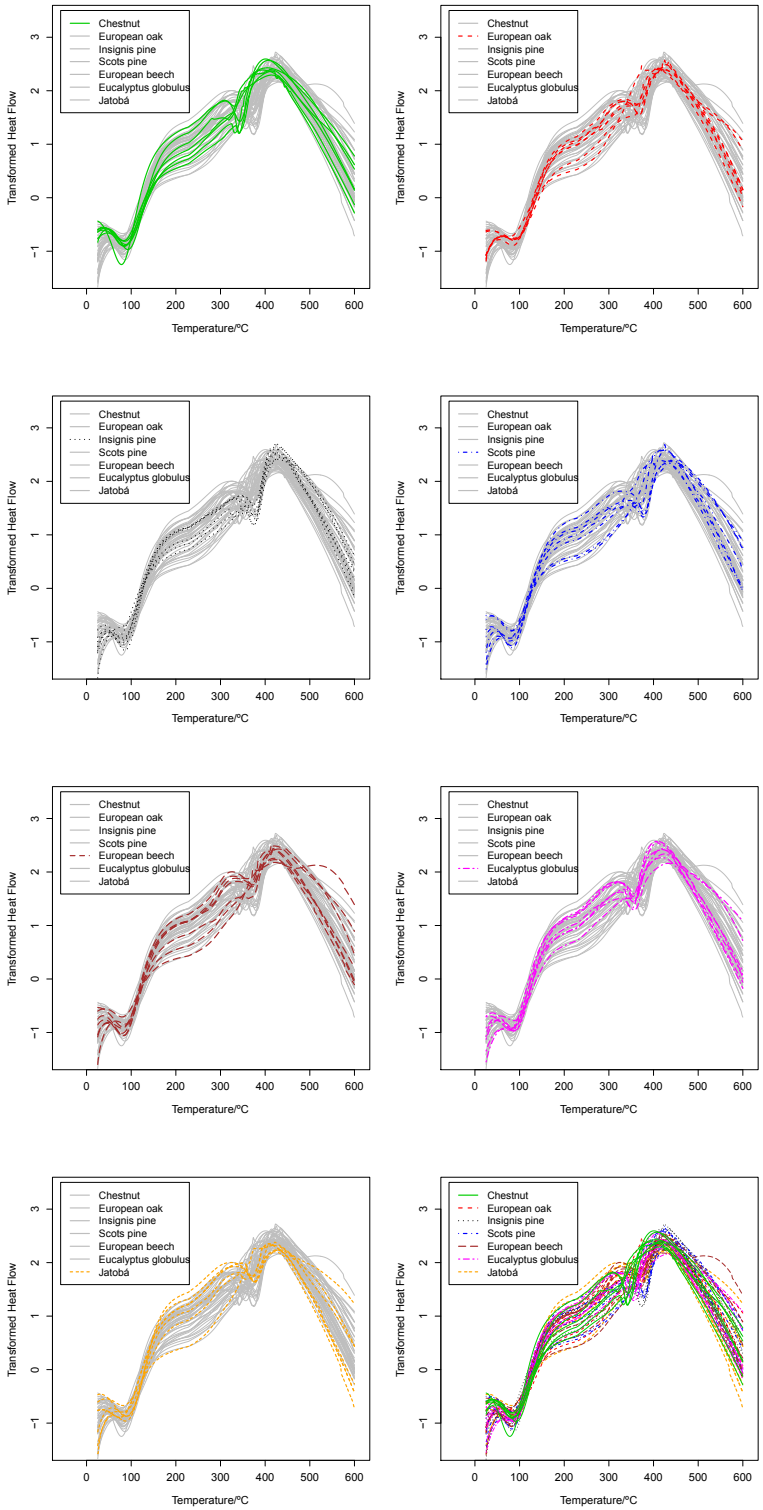


Figura 4.7: Location-scale transformed DSC curves.

# Capítulo 5

# Classification of wood micrographs by image segmentation

**Resumen:** El objetivo principal de este estudio consiste en clasificar especies de madera utilizando la micrografías obtenidas a 1500× aumentos mediante Microscopía Electrónica de Barrido (SEM) y procesadas por el método de segmentación de imágenes. Los resultados muestran que es posible observar diferencias entre especies en la textura de la madera, observada a 1500× aumentos. Las micrografías han sido procesadas de forma sencilla mediante la segmentación y reconocimiento de objetos para identificar las secciones transversales de las traqueidas pertenecientes a madera joven perteneciente a 7 especies de madera diferentes: *Fagus sylvatica*, *Castanea sativa*, *Juglans regia*, *Eucalyptus globulus*, *Hymenaea courbaril*, *Pinus silvestris* and *Pinus radiata*. Se ha analizado la forma, número y distribución de las traqueidas mediante cinco características: la circularidad, ortogonalidad, el número de traqueidas, la distancia entre traqueidas y área media de las mismas. Las muestras de madera, definidas por las características extraídas, se clasifican utilizando diferentes métodos estadísticos: el Análisis Discriminante Lineal (LDA), la Clasificación Cuadrática, la Regresión Logística, $K$ Nearest Neighbors o vecinos más próximos (KNN), Bayes Naive (NBC), Máquinas de Soporte Vectorial (SVM) y Redes Neuronales (NN). También se presenta un estudio comparativo basado en las características obtenidas por "gray level co-occurrence", siendo evidente la mejora que supone usar el método de segmentación propuesto. Finalmente, se muestra la posibilidad de utilizar el análisis fractal en el marco de este capítulo para completar la investigación.

**ABSTRACT:** The principal aim of this study is to classify wood species using scanning electron microscopy (SEM) micrographs obtained with 1500× magnification and processed by image segmentation. The results show that is possible to observe differences among species in the wood texture at this magnification. The micrographs have been processed in a simple way using segmentation and object recognition to identify the cross-secction tracheids belonging to earlywood of 7 different timber species: *Fagus sylvatica*, *Castanea sativa*, *Juglans regia*, *Eucalyptus globulus*, *Hymenaea courbaril*, *Pinus silvestris* and *Pinus radiata*. We have analyzed the shape, number and distribution of the tracheids using 5 features: circularity, rectangularity, number of tracheids, distance between tracheids and average area. The extracted features are classified using different statistical methods: Linear Discriminant Analysis (LDA), Quadratic classification, Logistic Regression, $K$ Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machines (SVM) and Neural Networks. A comparative study using gray level co-occurrence based features is also presented, being clear the improvement of using the segmentation method. Moreover, some additional results showing the possibility of using fractal analysis in this framework complete the research.

## 5.1. Introduction

Whether a piece of furniture made from an unfamiliar wood has to be restored or you are debating the authenticity of a particular board with a local lumberyard, a knack for identifying a piece of lumber is a useful skill. This skill becomes very important if we transfer these cases to the industry. Depending on the specie to which it belongs, a timber has a certain physical-chemical properties determining its industrial applications. Thus, given its properties, there are woods suitable for the development of flats, others for the manufacture of structural elements in a building, some for cabinet or furniture of different qualities, etc. Uses or applications that are not always interchangeable. In addition to the own properties of the wood, such as chemical resistance, tensile, bending, compression strength, hardness, elastic modulus, density, porosity, odor, etc., another important factor determining the application or not of a particular specie of wood is the price. Therefore, identifying types of wood becomes crucial to check a possible fraud as some timber traders tend to mix different types of wood to increase their profit margin (Khalid et al., 2008). It is also the case when the cheaper timber (similar at first glance) is directly sold as it were the more expensive one. Often the identification of the actual building material is the most daunting task for a furniture researcher, a construction foreman or even for

an archeologist (Hayek et al., 1990b). In fact, the identification of wood is one of the most difficult tasks to perform related with the technology of this material, due to the wide variety of species and anatomical heterogeneity of its elements. Wood identification can often be made on the basis of readily visible characteristics such as color, odor (usually caused by oils in the heartwood), density, presence of pitch, grain pattern, texture (depending on the size and distribution of its cell), type of transition from earlywood to latewood, presence of resin canals, or pores. The principal standard tool for macroscopic viewing of wood is 10× hand lens. Very often the identification can be performed using these lenses and some identification keys from books (Arno et al., 1993; Leavengood, 1998; Hoadley, 1990). This analysis is typical in the furniture industries and the wood panel production. However, many woods are impossible to tell apart without using a microscope. Sometimes a great deal of knowledge and laboratory equipment is needed to identify species (Arno et al., 1993; Leavengood, 1998; Hoadley, 1990; Guindeo Casasús et al., 1997; Lewis et al., 1994; Miller, 1999). But even having the best equipment, many times the performed analysis has a non-uniform accuracy due to the operator. Correctly identifying an unfamiliar wood sample out of thousands of possibilities requires close observation, and a thorough knowledge of wood and its properties. Training a skilled worker takes years, with the corresponding cost, and the industry trend for automatization has meant to dispense with the traditional trades. Accordingly, these workers are increasingly scarce. Therefore, the implementation of statistical models and automatic recognition methods of wood samples are justified and can be immediately useful. While there are various computational procedures to evaluate and rate the quality of a timber inspecting its defects by image processing techniques and spectral analysis (Cavalin et al., 2006; Fuentealba et al., 2004; Guindeo Casasús et al., 1997; Gu et al., 2009; Lampinen et al., 1995; Watanabe et al., 2010), these are not so generally used for species identification, although there are also several works addressing this problem (Khalid et al., 2008; Lavine et al., 2001; Lewis et al., 1994; Nuopponen et al., 2006; Piuri and Scotti, 2010; Yang et al., 1999). A first step in a classification problem is to choose a discriminant feature from which it will be possible to classify. In the case of wood species classification, this discriminant feature could be the output of an experimental technique that really differentiates between them. In the literature, wood samples are mainly classified based on the results of two techniques: image-based and spectrum-based processing systems. In Lewis et al. (1994), the Fourier Transform Raman (FTR) spectroscopy and Neural Network technology have been coupled for spectral feature extraction and non supervised classification. This represents the first time that both methodologies are combined. Later, Neural Networks and the

FTR spectra for hardwoods and softwoods to differentiate temperate woods from tropical woods were also used Yang et al. (1999). Genetic algorithms and principal component analysis were used to classify 98 Raman spectra of temperate softwoods, hardwoods and Brazilian and Honduran tropical woods Lavine et al. (2001). Recently, in Piuri and Scotti (2010), an automatic wood type classification system based on the analysis of the fluorescence spectra, using Nearest Neighbor classifiers, Linear and Quadratic classifiers, and Support Vectors Machines (SVM) is designed. Another alternative is to classify attending to the thermograms obtained by TGA (thermogravimetric analysis) Tarrío-Saavedra et al. (2011). These curves can be processed in a relatively simple way with functional analysis methods (Ferraty and Vieu, 2006; Ramsay and Silverman, 2005, 2002) and their shape is directly related to the wood composition. On the other hand, a method of classification of 20 types of tropical timber from image processing, using extracting textural wood features from wood images obtained with 10× magnification has been successfully tested in Khalid et al. (2008). They obtained a good classification proportion of 95 % using Neural Networks and test samples of 10 items. In Brandtberg (2002), different wood species using high spatial resolution infrared color aerial photographs (taken from the tree crowns) are classified. Nine different features of each image object are estimated, and transformed using principal component analysis (PCA). The accuracy, using the supervised grade of membership (GoM) model with cross-validation was 67 %. In Tou et al. (2009), macroscopic images belonged to six different tropical wood species, taken from cross-sections, are classified obtaining a success of 80 % with test samples of 60 items. They apply a rotational invariant method using the grey level co-occurrence matrices (GLCM) as the features, an energy value representing the similarity between the test sample and the template. However, it seems that the possibility of using micrographs with a bigger magnification as a source of data for statistical classification of wood species has not been enough studied yet. Is it possible to observe these differences among species in the wood texture at a magnification of 1500×, attending to the shape, number, and distribution of the tracheids? This work pretends to answer this question. Generally, wood is defined as the set of xylem tissues forming the trunk, roots and branches of woody plants, excluding the bark. The tubular cells size, shape and distribution, along with other anatomical elements such as wood radios, the presence of resin canals or vessels, etc., in addition to the variable proportion of its chemical components, define the different wood species and their properties (Alén et al., 1996; Gašparovič et al., 2009; Grønli et al., 2002; Guindeo Casasús et al., 1997; Miller, 1999; Müller-Hagedorn et al., 2003; Raveendran et al., 1996; Roberts, 1970; Wang et al., 2009; Yang et al., 1999). Also, the different wood types can be gen-

erally divided in two broad categories: softwoods or conifers (gymnosperms) and hardwoods (dicot angiosperms), which can be subdivided in boreal, austral and tropical hardwoods (Guindeo Casasús et al., 1997; Miller, 1999). The difference between softwood and hardwood, and also among species, is readily apparent when viewed under microscope magnification. For both hardwoods and softwoods, the side of the board to be examined is often discussed. In fact, wood has three distinct surfaces: the cross-section (end-grain), the tangential surface, and the radial surface. Moreover, trees growing in temperate areas with a winter season display rings. Growth rings consist of two separate layers where the first, called earlywood, is laid down at the beginning of the growing season and the second one, latewood, is formed toward the end. Earlywood is more porous than latewood (Arno et al., 1993). There are many differences in wood texture between these two zones (Arno et al., 1993; Schweingruber, 2007; Miller, 1999). Therefore, in this study, only the earlywood region is studied, defining earlywood as wood zones where double cell wall of the tracheids is smaller than the lumen (Schweingruber, 2007). In the present paper, some features corresponding to micrographs obtained by scanning electron microscopy (SEM) with 1500× magnification are used as discriminant characteristics. These images, taken in cross sections, can be processed in a relatively simple way using segmentation and object recognition to identify the elemental cells and analyze their shape, number and distribution. The extracted features are classified using Linear Discriminant Analysis (LDA), Quadratic classification, Logistic regression, $K$ Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machines (SVM) and Neural Networks. A comparative study using GLCM based features is also presented, showing the effectiveness of the segmentation method using these data. Moreover, some additional results showing the possibility of using fractal analysis in this context complete the research.

Accordingly, the objectives of this study are:

1. Checking that the correct classification of timber species (and also between softwoods and hardwoods) is feasible, using all the features obtained from the segmentation of micrographs taken with an electron microscope at 1500× magnification, where the tracheids structure can be observed.

2. Showing the advantages of image segmentation for this particular case of wood micrographs classification problem.

3. Evaluating the potential of supervised classification methods, such as LDA, Quadratic classification, Logistic regression, KNN, Naïve Bayes,

SVM and Neural Networks for distinguishing among European chestnut, European beech, eucalyptus, jatobá, walnut, and Scots and insignis pine on the basis of some features obtained by image segmentation; and additionally, for distinguishing between hardwoods and softwoods.

The content of the paper is as follows. In Section 2, the materials and experimental techniques used to obtain the micrographs are described. In Section 3, the image enhancement, the segmentation process and the classification methods used in the field of statistical analysis of multivariate data are explained. In Section 4, these methods are applied and the results are obtained and analyzed. Finally, Section 5 collects the main conclusions.

## 5.2. Experimental

Tests for 5 different hardwoods (European beech or *Fagus sylvatica*, chestnut or *Castanea sativa*, common walnut or *Juglans regia*, *Eucalyptus globulus* and jatobá or *Hymenaea courbaril*) and 2 softwoods (Scots pine $-$*Pinus silvestris*$-$ and insignis pine $-$*Pinus radiata*) are carried out. At least ten samples per each one of the above mentioned species, obtained from commercial wood of different trees are tested (see Table 5.1). The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the time of experimentation. All samples have been cut using a razor blade so that the cross section of wood can be studied. This is a common technique used when wood samples are classified by visual inspection using 10× hand lens (Arno et al., 1993; Leavengood, 1998; Hoadley, 1990). This method may very slightly affect the structure of the wood, but, in this work, we have wanted to conduct a timber classification based on standard methods for collecting wood samples, minimizing the preparation time, and thus adjusting to the industry requirements. However, this is a destructive method (although required samples are very small) and it is not applicable in case of artworks (or it must be applied very carefully). Each sample has approximately a prismatic shape with dimensions $0{,}5 \times 0{,}5 \times 0{,}1$ cm. The samples were dehydrated in a series of graded concentrations of ethanol and critical-point dried using liquid $CO_2$. The dried samples were then affixed on aluminum stubs with adhesives, covered with gold (BAL-TEC SCD004 sputter coater) and examined in a scanning electron microscope (Jeol JSM-6400), using a magnification equal to 1500×.

## 5.3. Image treatment methodology and statistical methods

### 5.3.1. Image Enhancement

The acquired images (micrographs) are enhanced for better accuracy of the results. The image is cropped for the important part of the picture. Sometimes there can be too much noise in the picture due to higher resolution in the microscope. Image noise is the random variation of color and brightness information in the image. This can occur from several sources including electrical sensor noise, photographic grain noise and channel errors. Noise can be of many types, like amplifier noise (Gaussian noise), Salt and Pepper noise, shot noise, quantization noise, etc. Image noise arising from a noisy sensor or channel transmission errors usually appears as discrete isolated pixel variations that are not spatially correlated. Pixels that are in error often appear visually to be markedly different from their neighbors (Huang et al., 1979). So the noise is reduced by applying a median filter by a 3-by-3 neighborhood to the gray scale image. Median Filter have been used to reduce image noise here. Median filter is a nonlinear digital noise filtering technique, developed by Tukey (1971). In one-dimensional form, the median filter consists of a sliding window encompassing an odd number of pixels. The center pixel in the window is replaced by the median of the pixels in the window (Huang et al., 1979). The intensity values are corrected so that 1 % of the pixels are saturated at low and high intensities. By this method the contrast of the picture is enhanced. This is further achieved by histogram equalization of the image. Many times the luminance histogram of the image is skewed towards either brighter or darker side. Histogram equalization is a process for which the histogram of the enhanced image is forced to be uniform. The method is useful in images with backgrounds and foregrounds that are both bright or both dark (Andrews et al., 1972; Hall et al., 1972; Hall, 1974). Figure 5.1 shows the effect of enhancement for a Scots pine micrograph.

### 5.3.2. Segmentation

Image segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). The goal is to analyze a part of the image only over the whole image, to infer more accurately. We have chosen here to concentrate on the shapes and distribution of tracheids in the pictures, to segment the image in such a way to get only those parts of the micrographs. There are many methods of segmenting an image. In the present paper,

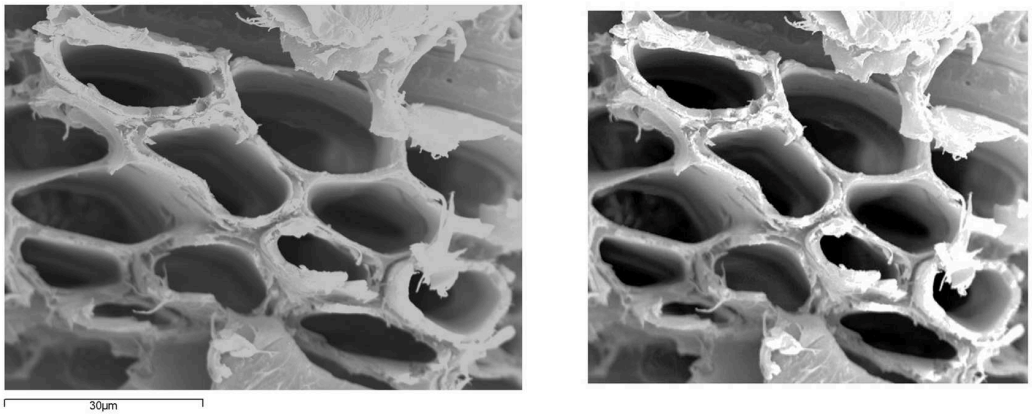


Figura 5.1: Effect of enhancement for a Scots pine micrograph.

Thresolding and Edge Detection segmentation methods have been used.

#### 5.3.2.1. Thresholding

The structure of the wood sample is analyzed by identifying the tracheids and their properties and patterns. Here, thresholding is used to get connected objects. As here the micrographs can be characterized as containing *holes* (tracheids) with reasonably uniform brightness against a background of different brightness, luminance can be used as a distinguished feature to segment the tracheids from the rest of the image. We have set an optimal value of the threshold depending on the brightness, and the tracheids are marked if their pixel values are less than that. The value of the threshold can be also automatically calculated (Shapiro and Stockman, 2001). The image is then inverted so that the important parts (tracheids) can be seen as white pixels. There can be a lot of noise coming in by the process of inversion. So, the small objects are removed from the image and then the closed boundaries are identified and filled with white pixels to get complete tracheids.

#### 5.3.2.2. Edge detection

Edge detection is an important way of image segmentation. Here, we identify the points in an image where the brightness changes abruptly. This is usually detected using difference of the pixels with their surrounding pixels. There are two major classes of differential edge detection: first order and second order derivative. For first order, some form of spatial first order differentiation is performed, and the resulting edge gradient is compared to a

threshold value. An edge is judged present if the gradient exceeds the threshold. For the second-order derivative class of differential edge detection, an edge is judged present if there is a significant spatial change in the polarity of the second derivative.

A sobel edge detector (a well-known first order edge detection method) has been used to find the edges of the tracheids. This operation is necessary because uneven darkness in the tracheids. Therefore, many times thresholding operation misses some parts of the them. Those parts need to be filled by detecting their boundaries using edge detection. The detected edges are dilated to get connected boundaries, then we fill the closed areas. The small objects are rejected to reduce noise. We use logical OR to sum two binary images. Now, this image can be used for extracting features from the tracheids. Figures 5.2 and 5.3 show the segmentation result for several micrographs of hardwood and softwood species, respectively (each row in these figures corresponds to each one of the considered species in the present paper).

### 5.3.3. Dilation

Dilation is a morphological operation. Morphological image processing is a type of processing in which the spatial form or structure of objects within an image is modified. With dilation, an object grows uniformly in spatial extent. We needed this because we are losing some of the boundary part of the tracheids in the time of thresolding as the boundary can have higher brightness than the inner part. So to restore the objects to the approximate original shape, dilation is applied.

### 5.3.4. Features

Now from this image, we identify and label the objects. The following features are extracted from the images:

1. Number of tracheids ($N$) detected in the image. This is a very important feature, as there are clearly different number of tracheids per image for different types of wood.

2. Average circularity of the tracheids. The circularity of the tracheids is measured as:
$$C = \frac{4\pi \times A}{P^2},$$
where $P$ denotes the object perimeter, and $A$ is the area of a tracheid, approximately measured as the number of pixels in the object. For a

perfectly round tracheid this should be 1, and less than 1 for any other kind of shape.

3. Average rectangularity of the tracheids.

   The rectangularity of the tracheids is measured as:

$$R = \frac{A}{A_r},$$

   where $A_r$ is the area of the surrounding rectangle. This index will be 1 for a perfect rectangle and less than 1 for other shape.

4. Average area per tracheid.

   This gives the measure of the size of the tracheids. It is a distinctive feature for the different kinds of wood micrographs.

5. Average distance between the tracheids.

   This is measured by the perimeter of the polygon formed by joining the centers of the tracheids in the image. This gives the information about the distribution and spacing between the tracheids in the micrograph. This is also a very distinctive feature for the images.

The objects are chosen on the basis of these features. As very small, or objects with very small circularity are unlikely to be tracheids, we reject them by setting some thresholds. Figure 5.4 presents an example of segmentation and some of the features extracted for a Scots pine micrograph.

### 5.3.5. Classification

A classification task usually involves training and testing data consisting of some observed instances. Each instance in the training set contains one target value or class label and several attributes or features. The goal of classification methods is to produce a model to predict class labels of data instances in the testing set for which only the attributes are known. In a first stage, each method compared in our study is validated through leave-one-out cross-validation. This is a technique widely used for the validation of an empirical model. It works by leaving out one instance (the testing sample); then a model is trained with the remaining samples and, finally, the developed model is used for the classification of the left out instance. This is repeated until all the micrographs have been left out once. As the data set available contain 101 samples, 100 samples are used for training and 1 sample

for testing. This process is repeated 101 times, and the percentages (measured as per one) of correct classification are calculated. Being more prone to lead to overfitting, i.e. minimum perturbation in the training set, leave-one-out cross-validation has a higher probability of resulting in mispredictions on external samples than 10-fold-cross-validation; however, the probabilities of wrong assignment inside the training samples are lower. This is an important fact to consider since compare directly the results obtained by both methods can be misleading. To complete the study, an external validation process is also carried out. There are many supervised classification methods performed to work with multivariate data, such as traditional Linear Discriminant Analysis (LDA), Quadratic classification and Logistic regression; besides advanced statistical methods belonging to Machine Learning as Support Vector Machines (SVM), $K$ Nearest Neighbors (KNN), Naïve Bayes, Classification Trees, Neural Networks or classification methods based on Adaboost algorithm. In this study, we have applied LDA, Quadratic, Logistic, Naïve Bayes, KNN, SVM and Neural Networks classifiers on the feature set, previously scaled. A comparative study is shown in the following section.

The most common classification model used in this study is the LDA, where a linear classifier is built among the different classes assuming normal densities and equal covariance matrices in the input data. For more information see Marcoulides and Hershberger (1997), Peña (2002), and Hernández Orallo et al. (2004).

If accepting the normality of the observations and the hypothesis of equal variances were not admissible, the procedure for solving the problem is to classify the observation in the group with maximum posterior probability. This is the case of Quadratic classification that use a second-order mapping of the input (Marcoulides and Hershberger, 1997; Peña, 2002).

The Logistic model is applied to a wide range of situations where the explicative variables do not have a multivariate normal distribution (McLachlan, 2004; Peña, 2002; Hernández Orallo et al., 2004). Considering only two classes ($C_1$ and $C_2$), the Logistic regression equation is the following:

$$\log \frac{p}{1-p} = \alpha + \beta' x,$$

or

$$p = \frac{\exp(\alpha + \beta' x)}{1 + \exp(\alpha + \beta' x)},$$

where $p = P(Y = C_1|x)$ is the posterior probability of $Y$ equal to $C_1$, $\log(\frac{p}{1-p})$ is the logit transformation, $x$ is the p-dimensional vector of predictor variables, $\beta$ is a vector of $p$ parameters and $\frac{p}{1-p}$ the odds ratio. The logit model can be generalized to more than two populations, i.e. for qualitative

response with more than two possible levels. If we suppose $G$ populations, then, defining $p$ as the probability that the observation $i$ belongs to the class $g$, it is possible to write:

$$p_{ig} = \frac{\exp(\beta_{0g} + \beta'_{1g}x_i)}{1 + \sum_{j=1}^{G-1} \exp(-\beta_{0j} - \beta'_{1j}x_i)},\ j = 1, \ldots, G-1$$

Therefore, we can say that the posterior probabilities, $p_{ig}$, satisfy a multivariate logistic distribution. The comparison between two categories is made in the usual way

$$\frac{p_{ig}}{p_{ij}} = \frac{\exp(\beta_{0g} + \beta'_{1g}x_i)}{\exp(\beta_{0j} + \beta'_{1j}x_i)}.$$

The Naïve Bayes classifier technique is based on the Bayes'theorem and is particularly suited when the dimensionality of the inputs is high. Given a set of variables, $X = \{x_1, x_2, ..., x_d\}$, we want to construct the posterior probability for the event $C_j$ among a set of possible outcomes $C = \{c_1, c_2, ..., c_d\}$. Using the Bayes rule (Hernández Orallo, 2004; Hill and Lewicki, 2007):

$$p(C_j|x_1, x_2, ..., x_d) \propto p(x_1, x_2, ..., x_d|C_j)p(C_j),$$

where $p(C_j|x_1, x_2, ..., x_d)$ is the posterior probability of class membership, i.e., the probability that $X$ belongs to $C_j$. Since Naïve Bayes assumes that the conditional probabilities of the independent variables are statistically independent, we can decompose the likelihood as a product of terms:

$$p(X|C_j) \propto \prod_{k=1}^{d} p(x_k|C_j)$$

and rewrite the posterior probabilities as:

$$p(C_j|X) \propto p(C_j) \prod_{k=1}^{d} p(x_k|C_j)$$

Using the Bayes'rule, we label a new case $X$ with a class level $C_j$ that achieves the highest posterior probability. Although the assumption that the predictor variables are independent is not always accurate, Naïve Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation. Furthermore, the assumption does not seem to greatly affect the posterior probabilities. Naïve Bayes can be modeled in several different ways, including normal, log-normal, gamma and Poisson density

functions. The Naïve Bayes classifier used in the present study assumes independence of the predictor variables, and a Gaussian distribution (given the target class) of metric predictors, i.e.

$$p(x_k|C_j) = \frac{1}{\sigma_{kj}\sqrt{2\pi}} \exp\left(\frac{-(x-\mu_{kj})^2}{2\sigma_{kj}}\right),$$

where $\mu_{kj}$ and $\sigma_{kj}$ are the corresponding means and standard deviations, respectively.

$K$ Nearest Neighbors is a simple nonparametric classification procedure that has been successful used with non-normal populations. It performs as follows (Hill and Lewicki, 2007; Peña, 2002):

1. Defining a measure of distance between points, normally Mahalanobis distance.

2. Calculating the distances from the test sample $x_0$ to the others points.

3. Selecting the $k$ nearest sample points to the one we intend to classify. Calculating the proportion of these $k$ points belonging to each of the populations. Classifying the point $x_0$ in the population corresponding to a higher points frequency from the $k$ points. In this study, the $k$ value has been selected minimizing the cross-validation misclassification error.

Support Vector Machine (SVM) is a classifier method developed by Vapnik and coworkers performing classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. They found that if a wide margin between the regions of distribution of sample points of different kinds exists, the mathematical model obtained as an optimal hyperplane will exhibit very good prediction ability, even if the dimension of the feature space was very high and the equation of this optimal hyperplane had to be expressed by many adjustable parameters (Vapnik, 1998; Hill and Lewicki, 2007; Chen et al., 2004). In Figure 5.5, the optimal hyperplane denotes the unique hyperplane having largest distances with the sample points of different classes. The sample points located on the border of the margin are called support vectors. It can be seen that the position of the optimal hyperplane is only decided by the support vectors.

The principle of SVM is much different from the others commonly used methods. The most important task in SVM is not dimension reduction, but dimension elevation using a kernel function to map the sample points of the input space into a feature space with higher dimensionality by nonlinear

transformation. In these high dimensional feature spaces, nonlinear separable sample points in the input space can become linearly separable with wide margin in the feature space, and a linear separation algorithm can be used to make a mathematical model with good prediction ability (Peña, 2002; Hernández Orallo, 2004; Vapnik, 1998; Hill and Lewicki, 2007; Chen et al., 2004).

For constructing an optimal hyperplane, SVM method uses an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, the SVM classification models can belong to two different groups: C-SVM and nu-SVM classification. In this study, the first one is used. So, the error function to be minimized in training is the following:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i,$$

subject to

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, \ldots, N,$$

where $C$ is the capacity constant, $w$ is the vector of coefficients, $b$ a constant and $\xi_i$ are parameters for handling non-separable data (inputs). The index $i$ labels the $N$ training cases. Note that $y_i$ are the class labels and $x_i$ the independent variables. The kernel $\phi$ is used to transform data from the input to the feature space. It should be noted that the larger the $C$, the more the error is penalized. Thus, $C$ should be chosen with care to avoid over-fitting. Different kernel types are used in the present study.

Additionally, a Neural Network method has been implemented. In particular, a single hidden layer perceptrons (feed-forward Neural Networks) classification method is used. For further information see Peña (2002), Hernández Orallo et al. (2004), Venables et al. (2002), Ripley (1994), and Hill and Lewicki (2007).

Before discussing the results, it is necessary to define what parameters were defined and the value thereof when we use the KNN, SVM and Neural Networks methods. Using KNN method, we obtained the optimal correct classification probability for 7 classes with $k = 2$ neighbors. In the case of SVM, 4 different kernels (linear, Gaussian, polynomial and sigmoid) were tested using the C-SVM algorithm. The optimal result was obtained using a polynomial kernel $((gamma \times u'v + coef)^{degree})$ with parameters $C = 36$ (testing from 1 to 100), $gamma = 0{,}03$ (testing from 0,01 to 1), $degree = 3$ (testing from 3 to 5) and $coef = 14$ (testing from 1 to 30). Finally, regarding Neural Networks, a single-hidden-layer neural network with 2 units, skip-layer con-

nections, initial random weights on $[-0{,}4, 0{,}4]$ and parameter of weight decay equal to 0,001. These optimal parameters were obtained using leave-one-out cross-validation. For classifying between 2 classes, the parameters obtained are the following: KNN ($k = 2$), SVM ($C = 12$, $gamma = 0{,}01$, $degree = 5$, $coef = 4$), Neural Networks (hidden-layer size=1, weight $decay = 0{,}01$ and initial random weights on $[-0{,}66, 0{,}66]$).

In the present study, the following R (free software) libraries are used: `nnet` (for Logistic regression and Neural Networks), `e1071` (for Naïve Bayes and SVM) and `MASS` (for LDA and Quadratic classification) R Development Core Team (2008). For applying the KNN classifier, the Matlab platform is used MATLAB (2010).

Figure 5.6 shows the flowchart of the classification process.

## 5.4. Results and discussion

In this section, the methods previously presented are applied to the features matrix to classify between different species and main groups (hardwoods and softwoods). First, the procedure for obtaining and scaling the data and, additionally, a brief descriptive analysis is shown.

### 5.4.1. Obtaining, scaling and descriptive analysis of features

The image analysis have been done on Matlab platform, using the image processing toolbox. We have used 7 different classes of wood as our test sample. Figures 5.2 and 5.3 shows some micrographs obtained using a scanning electron microscope and the resulting image after applying image enhancement, segmentation and dilation processes, as described in Section 5.3 of this study.

The features extraction, described in Section 5.3.4, is performed to the processed images. We have used a total of 101 micrographs, with variable number of samples for each group. So our feature matrix is of size $101 \times 5$. In Table 5.1, the number of samples per class is shown.

Data preparation is a very important step in Machine Learning. In fact, the scaling of data is a common practice in multivariate analysis, especially when techniques aimed at the prediction are used (Chen et al., 2004). In this study, each column of matrix features is centered (the mean of each column is subtracted from it) and then divided by its standard deviation. Using this standardization, the features variabilities are matched (Peña, 2002).

| Type of Wood | No of samples |
|---|---|
| Eucalyptus | 15 |
| European beech | 11 |
| Scots pine | 16 |
| European chestnut | 15 |
| Jatobá | 18 |
| Isignis pine | 15 |
| Walnut | 11 |

Tabla 5.1: Number of samples per wood class.

After obtaining the standardized features, we have used the metric multidimensional scaling (MDS) technique to obtain preliminary information about the data (Peña, 2002; Hill and Lewicki, 2007). The MDS method can be applied to many multivariate data, provided that the calculation of the distances or similarities be doable. After calculating the distance matrix corresponding to the data, the aim of this first descriptive analysis is to obtain information about the structure of the data, i.e., try to find out what elements present similar properties, if there are distinct groups, outliers, etc. The information provided by the distance matrix can be approximated by using two principal coordinates (corresponding to the two largest eigenvalues in the similarities matrix) (Peña, 2002). Figure 5.7 shows all the obtained samples on the basis of these two principal coordinates.

Using two dimensions has been appropriate because we achieve to represent a high percentage of data variability. In this case, the explained variability proportion using the two first eigenvalues is 84 % (Peña, 2002). It can be observed that the samples belonging to the same wood specie form groups. Some groups of species appear particularly distinct with respect to the others, such as insignis pine, jatobá, European chestnut, and especially eucalyptus. Eucalyptus samples displayed no overlap with samples of other kind of wood. The differences between eucalyptus and the other species are much more pronounced than any other. However, there are different groups or types showing some overlap. These are, above all, European beech, Scots pine and walnut. In particular, the overlapping between beech and Scots pine samples may hinder the classification of new samples belonging to both timber species. The MDS analysis shows that the chosen features characterize, in greater or lesser extent, the wood species.

### 5.4.2. Supervised classification results

We apply the different classification methods defined in Section 5.3.5. Traditional methods estimating the population parameters by statistical inference are used: LDA, Quadratic classification and Logistic regression. Moreover, other methods related to Machine Learning are applied: SVM, Naïve Bayes, KNN and Neural Networks. These last classifiers focused on the accuracy of prediction rather than on the interpretation of the models generating them. The correct classification probabilities obtained by the methods previously mentioned are presented in Table 5.2. The leave-one-out cross-validation process (described in Section 5.3.5) has been used to obtain these results and to compare the classifiers. In the next subsection, an external validation test completes the comparison between the different approaches. The probabilities are computed in two settings, classifying among the 7 different species and in the more general case of classifying into 2 different groups: hardwoods and softwoods.

Usually, we get better results when using Machine Learning methods, i.e., SVM classifiers, KNN and Neural Networks. High probabilities of good classification between 0,87 (SVM) and 0,89 (Neural Networks) are obtained using these methods for classifying between softwoods and hardwoods (Table 5.2). We also obtain good results when we want to classify among 7 different classes: the correct classification probabilities are between 0,78 (KNN) and 0,81 (SVM). Note that the results obtained when classifying between 2 classes are slightly better than those obtained in the case of 7 species. As pointed out previously, classifiers related to Machine Learning usually seem to work better than traditional methods, especially when classifying in 2 classes. For example, the LDA method is the worst method when 7 classes are used (0,72), but the difference with respect to the three Machine Learning techniques is even larger in the case of 2 classes (0,75 for LDA). Although the probabilities obtained by LDA are not low in absolute terms, better results are obtained by other nonlinear methods such as Logistic (0,80 with 7 classes) or Quadratic classification.

In general, the results in Table 5.2 show a relatively high probability of correct classification, especially if we take into account the heterogeneity of the wood and the results obtained in other studies (Brandtberg, 2002; Tou et al., 2009). Nevertheless, since the probability of misclassification is larger than zero, an interesting question is: what are the species tending to get confused using these features and methods? The answer is in the confusion matrices shown in Table 5.3 and 5.4.

Table 5.3 shows the confusion matrices corresponding to the SVM, Neural Networks and KNN methods applied to classify between hardwoods and

| | 7 Groups Classification | 2 Groups Classification |
|---|---|---|
| Methods | Prediction | Prediction |
| LDA | 0,72 | 0,75 |
| Quadratic classification | 0,77 | 0,78 |
| Logistic regression | 0,80 | 0,75 |
| Naïve Bayes | 0,79 | 0,75 |
| KNN | 0,78 | 0,88 |
| SVM | 0,81 | 0,87 |
| Neural Networks | 0,80 | 0,89 |

Tabla 5.2: Prediction probabilities obtained by each classification method and leave-one-out cross-validation, using features based in segmentation process. The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

softwoods. It can be observed that the hardwood samples generally are better predicted than softwoods (see SVM and Neural Networks methods). There is an exception, KNN classifier predicts correctly all the softwoods (the posterior probability is equal to 1). These little confusions may be due to the beech and Scots pine overlapping that can be observed below in Table 5.4.

| | | Estimated | |
|---|---|---|---|
| Methods | Actual | Hardwoods | Softwoods |
| SVM | Hardwoods | 0,90 | 0,10 |
| | Softwoods | 0,23 | 0,77 |
| Neural Networks | Hardwoods | 0,93 | 0,07 |
| | Softwoods | 0,20 | 0,80 |
| KNN | Hardwoods | 0,83 | 0,17 |
| | Softwoods | 0,00 | 1,00 |

Tabla 5.3: Probabilities of correct classification, using leave-one-out cross-validation, in 2 different classes (hardwoods and softwoods) obtained by SVM, Neural Networks and KNN. The probabilities are rounded using two significant figures.

Table 5.4 shows the confusion matrices corresponding to the SVM, Logistic regression, KNN and Neural Networks methods applied to classify between 7 different types of wood. The first two methods presented in Table 5.4 are representative of all others (except KNN), in addition, they produce the best

results. As the descriptive analysis had pointed, the posterior probability of eucalyptus is 1, i.e., all eucalyptus samples have been predicted correctly (see leave-one-out cross-validation in Section 5.3.5). Moreover, there is no confusion with any other wood species. Other distinct groups that have been observed using the MDS technique present high posterior probabilities (correct classification). This is the case of insignis pine (0,93), jatobá (0,89) and chestnut (0,80), using SVM or Logistic regression. The use of Neural Networks improves the insignis pine prediction (1,00), but the result with chestnut is worse (0,60). Moreover, the SVM method has been able to correctly classify walnut (0,82). This is a very good result, because according to the descriptive analysis previously shown, this specie presented an overlapping with other wood species. In general, beech and Scots pine are the most difficult species to predict. The beech samples can be confused with the Scots pine ones and Scots pine samples can be predicted as walnut or beech (see Table 5.3). The best results in this case corresponds to the use of the KNN method: this technique is able to correctly predict all the samples of beech wood, but at the cost of worse walnut predictions. Therefore, Table 5.4 shows three methods giving complementary information.

The results in Tables 5.2, 5.3 and 5.4 indicate that, overall, classifying between hardwoods and softwoods and, on other hand, between 7 different species, has been possible. Therefore, the existence of differences between wood species from the earlywood tracheids structure (defined using the 5 proposed features obtained at 1500× magnification) has been proven. On the other hand, the SVM method, particularly suitable for small samples (Chen et al., 2004) as in this case, as well as Neural Networks and KNN have shown the best prediction behavior.

#### 5.4.2.1. External validation

The validation scheme used in the previous section could be not enough to guarantee a proper generalizability of the outcomes and to get an accurate idea of the model performances on an actually external set. This is because Machine Learning classification methods, such as KNN, Neural Networks or SVM require to fix the values of some adjustable parameters (learning coefficients and number of hidden neurons, value of $k$ in KNN, learning parameters in SVM) which are chosen based on the minimum error in leave-one-out cross-validation, and the same cross-validation approach has been used to get an estimate of the predictive ability of the models. Therefore, the cross-validation samples may be not entirely external for the above mentioned models themselves.

Given these arguments, a new test set, completely external, is included

to properly evaluate the predictive ability of the models. Considering the available number of samples, the procedure is the following: in the hardwood and softwood (2 classes) classification case, an external test set (consisting of 30 samples) based on the Kennard-Stone intelligent criterion (Kennard and Stone, 1969) is selected. The adjustable model parameters are obtained using the training set and the 10-fold cross-validation error minimization procedure. On the other hand, for the 7 class problem, taking into account the reduced number of samples per class, a first set of samples (one per class) is randomly extracted as external validation set. The remaining samples are used as training set, and models choosing the parameters according to a 10-fold cross-validation procedure are built. This global procedure is repeated 100 times to ensure that, with high probability, each sample is included in the test set at least once.

Table 5.5 shows the results obtained in this framework. Compared with the previous leave-one-out cross-validation results (Table 5.2), and when classifying between 2 classes, a better performance for all the methods (except SVM) is observed. Considerable high probabilities of good classification (0,93) are obtained by KNN and Neural Networks. It is important to note that the third best method using external validation is now the Quadratic classification procedure, instead of SVM.

When the samples are classified according to the 7 different wood species, the results obtained by LDA, Logistic regression and Naïve Bayes methods using external validation (Table 5.5) are very similar to those obtained when using the leave-one-out cross-validation criterion (Table 5.2). Note that the result produced by the Quadratic classification approach improved slightly in this setting, giving a probability of correct classification equal to 0,80. Other Machine Learning methods, such as Neural Networks and KNN produced slightly worse results and, to a lesser extent, SVM (0,76). In short, the best correct classification percentages obtained by leave-one-out cross-validation are similar to those obtained using the external validation test.

#### 5.4.2.2. Additional studies

The study presented in the previous sections can be completed in different directions. In this section, we focus on two possible extensions of our research; on one hand, using standard GLCM features and, on the other hand, applying fractal analysis techniques. These two approaches are related to the acquisition of new attributes or variables from which to classify.

Firstly, we apply the same statistical methods to the standard GLCM features of the micrographs. The corresponding results are shown in Table 5.6, using leave-one-out cross-validation (similar results are obtained with

the external validation test described in Section 5.4.2.1). They are clearly worse than those obtained with the segmentation study (Table 5.2), proving the advantage of using the approach presented in this paper.

On the other hand, a possible extension of the present work could be using fractal analysis procedures in our research. The concept of fractal was first introduced by Mandelbrot (1983). Fractal refers to entities, especially sets of pixels, which display a degree of self-similarity at different scales. It has been successfully used in Gonzales-Barron and Butler (2008a) to study the capability to accurately describe the surface roughness of bread crumb. In the same context, in Gonzales-Barron and Butler (2008b), the relationship between different features and panelists'perception of bread crumb is analyzed. In both papers, the study is carried out computing the fractal dimension (FD) of digital images of bread crumb using different methods: fractional Brownian motion (FBM) method, frequency domain method, box-counting (BC) method, morphological fractal method, mass fractal method and random walks method. Fractal analysis have been also applied in medical image applications. For example, in Chen et al. (2005), the FD computed using the FBM method is used to classify breast ultrasond images. Recently, in Fenghu et al. (2010), some applications of fractal theory in wood science are proposed.

Following the lines in the previous papers, we obtained the FD of our 101 images, using the FBM and BC approaches. A detailed description of these two methods can be found in Gonzales-Barron and Butler (2008a), Chen et al. (2005), Russ (2002), and Gonzales-Barron and Butler (2008a) and Buczkowski et al. (1998), respectively. The means and standard deviations of the FD's for each one of the 7 wood species are given in Table 5.7. As in Gonzales-Barron and Butler (2008a), we compute the correlation coefficients between the 5 original texture features and the fractal dimensions determined with the FBM and BC methods. Unlike in Gonzales-Barron and Butler (2008a), no strong correlations are observed in our case. Therefore, the FD vectors are directly used as two extra features, jointly with the 5 original ones, in the classification procedures. Table 5.8 shows the percentages of correct classification obtained when the FD's are included as new features, using the external validation framework described in Section 5.4.2.1. The results presented in Table 5.8 correspond to three scenarios, depending on the features considered in the classification approaches: the 5 original features jointly with the FD computed by the FBM method (denoted by FBM in the table), the 5 original features jointly with the FD computed by the BC method (denoted by BC in the table), and the 5 original features jointly with the FD computed by the FBM and BC methods (denoted by FBM+BC in

the table). As in previous tables, the results are divided in 7 and 2 classes.

The results indicate that, in general, including the FD computed with the FBM method improves the classification process when the aim is classifying in 7 classes. Adding this fractal feature (FBM), the best classification probability is 0,83, corresponding to the Naïve Bayes method. Moreover, the results obtained by Logistic regression (0,81), SVM (0,81), KNN (0,77) and Neural Networks (0,78) are clearly better than those obtained without including in the feature set the FD by the FBM method (see Table 5.5). This improvement is mainly due to the existence of a minor confusion in predicting classes corresponding to beech and Scots pine samples. Nevertheless, when we want to classify between hardwoods and sofwoods, the percentages of correct classification, having added the feature related to the fractal dimension by FBM (and selecting again a new test sample by Kennard Stone method) does not improve the previous results (see Table 5.5). The only exception is the correct classification probability improvement obtained by SVM (0,90). On the other hand, the percentages of correct classification for the statistical classification techniques, adding the fractal dimension computed by the BC method to the previous 5 extracted features (columns BC in Table 5.8) show that, generally, including only the BC dimension feature does not improve the classification process. Finally, the last two columns in Table 5.8 present the percentages of correct classification for each statistical classification technique, adding the two proposed fractal dimensions to the original 5 feature data set. It is observed that, in this setting, the best classification probability (among 7 classes) is 0,83, corresponding to the Naïve Bayes method. Moreover, a maximum correct classification probability equal to 0,93 by KNN and Neural Networks methods is obtained when the aim is classifying between hardwoods and softwoods. In general, compared with the case of just using the 5 original texture features, if the two additional fractal dimensions are used jointly with them, the resulting probabilities of correct classification for the 7 class case are improved, while similar probabilities of correct classification are obtained for classifying into 2 groups.

## 5.5. Conclusions

In the present paper, micrographs obtained by SEM with 1500× magnification, taken in cross sections, have been processed in a simple way using segmentation and object recognition to identify tracheids corresponding to the earlywood of 7 different timbers. Segmentation process has allowed to extract 5 features related to the shape, number, area and distribution of this type of cells.

Classifying wood species using these 5 features extracted from SEM micrographs with 1500× magnification has been possible. High correct classification probabilities have been obtained when we want to discern between hardwoods and softwoods (0,89 using leave-one-out cross-validation, and 0,93 using an external validation test) and among 7 different wood species (0,81 using one-leave-out cross-validation, and 0,80 using an external validation test), taking into account the heterogeneity of wood. The results show that it is possible to observe differences among species in the wood texture at this magnification (1500×).

Observing the shape, number, area and distribution of the earlywood tracheids, eucalyptus is very different to the others species. On the other hand, beech can be confused with Scots pine using these 5 features.

Fractal analysis has been successfully used in this work. We obtained the FD of our 101 images using the FBM and BC approaches. Adding to the original 5 features these fractal dimensions, the best classification probability (among 7 classes) is 0,83, corresponding to the Naïve Bayes method. Moreover, a maximum correct classification probability equal to 0,93 by KNN and Neural Networks methods, when we want to classify between hardwoods and softwoods, is obtained. If the two fractal dimensions are added to the 5 texture features considered in this research, a clear improvement in the resulting probabilities of correct classification for the 7 class case is obtained, while similar results are obtained if the aim is classifying into 2 groups. A more complete study including fractal analysis would be of great interest in this context, for example, determining the FD using some other methods and using these vectors as new features. This will be analyzed in a future research.

Classifiers related to Machine Learning seem to work better than traditional methods, especially when we want to classify between 2 classes. The drawback when SVM or Neural Networks classifiers are used is the relatively high computing time.

A comparative study using gray level co-occurrence (GLCM) based features is also presented, showing the effectiveness of the image segmentation method, at least with this data set. In fact, the correct classification probabilities obtained using segmentation are much higher.

## 5.6. Acknowledgments

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

thank two anonymous referees for constructive comments that improved the presentation of this article.

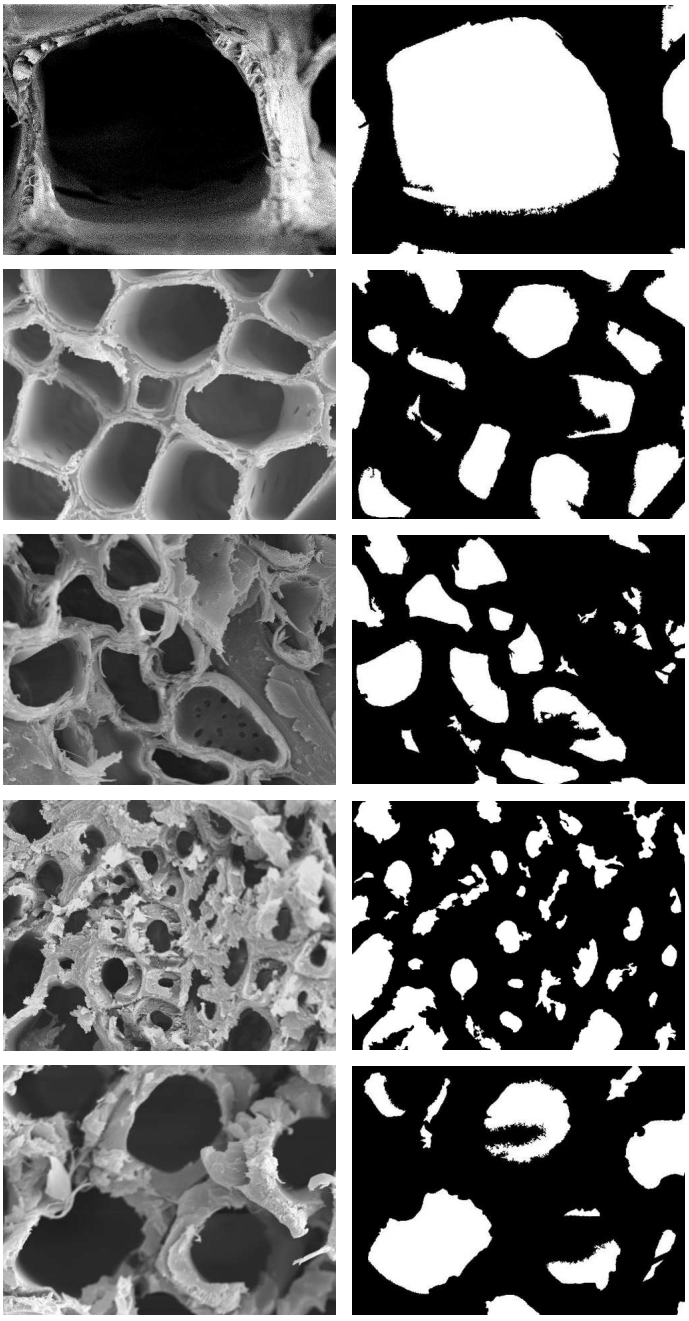

Figura 5.2: Segmentation process for hardwood species. Row 1: eucalyptus. Row 2: beech. Row 3: chestnut. Row 4: jatobá. Row 5: walnut.

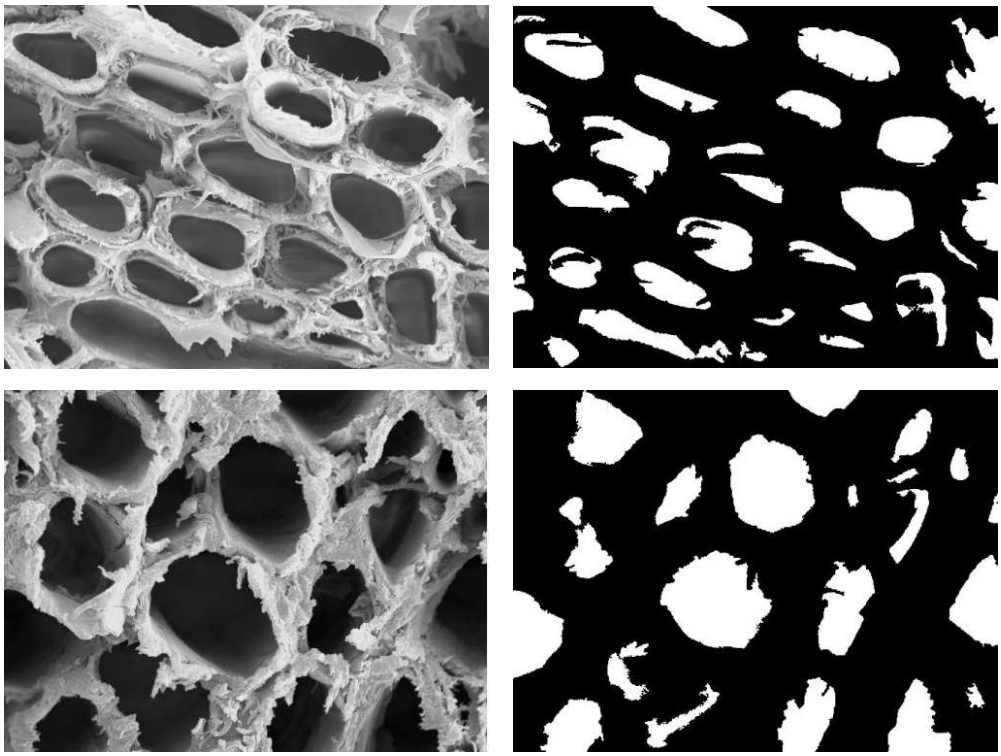

Figura 5.3: Segmentation process for softwood species. Row 1: Scots pine. Row 2: insignis pine.

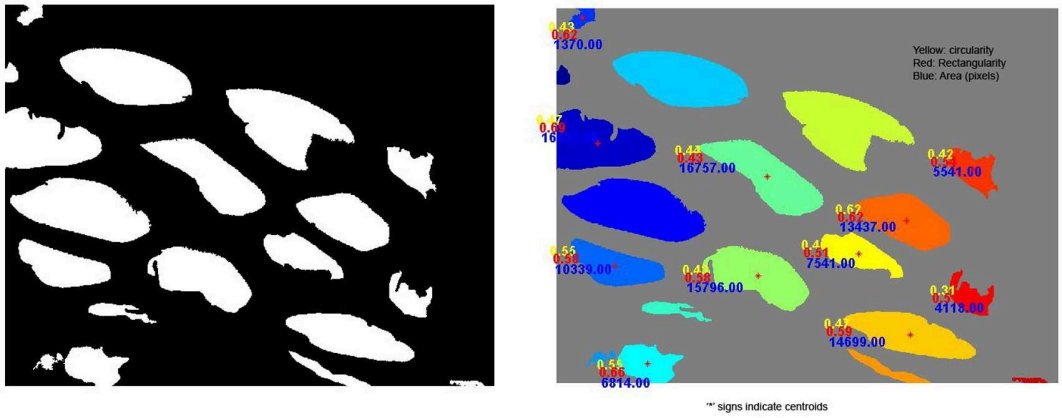


Figura 5.4: Segmentation and some extracted features for a Scots pine micrograph.

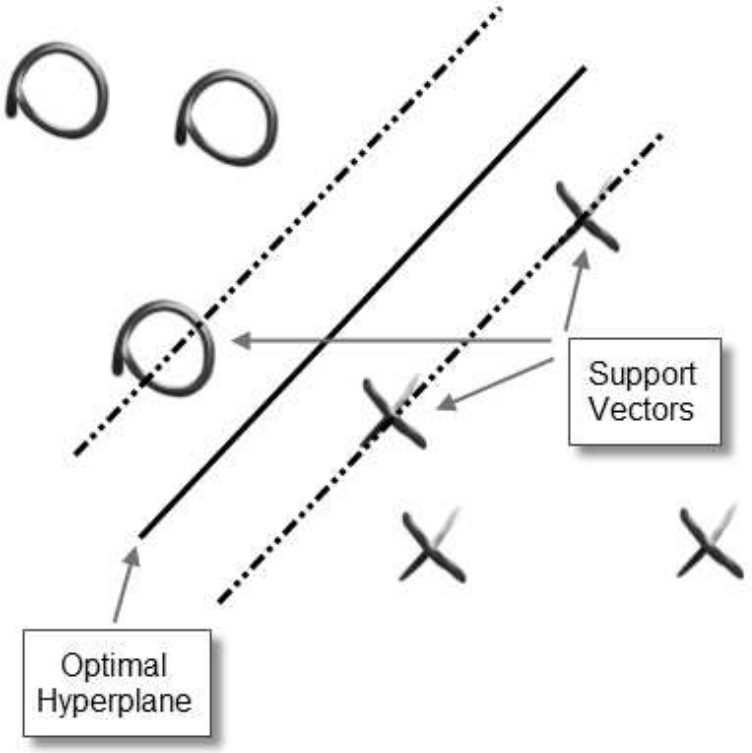


Figura 5.5: Optimal hyperplane and large margin in SVM.

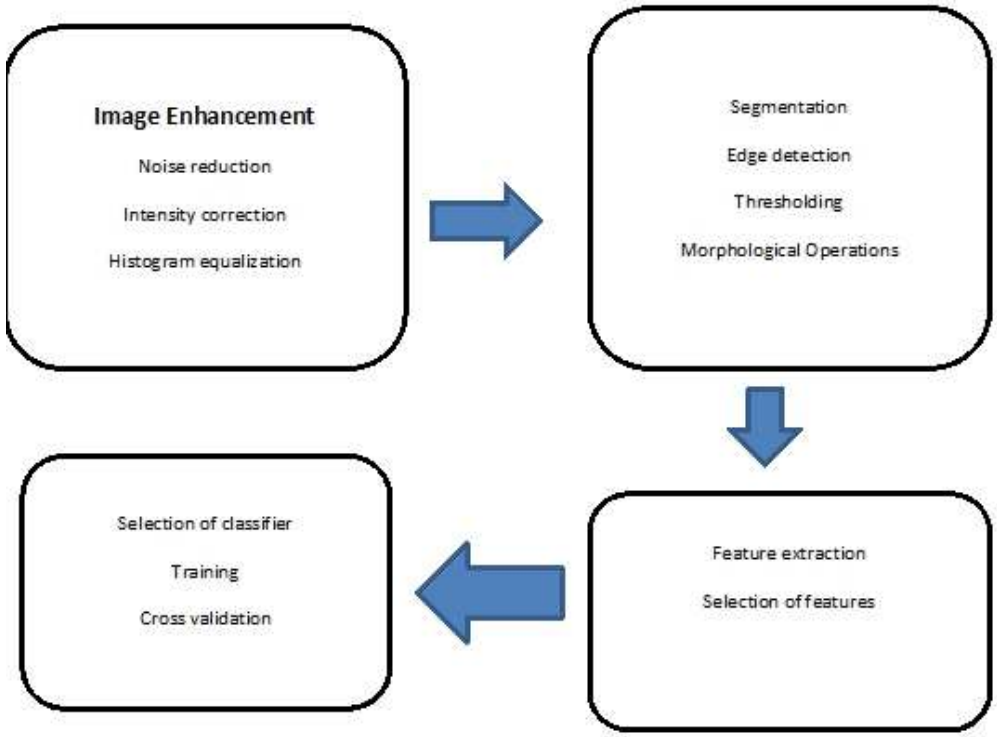


Figura 5.6: Flowchart of the classification process.

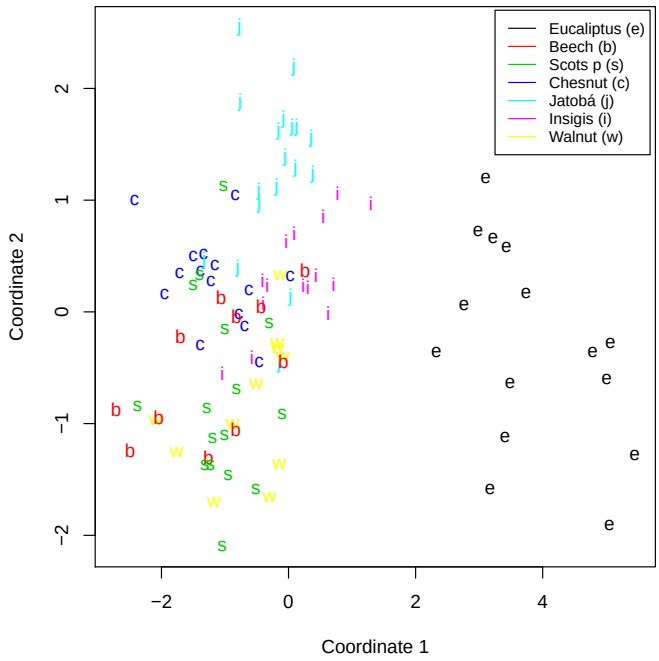


Figura 5.7: Samples from different species of wood, using MDS with two principal coordinates.

| | | Estimated | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Methods | Actual | Eucal. | Beech | Scots P. | Ches. | Jat. | Ins. P. | Walnut |
| SVM | Eucal. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 0,64 | 0,18 | 0,09 | 0,00 | 0,00 | 0,09 |
| | Scots P. | 0,00 | 0,06 | 0,63 | 0,13 | 0,00 | 0,06 | 0,13 |
| | Ches. | 0,00 | 0,07 | 0,07 | 0,80 | 0,00 | 0,00 | 0,07 |
| | Jat. | 0,00 | 0,11 | 0,00 | 0,00 | 0,89 | 0,00 | 0,00 |
| | Ins. P. | 0,00 | 0,00 | 0,07 | 0,00 | 0,00 | 0,93 | 0,00 |
| | Walnut | 0,00 | 0,00 | 0,09 | 0,18 | 0,00 | 0,00 | 0,73 |
| Logistic R. | Eucal. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 0,45 | 0,45 | 0,00 | 0,09 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,06 | 0,63 | 0,13 | 0,00 | 0,06 | 0,13 |
| | Ches. | 0,00 | 0,07 | 0,07 | 0,80 | 0,00 | 0,00 | 0,07 |
| | Jat. | 0,00 | 0,06 | 0,06 | 0,00 | 0,89 | 0,00 | 0,00 |
| | Ins. P. | 0,07 | 0,00 | 0,00 | 0,00 | 0,00 | 0,93 | 0,00 |
| | Walnut | 0,00 | 0,00 | 0,18 | 0,00 | 0,00 | 0,00 | 0,82 |
| KNN | Eucal. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Beech | 0,00 | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,38 | 0,63 | 0,13 | 0,00 | 0,06 | 0,13 |
| | Ches. | 0,00 | 0,13 | 0,00 | 0,87 | 0,00 | 0,00 | 0,00 |
| | Jat. | 0,00 | 0,06 | 0,06 | 0,11 | 0,78 | 0,00 | 0,00 |
| | Ins. P. | 0,00 | 0,13 | 0,07 | 0,00 | 0,00 | 0,80 | 0,00 |
| | Walnut | 0,00 | 0,09 | 0,36 | 0,09 | 0,09 | 0,00 | 0,36 |
| Neural | Eucal. | 1,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 |
| Networks | Beech | 0,00 | 0,64 | 0,18 | 0,18 | 0,00 | 0,00 | 0,00 |
| | Scots P. | 0,00 | 0,00 | 0,69 | 0,13 | 0,00 | 0,06 | 0,13 |
| | Ches. | 0,00 | 0,13 | 0,07 | 0,60 | 0,07 | 0,00 | 0,13 |
| | Jat. | 0,00 | 0,00 | 0,06 | 0,06 | 0,89 | 0,00 | 0,00 |
| | Ins. P. | 0,00 | 0,00 | 0,00 | 0,00 | 0,00 | 1,00 | 0,00 |
| | Walnut | 0,00 | 0,09 | 0,09 | 0,09 | 0,00 | 0,00 | 0,73 |

Tabla 5.4: Probabilities of correct classification, using leave-one-out cross-validation, in 7 different classes (chestnut, walnut, insignis pine, Scots pine, beech, eucalyptus and jatobá) obtained by SVM, Logistic regresion, KNN and Neural Networks. The probabilities are rounded using two significant figures.

| | 7 Groups Classification | 2 Groups Classification |
|---|---|---|
| Methods | Prediction | Prediction |
| LDA | 0,73 | 0,80 |
| Quadratic classification | 0,80 | 0,90 |
| Logistic regression | 0,77 | 0,83 |
| Naïve Bayes | 0,79 | 0,87 |
| KNN | 0,73 | 0,93 |
| SVM | 0,76 | 0,87 |
| Neural Networks | 0,73 | 0,93 |

Tabla 5.5: Prediction probabilities by each classification method and an external validation test, using features based in segmentation process. The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

| | 7 Groups Classification | 2 Groups Classification |
|---|---|---|
| Methods | Prediction | Prediction |
| LDA | 0,41 | 0,67 |
| Quadratic classification | 0,37 | 0,59 |
| Logistic regression | 0,37 | 0,69 |
| Naïve Bayes | 0,36 | 0,53 |
| KNN | 0,39 | 0,78 |
| SVM | 0,37 | 0,69 |
| Neural Networks | 0,47 | 0,73 |

Tabla 5.6: Prediction probabilities obtained by each classification method and leave-one-out cross-validation, using GLCM as the features. The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

| | FBM | | BC | |
|---|---|---|---|---|
| Species | Mean | Standard Deviation | Mean | Standard Deviation |
| Eucalyptus | 2,986 | 0,0019 | 1,769 | 0,0171 |
| Beech | 2,988 | 0,0007 | 1,840 | 0,0164 |
| Scots P. | 2,986 | 0,0010 | 1,841 | 0,0110 |
| Chesnut | 2,985 | 0,0018 | 1,839 | 0,0125 |
| Jatobá | 2,985 | 0,0030 | 1,831 | 0,0233 |
| Insignis P. | 2,986 | 0,0014 | 1,841 | 0,0106 |
| Walnut | 2,983 | 0,0019 | 1,803 | 0,0354 |

Tabla 5.7: Means and standard deviations of FD computed with the FBM and BC methods for the 7 wood species.

| | FBM | | BC | | FBM+BC | |
|---|---|---|---|---|---|---|
| Methods | 7 Clas. | 2 Clas. | 7 Clas. | 2 Clas. | 7 Clas. | 2 Clas. |
| LDA | 0,77 | 0,77 | 0,75 | 0,80 | 0,75 | 0,90 |
| Quadratic classification | 0,75 | 0,87 | 0,76 | 0,87 | 0,74 | 0,87 |
| Logistic regresion | 0,81 | 0,77 | 0,71 | 0,87 | 0,76 | 0,90 |
| Bayes Naïve | 0,83 | 0,80 | 0,78 | 0,87 | 0,83 | 0,87 |
| KNN | 0,77 | 0,87 | 0,75 | 0,80 | 0,75 | 0,93 |
| SVM | 0,81 | 0,90 | 0,72 | 0,87 | 0,79 | 0,90 |
| Neural Networks | 0,78 | 0,80 | 0,69 | 0,90 | 0,80 | 0,93 |

Tabla 5.8: Prediction probabilities obtained by each classification method an external validation test, using the 5 original features jointly with the FD computed by the FBM method (FBM), the 5 original features jointly with the FD computed by the BC method (BC), and the 5 original features jointly with the FD computed by the FBM and BC methods (FBM+BC). The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

# Capítulo 6

# Conclusiones generales y líneas futuras de investigación

El propósito de este capítulo es discutir y resumir, conjunta y brevemente, las conclusiones obtenidas en cada uno de los capítulos presentados. Finalmente, se enumeran una serie de líneas futuras de investigación, relacionadas con los estudios que se han expuesto en esta tesis doctoral.

## 6.1. Conclusiones generales

En primer lugar, de forma general, se puede afirmar que el uso de técnicas estadísticas relacionadas con el Análisis de Datos Funcionales (FDA), el análisis multivariante, el remuestreo, la regresión y el diseño de experimentos, aplicadas a datos térmicos como las curvas termogravimétricas (TG) y calorimétricas (DTG), aporta una información muy valiosa para el estudio de materiales, en particular para tareas de clasificación supervisada y detección de diferencias en casos difíciles como los mostrados en este trabajo. Los materiales estudiados han sido micro-nanocompuestos de matriz epoxídica y madera de uso industrial.

Con respecto a la Parte I de la presente tesis, se dio pertinente respuesta a los objetivos principales marcados en 1.4:

Se evaluó el efecto de la adición de humo de sílice en la estabilidad térmica de una resina epoxi mediante un test ANOVA de carácter funcional. Cuando se incrementa la cantidad en humo de sílice añadida, de 0 a 10 wt % o de 0 a 20 wt %, la estabilidad térmica de la resina epoxi pura presente en la muestra se incrementa en el rango de temperaturas correspondiente al segundo escalón o etapa de degradación, siendo el cambio estadísticamente significativo. Este hecho puede resultar indicativo de la interacción entre las

fases orgánica e inorgánica dentro de la muestra, teniendo en cuenta que los datos analizados son las curvas TG correspondientes a, únicamente, la resina epoxi existente dentro de cada compuesto formado. La interacción descrita puede estar relacionada con la aparición de una interfase orgánica-inorgánica, ya apuntada en Tarrío-Saavedra et al. (2010a).

Asimismo, para realizar un completo estudio de la estabilidad térmica de un material y su variación con respecto a la modificación de un factor (la cantidad de humo de sílice añadida), es necesario estudiar la velocidad de degradación térmica. Una menor velocidad de degradación térmica, a una temperatura dada, indica una mayor estabilidad térmica del material. La velocidad de degradación térmica se ha estudiado mediante las curvas DTG. Los resultados obtenidos mediante la aplicación del test ANOVA funcional y de las pruebas pareadas sostienen que la velocidad de degradación térmica varía con significación estadística al añadir una cantidad creciente de humo de sílice. La velocidad de pérdida de masa disminuye al comienzo de la segunda etapa de degradación, que se manifiesta mediante el pico exotérmico principal en las curvas DTG. Por otro lado, la temperatura a la cual se produce la mayor pérdida de masa apenas sufre cambios, independientemente de la proporción de carga añadida; sin embargo, el módulo de la máxima velocidad de pérdida de masa disminuye al añadir un 20 wt % de humo de sílice. Con respecto a la velocidad de degradación térmica, se puede decir que ésta es diferente dependiendo de la cantidad de humo de sílice añadida, mediante la aplicación del ANOVA funcional y de las comparaciones dos a dos (pairwise).

En los grupos correspondientes al 10 y 20 wt % de sílice, las curvas reescaladas y sus derivadas son muy similares. Las diferencias son mayores en la última etapa de degradación, aunque también se aprecian diferencias significativas en la segunda etapa de degradación térmica. Finalmente, la adición de 20 wt % de sílice puede causar una disminución en la estabilidad térmica a temperaturas muy altas, debido a la propia morfología del material. Por consiguiente, si el objetivo buscado es el aumento de la estabilidad térmica, economizando en la medida de lo posible los costes derivados del material, una adición del 10 wt % en humo de sílice podría ser una buena solución.

El estudio que aquí se presenta, mejora en muchos aspectos y complementa el correspondiente a Tarrio-Saavedra et al. (2008), gracias a la aplicación de técnicas estadísticas de caracter funcional.

En la Parte II, se ha probado, fundamentalmente, el desempeño de diferentes métodos de clasificación supervisada funcional y multivariante para identificar especies de madera a partir de sus curvas TG, de sus curvas DSC y de sus micrografías SEM tomadas a 1500× aumentos. Se han compara-

do diferentes planteamientos usando para ello datos reales, además de un completo estudio de simulación de las curvas TG. La simulación se efectuó a partir de los parámetros de un modelo de regresión, resultado de la suma de 4 componentes logísticas generalizadas. Se pueden obtener conclusiones similares a partir de ambos estudios, comprobándose que el intervalo de temperaturas donde se alcanza un porcentaje de clasificación correcta más alta se corresponde con los intervalos de degradación de sus 3 constituyentes fundamentales, especialmente con el de la hemicelulosa. Se ha observado también que la utilización, por vez primera, de micrografías que describen la estructura de las traqueidas de la madera joven, y su posterior procesado (mejora de la imagen, segmentación), es una alternativa factible si se la compara con otras bases de datos (espectros, datos térmicos, fotografías, etc.). Se demuestra, en definitiva, que la clasificación supervisada de especies de madera es posible mediante la aplicación de las técnicas estadísticas descritas a sus correspondientes micrografías SEM (a las 5 características representativas extraídas de cada una) y curvas TG.

Por último, es importante subrayar que los trabajos de investigación de los que consta la Parte II se llevaron a cabo teniendo en cuenta algunas especies de madera específicas, pero las metodologías aquí descritas se podrían aplicar a otras especies diferentes o, incluso, a materiales diferentes de la madera. La aplicación de los procedimientos aquí descritos en un amplio espectro de materiales, tanto en el ámbito industrial como académico, representa, a la vez, una motivación y la posible extensión futura de este trabajo.

Una gran parte de los enfoques mostrados en los Capítulos 3, 4 y 5 muestran alternativas viables para la clasificación supervisada de muestras de especies de madera con interés industrial.

Con respecto al Capítulo 3, entrando más en detalle, se enumeran las siguientes conclusiones:

La principal contribución del estudio de simulación realizado es la posibilidad de diseñar escenarios con diferentes valores para las varianzas y covarianzas de las curvas TG artificiales. Este estudio de simulación permite estudiar el comportamiento de los métodos propuestos en situaciones más desfavorables que las obtenidas experimentalmente. De hecho, es muy complicado obtener una muestra totalmente representativa de un conjunto de especies de madera, pues se presentan diferencias no sólo dentro de los árboles de la misma especie, sino también dentro del mismo árbol. Esto puede ser debido a las muy variadas condiciones ambientales soportadas, a los diferentes tipos de suelo, a la edad del árbol, etc. Dada la gran heterogeneidad de un material como la madera, la simulación de escenarios con distintos grados de variabilidad y dependencia (entre parámetros de una misma com-

ponente logística) no sólo representa una alternativa eficaz sino que es un procedimiento necesario para obtener unos resultados fiables.

Se han considerado tres diferentes bases de datos: las curvas TG, los parámetros obtenidos a partir del ajuste a cada curva TG de un modelo de regresión compuesto por la suma de funciones logísticas generalizadas, y las componentes principales de las curvas TG (que representan el 99 % de la variabilidad de los datos). Es decir, se utilizan dos bases de datos novedosas en el campo de la Quimiometría como son los propios datos funcionales obtenidos y los parámetros del modelo (que representa un nuevo y sencillo método de extracción de características representativas de cada curva), comparándose los resultados con los obtenidos a partir de una base de datos tradicional en ese mismo campo, las puntuaciones (o "scores") calculadas mediante PCA. Por otra parte, se utilizaron dos procedimientos de validación diferentes: la validación externa basada en la selección aleatoria de las muestras de entrenamiento y de test y, por otro lado, un proceso de validación cruzada tipo "leave-one-out", LOO. El empleo de ambos procedimientos es necesario pues cada uno de ellos tiene sus inconvenientes: la estimación de los errores de clasificación por LOO puede presentar una gran variabilidad, mientras que el método de validación externa empleado puede presentar cierto sesgo, como ya se ha apuntado en el Capítulo 3.

En general, se obtuvieron las mayores probabilidades de clasificación correcta utilizando los parámetros logísticos como conjunto de datos y el Análisis Discriminante Lineal (LDA) como método de clasificación. En este caso, la aplicación de los métodos Bayes Naive (NBC), Máquinas de Soporte Vectorial (SVM) y Redes Neuronales (NN) ha producido resultados competitivos. Se obtuvieron proporciones de clasificación correcta menores al aplicar los métodos de clasificación multivariante descritos a las puntuaciones de las componentes PCA obtenidas a partir de las curvas TG.

Respecto al método no paramétrico funcional kernel basado en el estimador de Nadaraya-Watson, K-NFDA, los resultados fueron generalmente peores, aunque se consiguieron altas probabilidades de clasificación correcta cuando las curvas TG artificiales se generaron en escenarios definidos por una varianza pequeña y bajo la hipótesis de independencia. Téngase en cuenta también que el método K-NFDA necesita de un tiempo de cálculo más corto que los demás métodos (principalmente debido al tiempo que se pierde en la extracción de características representativas a partir de cada curva). Es por tanto, en estos escenarios, una alternativa más que viable.

En el caso de la validación externa, se obtuvieron altas probabilidades de clasificación correcta en todos los escenarios considerados, incluso en las condiciones más desfavorables, relacionadas con un alto grado de variabilidad

y de dependencia, y con un tamaño de muestra diferente para cada especie simulada. La aplicación de los métodos LDA, NBC y SVM han producido porcentajes de clasificación correcta superiores al 80 %. Por otro lado, la metodología propuesta ha resultado robusta cuando se varía el número de muestras pertenecientes a cada especie dentro del conjunto de las muestras de entrenamiento.

Teniendo en cuenta todos estos resultados, se recomienda la aplicación del método LDA a los parámetros obtenidos por el ajuste logístico para abordar este tipo de problemas de clasificación. Este método ha proporcionado un buen comportamiento general, tanto en la aplicación inicial a datos reales como en las simulaciones, independientemente de la magnitud de varianza o covarianza de los datos (en otras palabras, sin importar la heterogeneidad de las muestras reales). Además, necesita de un tiempo de cálculo menor para su aplicación que otros enfoques multivariantes mostrados en el Capítulo 3.

Las conclusiones del Capítulo 4 están en relación directa con los objetivos mostrados en el Capítulo 1:

Se ha evaluado el potencial de los métodos no paramétricos funcionales de análisis discriminante para la clasificación de grupos genéricos amplios (maderas blandas y duras) y especies de madera. Así, se ha probado factible clasificar diferentes especies de madera usando las curvas TG como característica discriminante, obteniéndose un porcentaje de clasificación correcta de hasta el 90 %, que mejora sensiblemente a aquéllos calculados en el Capítulo 3 por el método K-NFDA. De hecho, las probabilidades de clasificación correcta, obtenidas mediante métodos funcionales, han resultado ser más altas que cualquier resultado obtenido en el Capítulo 3, por cualquier método de clasificación aplicado a las curvas TG experimentales. Esto es debido a la gran mejora que supone el analizar las curvas TG en diferentes rangos de temperatura y no, únicamente, emplear las curvas TG completas. En conclusión, la descomposición de determinados contituyentes de la madera ayudan más que otros a distinguir entre especies. También se han obtenido resultados altamente satisfactorios si se quiere distinguir entre maderas blandas o coníferas, frondosas boreales y otras frondosas (tropicales y australes); de hecho, los resultados alcanzados (94 % de clasificación correcta) son comparables a los obtenidos mediante el procesado de imágenes y espectros en otros trabajos relativos al tema en cuestión.

En cambio, las curvas DSC obtenidas en un analizador simultáneo SDT como el descrito, mostraron un menor poder discriminante que las curvas TG. Esto puede estar relacionado con la menor sensibilidad y resolución de las curvas DSC obtenidas mediante este aparato, especialmente si las comparamos con las obtenidas mediante un dispositivo DSC puro. Esto se debe,

entre otras cosas, a la posición en la que se encuentran los termopares. Sin embargo, se han obtenido muy buenos resultados (una proporción de clasificación correcta del 80 %) utilizando este tipo de curvas para distinguir entre 3 grandes grupos: frondosas boreales, otras frondosas (tropicales y australes) y coníferas. También se han obtenido resultados más que aceptables al distinguir entre determinadas especies (castaño, jatobá y eucalipto).

En el caso de las curvas TG, los resultados de clasificación obtenidos en cada intervalo estudiado se han relacionado con los intervalos de temperaturas a los que se degradan la celulosa, la lignina y la hemicelulosa en una atmósfera inerte. Se ha observado que los rangos de temperatura para los que se alcanza una proporción mayor de clasificación correcta, se corresponden básicamente con los intervalos de degradación térmica de los tres constituyentes principales de la madera (obtenidos mediante el estudio de la bibliografía del Capítulo 4), en mayor medida con el intervalo correspondiente a la hemicelulosa. Sin embargo, en lo que se refiere a la utilización de las curvas DSC, el intervalo de temperaturas en el cual se obtiene una mayor proporción de clasificación correcta se corresponde principalmente con el rango de degradación de la celulosa y la lignina, eso sí, obteniendo unos resultados sensiblemente peores que los obtenidos mediante el estudio de las curvas TG.

Si se observa la comparación de los diversos métodos de clasificación funcionales no paramétricos, aquéllos basados en el estimador de Nadaraya-Watson (denominados en el Capítulo 4 por K-NPFDA y KNN-NPFDA) proporcionan una probabilidad de clasificación correcta mayor que aquéllos basados en el algoritmo Adaboost. Además, presentan la ventaja de obtener resultados en un período de tiempo más corto. Esta es una de las razones por las que en el Capítulo 3 se ha empleado el clasificador K-NPFDA (así llamado en el capítulo 4; K-NFDA en el Capítulo 3).

Se demuestra, por último, el alto rendimiento que proporciona la transformación efectuada a las curvas antes de la aplicación de los diferentes métodos de clasificación.

En el Capítulo 5, en contraposición al uso de datos térmicos para tareas de clasificación, ya sea desde una perspectiva funcional o multivariante, se ha propuesto una base de datos alternativa: las micrografías obtenidas por Microscopía Electrónica de Barrido (SEM). El uso de este tipo de datos presenta diversas dificultades como es la selección del número de aumentos, la necesidad de llevar a cabo los procesos de mejora de la imagen y segmentación (reducción a dos tipos de píxeles de diferente tonalidad, blanco y negro) y la elección y extracción de características a partir de las imágenes segmentadas.

Primeramente, hay que señalar que ha resultado del todo factible la clasi-

ficación supervisada de especies de madera a partir de 5 características extraídas a partir de las micrografías SEM tomadas a 1500× aumentos, en secciones transversales a la dirección de las traqueidas, en zonas correspondientes a la madera joven. Gracias a los procesos de mejora y segmentación de las imágenes ha sido posible extraer esas 5 características, representativas de cada micrografía, donde se ha pretendido definir la geometría, distribución y tamaño de las traqueidas. De esta forma, y aplicando los distintos métodos de clasificación a los vectores de características que definen cada imagen, se han alcanzado probabilidades de clasificación correcta relativamente altas cuando se pretende distinguir entre maderas frondosas (duras) y coníferas o maderas blandas: 0,89 utilizando un procedimiento LOO y 0,93 mediante un proceso de validación externa. Si se pretende clasificar entre 7 especies de madera diferentes, teniendo en cuenta la alta heterogeneidad extructural de la madera, las probabilidades de clasificación correcta obtenidas siguen siendo altas: 0,81 usando LOO, y 0,80 mediante un proceso de validación externa. En definitiva, se puede decir que las características extraídas de las micrografias SEM tomadas a 1500× aumentos, definen y diferencian, en gran medida, unas especies de madera de otras.

En lo concerniente a los métodos de clasificación, se ha podido observar que las probabilidades de clasificación correcta más altas se obtienen, por lo general, empleando métodos pertenecientes al denominado aprendizaje máquina (SVM, NN). Esto último contrasta con los resultados obtenidos a partir de las curvas TG dentro del Capítulo 3, en el que métodos tradicionales como el LDA proporcionaban, en general, las proporciones de clasificación correcta más altas.

En el Capítulo 5 también se intrudujo con éxito el concepto de dimensión fractal a través de dos de sus estimadores: "Fractal Brownian Motion" (FBM) y "Box-Counting" (BC). En particular, se añadió a las 5 características que definen cada una de las 101 micrografías estudiadas, la dimensión fractal calculada mediante el estimador BC. El resultado fue el incremento de la proporción de clasificación correcta: el mejor resultado en el caso de la clasificación entre 7 clases diferentes presentó un valor de 0,83, usando para ello el método Bayes Naive. Además, también se obtuvieron mejores resultados cuando se pretende clasificar una muestra dentro de dos grandes grupos, frondosas y coníferas, aplicando los métodos $k$-Nearest Neighbors o $k$-vecinos más próximos ($k$-NN) y NN (0,93). Si, adicionalmente se añade a la base de datos la estimación BC, se obtiene un significativo aumento de la proporción de clasificación correcta en el caso de las 7 clases, mientras que en el problema de 2 clases los resultados son similares. El cómputo de nuevos estimadores de la dimensión fractal de imágenes puede ser una alternativa

interesante a la extracción de características relacionadas con la geometría, e incluso una alternativa al uso de las puntuaciones correspondientes al Análisis de Componentes Principales (PCA) en el caso particular de curvas.

Es de destacar la gran diferencia existente entre la madera de eucalipto y el resto de maderas estudiadas, teniendo en cuenta la geometría, distribución y tamaño de las traqueidas pertenecientes a zonas de madera joven. Por el contrario, observando las mismas 5 características representativas, las maderas de pino rojo y haya son las que presentan las mayores similitudes y, por tanto, mayor confusión una vez aplicados los clasificadores propuestos.

Por último, se ha demostrado la efectividad del proceso de segmentación frente a otros procesos de tratamiento de imágenes y extracción de características como es el llamado "gray level co-occurrence" (GLCM).

## 6.2. Líneas futuras de investigación

En esta sección, y para finalizar, se enumeran brevemente una serie de líneas futuras de investigación que arrancan de los trabajos que aquí se presentan y describen. Algunas de ellas representan estudios recientemente enviados a revistas, mientras que otros están todavía en diversos estados de elaboración.

1. Dado que, en la Parte II de la presente tesis, se obtuvieron probabilidades de clasificación correcta relativamente bajas utilizando las curvas DSC como base de datos, se ha propuesto la utilización de una técnica similar, la Calorimetría Diferencial de Barrido bajo condiciones de alta Presión (PDSC), utilizada principalmente en el análisis y caracterización de combustibles. Esta es una técnica experimental relativamente novedosa. En este trabajo, admitido para su revisión en la revista *Analytica Chimica Acta*, se aborda, por tanto, el mismo problema expuesto en la Parte II pero utilizando como característica discriminante las propiedades termo-oxidativas de las maderas. Mediante esta aplicación se han obtenido resultados muy satisfactorios, competitivos con respecto a los que se obtuvieron mediante la utilización de curvas TG y el procesado de imágenes y espectros. Se han aplicado, igualmente, métodos de clasificación supervisada multivariantes y funcionales (que representan técnicas novedosas en el campo de la Quimiometría). En este trabajo también se ha dado un paso más en lo que se refiere al proceso de extracción de características representativas de las curvas. Por un lado se calculan las puntuaciones resultantes de análisis PCA (se obtiene el espacio de las componentes principales correspondiente

a la muestra de entrenamiento, para acto seguido proyectar la muestra de test en dicho espacio; de este modo se evitan sesgos positivos en la estimación de las probabilidades de clasificación correcta) y se emplean las puntuaciones correspondientes a la aplicación de la técnica "Partial Linear Squares" (PLS), masivamente utilizada durante los últimos 10 años en el campo de la Quimiometría. Y, por otro lado, se proponen dos nuevos métodos de extracción de características:

- Utilización de los parámetros de un modelo de regresión no lineal basado en la ecuación de Arrhenius, ajustado a las curvas PDSC, como características.
- Cálculo de 6 estimadores diferentes de la dimensión fractal aplicada a series de tiempo.

Los resultados obtenidos hacen de este enfoque una alternativa viable para la clasificación de la madera.

2. La aplicación de los procesos de mejora y segmentación de imágenes, así como la extracción de características y la estimación de la dimensión fractal se postulan como herramientas muy útiles para llevar a cabo el control de calidad de piezas industriales, usando como datos la toma de imágenes en distintos formatos. La identificación de fallos en cigüeñales mediante la toma de imágenes se presentará como un caso particular de estudio.

3. Los procesos de segmentación y extracción de características, descritos en el Capítulo 5, serán de gran utilidad en dos futuros proyectos de investigación que se llevarán a cabo con el "Laboratoire Matière et Systèmes" de la "Université Paris Diderot, Sorbonne Paris Cité". En uno de ellos se pretende estimar la relación entre la geometría, número y distribución de nanopartículas magnéticas con respecto al tiempo de aplicación de un campo magnético. Las aplicaciones últimas de este proyecto incipiente tendrían relación con el transporte de medicamentos dentro del cuerpo humano. En el otro, la tarea consiste en medir el número y estimar la distribución de microesferas de copolímero en función del tiempo en el que se aplica una determinada presión, con el objetivo de relacionar los resultados con las propiedades reológicas de las mismas, también medidas a través del tiempo y bajo presión.

4. Otra futura línea de investigación nace de la toma de datos térmicos (TG, DSC) y la aplicación de técnicas FDA. Habiendo aplicado ya

técnicas de diseño de experimentos y clasificación, el empleo de la regresión funcional para la resolución de problemas en el ámbito de la Química Analítica y la Ciencia de Materiales representa el siguiente paso. De hecho, sería de utilidad inmediata la aplicación de modelos de regresión con variable respuesta y regresora funcionales. El objetivo sería la estimación de la curva DSC, obtenida en unas determinadas condiciones experimentales, a partir de la curva TG correspondiente al mismo material ensayado en esas condiciones, o viceversa. Dado el coste de los dispositivos de medida DSC y TG, disponer de un modelo que pudiera proporcionar fielmente este tipo de estimaciones supondría un ahorro significativo en tiempo y dinero, aparte de proporcionar una gran flexibilidad de trabajo a los investigadores del campo del Análisis Térmico.

5. Finalmente, la aplicación de las técnicas de diseño de experimentos y regresión funcionales al estudio del fallo a fatiga de materiales metálicos representa también un proyecto en fase de gestación.

# Parte III

# Apéndices: publicaciones en su versión original

# Apéndice A

# Application of functional ANOVA to the study of thermal stability of micro-nano silica epoxy composites

Chemometrics and Intelligent Laboratory Systems 105 (2011) 114–124

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Application of functional ANOVA to the study of thermal stability of micro–nano silica epoxy composites

Javier Tarrío-Saavedra [a], Salvador Naya [a], Mario Francisco-Fernández [b,*], Ramón Artiaga [a], Jorge Lopez-Beceiro [a]

[a] University of A Coruña, Higher Polytechnic School, Campus de Esteiro, Ferrol 15403, Spain
[b] University of A Coruña, Faculty of Computer Science, Campus de Elviña, s/n, A Coruña 15071, Spain

ARTICLE INFO

*Article history:*
Received 9 September 2010
Received in revised form 17 November 2010
Accepted 18 November 2010
Available online 26 November 2010

*Keywords:*
Functional data
Functional ANOVA
TGA
Silica nanocomposites

ABSTRACT

The main purpose of this work is to use a new technique that combines functional data analysis and design of experiments, functional ANOVA for a one way treatment, to measure the influence of adding fumed silica on the thermal degradation of an epoxy resin. To achieve this, a design of experiments with a treatment factor (the amount of fumed silica) at three different levels (0, 10 and 20 wt.%) is performed. The data are obtained through the use of Thermogravimetric Analysis (TG), resulting in five degradation curves per level. The functional ANOVA uses all the information of each curve or functional data. The results obtained using this methodology with the TG rescaled data and their derivatives (DTG) indicate that the amount of fumed silica significantly affects the thermal stability of the compound. These facts may be indicative of the interaction between the organic phase and the inorganic particles. In addition, pairwise comparisons using the functional ANOVA method and a bootstrap distance based test are carried out to discern which factor levels provide different ways of degradation.



## 1. Introduction

Although epoxy resins are widely used thermostable polymers, their use is somewhat limited due to the high stiffness, caused by the dense cross-linking structure of these materials. However, their mechanical properties can be improved by the addition of inorganic particles [1]. The shape, volume, size, surface characteristics and dispersion of particles within the matrix determine the mechanical properties of resulting composites [2–4]. The nano-composites with organic matrix and inorganic fillers have proven capacity of providing simultaneous increases in properties such as thermal stability, flame retardation, glass transition temperature and dimensional stability, as well as the decrease of the dielectric constant [5–7].

In the present work, we perform an experimental design to evaluate the effect of the addition of fumed silica on the thermal degradation of the resulting material. The fumed silica epoxy-resin composites are prepared and characterized by Thermogravimetric Analysis (TG) and Differential Thermogravimetric Analysis (DTG), usual techniques in assessing the thermal stability of a material [8]. A non-conventional epoxy resin based on trimethylolpropane (TMP), particularly suitable for the manufacture of composite materials, is chosen. Moreover, fumed silica used is a byproduct of the manufacture of silicon and ferrosilicon. It is produced at the top of the melting furnaces, thus its production method is different from the conventional processes for synthetic $SiO_2$ [9]. Due to the special characteristics of the production method, the fumed silica used has a variable purity, depending on the operating conditions in the furnaces. In any case, the silica weight ratio is never less than 95%. It is also variable in particle size. In fact, fumed silica consists of nano- and micro-particles, taking into account the Schadler approach [10] (diameter $<100$ nm implies that it is a nano-particle). This particular size distribution suggests a possible combination of micro and nano effects.

To perform the analysis previously mentioned, a new statistical method, called one-way functional ANOVA, is applied. This procedure allows to test the possible differences in responses according to the treatments used, considering that the data are functions or curves. In our case, the silica content in each sample, with three levels (0, 10 and 20 wt.%, weight percentage, of fumed silica) is chosen as the treatment factor or explanatory variable. Five experiments or replicates for each level are considered, which gives a balanced design. The number of replicas selected is set to reach an acceptable compromise between the adequate representation of the variability within each level and the total experimental time required. The response variables or dependent variables are functions, where each one is a curve representing the mass of material depending on the temperature at which it is subjected to. To obtain them, a constant increase in temperature of 10 °C/min is scheduled. All the curves decrease because the material degrades, or loses mass, by increasing the temperature to which it is subjected to. Therefore, each curve represents the particular way of degradation of each sample tested.

* Corresponding author. Tel.: +34 981167000x1222; fax: +34 981167160.
*E-mail address:* mariofr@udc.es (M. Francisco-Fernández).

0169-7439/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2010.11.006

Our analysis allows to answer questions like, will the way of degradation be different for different levels of the factor *amount of silica*? or, can it be said that the thermal stability of the material increases or decreases with statistical evidence?

On the other hand, an important part of this study is to determine the degree of interaction between fumed silica and epoxy resin, related to organic–inorganic interphase. This region is defined as starting at the point where the filler differs from the rest of the load and it finishes at the point of the matrix in which its properties are the same as in the rest of the matrix [10]. The existence of this interphase affects some properties, such as thermal stability and glass transition temperature. In fact, the variation of these properties can be taken as an index of its existence [8].

To be able to compare all the data conveniently, the mass of each sample is expressed as a percentage of the initial amount [11]. Thus, all curves start with a value of 100% in the vertical axis.

The content of the paper is as follows. In Section 2, a comprehensive review on the issue of functional data analysis (FDA) is presented, and the statistical method used in our research is described. In Section 3 the experimental process carried out to obtain the data is described. Section 4 presents a descriptive analysis of the data under consideration as well as the data preprocessing needed to apply the statistical methods. The results and discussion are included in Section 5, while Section 6 collects the main conclusions.

## 2. Statistical methods

In this Section, the main statistical technique used in our analysis is briefly described. This method takes advantage of the functional nature of the data under consideration, producing more reliable results. First, a comprehensive review about FDA is presented.

### 2.1. Functional data analysis (FDA)

FDA deals, in general, with experiments whose data and/or results are curves. According to [12], we could say that a random variable $X$ is called a functional variable if it takes values in an infinite dimensional space (or functional space normed or semi-complete normed) and therefore, an observation $x$ of $X$ is called a functional data and, in addition, a functional dataset $x_1,\ldots,x_n$ is the observation of $n$ functional variables $X_1,...,X_n$ identically distributed as $X$. The functional data, also called longitudinal data, turn out to be associated with processes continuously monitored in time. That is, when a variable is measured on a discrete and finite set of arranged values, considering that this variable follows a continuous functional relationship. A special case is when the functional variable $X$ belongs to a Hilbert space, as it is the case of continuous functions on an interval [13]. This is the case of the TG curves used in the present work, $X \in L_2([0,T])$ [14]. In addition, FDA often makes use of the information in the slopes and curvatures of curves, reflected in their derivatives. This is the case of the TG derivative curves (DTG) also used in our research.

This relatively new research field has received a lot of attention by the scientific community over the last two or three decades, although the study of probabilistic tools for infinite dimensional variables started in the beginning of the 20th century [15]. Nevertheless, lately, the interest in FDA methods has considerably increased, since the technological progress allows collecting observations of infinite dimensional objects. The books by Ramsay and Silverman [13] and that of Ferraty and Vieu [12] are good introductory texts about this kind of data. On the other hand, the recent monograph of Ferraty and Romain [16] presents the latest progress on this topic. Many databases belonging to very different branches of science are likely to be treated as functional data: econometrics, medicine, environmetrics, geophysics, biostatistics and, of course, chemometrics. Apart from the examples studied in [13] or in [12], there are many other datasets where functional analysis techniques are successfully applied. Many of them are treated in the special issues presented in [15,17–19]. For example, climatologic and environmetrical curves are studied in [20], economical curves in [21], spectrometric curves in [22,23], geophysics curves are analysed in [24], and biostatistics datasets are treated as functional data in [22].

Given the wide range of databases that can be studied as functional data, and given the technical possibility of treating them, a considerable effort is being made for adapting the standard statistical methods to the functional context. For example, principal component analysis for functional data is studied in [25]. Regression models with functional covariates (and scalar or functional response) are analysed in [26–30]. Functional data classification is other important field in FDA [31]. This terminology includes supervised and unsupervised classification. Finally, in relation with the method used in the present paper, the procedure proposed in [23] is particularly interesting.

Some of the references listed previously use nonparametric methods (kernel estimation, wavelets, splines, …) to analyse functional data, for example, [22,29,31]. In this framework, it is important to stress the monograph by Ferraty and Vieu [12], where many of these two authors' contributions to the nonparametric estimation with functional data are summarized. Nonparametric techniques do not assume, in general, any parametric shape for the functions to be estimated or a specific distribution for the variables under consideration. In this sense, nonparametric functional data techniques are powerful and flexible tools.

When working with functional data, previously to the application of a specific statistical procedure, sometimes, a data preprocessing should be done to prepare the datasets to be analysed. For example, in practice, sample curves are usually observed in a finite set of points. These points can be not equally spaced or can be different for the observed curves. In this case, the usual step is using a smoothing process, representing the functions in a proper functional basis, for example (using $b$-splines, for instance), to try to recover the original functional relationship. On the other hand, if the curves show a similar pattern, but present variations in their range or in their domain, transformations of the curves can be useful to align special features or to minimize variability. Different procedures are proposed for this task in the literature in many fields. *Marker registration* involves identifying the timing of specified features in the curves, and then transforming the argument of the curves so that these marker events occur at the same moment. A comprehensive reference is [32]. *Time warping*, a term mainly used in the engineering literature, is applied in the procedure proposed in [33], while its statistical aspects are studied in [34]. These methods, however, can be sensitive to errors in feature location, and these features may even be missing in some curves. Moreover, substantial variation in the argument of the curves may remain between widely separated markers. Under the name of *curve registration*, in [35], a technique that does not require markers is developed, and in [36] the Silverman's method is extended by using a flexible smooth monotone transformation family developed in [37]. These approaches involve defining an entire warping curve for each observation, and they can be somewhat complex to program. In [38], a local nonlinear regression technique, computationally convenient, is described for identifying the smooth monotone transformations.

In the analysis of the real curves presented in the next sections, preliminary to the application of the functional ANOVA test, some approaches related with those previously described are used. On one hand, each curve is written as a linear combination of the elements of a functional basis. To select a proper basis, $b$-splines and penalized $b$-splines fits are used. On the other hand, in line with the results pursued with the registration of the curves, in our case, a simple rescaling of the curves can give a clue of possible existence of a chemical interaction between resin and inorganic. A detailed description about this is given in Section 4.

### 2.2. Functional ANOVA

When the data are functional, an alternative to the classical Analysis of Variance (ANOVA) is the named functional ANOVA (FANOVA) [39].

The covariates are factors while the response is functional. This technique, compared with the classical one, has the advantage of using all the information in the curves, instead of some specific values on them.

Following the nomenclature of [39], each functional data can be written as $X_{ij}(t)$, where $t$ usually represents the time, with $t\in[a,b]$, $i$ is the subscript that indicates the level of factor and $j$ the replication number ($j=1,2,\ldots,n_i$ and $i=1,2,\ldots,k$). Variables $X_{ij}(t)$ can be considered as $k$ independent samples of trajectories drawn from $L_2$-processes $X_i$, $i=1,\ldots,k$.

In the present study, as the variable of interest, *the mass of the sample*, is evaluated originally at every second, the data can be considered functional. In this case, the temperature is directly proportional to the time (10 °C/min), and therefore, $t$ can be considered as the values of the temperature instead of the time.

The mean for each level or independent sample is given by $E(X_i(t))=m_i(t)$, while the covariance between two specific values of a curve, $Cov(X_i(s),X_i(t))$, in the most restrictive case (existence of heteroscedasticity), can be estimated by $K_i(s,t)$:

$$K_i(s,t)=\sum_{j=1}^{n_i}\frac{\left(X_{ij}(s)-\overline{X}_{i\cdot}(s)\right)\left(X_{ij}(t)-\overline{X}_{i\cdot}(t)\right)}{n_i-1} \tag{1}$$

We want to test:

$$H_0: m_1=m_2=\ldots=m_k \tag{2}$$

The statistic implemented by [39] to test Eq. (2) is as follows:

$$V_n=\sum_{i<j} n_i\|\overline{X}_{i\cdot}-\overline{X}_{j\cdot}\|^2 \tag{3}$$

The use of Eq. (3) avoids the requirement of the hypothesis of homoscedasticity in the usual ANOVA.

Assuming:

1. $n_i, n\to\infty$ in such a way that $n_i/n\to p_i>0$ for $i=1,\ldots,k$,
2. and the observations $X_{ij}(t)$ with $j=1,\ldots,n_i$, corresponding to independent samples of size $n_i$ from $k$ $L_2$-processes with mean zero and covariance $Cov(X_i(s),X_i(t))$.

It can be proved that the asymptotic distribution of $V_n$, under $H_0$, coincides with that of the statistic

$$V=\sum_{i<j}^{k}\|Z_i(t)-C_{ij}\cdot Z_j(t)\|^2, \tag{4}$$

where $C_{ij}=(p_i/p_j)^{1/2}$ and $Z_1(t)$, ..., $Z_j(t)$ are independent Gaussian processes with mean zero and covariance $Cov(X_i(s),X_i(t))$.

To apply the test, if the $n_i$ are large enough, hypothesis $H_0$ is rejected, at a level $\alpha$, whenever $V_n>V_\alpha$ where $P_{H_0}(V>V_\alpha)=\alpha$.

In a practical situation, the distribution of $V$ under the null hypothesis can be approximated by applying a parametric bootstrap and the Monte Carlo method. This allows to estimate the value of the $\alpha$-quantile, $V_\alpha$.

In this case, the use of the parametric bootstrap is justified because the distribution of $V$ is a complicated function of $k$ Gaussian processes. This bootstrap procedure consists of resampling from these Gaussian distributions, but replacing their parameters with the maximum likelihood estimators calculated from the original sample. A detailed description of this procedure with the data analysed in this paper can be found in Section 5.

In the previous procedure the word *functional* in the term functional ANOVA refers to the functional nature of the data. A similar approach to the one used here has been already considered in [13,40], but there the problem is formulated from a regression point of view and a (pointwise) $F$ statistic at each $t$ is proposed. However, in the same functional context as here, there are several works developing an $F$ test defined by a unique value. In [41], a HANOVA (high dimensional ANOVA) test, relying on wavelet thresholding techniques, which limits the size of each curve (discretized curves) to satisfy some asymptotic assumption, is proposed. An $F$ test for choosing among two nested functional linear models is developed in [42]. A discussion of the ways in which the functional ANOVA can be treated is shown in [43]. Other interesting reference is [23], where standard approaches based on factorial analysis for comparing groups of multivariate data are extended to the infinite-dimensional framework. One of the differences between this and other tests and the one applied in the present article is in the discretization process. As stated in [39], the test applied here is purely functional in the sense that our test statistic is a functional of the sample trajectories and its motivation is also given in functional terms. Since it is conceivable that the increasing interest on functional data could lead to measurement devices whose outputs provide *true* functions, with *analytical* expressions (obtained maybe by nonparametric smoothing) instead of finite dimensional approximations, our procedure could give more reliable results.

It is very important to note that the term *functional* ANOVA is also used for different models to the one used here by some authors. Specifically, they employ these words when a multivariate function is represented by a decomposition in terms of functions of fewer variables, linked with non-linear models. In that context, there are also substantial works where functions of the predictors are estimated and tested. Several applications of this methodology can be seen in the following papers. In [44], analysis of variance type models are considered for a regression, conditional probability, density and hazard functions using polynomial splines to model the main effects and interaction components. In [45], the logarithm of the relative risk function in a proportional hazards model involving one or more possibly time-dependent covariates is treated as a sum of a constant, main effects and selected interaction terms. Moreover, in [46], a methodology for modeling covariate effects on the time-to-event data is developed using polynomial splines. Therefore, it is important to point out that the statistical community is (curiously) using the same wording (FANOVA) for two very different models.

## 3. Materials and instrumental methods

An epoxy resin matrix consisting of two components is used. It is based on the diglycidyl ether of trimethylolpropane, Triepox GA, manufactured by Gairesa, SA. The curing agent used is an aromatic amine, 1,3-benzenedimethanamine, supplied by Aldrich. Triepox GA is a highly thixotropic resin that also possesses a low density and the ability to cure at room temperature in the absence of plasticizers or additives.

The fumed silica has been provided by Ferroatlántica I + D, Spain. It is obtained as a byproduct in the production of silicon in electrical melting furnaces. This process involves the reduction of high purity quartz, at temperatures above 1800 °C. Fumed silica is formed when SiO gas, resulting from the reduction of quartz, is mixed with oxygen at the top of the furnace, resulting in the production of spherical particles of silica. It is a fine powder varying in colour from nearly black to slightly off-white, according to their carbon content. Its average particle size is 0.15 μm and 41.9% of the particles have a diameter less than 0.2 μm, as shown in Table 1. The surface area is about 20 $m^2\ g^{-1}$.

As regards to the chemical composition, fumed silica consists of variable purity amorphous $SiO_2$. Table 2 shows the main physical properties and composition.

The samples are prepared for contents of 0, 10, and 20 wt.% of fumed silica. Both resin and hardener are mixed in a stoichiometric ratio. To obtain the compounds corresponding with 10 and 20 wt.%, the silica and resin mixtures are stirred for 15 min in order to obtain a distribution as uniform as possible. Then, an ultrasonic treatment is applied for 5 min at room temperature to disperse the silica agglomerates. The paste thus obtained is poured into a silicone mold with cavity dimensions of

**Table 1**
Particle size distribution in the fumed silica.

| Diameter (μm) | Mass (%) |
|---|---|
| 50–100 | 1.2 |
| 20–50 | 2.0 |
| 10–20 | 0.2 |
| 5–10 | 0.5 |
| 2–5 | 1.4 |
| 1–2 | 1.5 |
| 0.5 − 1 | 8.3 |
| 0.2 − 0.5 | 43.0 |
| <0.2 | 41.9 |

$0.8 \times 4 \times 30$ mm. In this area, the samples are cured at room temperature for 24 h and then a post-curing is applied at 90 °C for 2 h.

The TG experiments are carried out using a thermo-balance STA 1500, Rheometric Scientific. All samples are subjected to a heating ramp of 10 °C/min in a temperature range between 20 and 600 °C. All experiments are performed under nitrogen atmosphere, maintaining an air flow of 50 mL $min^{-1}$.

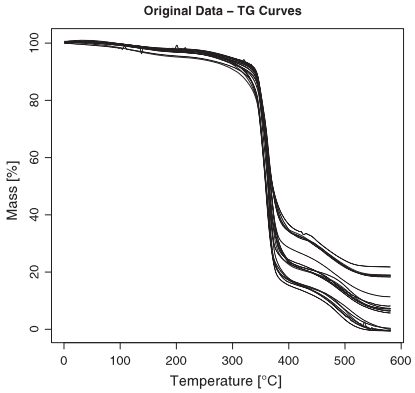


**Fig. 1.** Experimental data: starting TG curves.

## 4. Descriptive analysis and data preprocessing

In order to obtain further information from the experiments performed, three different datasets are considered and analysed: the TG curves directly obtained from the experimentation, the rescaled TG curves (used to find the true way of the epoxy matrix degradation within the composite, obtained by removing the mass at the end of each experiment, fumed silica, since the fumed silica is not degraded), and the derivatives with respect to the temperature of TG curves (DTG).

To investigate the amount of silica influencing on the thermal stability of the composites, 19 experiments are conducted: 7 with unloaded epoxy resin, 7 with 10 wt.% silica content, and 5 with 20 wt.% (see Fig. 1). Each experiment corresponds to a functional datum which represents the mass of the sample in functional relation with the temperature at which it is carried out. As already indicated, each sample is heated to a rate of 10 °C/min with temperatures ranging from 20 to 600 °C. At the end of each 600 °C trial, the organic phase (epoxy resin) is completely degraded, leaving only the added mass of fumed silica, which is much more heat resistant. It is important to note that the mass of the sample is represented in percentage, that is, the initial mass is assigned to the 100% value.

Originally, each curve consists of a variable number of points around 3480, one per second, depending on the environment temperature at which the testing machine is. To facilitate further calculations, without losing information, 581 points are chosen within each experimental curve, one for each Celsius degree, in a range of temperatures from 20 °C to 600 °C.

The first step in our study is to find a proper functional basis to write each curve as a linear combination of the elements of it, and to achieve a smooth functional relationship. Therefore, each functional datum is represented, discretized, by a finite basis, so that an explicit form for the function is obtained [12,13,47],

$$y(t) = X(t) + \varepsilon(t) = \sum_{k=1}^{n} c_k \phi_k(t) + \varepsilon(t), \tag{5}$$

where $\phi_k$, with $k \in \mathbb{N}$, is a set of known and independent functions, such that any function can be approximated by the linear combination of $K$ of these (elements of the basis). On the other hand, $\varepsilon(t)$ is the experimental error non-explained by the adjusted model.

Two procedures to select an appropriated basis are tested: *b*-splines and penalized *b*-splines. Both techniques provide bases with the required flexible structure. On the other hand, Fourier bases would not be as appropriate here as the data do not show an apparent periodic trajectory. Moreover, given the already smooth appearance of the original data, it seems reasonable not to test Wavelets bases.

Generally speaking, a *b*-spline is a spline (a function defined piecewise by polynomials) that has minimal support with respect to a given degree, smoothness, and domain partition. It can be written as a linear combination of normalized *b*-splines blending functions and some control points.

To avoid very variable fits, penalized *b*-splines can be used. The idea of this procedure is to fit a *b*-spline, but penalizing the variance of the fit represented by the second derivative of its density function. Specifically, the amount of residual fitting for a penalized *b*-spline basis respond to the expression

$$PRSS = \|Y - X\beta\|^2 + \lambda \int \left(f''(x)\right)^2 dx, \tag{6}$$

where $Y$ is the data to be fitted, $X\beta$ is the *b*-spline fitting, $f''(t)$ is the second derivative of the fit, and $\lambda$ the smoothing parameter (which penalizes the second derivative, that is, it restricts the internal variance the fit may have). More information can be found in Ferraty and Vieu's work [12], or in [13,48].

To choose the number of elements in the optimal basis, the Generalized Cross Validation (GCV) criterion is applied. Bases with number of elements ranging between 4 and 480 are tested. The number of elements corresponding to the minimum GCV for each functional datum is chosen. Finally, the number of elements in the basis is selected as the minimum of the minima computed for each one of the 19 GCV expressions (each one for each functional datum). This prevents against the risk of oversmoothing.

**Table 2**
Physical properties and chemical composition of fumed silica.

| | |
|---|---|
| Moisture 110 °C | 0.50% |
| Loss on ignition at 1000 °C | 2.78% |
| Real density | 2.26 g $cm^{-3}$ |
| Apparent density | 0.66 g $cm^{-3}$ |
| $SiO_2$ | + 95% |
| CaO | 0.68% |
| MgO | 0.22% |
| $Na_2O$ | 0.10% |
| $K_2O$ | 0.22% |
| Cl | 0.006% |
| $SO_4$ | 0.076% |

The result of minimizing the GCV criterion for one of the original curves can be seen in Fig. 2.

Table 3 shows the number of elements in the optimal basis according to the GCV criterion. It shows that for *b*-splines, the optimal GCV is smaller than that obtained for penalized *b*-splines, but the number of bases is too large, with the risk of interpolating the data. However, as observed in that table, an acceptable GCV is obtained with a basis of 80 elements. In that case, a smooth fit without departing from the path of the original data is obtained. A similar GCV is obtained in the case of penalized *b*-splines with a basis of 80 elements. Fig. 2 (right panel) corroborates these arguments. Additionally, in Fig. 2 (left panel), it can be observed that the GCV decreases sharply for a given number of elements in the basis. This supports the decision to opt for a smaller basis, corresponding to a number of elements closer to the beginning of this drop. If a smaller number of elements in the base were selected, an unacceptable model error would be possibly obtained. In that case, the fits would be far from experimental data in the steep slope changes, where it is very important that the data are faithfully reproduced by the fits. Therefore, a four order penalized *b*-spline basis with 80 elements is selected. The fit is perfect, very slightly smoother than that obtained with a *b*-splines basis. In addition, the small number of data does not cause a large computational cost.

The smoothed functional data using a penalized *b*-spline basis of 80 elements can be seen in Fig. 3.

**Table 3**
Number of elements in the optimal basis according to the GCV criterion.

| Basis | Optimal GCV | No of elements in the basis | GCV (with 80 elements in the basis) |
|---|---|---|---|
| *b*-splines | $2.0 \cdot 10^{-7}$ | 375 | $3.1 \cdot 10^{-3}$ |
| Penalized *b*-splines ($\lambda = 0.5$) | $3.1 \cdot 10^{-4}$ | 182 | $6.4 \cdot 10^{-3}$ |

### 4.1. Descriptive analysis of rescaled TG data

Each level of the factor appears in Fig. 3 relatively well differentiated from the others. The differences are especially found at high temperatures. The resin is degraded gradually. The resin without fumed silica is completely degraded while there is still around 10 and 20 wt.% of mass for groups with 10 to 20 wt.% of fumed silica.

At a first glance three steps in the curves can be observed. The first, not always perceptible, corresponds to the loss of volatiles and moisture (located at temperatures around 100 °C). The second step is singularly important, as it shows the main degradation process. When going from 0 to 10 wt.% of fumed silica content, the mass remaining at the beginning of this step is slightly larger at a given temperature (e.g. around 320 °C). This means that, apparently, the thermal stability of the resulting composite material increases. However, when moving from 10 to 20 wt.% of fumed silica, this increment is much lighter. The third step corresponds to the disappearance of the carbon residue resulting from the previous reaction. Here we can see that the differences are very evident among the different factor levels, mainly because the fumed silica is not degraded at these temperatures.

If we look at the original TG curves, it is clear that the addition of fumed silica to an epoxy resin leads to differences in the path of degradation. In turn, since the curves are riding on each other by increasing the amount of silica added, it is also evident that its thermal stability increases (a less quantity of mass is lost at the same temperature). This is mainly because a material that does not degrade in the temperature range tested is being added: fumed silica. Are there other reasons?

One might then ask oneself these questions: does fumed silica really interact with epoxy resin forming an interphase?, does the addition of inorganic phase influence in the way of degradation of the epoxy resin put into each sample?, and what is the best way to study the degradation of the neat resin (without taking into account the added inorganic matter)?

The answer lies in the rescaling of the data [8]. First, the mass at the end of the experiment is subtracted for calculating the TG curve only for the epoxy resin. This gives its way of degradation. This mass corresponds to the real mass of fumed silica added, that does not degrade itself. Then, each curve is rescaled, so that the initial value corresponded to 100% of the sample and the end to 0%, according to the degradation curve with epoxy resin. Once all the data obtained are rescaled, if significant differences in TG curves between factor levels are observed, a clue of the possible existence of a chemical interaction between resin and inorganic fillers will have been discovered.

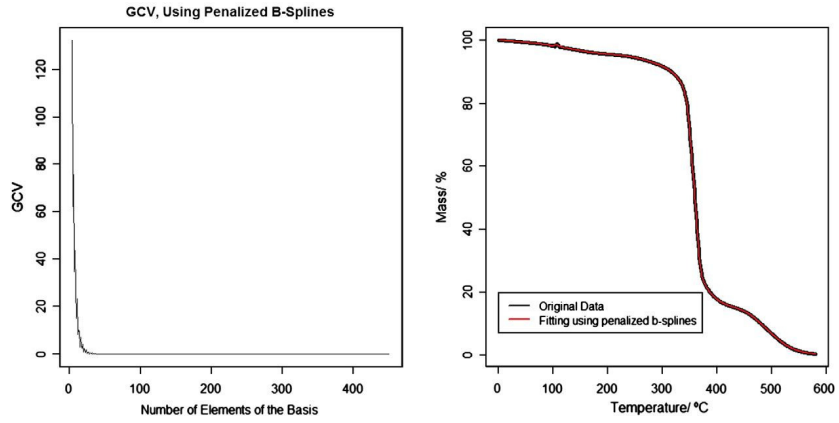


**Fig. 2.** Left panel: GCV versus the number of elements in a penalized *b*-spline basis, for a given functional data. Right panel: Experimental datum (epoxy resin) and fitting with a penalized *b*-spline basis with 80 elements.

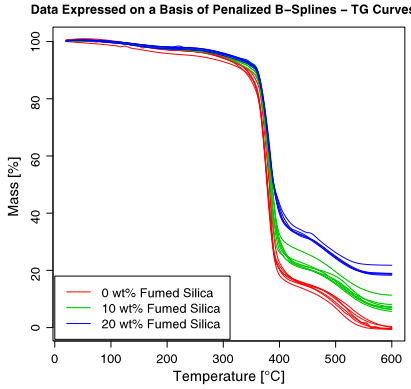


**Fig. 3.** Functional data smoothed with a penalized *b*-spline basis of 80 elements.

The curves of Fig. 4 represent the way of degradation of the epoxy resin (neat epoxy resin) in each sample of composite material. Eliminating mathematically the inorganic mass proportion (inert mass), it is noted that the curves corresponding to 10 and 20 wt.% levels are more similar to the way of degradation of the epoxy resin without silica. If it were possible to prove that at least one of the means of a group is different from the others, it would be possible to prove the existence of an organic–inorganic interphase. In fact, in [49], dynamic mechanical tests (DMA) performed on the same material already suggested the existence of this interphase. Therefore, the results of the present study could support the DMA ones.

For simplicity, a balanced design is intended to be performed. Therefore, two functional data for the 0 wt.% level and two others for the 10 wt.% level should be rejected. In fact, there are some slight differences in the experimental conditions for some experimental data. Therefore, it would be interesting to explore their depth.

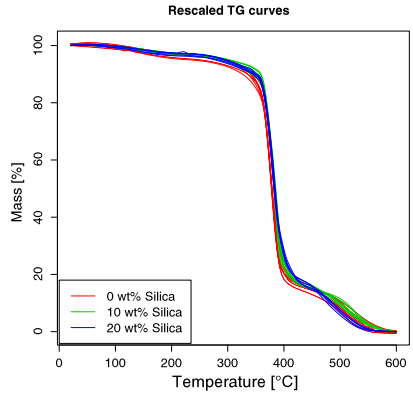


**Fig. 4.** Functional data: rescaled TG curves.

**Table 4**
Depths for the 7 samples of epoxy resin without fumed silica according to the 3 criteria. Rescaled data.

| Depth | Sam. 1 | Sam. 2 | Sam. 3 | Sam. 4 | Sam. 5 | Sam. 6 | Sam. 7 |
|---|---|---|---|---|---|---|---|
| Freiman–Muniz | 0.766 | 0.757 | 0.746 | 0.766 | 0.718 | 0.799 | **0.661** |
| Mode | 0.645 | **0.345** | 0.480 | 1.000 | 0.725 | 0.910 | **0.000** |
| Random projections | 0.247 | 0.233 | **0.215** | 0.296 | 0.251 | 0.262 | **0.210** |

Depth is a concept that explains how central a datum is with respect to a set of points belonging to a population. Following this criterion, different functional data, from a sample of a given population, can be sorted: deeper curves are identified as closer to the center (the deepest datum is defined as the median) [47].

The depths given by Fraiman and Muniz (FM) [50], called the Median depth, the Mode depth (the deepest point would be equivalent to the mode of the data) [47], and the depth of Random Projections (RP) [51] are calculated for the groups of 0 and 10 wt.%, separately. We are interested in reducing the number of data just for these groups. The results appear in Tables 4 and 5.

The curves chosen as less deep are the sample 7 (for the 0 wt.% in silica group) and the sample 8 (for 10 wt.%). In the case of sample 7, the three criteria coincide, while for the sample 8, there is an overlap in the criteria RP and Mode, being the second less deep datum according to FM criterion. Note the numbers in bold in the tables.

It is still necessary to eliminate another curve in these two groups. Sample 3 is chosen in the 0 wt.% group. This curve is the less deep datum according to the RP criterion and the third less deep according to the Mode and FM criteria. In the case of 10 wt.%, RP and Mode criteria select sample 13. So, this time, sample 13 is selected. It is worth noting that the curves identified as less deep correspond to that obtained with slightly different experimental conditions (a smaller quantity of experimental mass, not homogeneous dispersion of the load in the matrix and, therefore, slightly different values in 10 wt.% curves). Then, those data selected as less deep, that is, samples 3 and 7 for 0 wt.%, and samples 8 and 13 for 10 wt.% of fumed silica, are finally removed.

With 15 functional data, 5 per level, mean and median from the functional data are calculated. Fig. 5 shows the mean of each group (0, 10 and 20 wt.%) jointly with the confidence bands obtained using the bootstrap naive approach. In Fig. 6, the medians and the corresponding confidence bands, calculated using a smoothed bootstrap method with a smoothing parameter $h = 0.07$ are presented. Smoothed bootstrap is mainly used in cases where the simple bootstrap does not provide a clear picture of what may be the confidence interval, requiring the addition of an additional random component (to fill gaps, obtaining confidence bands). The mean is given by:

$$\overline{X}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t), \tag{7}$$

and the medians are computed using the depth concept of [50].

Figs. 5 and 6 show that there is a greater overall variability in the data with the epoxy resin alone. The variability decreases slightly for

**Table 5**
Depths for the 7 samples of epoxy resin and 10 wt.% silica according to the 3 criteria. Rescaled data.

| Depth | Sam. 8 | Sam. 9 | Sam. 10 | Sam. 11 | Sam. 12 | Sam. 13 | Sam. 14 |
|---|---|---|---|---|---|---|---|
| Freiman–Muniz | **0.701** | **0.665** | 0.760 | 0.704 | 0.747 | **0.761** | 0.873 |
| Mode | **0.000** | 0.249 | 0.301 | 0.562 | 0.687 | **0.070** | 1.000 |
| Random Projections | **0.202** | 0.242 | 0.239 | 0.258 | 0.277 | **0.202** | 0.294 |

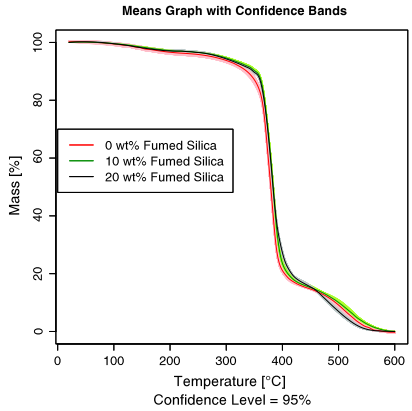


**Fig. 5.** Means of each group of data, with confidence bands developed using bootstrap.

the 10 wt.% group and, finally, is much smaller for 20 wt.% (this can be seen in the confidence bands of mean and median). This trend may be due to the heterogeneity of the samples or to the learning effect of the operator. In fact, the first samples tested correspond to 0 wt.%, and the last ones to 20 wt.% (those with less variability). In the case of 0 wt.%, another possible reason to take into account is the testing of samples with different moisture contents.

Comparing the statistics (mean and median) of the rescaled curves with those of the original data, one can see that the differences between the curves are substantially lower in the rescaled case. Still, in Figs. 4 and 5, different ways are observed according to the group, with very little dispersion. The differences are especially conspicuous between the groups 0 and 10 wt.%, or the 20 wt.% class. On the other hand, the differences are slighter between 10 wt.% and 20 wt.% levels; they seem mainly differ in the last step of degradation. In the region of the second step, the most important, the mean of the 10 and 20 wt.% groups are significantly above the mean of the 0 wt.% group. As a result, the thermal stability increases when going from a level of 0 wt.% fumed silica to any other. The difference is especially noticeable at the top of the second step, which is the area where the thermal stability of a material is commonly evaluated. However, smaller differences are observed between the degradation pathways for 10 and 20 wt.% groups. On the other hand, some differences are observed in the third step: the 20 wt.% curves start located at a higher level to then fall much more sharply than those of other groups. Definitely, the addition of 20 wt.% of fumed silica causes a decrease in thermal stability in the last stage of the third step of degradation. In this last stage only a char residue exists. The limits of the confidence bands (especially for the means, which are those to be used to build the statistic of functional ANOVA) could suggest that these findings, done from a descriptive analysis of the data, have statistical significance. This is shown in Section 5.

### 4.2. Descriptive analysis of DTG curves

The study of the Differential Thermogravimetric Analysis (DTG) curves is very common in thermal analysis. Through the study of the derivative, apart from supporting the previous results, we intend to complete the study by tackling an important concept: the degradation rate of the samples. Does the amount of fumed silica significantly affect the degradation rate of the epoxy? To answer this question, the derivatives of the rescaled TG data using the statistical package R [52] are calculated (Fig. 7). Then, each data is fitted using a penalized splines basis consisting of 60 elements, obtaining a GCV = 0.006 and a spar coefficient equal to 0.35. Spar is the integrated squared derivative of order 2 which controls the amount of smoothing. It is related with lambda parameter. The optimum is obtained using a basis of 124 elements, giving $GCV = 9.32 \cdot 10^{-6}$, but due to the same reason as in the rescaled data, a smaller basis is chosen. The parameters are selected in order to smooth the data conveniently in the derivative, with much more noise than that in the original data.

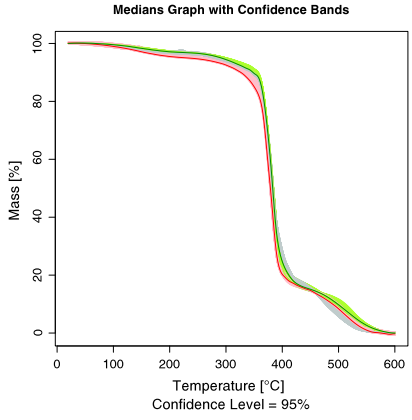


**Fig. 6.** Medians of each group of data, with confidence bands developed using smoothed bootstrap.

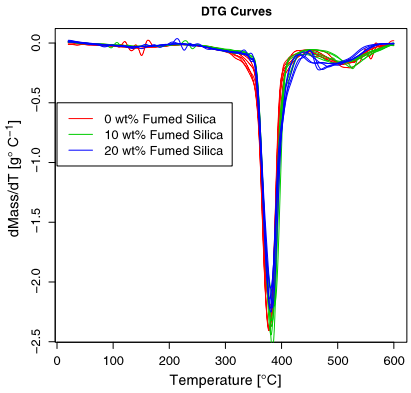


**Fig. 7.** DTG curves for the different groups.

In the case of the DTG curves, the semi-metric given by:

$$d\left(X_i, X_j\right) = \sqrt{\int_a^b \left(X_i'(t) - X_j'(t)\right)^2 dt}$$

is used.

The mean of the DTG curves are plotted in Fig. 8.

As in the case of not derived rescaled data, differences in mean between groups seem to be small.

The main differences between groups are observed at the beginning and the end of the main degradation process (Figs. 7 and 8). At the beginning of the main degradation process, the degradation rate in 0 wt.% group is significantly greater than in groups of 20 and 10 wt.%. The DTG mean curves of these two groups are over the first one. Nevertheless, the differences between the degradation rate in 10 and in 20 wt.% groups are smaller. On the other hand, at the end of the main degradation process, the degradation rate in 0 wt.% group is slightly lower than in 20 and 10 wt.% levels. However, the differences are slighter between 10 wt.% and 20 wt.% groups. Additionally, at the beginning of the last degradation process, the degradation rate in the 20 wt.% group is significantly greater than in other groups. The addition of 20 wt.% of fumed silica accelerates the degradation of the epoxy resin at very high temperatures. This is consistent with the conclusions obtained using the rescaled data without deriving. It may be due to the confinement effect of silica agglomerates on the resin, or even to the fact of better heat transmission within the sample through thinner layers of resin. Finally, the temperature at which the maximum mass loss rate occurs, hardly varies. Although it decreases slightly in module for the group of 20 wt.%.

In Fig. 9, the functional variance in each one of the three groups of both the TG and DTG curves are shown. It can be observed that they are very small throughout the range, compared with the values of the functional means (Figs. 5 and 8). Therefore, the variability per group is properly represented with just 5 curves per level, optimizing the total experimental time required.

## 5. Results and discussion

Once the data to be analysed were described in the previous Section, now, the statistical methods presented in Section 2 are applied to those data in order to measure statistically the influence of adding fumed silica on the thermal degradation of an epoxy resin. The results obtained also give statistical significance to some of the conclusions derived in the previous descriptive analysis.

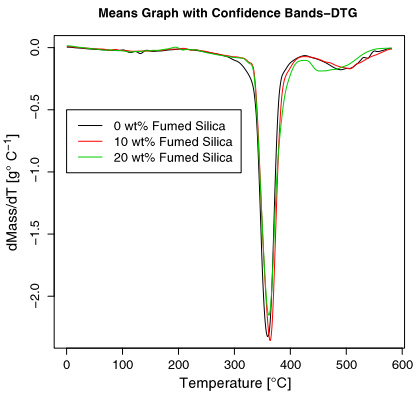


**Fig. 8.** DTG mean curves for each group.

The FANOVA test described in Subsection 2.2 is applied to the rescaled TG curves and also to the DTG curves. The null hypothesis to be tested is

$$H_0 : m_1 = m_2 = m_3,$$

where $m_i$ is the mean of the functional data within each of the three levels studied belonging to the factor *amount of fumed silica*.

First, the statistic $V_n$ given in Eq. (3) is calculated. The number of replicates for each one of the three levels (0, 10 and 20 wt.% of fumed silica), $n_i$, is equal to 5 (balanced design). As explained in Section 2.2, under the null hypothesis, the asymptotic distribution of $V_n$ coincides with that of the statistic $V$, given by Eq. (4). The distribution of $V$, under $H_0$, can be approximated by using a parametric bootstrap and Monte Carlo with the following steps:

1. The variance–covariance matrix (assuming heteroscedasticity) for each level is estimated, computing $K_i(s,t)$ in Eq. (1), where $s$ and $t$ are two given moments within each functional data.
2. Starting from the original sample, matrices $K_i(t_p, t_q)_{1\le p,q\le m}$ are computed. Next, $B=2000$ bootstrap resamples by level, following a normal distribution with zero mean and covariance matrix given by $K_i(t_p, t_q)_{1\le p,q\le m}$, are generated.
3. Therefore, 2000 values, $Z_{il}^* = (Z_{il}^*(t_1),\ldots,Z_{il}^*(t_m))$, by level are obtained, with $l=1,\ldots,2000$ and $i=1,2,3$. These values approximate the $Z_i(t)$ continuous paths or trajectories with discrete versions, evaluated in a grid $a\le t_1,\ldots,t_m\le b$.
4. Finally, 2000 values of the expression

$$\hat{V}_l = \sum_{i<j} \left\| Z_{il}^*(t) - C_{ij}\cdot Z_{jl}^*(t) \right\|^2.$$

   are calculated. These values approximate the distribution of $V$ when $H_0$ is true.
5. Then, it is possible to estimate the $\alpha$-quantile (denoted by $V_\alpha$) such that, $P(V>V_\alpha)=\alpha$, under $H_0$, for any $\alpha$.
6. If $V_n>V_\alpha$, the test is significant, and at least one of the functional means is different from the others.

The result of the application of this procedure in the case of the rescaled TG curves is the following: $V_n = 34470.8$ and $V_\alpha = 1235.671$. Therefore, $V_n \gg V_\alpha$. The test is highly significant. At least one of the functional means is different from the others. This agrees with the arguments explained in the descriptive analysis of the data (Section 4.1). The addition of silica causes changes in the functional means of the rescaled TG curves. The thermal stability of epoxy resin, which forms part of the composite material, undergoes a highly significant statistical increment with the addition of an increased amount of fumed silica (it supports higher temperatures before degrading). It could be said that its particular path of degradation has a different shape with the addition of fumed silica, at least for one level. This was the main aim of our research. This is the indicator of an interaction between the epoxy resin and fumed silica, which could result in the creation of a matrix–filler interphase. This result is supported by the DMA tests performed with the same material [49]. At least one level is different to the others, but what levels are really different?

However, once it is statistically proved that these differences exist, the next step of our research is trying to find out what groups are really different. This kind of analysis would correspond to the named *post-hoc analysis* in the context of design of experiments. To our knowledge, there are no specific procedures for this task in the setting of functional data. Nevertheless, since only three groups are considered here, three pairwise comparisons, using the same functional ANOVA method, can be used as a first attempt to tackle this problem. To correct the problem

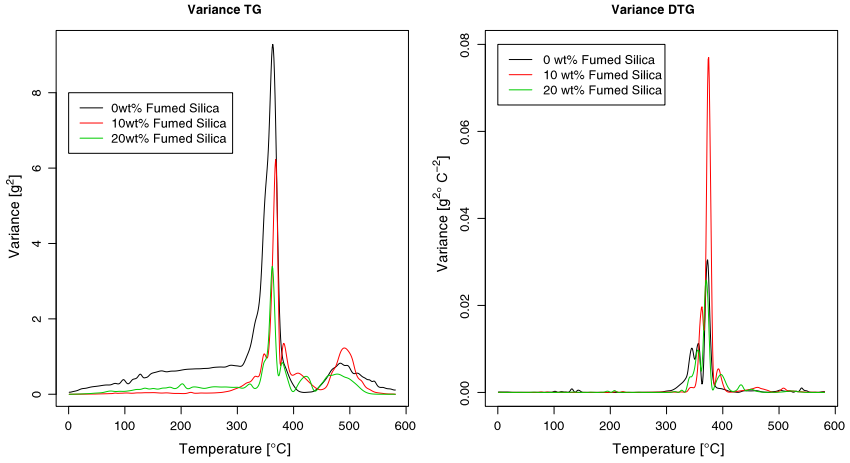


**Fig. 9.** Functional variance curves.

of multiple testing, a Bonferroni correction is used [53]. The idea behind this approach is to consider a new significance level, $\alpha_{\text{Bonf}} = \frac{\alpha}{J}$, $J$ being the number of groups to be compared ($J=3$, in our case), and compute individual tests using this new level. Table 6 shows the results of all pairwise comparisons with the functional ANOVA test using $\alpha=0.05$ and $\alpha_{\text{Bonf}} = 0.05/3 \approx 0.015$. As it can be observed, according to these tests, the three groups are significant different.

It is possible that the previous approach can suffer of lack of power, due to the small sample size in each one of the applications of the test (just 10 curves). To solve this drawback, additionally to the previous proposal, we perform a Tukey-like test specially designed for this situation of functional data. Following analogous ideas as those used in the Tukey test, the method proposed consists in computing the differences of pairwise curve means, using the $L_2$ distance, and comparing these values with the corresponding quantile of the distribution of the maximum distance between the means of the groups, under the null hypothesis of no difference between groups. The distribution under the null hypothesis is approximated using a bootstrap procedure. Assuming that there are no differences between groups, resamples (with replacement) from the original 15 curves are obtained. These are divided in three subsets and the maximum of the distance between the pairwise curve means of these groups is obtained. This process is repeated a large number $B$ ($B=1000$ in our case) of times and the corresponding quantile is computed. The result of the application of this procedure to our data with $\alpha=0.05$ is shown in Table 7.

The previous naive bootstrap test could be affected by possible outliers due to the small sample size, so that, we also perform a parametric bootstrap. The only difference with respect to the naive method is in the resampling process. Under the null hypothesis, the resamples are now obtained as the sum of the mean curve (computed with 15 original curves) and an error curve generated from a Gaussian distribution with zero mean and variance–covariance matrix estimated form the original 15 curves. In this case, the curves as treated as a 581-dimensional vector (the number of values where the curves are evaluated) in order to generate new curves. The quantiles obtained with this parametric bootstrap approach are slightly different to those shown in Table 7, but the conclusion is the same: the three groups are statistically different.

In addition, other competitive models like unfunctional ANOVA models based on discretized curves (MANOVA test) and the pointwise $F$ test presented in [13] were applied to the TG and DTG curves. In the first case, a MANOVA test (Wilks) was performed. For this, the curves were discretized just using 12 points per curve, because the sample size here is 15. Depending of the election of these points, the results of the test were different. Moreover, the relatively small sample size can make that important information could get lost in the process of discretization. In this sense, our test takes the full functional nature of the curves into account and more reliable results are obtained. On other hand, a pointwise $F$ test was applied, obtaining similar results to the functional test performed in the present paper.

In the case of the DTG curves, the application of the FANOVA test gives as results: $V_n=57.37$ and $V_\alpha=3.58$. Therefore $V_n \gg V_\alpha$. The resulting test is highly significant. At least one of the functional means is different from the others. Therefore, the arguments explained in the descriptive analysis have a statistically significance. The addition of silica causes changes in the functional means of the derivatives of the rescaled TG curves. The thermal stability of epoxy resin experienced a statistically significant change with an increasing amount of fumed silica. It can be said that the rate of degradation really experiences the

**Table 6**
Pairwise comparisons using the functional ANOVA test with TG curves.

| Groups compared | $V_{0.05}$ | $V_{0.015}$ | $V_n$ | Result |
|---|---|---|---|---|
| 0 wt.% − 10 wt.% | 1506.59 | 2252.66 | 13654.93 | Significant |
| 0 wt.% − 20 wt.% | 1439.34 | 2135.89 | 16199.65 | Significant |
| 10 wt.% − 20 wt.% | 774.42 | 1121.88 | 4616.23 | Significant |

**Table 7**
Pairwise comparisons using a Tukey-like bootstrap test with TG curves.

| Groups compared | Mean distance | Quantile | Result |
|---|---|---|---|
| 0 wt.% − 10 wt.% | 0.058 | 0.038 | Significant |
| 0 wt.% − 20 wt.% | 0.103 | 0.038 | Significant |
| 10 wt.% − 20 wt.% | 0.045 | 0.038 | Significant |

changes indicated in the descriptive study, at least for a level. This result is also an indicator of some interaction between the epoxy resin and fumed silica, which could result in the creation of a matrix–filler interphase, like in the rescaled data.

Following the same steps as before, the functional ANOVA pairwise comparisons are applied to distinguish which groups (0, 10 and 20 wt.%) are really different, that is, what levels of silica provide different degradation rates (Table 8). Additionally, the previously proposed Tukey based test is performed to complete the study (Table 9).

Observing the results, we could say that the degradation rate for each group is significantly different from the others.

## 6. Conclusions

In this paper, a functional ANOVA (FANOVA) test is used to evaluate the effect of adding fumed silica on thermal degradation of an epoxy resin. Three different levels (0, 10 and 20 wt.%) of amount of fumed silica are considered. The procedure is applied to rescaled TG curves and the derivatives of the rescaled TG. Previously to the application of the test, some data preprocessing is done to prepare the datasets to be analysed. Statistical techniques like penalized *b*-splines or the concept of depth are used for this purpose.

All tests performed have resulted highly significant. Therefore, it can be said that the addition of fumed silica affects the way of degradation of epoxy resin involved in the sample, at least in one group. Moreover, by performing pairwise comparisons using the functional ANOVA method and a bootstrap distance based test, it can be said that the way of degradation for each group is significantly different from the others.

As one moves from 0 to 10 wt.% or 20 wt.% of fumed silica, the thermal stability of pure epoxy resin increases at temperatures corresponding to the second degradation step. This is an indication of the interaction between the organic and inorganic phases, with statistical significance.

Regarding the thermal degradation rate, it can be said that this is different with the addition of fumed silica, by applying functional ANOVA and pairwise comparisons. The mass loss rate decreases at the beginning of the main degradation process. The temperature at which the maximum mass loss rate occurs hardly changes, independently on the filler content, but the maximum mass loss rate module decreases with the addition of 20 wt.% of fumed silica.

In the groups of 10 and 20 wt.% of fumed silica, the rescaled curves and their derivatives are very similar. They mainly differ in the last degradation process, but significant differences in the second degradation step are also appreciated. The addition of 20 wt.% in silica can cause a decrease in thermal stability at very high temperatures, because of the morphology of the material itself.

It would be interesting, in a future research, studying some extensions of this work, analysing the possible influence of more factors (for example, the particle size distribution) in the response variable.

**Table 8**
Pairwise comparisons using the functional ANOVA test with DTG curves.

| Groups compared | $V_{0.05}$ | $V_{0.015}$ | $V_n$ | Result |
|---|---|---|---|---|
| 0 wt.% − 10 wt.% | 4.135 | 7.051 | 25.348 | Significant |
| 0 wt.% − 20 wt.% | 2.887 | 4.769 | 21.78 | Significant |
| 10 wt.% − 20 wt.% | 3.752 | 5.763 | 10.24 | Significant |

**Table 9**
Pairwise comparisons using a Tukey-like bootstrap test with DTG curves.

| Groups compared | Mean distance | Quantile | Result |
|---|---|---|---|
| 0 wt.% − 10 wt.% | 0.0017 | 0.0016 | Significant |
| 0 wt.% − 20 wt.% | 0.0044 | 0.0016 | Significant |
| 10 wt.% − 20 wt.% | 0.0029 | 0.0016 | Significant |

## Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 (ERDF included) and by Xunta de Galicia, Grant PGIDIT07PXIB105259PR. The authors wish to express special thanks to Manuel Febrero Bande and Aldana González Montoro, for their valuable comments. In addition, the authors thank two anonymous referees for constructive comments that improved the presentation of this article.

## References

[1] M. Harsch, J. Karger-Kocsis, M. Holst, Influence of fillers and additives on the cure kinetics of an epoxy/anhydride resin, Eur. Polym. J. 43 (2007) 1168–1178.
[2] A. Lee, J.D. Lichtenhan, Thermal and viscoelastic property of epoxy-clay and hybrid inorganic–organic epoxy nanocomposites, J. Appl. Polym. Sci. 73 (1999) 1993–2001.
[3] S. Mehta, F.M. Mirabella, K. Rufener, A. Bafna, Thermoplastic olefin/clay nanocomposites: morphology and mechanical properties, J. Appl. Polym. Sci. 92 (2004) 928–936.
[4] F. Shao-Yun, F. Xi-Qiao, B. Lauke, M. Yiu-Wing, Effects of particle size, particle/matrix interface adhesion and particle loading on mechanical properties of particulate-polymer composites, Compos. Pt. B Eng. 39 (2008) 933–961.
[5] H. Zhang, Z. Zhang, K. Friedrich, C. Eger, Property improvements of in situ epoxy nanocomposites with reduced interparticle distance at high nanosilica content, Acta Mater. 54 (2006) 1833–1842.
[6] M. Preghenella, A. Pegoretti, C. Migliaresi, Thermo-mechanical characterization of fumed silica-epoxy nanocomposites, Polymer 46 (2005) 12065–12072.
[7] A. Yousefi, P.G. Lafleur, R. Gauvin, Kinetic studies of thermoset cure reactions: a review, Polym. Compos. 18 (1997) 157–168.
[8] J. Tarrio-Saavedra, J. López-Beceiro, S. Naya, R. Artiaga, Effect of silica content on thermal stability of fumed silica/epoxy composites, Polym. Degrad. Stab. 93 (2008) 2133–2137.
[9] A. Mohammad, G.P. Simon, Rubber–clay nanocomposites, in: M. Yiu-Wing, Y. Zhong-Zhen (Eds.), Polymer Nanocomposites, Woodhead Publishing Limited, 2006.
[10] L.S. Schadler, Polymer-based and polymer-filled nanocomposites, in: P.M. Ajayan, L.S. Schadler, P.V. Braun (Eds.), Nanocomposite Science and Technology, Wiley-VCH, Weinheim, 2003, pp. 77–135.
[11] R.B. Prime, Thermosets, in: E. Turi (Ed.), Thermal Characterization of Polymeric Materials, Second edition, Academic Press, San Diego, 1997, pp. 1380–1766.
[12] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis: Theory and Practice, Springer-Verlag, New York, 2006.
[13] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer-Verlag, New York, 2005.
[14] S. Naya, Nuevas aplicaciones de la estimación paramétrica y no paramétrica de curvas al análisis térmico, Ph.D. thesis, University of A Coruña, Spain, 2003.
[15] W. González Manteiga, P. Vieu, Statistics for functional data, Comput. Stat. Data Anal. 51 (2007) 4788–4792.
[16] F. Ferraty, Y. Romain, The Oxford Handbook of Functional Data Analysis, Oxford University Press, 2010.
[17] M.J. Valderrama, An overview to modelling functional data, Comput. Stat. Data Anal. 22 (2007) 331–334.
[18] F. Ferraty, High-dimensional data: a fascinating statistical challenge, J. Multi. Anal. 101 (2010) 305–306.
[19] A.A. Tsiatis, M. Davidian, Joint modeling of longitudinal and time-to-event data: an overview, Stat. Sin. 14 (2004) 793–818.
[20] S. López-Pintado, R. Romo, Depth-based inference for functional data, Comput. Stat. Data Anal. 51 (2007) 4957–4968.
[21] E. del Barrio, J. Cuesta-Albertos, R. Fraiman, C. Matran, The random projection method for goodness of fit for functional data, Comput. Stat. Data Anal. 51 (2007) 4814–4831.
[22] A. Antoniadis, T. Sapatinas, Estimation and inference in functional mixed effects models, Comput. Stat. Data Anal. 51 (2007) 4793–4813.
[23] F. Ferraty, P. Vieu, S. Viguier-Pla, Factor based comparison of groups of curves, Comput. Stat. Data Anal. 51 (2007) 4903–4910.
[24] D. Nerini, B. Ghattas, Classifying densities using functional regression trees: application in oceanology, Comput. Stat. Data Anal. 51 (2007) 4984–4993.
[25] N. Locantore, J.S. Marron, D.G. Simpson, N. Tripoli, J.T. Zhang, K.L. Cohen, Robust principal component analysis for functional data (with discussion), Test 8 (1999) 1–74.
[26] H. Cardot, F. Ferraty, P. Sarda, Functional linear model, Stat. Probab. Lett. 45 (1999) 11–22.
[27] A. Cuevas, M. Febrero, R. Fraiman, Linear functional regression: the case of fixed design and functional response, Can. J. Stat. 30 (2002) 285–300.
[28] F. Ferraty, P. Vieu, The functional nonparametric model and application to spectrometric data, Comput. Stat. 17 (2002) 545–564.
[29] H. Cardot, C. Crambes, A. Kneip, P. Sarda, Smoothing spline estimators in functional linear regression with errors-in-variables, Comput. Stat. Data Anal. 51 (2007) 4832–4848.
[30] S. Dabo-Niang, N. Rhomari, Kernel regression estimate in a banach space, J. Stat. Plan. Infer. 131 (2009) 1421–1434.
[31] F. Ferraty, P. Vieu, Curves discrimination: a nonparametric functional approach, Comput. Stat. 44 (2003) 161–173.

[32] F.L. Bookstein, Morphometric Tools for Landmark Data: Geometry and Biology, Cambridge University Press, Cambridge, United Kingdom, 1991.
[33] H. Salkoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Trans. Acoust. 26 (1978) 43–49.
[34] A. Kneip, T. Gasser, Convergence and consistency results for self-modeling nonlinear regression, Ann. Stat. 16 (1988) 82–112.
[35] B.W. Silverman, Incorporating parametric effects into functional principal components analysis, J. R. Stat. Soc. Ser. B 57 (1995) 673–689.
[36] J.O. Ramsay, X. Li, Curve registration, J. R. Stat. Soc. Ser. B 60 (1998) 351–363.
[37] J.O. Ramsay, Estimating smooth monotone functions, J. R. Stat. Soc. Ser. B 60 (1998) 365–375.
[38] A. Kneip, X. Li, K.B. MacGibbon, J.O. Ramsay, Curve registration by local regression, Can. J. Stat. 28 (2000) 19–29.
[39] A. Cuevas, M. Febrero, R. Fraiman, An anova test for functional data, Comput. Stat. Data Anal. 47 (2004) 111–122.
[40] J.O. Ramsay, G. Hooker, G. S., Functional Data Analysis with R and Matlab, Springer, New York, 2009.
[41] J. Fan, S.K. Lin, Functional anova models for proportional hazards regression, J. Amer. Stat. Assoc. 93 (1998) 1007–1021.
[42] Q. Shen, J.J. Faraway, An *F* test for linear models with functional responses, Stat. Sin. 14 (2004) 1239–1257.
[43] B.A. Brumback, J.A. Rice, Smoothing spline models for the analysis of nested and crossed samples of curves, J. Amer. Stat. Assoc. 93 (1998) 961–994.
[44] C.J. Stone, M.H. Hansen, C. Kooperberg, Y.K. Truong, Polynomial splines and their tensor products in extend linear modeling, Ann. Stat. 25 (1997) 1371–1470.
[45] J. Huang, C. Kooperberg, C.J. Stone, Y.K. Truong, Functional anova models for proportional hazards regression, Ann. Stat. 28 (2000) 961–999.
[46] A. Kawaguchi, K. Yonemoto, Y. Tanizaki, Y. Kiyohara, T. Yanagawa, Y.K. Truong, Application of functional anova models for hazard regression to the Hisayama data, Stat. Med. 27 (2008) 3515–3527.
[47] A. Cuevas, M. Febrero, R. Fraiman, On the use of the bootstrap for estimating functions with functional data, Comput. Stat. Data Anal. 51 (2006) 1063–1074.
[48] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis, Springer-Verlag, New York, 2002.
[49] J. Tarrío-Saavedra, J. López-Beceiro, S. Naya, C. Gracia, R. Artiaga, Controversial effects of fumed silica on the curing and thermomechanical properties of epoxy composites, Express Polym. Lett. 4 (2010) 382–395.
[50] R. Fraiman, G. Muniz, Trimmed means for functional data, Test 10 (2001) 419–440.
[51] A. Cuevas, M. Febrero, R. Fraiman, Robust estimation and classification for functional data via projection-based depth notions, Comput. Stat. 22 (2006) 481–496.
[52] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org 2008.
[53] R.G. Miller, Simultaneous Statistical Inference, Springer-Verlag, New York, USA, 1991.

# Apéndice B

# A comprehensive classification of wood from thermogravimetric curves

ARTICLE IN PRESS
CHEMOM-02540; No of Pages 14
Chemometrics and Intelligent Laboratory Systems xxx (2012) xxx–xxx
Contents lists available at SciVerse ScienceDirect
Chemometrics and Intelligent Laboratory Systems
journal homepage: www.elsevier.com/locate/chemolab

# A comprehensive classification of wood from thermogravimetric curves

Mario Francisco-Fernández [a,*], Javier Tarrío-Saavedra [b], Abhirup Mallik [c], Salvador Naya [b]

[a] University of A Coruña. Faculty of Computer Science, Campus de Elviña, s/n, A Coruña 15071, Spain
[b] University of A Coruña. Higher Polytechnic School. Campus de Esteiro, Ferrol 15403, Spain
[c] Indian Institute of Technology. Department of Mathematics, Kharagpur 721302, India

ARTICLE INFO

*Article history:*
Received 16 March 2012
Received in revised form 5 June 2012
Accepted 4 July 2012
Available online xxxx

*Jel classification:*
2000 MSC: 62H30
2000 MSC: 62G99
MSC 62P30

*Keywords:*
Wood
Thermal analysis
Functional data analysis
Supervised

ABSTRACT

Wood is one of the most complicated materials to be classified in different classes or species. In this paper, the thermogravimetric (TG) curves of 49 wood samples are used to classify them in 7 predetermined species. Different functional and multivariate statistical supervised classification methods are used for this task: a nonparametric Nadaraya–Watson kernel functional estimator (K-NFDA), using the complete TG curves, and multivariate supervised classification approaches, such as linear discriminant analysis (LDA), *k* Nearest Neighbors (*k*-NN), Naïve Bayes (NBC), Neural Networks (NN), and Support Vector Machines (SVM). Before applying the multivariate techniques, the TG curves are discretized using principal component analysis (PCA) or fitting a four-component generalized logistic model. The results show that the classical method of LDA using the logistic parameters had the best performance, although high correct classification percentages were also obtained with the rest of the approaches. The work is completed with a comprehensive simulation study, comparing the classification techniques in different scenarios. Artificial TG curves are generated using the logistic model and additional conclusions on wood classification are established. Due to the heterogeneity of wood, this simulation study is very useful to describe worst-case scenarios and to assess more accurately the proposed classification methodologies.



## 1. Introduction

The discrimination of a specific material in different classes represents an important practical problem with direct industrial applications. When the material under consideration is wood, the correct classification into different species becomes an extremely difficult task. The reasons are complex compositions of this material, the wide variety of existing species and the anatomical heterogeneity of its elements. Depending on wood species, a timber presents different physical–chemical properties determining its industrial applications and price. Wood identification is often made on the basis of readily visible characteristics such as color, odor, density, presence of pitch, or grain pattern which may result error arising from human-bias. To get a more accurate classification, in case of particular difficulties, it is essential to use microscopy techniques, physical hardness tests and chemical analyses [1–3]. These kinds of analyses are useful in the furniture industry, the wood panel production, or even in archeology, where it is crucial to know the kind of wood used to combat fraud [4–7]. Therefore, the implementation of quantitative models and automatic wood sample recognition methods are justified and can be of immediate application in these fields.

In this paper, the thermogravimetric (TG) curves of different wood samples are used to classify them in different species. The TG curves explain the mass loss when the temperature is increased. They are the result of applying a common technique, called thermogravimetry, belonging to the thermal analysis of materials [8]. Different multivariate techniques have been applied to thermograms [9–11]. However, to our knowledge, the present research is the first of its kind, where different classification methods are evaluated and compared using experimental and simulated TG curves.

Fig. 1 shows the 49 TG curves of the wood samples (7 samples per class of wood) used in our analysis (see Section 3.1 for a description on the wood species and their corresponding TG curves employed in this research). A particular wood species is highlighted in each panel. The last panel (in row 4, column 2) shows all the TG curves.

The functional nature of these curves suggests the use of functional supervised classification methods for this task. On the other hand, if the curves are properly discretized, multivariate classification techniques can also be applied here. Different classification approaches in these two settings are compared in this work. Specifically, classical multivariate supervised classification methods, such as linear discriminant analysis (LDA), some machine learning techniques (including Naïve Bayes (NBC), *k* Nearest Neighbors (*k*-NN), Neural Networks (NN) or Support Vector Machines (SVM)), and the nonparametric Nadaraya–Watson kernel functional (K-NFDA) method are used and compared. Regarding the multivariate methods, an important problem arising is

* Corresponding author. Tel.: +34 981167000x1222; fax: +34 981167160.
*E-mail address:* mariofr@udc.es (M. Francisco-Fernández).

0169-7439/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2012.07.003

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

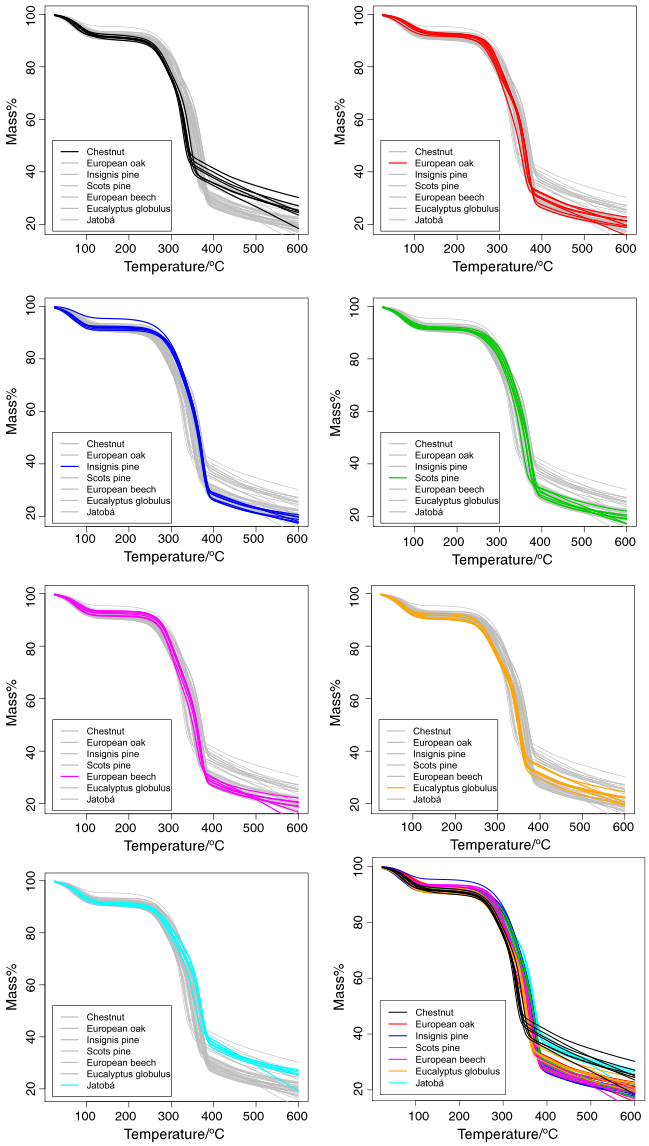


**Fig. 1.** TG curves of the wood samples (7 per class) used in the analysis. Each particular wood species is highlighted in the corresponding panel. In the last panel (row 4, column 2) all the curves are presented.

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

how to discretize the curves. In other words, what features are actually representative of each obtained TG curve? In the present study, two approaches are followed. On the one hand, using principal component analysis (PCA), some components of the TG curves which are together explaining most of the variation of the data are selected. We set a cut-off for the percentage of variation which we wish to achieve. The rest of the components of the curves are then neglected. Note that, in practice, many traditional methods including discriminant analysis cannot be applied directly when the number of components exceeds the number of observations and, therefore, this fact must be taken into account in the process of selecting the principal components. On the other hand, a new nonlinear regression model is proposed to fit the TG curves and to extract some representative features from them. This model can be written as the sum of four generalized logistic components, one per principal constituent of wood (hemicellulose, cellulose and lignin) and one component corresponding to the water involved. The parameters obtained from the fit of each TG curve (4 for each component, 16 in total) are used as a vector of features, ensuring the representativity of them.

In our research, tests for seven different wood species and seven samples per class are performed. This can be considered a moderate sample size and, while important conclusions can be deduced from this analysis, a more comprehensive study to compare the different classification methods would require of samples collected in different scenarios. Nevertheless, this is a problem from a practical point of view. The cost of each thermogravimetric test is not negligible, both in time and money. The time spent on sample preparation must be added to the duration of each test. In addition, while performing the tests, the device is occupied, being impossible its use in other applications. Moreover, due to the extreme heterogeneity of wood (there are differences in wood according to the tree from which it is extracted, or even between different parts of the same tree), it is very difficult to get a fully representative sample. For this reason, apart from using the real wood samples, a complete simulation study to compare the different approaches is carried out. The TG curves are mimicked using the previously mentioned generalized logistic model. Artificial TG curves are simulated using the parameters obtained from fitting the generalized logistic mixture model to the real curves, considering different covariance matrices and sample sizes. This represents a general proposal that can be applied not only in the context of the present paper. Representative samples in less time and in a variety of scenarios can be then produced. An alternative way to simulate TG curves, based on the Arrhenius model, was presented in [12]. They focused on the thermal behavior of poly(methl methacrylate) (PMMA) and described an algorithm to simulate the overall weight loss of PMMA in any given experimental condition.

Some statistical classification methods have been previously compared in different papers. For example, in [13], the K-NFDA method (also used in the present work) was defined and compared with several existing curve discrimination techniques. This comparison was performed using two real data sets and through simulations. The same real data sets were used in [14] to compare their functional segment discriminant analysis proposal with standard classification methods. Some simulation experiments complete that paper. Working with microarray data, in [15], 21 classification methods were compared on 7 data sets. With these (and other) studies in mind, an important aim of the present paper lies in comparing the performance of different classification methods applied to functional data, but focusing on the practical problem of wood species discrimination, using their TG curves. The comparison of nonparametric functional methods and classical and machine learning multivariate techniques in this framework represents a relevant challenge in the wood industry. It is also important to stress that the proposed logistic model to extract representative features from the TG curves is also a novel approach with physical–chemical interpretations in this context. Moreover, it can be used to generate artificial TG curves, making it possible to extend the comparison in a simple way to different scenarios. Note that statistical simulation studies are unusual in thermal analysis, being an exception the work of [16]. Therefore, the simulations experiments carried out in this research are a novelty in themselves in this framework, giving the opportunity not only of establishing a comparison among the different classification techniques, but also to extract some conclusions related with wood discrimination from a chemometric point of view.

Accordingly, the objectives of the present study are, on the one hand, comparing different methods to classify seven commercial wood species (European chestnut, European beech, eucalyptus, jatobá, European oak, Scots and insignis pine) on the basis of their TG curves. These methods include nonparametric functional techniques, using the complete TG curves, and multivariate supervised classification approaches, such as LDA, *k*-NN, NBC, NN, and SVM, using some components of the curves selected by PCA or fitting a generalized logistic mixture model. Additionally, these classification methods applied in a functional context are compared through a comprehensive simulation study. The synthetic curves used to test the different methods try to imitate the TG curves in different scenarios.

The structure of the paper is as follows. In Section 2, the functional and multivariate data classification methods used in the present research are explained. In Section 3, these methods are applied to the real wood samples using their TG curves, analyzing the results obtained. Previously, in this section, the materials and experimental techniques used to obtain the TG curves are also described. In Section 4, the performance of the different classification techniques are compared in different scenarios through a simulation study. Finally, Section 5 collects the main conclusions.

## 2. Classification techniques

The process of assigning an observation to one of several predetermined groups is called supervised classification. The principal aim is to obtain a discriminant function summarizing the information contained in the observations (curves or multidimensional vectors, in the framework of the present paper), and using this function to classify a new observation in one of the given groups. In other words, given a learning sample consisting of observed curves from known groups, our issue is to predict the group membership of a new incoming curve. In the statistical literature, several multivariate and functional methods have been developed to address supervised classification problems. In the next sections, the approaches used in our experiments are briefly described. The free statistical software R [17] was used to implement these methods. The specific R packages and libraries employed for this task, as well as the range of values used to select the optimal parameters in each case, are cited in the corresponding description of each method. This can demonstrate to practitioners how to apply the different approaches.

### 2.1. Linear Discriminant Analysis

The classical Linear Discriminant Analysis (LDA), proposed by Fisher [18], is one of the most widely used techniques for data classification. It works by maximizing the ratio of between-class variance to the within-class variance, being optimal under the assumptions of Gaussian likelihood and equal covariance matrices between groups. There has been many modifications proposed to this method, such as, Orthonormal LDA [19], Non parametric LDA [20], etc. But in our case, we have found the performance of classical LDA satisfactory. LDA has been successfully applied for classification and pattern recognition in fields as diverse as engineering, economics, computing science, biology, etc. Related to the topic of the present paper, there are some interesting works where LDA is applied to the features extracted from wood micrographs [21] and fluorescence spectra [22–24]. The R `MASS` library [25] was used to apply the LDA in our

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

experiments. Specifically, the function named `lda` is designed to perform this classification method.

### 2.2. Naïve Bayes

Naïve Bayes classifier (NBC) is a supervised multivariate classification technique based on the Bayes theorem. Using this rule, we intend to calculate the posterior probability that a sample belongs to a particular class (from a group of possible classes), given the feature values that define the sample. Then, the class of a test sample is estimated using the largest posterior probability obtained. NBC assumes that the conditional probabilities of the independent variables are statistically independent and, therefore, the posterior probabilities can be rewritten in a simpler way. The function `naiveBayes` included in the R `e1071` library [26] was used in our analysis to apply the NBC method. Many papers showing the application of this technique in a variety of classification problems can be found in the scientific literature. For instance, in computer science, addressing the problem of classifying E-mails in spam and non-spam; in medicine, to solve medical diagnosis; in acoustics, performing automatic classification of sound and voice, or in image classification. Some references are, for example, [27,28]. For the particular case of wood classification, we can cite the studies of Mallik et al. and Gasson et al., where NBC technique is applied to features extracted from wood micrographs at different magnification [21,29].

### 2.3. k Nearest Neighbors

$k$ Nearest Neighbors ($k$-NN) is a multivariate nonparametric supervised classification method introduced by Fix and Hodges [30]. It can also be applied in populations where the assumption of normality is not required. In principle, this represents an advantage over the LDA method, which is not optimal for non Gaussian populations. The basic idea of $k$-NN is the following: the class of a given sample will be the most repeated class corresponding to the surrounding neighbors. The $k$-NN procedure starts choosing an appropriate distance (mainly the Euclidean or Mahalanobis distances) between samples, represented by vectors of features. Then, the distances between the test sample, $x_0$, and the other samples are calculated. The $k$ nearest samples to those we want to classify are selected. Next, the proportion of these $k$ samples belonging to each of the studied populations is calculated. Finally, the sample $x_0$ is classified within the population corresponding to the highest existing frequency. Among the different methods available for choosing the value of $k$, the minimization of the cross-validation error is one of the most used. A weighted version of $k$-NN was used in the present study. Each $k$ nearest observation is weighted according to its distance to the new observation that we want to classify [31]. The distances are transformed by the application of some type of kernel (Gaussian, triangular, rectangular, Epanechnikov, etc). Then, the winning class is chosen as the class estimation which shows a weighted majority of the $k$ nearest neighbors. The model parameters, number of nearest neighbors and type of kernel, were obtained by the inner loop corresponding to a double cross-validation process (see Section 3.2). For this purpose, `kknn` library [32] was used. Specifically, the function `kknn` of that package is designed to apply this method. We fix the range of possible values for the number of nearest neighbors between 1 and 25, while rectangular, triangular, Epanechnikov, Gaussian and rank kernels were tested. The $k$-NN technique have been successfully applied in diverse fields, such as chemistry, biology, medicine, computer science, genetics and material science, identifying wood species [33,21–24].

### 2.4. Support Vector Machines

Support Vector Machines (SVM) is a comparatively new machine learning technique developed by Vapnik and co-workers [34]. SVM is a non-probabilistic classifier. It works by constructing a hyperplane or a set of hyperplanes maximizing its distance from the nearest data point on each side, therefore achieving largest separation. This hyperplane is called the maximum margin hyperplane. For non linear cases the hyperplane is constructed by a nonlinear kernel function in place of dot products. Homogeneous polynomial and Gaussian radial basis functions are mostly used kernels. Note that in the original form, SVM are binary classifiers, i.e. they discriminate between two classes. In a situation like the one studied in this research, where there are 7 possible classes, a common procedure to overcome this drawback is to turn one multi-class problem into several two-class problems. In the present paper, a C-SVM model and a Gaussian kernel were tested. The C (cost of constraints violation) and gamma parameters were obtained by the inner loop of a double cross-validation procedure (see Section 3.2). Values between 1 and 21 for the C parameter, and between 0.001 and 0.051 for the gamma parameter were used in the present research. The `svm` function of the R package `e1071` (related to the `libsvm` library) [26] was employed to apply SVM. It is important to point out that the `libsvm` library uses one-against-one classifications for multiclass problems with $k$ levels, $k>2$, in which $n$ choose $k$ binary classifiers are trained. The class which has won most number of times is chosen as winner class. During recent years SVM is being successfully used in classification [35], text and data mining [36] and image retrieval [37] problems. Regarding wood classification problems, SVM was applied in [21,22].

### 2.5. Neural Networks

Neural Networks (NN) is a machine learning technique that is motivated to imitate the structural and functional aspects of biological neural networks [38]. In most cases, NN consists of a network of artificial neurons, which changes its structure based on external or internal information flowing through the network at learning phase. NN are becoming widely popular because of their wide area of applicability such as functional approximation, data modeling, classification [39,40], robotics, etc. NN are computationally demanding, but they are very useful because they provide flexibility of choosing various learning algorithms and cost functions, thus offering superior control over learning rate of the model and its accuracy. For our purpose, we have used a single hidden layer feed forward network to train and test our dataset. The model parameters, in this case, number of units in the hidden layer and weight decay, were obtained by the inner loop corresponding to a double cross-validation process (see Section 3.2). We used the R `nnet` library [25] to implement this classification method, selecting a range between 1 and 3 for the number of units in the hidden layer, and between 0.00001 and 0.001 for the weight decay. In relation with the topic of this paper, there are several interesting studies where NN was applied to classify wood species [21,24,41].

### 2.6. Nonparametric functional data analysis

Statistics for functional data is a recent field of research popularized by the monographs of [42,43] or [44]. Assuming that a set of $n$ curves $X_i = X_i(t)$, $i = 1, \ldots, n$, have been observed, each one belonging to a known class $g$, with $g \in \{0, 1, \ldots, G\}$, and given a new curve, $x = x(t)$, we are interested in classifying $x$ in one of the classes $\{0, 1, \ldots, G\}$. This problem is solved estimating the posterior probability that $x$ belongs to a class $g$, for each $g \in \{0, 1, \ldots, G\}$, and selecting the class with largest estimated posterior probability. It is clear that this problem can be tackled from a regression point of view and, therefore, the issue of estimating those posterior probabilities is equivalent to that of estimating the corresponding regression functions.

Taking this account and given the functional nature of the data, the posterior probability estimator that $x$ belongs to a class $g$, with $g \in \{0, 1, \ldots, G\}$, used in the present work is the functional Nadaraya–Watson

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

ARTICLE IN PRESS

M. Francisco-Fernández et al. / Chemometrics and Intelligent Laboratory Systems xxx (2012) xxx–xxx 5

nonparametric kernel method (K-NFDA) shown in (1), and defined in [13]:

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

where $Y_i$ are the corresponding classes of the observed curves $X_i$, with $i=1, \ldots, n$. In Eq. (1), $I_{\{\cdot\}}$ is the indicator function, the parameter $h$ is the bandwidth or smoothing parameter, K is a kernel function, and $\|\cdot\|$ is a semi-norm. The kernel function is typically a density function chosen by the user (in our experiments, we used the Gaussian kernel) and $h$ regulates the amount of smoothing to be used. Although the choice of the kernel function is of secondary importance, the smoothing parameter plays a crucial role in kernel estimation. We chose $h$ as the value that minimizes the probability of misclassifying a future observation (it is selected according to the cross-validation method, searching a bandwidth in a grid of values ranging between 0.001 and 2). Finally, we used as semi-norm, the $L_2$ norm, measuring the differences between curves by:

$$d\left(X_i, X_j\right) = \int_a^b \left(X_i(t) - X_j(t)\right)^2 dt, \quad (2)$$

where $[a, b]$ is the interval studied. The R Packages `fda` [45] and `fda.usc` [46] were mainly used to perform the classification applying nonparametric functional analysis.

## 3. Classification of wood samples

In this Section, we describe the wood data used in our analysis and the experimental process to obtain their TG curves. Then, the statistical methods previously presented are applied to these TG curves to classify the corresponding timber samples indifferent species. Some programs coded in R using the specific packages and functions cited in the previous section were employed to compare the approaches, using the real data and through simulations.

### 3.1. The data

Tests for five different hardwoods (European beech or *Fagus sylvatica*, European oak or *Quercus robur*, chestnut or *Castanea sativa*, *Eucalyptus globulus* and jatobá or *Hymenaea courbaril*) and two softwoods (Scots pine—*Pinus silvestris*—and insignis pine—*Pinus radiata*) were carried out. Seven samples per each one of the above mentioned species, obtained from wood of different trees were tested. The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the experimentation time. The samples were not dried in order to avoid disturbance to their structure and composition as much as possible, and to test the automatic classification methods with a minimal sample preparation, under the worse conditions.

The test was performed on a SDT 2960 TA Instruments thermo balance. This apparatus provides TG curves used in the classification analysis. A heating ramp of 20 °C·min$^{-1}$ was applied in the range from 20 to 600 °C, at a rate of 50 mL·min$^{-1}$ of N2 [8]. The nitrogen was purged for 10 min, before starting the heating program for establishing an inert environment. The used heating rate was chosen to obtain a proper balance between time test and resolution [8]. It aims to assess the discriminatory power of the resulting curves, using the minimum experimental time. The sample mass chosen was between 6 and 8 mg. Alumina crucibles were used.

### 3.2. Results and discussion

Fig. 1 in Section 1 shows the 49 TG curves obtained (7 per class). A descriptive analysis of these data was shown in [33]. Different trends are observed for almost all types of wood. Note also that the TG curves of some species tend to overlap with other species. This problem arises because of their similar densities, hardness and mechanical properties. More specifically, the shape of the TG curves obtained in a pyrolysis test (as it is the case of the present work) is directly related to wood composition [47–53]. In fact, wood degradation in an inert atmosphere is dominated by the degradation behavior of its three main components (cellulose, lignin, and hemicellulose) as was reported in [47,49]. Cellulose represents about 40 and 60% in the overall weight of dry wood (23–33% of the mass of softwoods), 23–33% of lignin in softwoods (16–25% in hardwoods), and 25–35% of hemicellulose [3,51]. The proportion of each wood component varies depending on the species, to a greater or lesser extent [50–53]. Therefore, it is expected that the TG curves are different depending on the type of wood to which they belong. Furthermore, effects on the shape of the TG curves, such as the existence of other components as ethanol extractives [54] and even the origin of lignin and hemicellulose are not absolutely negligible. Differences in the pyrolysis of lignin and hemicellulose depending on whether these come from softwoods or hardwoods, or even from different species, were also observed [52–55]. Considering all these studies, it seems reasonable to suggest that TG analysis has the potential to be useful to discriminate between classes of wood species. Moreover, it is important to note that the hemicellulose, cellulose and lignin decompose in temperatures ranging between 200–260 °C, 240–350 °C, and 280–500 °C, respectively [47–53].

As observed in Fig. 1, each curve represents a functional data and, therefore, it is quite natural to use functional classification methods (see Section 6) to carry out the classification process. On the other hand, after discretizing the TG curves, the multivariate classification methods described in the previous section can also be applied. As pointed out in the Introduction section, two methods were employed for this process. First, using PCA to select the curve components explaining most of the variation of the data (99% in our case). Second, fitting a generalized logistic model to the TG curves and using the parameters of this model as representative curve features. The proposed model to be fitted consists of a mixture of 4 generalized logistic functions related to the wood main components, i.e. cellulose, hemicellulose, lignin and water:

$$Y(t) = \sum_{i=1}^{4} c_i \left(1 - \frac{1}{\left(1 + \tau_i \cdot e^{-b_i \cdot (t-m_i)}\right)^{1/\tau_i}}\right) \quad (3)$$

where the $c_i$ parameter is related to the mass involved in the degradation process, $b_i$ is related to the decomposition rate or rate of change, $\tau_i$ accounts for the asymmetry, $m_i$ represents the temperature at the maximum rate of change, $t$ is the temperature, and $Y$ the fitted TG curve. The optimal fittings were obtained by minimizing the average squared error (ASE), using the Nelder–Mead algorithm [56]. All the resulting fits presented coefficient of determination higher than 0.999. An example of the goodness-of-fit of this model can be observed in Fig. 2, where an original TG curve of beech and the corresponding fit using Eq. (3) is presented. The four logistic components of this fit are also included in this plot. Using these 16-dimensional vectors (or the corresponding vectors obtained with the PCA approximation), it is possible to apply the multivariate supervised classification methods previously described to perform the curve classification.

In the case of the nonparametric functional approach, after a careful examination of the TG curves, it was found that the curves can be better discriminated if they were standardized, using a linear transformation on the original curves based on their mean and variances [57,33].

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

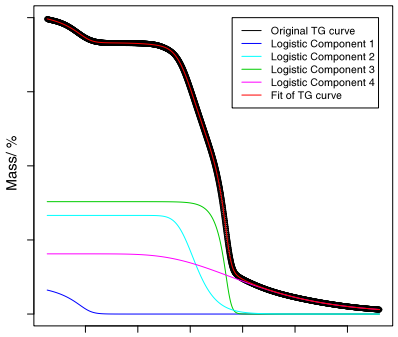


**Fig. 2.** Original beech TG curve and the corresponding logistic fit.

Two different procedures were used to validate the different approaches, cross-validation and external validation. Regarding the first proposal, a double cross-validation process was performed [58]. It consists of an inner loop where the optimal parameters corresponding to the classification methods (SVM, NN, *k*-NN, etc.) were obtained, and an external loop to assess the prediction error in each case. This methodology was used to avoid the bias resulting from the use of the same sample to estimate the parameters of each classification method, and to obtain the misclassification errors [58]. In the present study, the inner and external loops were consisted of two leave-one-out cross-validation procedures. This process works by leaving out one TG curve; then a model is trained with the remaining thermograms and, finally, the developed model is used for the classification of the left out TG curve. This is repeated until all the curves have been left out once. As the dataset available contain 49 samples, 48 samples were used for training and 1 sample for testing. It is important to note that the model is trained by another leave-one-out procedure using the 48 remaining curves (inner loop). This process was repeated 49 times, and the percentages (measured as per one) of correct classification were calculated. Fig. 3 shows a kind of flowchart describing the validation process previously described for the multivariate methods. Similar steps are followed when the K-NFDA method is validated through leave-one-out cross-validation, but obviating the feature extraction step.

The error estimates obtained by leave-one-out cross-validation can present a relatively high variance, although without bias. On other hand, multiple cross-validation (e.g. 10-fold) is about ten times faster and presents less variability in the error estimations [59,58], although a bias can occur because the model is based on a data set that is smaller than the real data set, giving slightly pessimistic misclassification error estimations. For avoiding these disadvantages of leave one out cross validation as far as possible, and for completing the results obtained by this technique, an external validation procedure was performed. Considering the relatively small number of samples per class, a first set of samples (one per class) is randomly extracted as external validation set. The remaining samples are used as training set to obtain the classification model parameters, according to a 10-fold cross-validation in this case. This global procedure is repeated 100 times to be sure that, with high probability, each sample is included in the test set at least once. Fig. 4 shows a similar flowchart to that presented in Fig. 3, but now for the external validation process.

Tables 1 and 2 shows the misclassification errors obtained by the different methods using leave-one-out cross-validation and the external validation procedure, respectively. In the second framework, the standard deviations are also included between brackets in Table 2.

The results obtained by cross-validation and external validation are generally very similar. This fact supports the validity of the obtained error estimations. In general, Tables 1 and 2 show a relatively high probability of correct classification, especially taking into account wood heterogeneity and the results obtained in other studies [60,61]. Therefore,

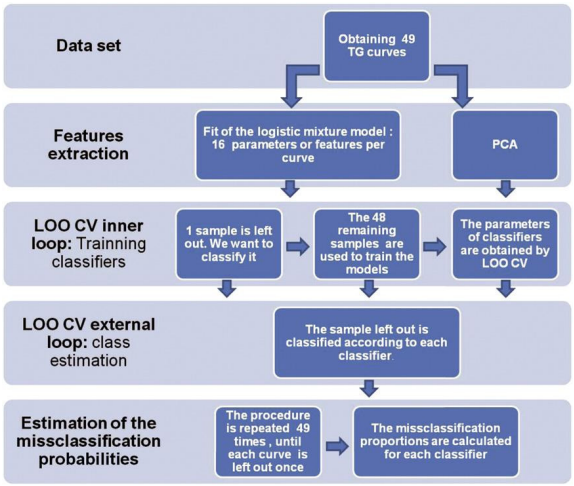


**Fig. 3.** Flowchart of the leave-one-out cross-validation process for the multivariate approaches.

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

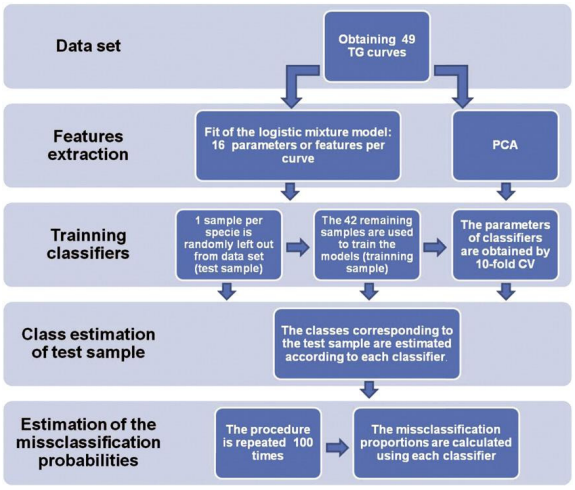


**Fig. 4.** Flowchart of the external validation process for the multivariate approaches.

classification using the logistic parameters or PCA has been shown feasible. It is interesting to note that SVM, LDA and K-NFDA methods gave the highest probabilities of good classification, but LDA and K-NFDA required less computing times. Since the probability of misclassification is larger than zero, an interesting question is: which are the species tending to get confused using these features and methods? The answer is in the confusion matrices shown in Tables 3 and 4.

Table 3 shows the confusion matrices corresponding to LDA, *k*-NN, SVM and K-NFDA, applied to classify between 7 different types of wood, when using PCA to discretize the TG curves in the multivariate classification approaches. Table 4 shows the same information, but when the features of the TG curves are selected by fitting the generalized logistic model given in Eq. (3) (K-NFDA is not included in Table 4. Instead, the results for NBC are presented).

Tables 3 and 4 show that, among the 7 species studied, Scots pine is the most difficult species to be classified. When we classify using the logistic parameters or the raw TG curves, there are some confusion between Scots pine and insignis pine, and between oak and beech. This may be due to their similar chemical, physical and mechanical properties. It is also interesting to observe that jatobá samples can be classified as Scots pine samples and vice versa when the features obtained by PCA are used. Misclassification errors in species like oak vary considerably depending on the method and the features used as dataset. This should be taken into account for further implementation in the industry.

## 4. Simulation study

This section shows a simulation study comparing the classification methods previously applied to the real TG curves. Using the parameters obtained from fitting the generalized logistic model (3) to the real TG curves, artificial curves imitating the real ones are generated. This allows to compare the classification approaches and, additionally, to establish conclusions on wood classification in different scenarios, saving time and money. Note that the extreme heterogeneity of wood makes it very difficult to perform a comprehensive comparison of the different techniques with real samples. So, the generation of artificial TG curves, mimicking the real ones, in different scenarios represents a useful tool in this setting. The analysis of the simulation results, jointly with the conclusions derived from the experiments with the real samples, could give clues of a general good approach to be recommended to practitioners in this field.

**Table 1**
Misclassification errors obtained by each classification method and leave-one-out cross-validation. The multivariate classification methods were tested using PCA and the generalized logistic fits.

| Classification | Misclassification errors | |
|---|---|---|
| Methods | PCA | Logistic |
| LDA | 0.18 | 0.14 |
| NBC | 0.33 | 0.27 |
| *k*-NN | 0.27 | 0.39 |
| SVM | 0.16 | 0.24 |
| NN | 0.24 | 0.24 |
| K-NFDA | 0.22 | |

**Table 2**
Misclassification errors obtained by each classification method and the external validation process. The multivariate classification methods were tested using PCA and the generalized logistic fits. Standard deviations are included between brackets.

| Classification | Misclassification errors | |
|---|---|---|
| Methods | PCA | Logistic |
| LDA | 0.20 (0.12) | 0.18 (0.14) |
| NBC | 0.29 (0.15) | 0.28 (0.17) |
| *k*-NN | 0.27 (0.15) | 0.39 (0.16) |
| SVM | 0.17 (0.12) | 0.23 (0.15) |
| NN | 0.34 (0.17) | 0.35 (0.17) |
| K-NFDA | 0.26 (0.15) | |

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

**Table 3**
Probabilities of correct classification, using leave-one-out cross-validation, in 7 different classes (chestnut, eucalyptus, beech, jatobá, insignis pine, oak and Scots pine) obtained by LDA, *k*-NN, SVM and K-NFDA, when using PCA to discretize the TG curves in the multivariate classification approaches. The probabilities are rounded using two significant figures.

| Methods | Actual | Estimated | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Oak | Beech | Chesn. | Eucal. | Jat. | Insig. P. | Scots P. |
| LDA | Oak | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 | 0.29 |
| | Insignis P. | 0.14 | 0.29 | 0.00 | 0.00 | 0.00 | 0.57 | 0.00 |
| | Scots P. | 0.14 | 0.00 | 0.00 | 0.00 | 0.43 | 0.00 | 0.43 |
| *k*-NN | Oak | 0.71 | 0.29 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.14 | 0.00 | 0.86 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.00 | 0.00 | 0.00 | 0.00 | 0.57 | 0.00 | 0.43 |
| | Insignis P. | 0.00 | 0.00 | 0.29 | 0.00 | 0.00 | 0.71 | 0.00 |
| | Scots P. | 0.00 | 0.14 | 0.00 | 0.14 | 0.43 | 0.00 | 0.29 |
| SVM | Oak | 0.86 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.14 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 | 0.29 |
| | Insignis P. | 0.00 | 0.14 | 0.14 | 0.00 | 0.00 | 0.71 | 0.00 |
| | Scots P. | 0.14 | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 | 0.57 |
| K-NFDA | Oak | 0.57 | 0.14 | 0.00 | 0.29 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.29 | 0.71 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.00 | 0.00 | 0.86 | 0.14 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.00 | 0.14 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.29 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.14 | 0.71 |

**Table 4**
Probabilities of correct classification, using leave-one-out cross-validation, in 7 different classes (chestnut, eucalyptus, beech, jatobá, insignis pine, oak and Scots pine) obtained by LDA, *k*-NN, NN and NBC, when the features of the TG curves are selected by fitting the generalized logistic model. The probabilities are rounded using two significant figures.

| Methods | Actual | Estimated | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Oak | Beech | Chesn. | Eucal. | Jat. | Insig. P. | Scots P. |
| LDA | Oak | 0.57 | 0.29 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.14 | 0.00 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.86 | 0.14 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.71 |
| *k*-NN | Oak | 0.71 | 0.29 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.14 | 0.29 | 0.29 | 0.29 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.14 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.14 | 0.00 | 0.00 | 0.00 | 0.71 | 0.14 | 0.00 |
| | Insignis P. | 0.14 | 0.14 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.86 | 0.00 |
| SVM | Oak | 0.71 | 0.14 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.14 | 0.00 | 0.00 | 0.86 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 | 0.29 |
| | Insignis P. | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 | 0.57 | 0.29 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.57 | 0.43 |
| NBC | Oak | 0.71 | 0.14 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.43 | 0.57 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Chesnut | 0.14 | 0.00 | 0.86 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 0.86 | 0.14 | 0.00 | 0.00 |
| | Jatobá | 0.29 | 0.00 | 0.00 | 0.00 | 0.71 | 0.00 | 0.00 |
| | Insignis P. | 0.14 | 0.00 | 0.00 | 0.00 | 0.00 | 0.71 | 0.14 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.29 | 0.71 |

### 4.1. Data-generating process

The generation of synthetic TG curves is an important point of this simulation study. The process is the following:

1. The generalized logistic model, given in Eq. (3), is fitted to each real TG curve, obtaining a 16-dimensional vector of parameters for each TG curve:

   $$((c_1, b_1, \tau_1, m_1), (c_2, b_2, \tau_2, m_2), (c_3, b_3, \tau_3, m_3), (c_4, b_4, \tau_4, m_4)).$$

   After these fits, 49 16-dimensional vectors (7 for each kind of wood) are obtained.
2. For the *r*th class of wood ($r = 1, 2, \ldots, 7$), $n_r$ 16-dimensional vectors, ($x_1$, $y_1$, $z_1$, $t_1$, ..., $x_4$, $y_4$, $z_4$, $t_4$), are generated from a multivariate normal distribution, $N_{16}(\mu_{(r)}, \Sigma_{(r)})$, where $\mu_{(r)}$ is the sample mean and $\Sigma_{(r)}$ represents a variance–covariance matrix, both computed from the 7 vectors of parameters of the *r*th class of wood, obtained from the logistic fits. We establish the condition of simulating a new vector if any of the components of the generated vector is negative.
3. Finally, we define,

   $$c_j = \frac{x_j}{\sum_{l=1}^{4} x_l} 100, b_j = y_j, \tau_j = z_j, m_j = t_j, \quad j = 1, 2, \ldots, 4,$$

   and using the logistic model (3), we obtain the artificial TG curves ($n_r$ for each class of wood, with $r = 1, 2, \ldots, 7$).

The variance–covariance matrix $\Sigma_{(r)}$ for the *r*th kind of wood, $r = 1, 2, \ldots, 7$ is defined by:

$$\Sigma_{(r)} = \begin{pmatrix} \Sigma_{(r),1} & 0 & 0 & 0 \\ 0 & \Sigma_{(r),2} & 0 & 0 \\ 0 & 0 & \Sigma_{(r),3} & 0 \\ 0 & 0 & 0 & \Sigma_{(r),4} \end{pmatrix},$$

with $\Sigma_{(r),j}$ a variance–covariance matrix of the *j*th component ($j = 1, 2, \ldots, 4$) of the logistic model. In our studies, we use two parameters, $\alpha$ and $\beta$, to control the amount of variance and the degree of the dependence, respectively, in the matrix $\Sigma_{(r),j}$. Specifically, we consider

$$\Sigma_{(r),j} = \alpha\Sigma^{(1)}_{(r),j} + \beta\Sigma^{(2)}_{(r),j},$$

with

$$\Sigma^{(1)}_{(r),j} = \begin{pmatrix} \sigma^2_{(r),1,j} & 0 & 0 & 0 \\ 0 & \sigma^2_{(r),2,j} & 0 & 0 \\ 0 & 0 & \sigma^2_{(r),3,j} & 0 \\ 0 & 0 & 0 & \sigma^2_{(r),4,j} \end{pmatrix}$$

and

$$\Sigma^{(2)}_{(r),j} = \begin{pmatrix} 0 & \sigma_{(r),12,j} & \sigma_{(r),13,j} & \sigma_{(r),14,j} \\ \sigma_{(r),21,j} & 0 & \sigma_{(r),23,j} & \sigma_{(r),24,j} \\ \sigma_{(r),31,j} & \sigma_{(r),32,j} & 0 & \sigma_{(r),34,j} \\ \sigma_{(r),41,j} & \sigma_{(r),42,j} & \sigma_{(r),43,j} & 0 \end{pmatrix}.$$

Therefore,

$$\Sigma_{(r),j} = \begin{pmatrix} \alpha\sigma^2_{(r),1,j} & \beta\sigma_{(r),12,j} & \beta\sigma_{(r),13,j} & \beta\sigma_{(r),14,j} \\ \beta\sigma_{(r),21,j} & \alpha\sigma^2_{(r),2,j} & \beta\sigma_{(r),23,j} & \beta\sigma_{(r),24,j} \\ \beta\sigma_{(r),31,j} & \beta\sigma_{(r),32,j} & \alpha\sigma^2_{(r),3,j} & \beta\sigma_{(r),34,j} \\ \beta\sigma_{(r),41,j} & \beta\sigma_{(r),42,j} & \beta\sigma_{(r),43,j} & \alpha\sigma^2_{(r),4,j} \end{pmatrix},$$

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

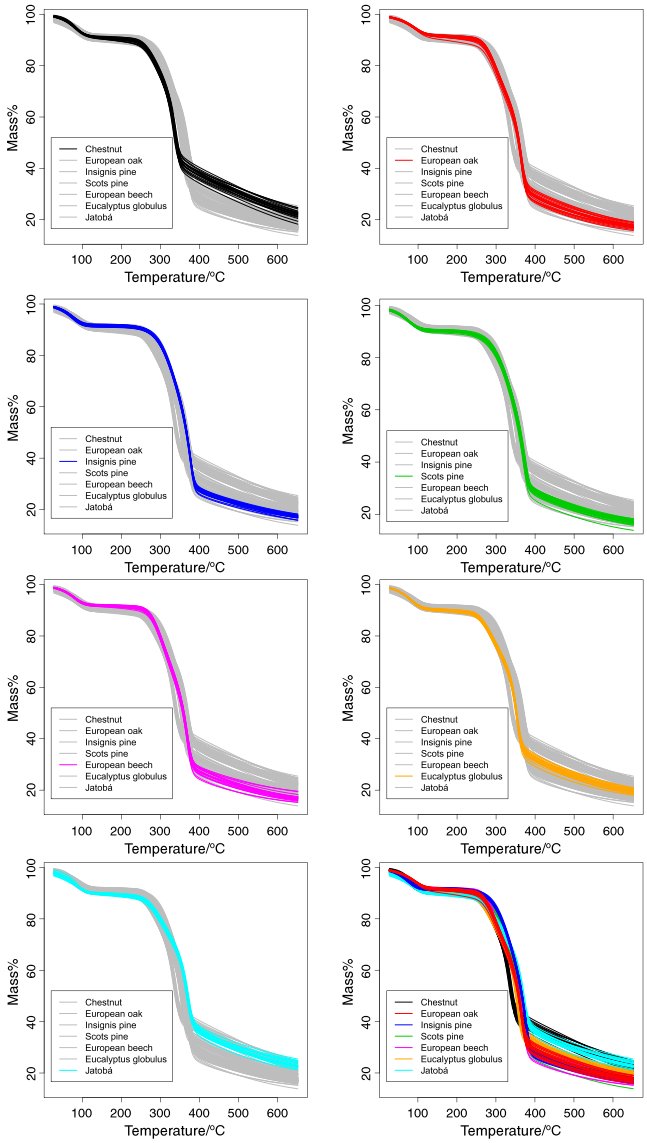


**Fig. 5.** Artificial TG curves for $\alpha = 0.05$, $\beta = 0$. Each particular wood species is highlighted in the corresponding panel. In the last panel (row 4, column 2) all the curves are presented.

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

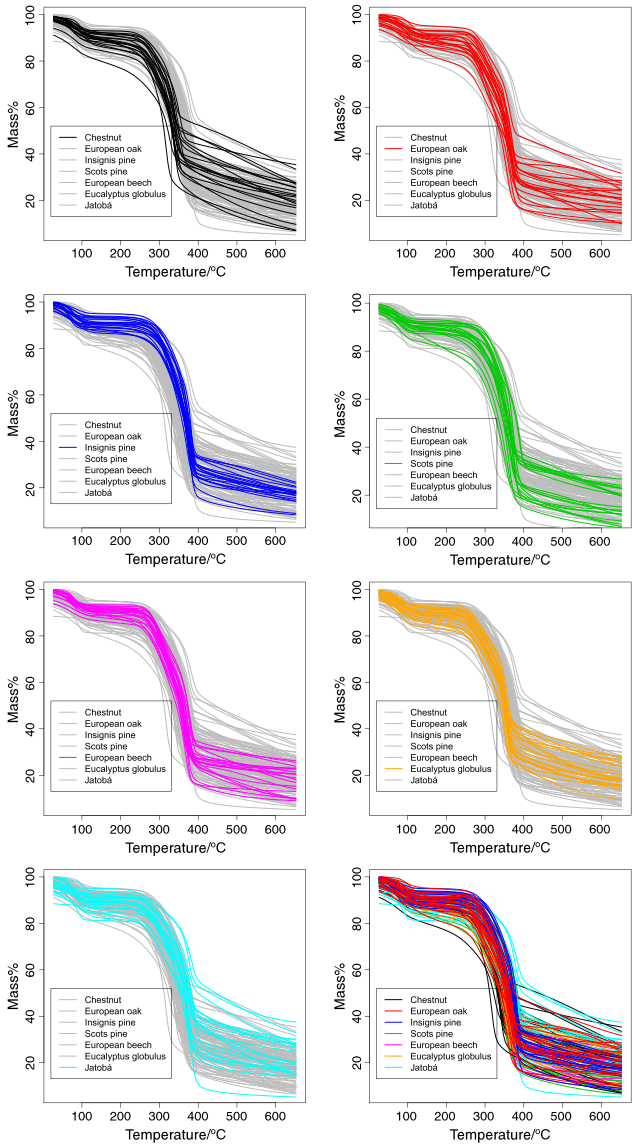


**Fig. 6.** Artificial TG curves for $\alpha = 2$, $\beta = 0.5$. Each particular wood species is highlighted in the corresponding panel. In the last panel (row 4, column 2) all the curves are presented.

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

where the matrix

$$\begin{pmatrix} \sigma^2_{(r),1,j} & \sigma_{(r),12,j} & \sigma_{(r),13,j} & \sigma_{(r),14,j} \\ \sigma_{(r),21,j} & \sigma^2_{(r),2,j} & \sigma_{(r),23,j} & \sigma_{(r),24,j} \\ \sigma_{(r),31,j} & \sigma_{(r),32,j} & \sigma^2_{(r),3,j} & \sigma_{(r),34,j} \\ \sigma_{(r),41,j} & \sigma_{(r),42,j} & \sigma_{(r),43,j} & \sigma^2_{(r),4,j} \end{pmatrix}$$

is the sample variance–covariance matrix of the $j$th component of the logistic model, $j=1, \ldots, 4$. Note that a value of $\beta=0$ indicates that the logistic model parameters are independent.

Different values of $\alpha$ and $\beta$ were used in the simulation study. Each simulation setting was repeated $B=1000$ times and results were obtained by averaging over the $B$ replicates. As an example, Figs. 5 and 6 show the artificial TG curves in two scenarios, $\alpha=0.05$, $\beta=0$, and $\alpha=2$, $\beta=0.5$, respectively. Note the similarity with the original curves in Fig. 1, but with different variability between the curves in both scenarios.

### 4.2. Results

In the first part of the study, 140 artificial TG curves (20 of each class) were simulated. Table 5 shows the probabilities of misclassification (averages over the 1000 replicates) obtained using leave-one-out cross-validation in a variety of scenarios. Three datasets were considered: the raw TG curves, the regression parameters resulting from the application of the logistic mixture model, given in Eq. (3), and the values obtained using PCA, explaining of the 99% of the total variance. Nine scenarios were chosen, corresponding to different values of $\alpha$ and $\beta$.

Table 5 shows that the best results were obtained using the logistic parameters as dataset and LDA as the classification method. Successful classifications were performed applying LDA on the above mentioned dataset in all the studied scenarios: the worst result was obtained when $\alpha=4$ and $\beta=0$ (probability of misclassification equal to 0.26), and the best for $\alpha=0.05$ and $\beta=0$ (probability of correct classification equal to 1). As expected, the smaller value of $\beta$ (less variance), the higher probabilities of correct classification. The application of SVM, NBC and NN produced competitive results. In fact, SVM gave the lower probability of misclassification in the case of scenarios with a high variance and covariance, such as $\alpha=2, \beta=2$ (applied to the logistic parameters). The results of the K-NFDA method were generally worse than those obtained by LDA, SVM, NN and NBC, and similar to those produced by $k$-NN. Nevertheless, the probabilities of correct classification with this method were very high for the cases of independence ($\beta=0$) and small variances, needing a shorter computing time than the other methods.

Fig. 7 shows boxplots of the misclassification errors over the 1000 replicas for the different classification methods, using leave-one-out cross-validation. Four scenarios are included in this plot, $\alpha=0.25$ and $\beta=0$, $\alpha=1$ and $\beta=0$, $\alpha=4$ and $\beta=0$, and $\alpha=2$ and $\beta=1$. For the multivariate approaches, boxplots using the parameters selected by the logistic fits (denoted by 'P') and by PCA (denoted by 'PC') are shown in this figure.

An external validation procedure, based on a random selection of the test and training samples, was proposed to supplement the results obtained by cross-validation. 1000 curves were simulated. The number of samples of each species was assigned by simulating a multinomial distribution (with equal probabilities to each one of the 7 classes or species). Then, a test and a training sample were randomly selected, and the misclassification probability was computed. This process was repeated 50 times. A similar procedure was carried out, for example, in [13] or [14]. In this part of the study, we used the raw artificial curves to validate the K-NFDA method, and the corresponding parameters obtained from the logistic fits, in the case of the multivariate approaches. A 10% of the total sample was selected as the training sample, while the remaining 90% corresponded to the test sample. The use of relatively small training samples would be in line with the frameworks usually found in real situations. The results shown in Table 6, obtained by external validation, are quite similar to the probabilities of misclassification when using the cross-validation procedure.

Additionally, samples using a multinomial distribution with non-equal probabilities for the different species were generated. We used probabilities of 0.3 for Scots pine and beech, and 0.08 for the remaining ones. The results are presented in Table 7.

Table 7 shows the robustness of the classification methods studied in this paper when the number of samples corresponding to each species and used in the training sample is changed.

Fig. 8 shows boxplots of the misclassification errors over the 50 replicas for the different classification methods in the external validation experiments. Four scenarios are included in this plot, $\alpha=0.25$ and $\beta=0$, $\alpha=1$ and $\beta=0$, $\alpha=4$ and $\beta=0$, and $\alpha=2$ and $\beta=1$. The frameworks of equal probabilities of generating a curve of each group (denoted by 'I') and non-equal probabilities (denoted by 'NI') are shown in this figure.

In general, the conclusions obtained from the simulations are similar to those in the initial application in scenarios of small variance and covariance. When these parameters increase, the best methods are LDA, SVM and NBC. In general, the multivariate methods work better when the parameters obtained with the generalized logistic model are used. Therefore, our general recommendation with this kind of data and in this framework could be to use LDA or SVM, using the parameters selected with the logistic model. They provide general good results no matter the degree of variance or covariance of the data (in other words, no matter the heterogeneity of the real samples). On the other hand, although the K-NFDA did not provide very good results in the simulations, this method could be useful if the interest is to find the temperature ranges where the highest probability of correct classification is reached. This problem is also interesting in this setting from a practical point of view [33].

**Table 5**
Misclassification probabilities corresponding to the simulation study, applying leave-one-out cross-validation to the raw TG curves, the logistic regression parameters and the PCA parameters (99% of total variance) of the TG curves. The probabilities are rounded using two significant figures.

| Scenarios | | PCA of TG | | | | | Logistic parameters | | | | | Raw TG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | LDA | NBC | $k$-NN | SVM | NN | LDA | NBC | $k$-NN | SVM | NN | K-NFDA |
| 0.05 | 0 | 0.00 | 0.03 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| 0.25 | 0 | 0.03 | 0.09 | 0.09 | 0.06 | 0.07 | 0.01 | 0.01 | 0.08 | 0.01 | 0.01 | 0.12 |
| 0.5 | 0 | 0.10 | 0.15 | 0.17 | 0.14 | 0.15 | 0.02 | 0.03 | 0.16 | 0.05 | 0.04 | 0.22 |
| 1 | 0 | 0.20 | 0.24 | 0.27 | 0.24 | 0.26 | 0.08 | 0.10 | 0.27 | 0.11 | 0.11 | 0.32 |
| 2 | 0 | 0.32 | 0.35 | 0.39 | 0.37 | 0.38 | 0.17 | 0.19 | 0.37 | 0.19 | 0.21 | 0.41 |
| 4 | 0 | 0.45 | 0.49 | 0.53 | 0.51 | 0.50 | 0.26 | 0.27 | 0.45 | 0.27 | 0.31 | 0.51 |
| 2 | 0.5 | 0.32 | 0.35 | 0.39 | 0.37 | 0.38 | 0.17 | 0.19 | 0.36 | 0.18 | 0.21 | 0.42 |
| 2 | 1 | 0.31 | 0.33 | 0.37 | 0.35 | 0.37 | 0.18 | 0.19 | 0.34 | 0.17 | 0.21 | 0.42 |
| 2 | 2 | 0.21 | 0.24 | 0.25 | 0.22 | 0.26 | 0.12 | 0.14 | 0.19 | 0.07 | 0.11 | 0.37 |

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

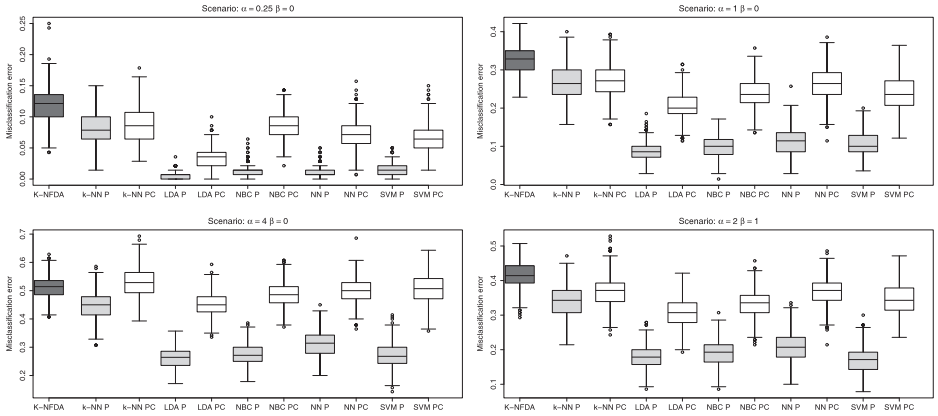


**Fig. 7.** Misclassification errors over 1000 replicas using leave-one-out cross-validation. For the multivariate approaches, boxplots using the parameters selected by the logistic fits (denoted by 'P') and by PCA (denoted by 'PC') are shown.

## 5. Conclusions

In the present paper, the performance of different functional and multivariate classification methods to classify wood species, using their TG curves, has been tested. The different approaches have been compared using real data and also validated through a comprehensive simulation experiment. Similar conclusions can be derived from both studies. They have shown that the classification of wood species is possible by applying these statistical techniques to their corresponding TG curves.

The main contribution of the simulation study in the present research is the possibility to design scenarios with different values of the variances and covariances of the artificial TG curves. This allows studying the behavior of the methods proposed in more unfavorable situations than those obtained experimentally. These situations are indeed very likely, given the high heterogeneity of a material like wood (even within the same species). Three different databases have been considered: the raw TG curves, the parameters obtained from fitting a generalized logistic mixture regression model to each TG curve, and the principal components of the TG curves (accounting for the 99% of the variability). Two different validation procedures were used: external validation based on the random selection of the training and test samples and leave-one-out cross-validation. In general, the higher probabilities of correct classification were obtained using the logistic parameters as dataset and LDA as classification method. In this case, the application of NBC, SVM and NN produced competitive results. Higher misclassification probabilities were obtained when applying the multivariate classification methods to the features selected with PCA. Regarding the K-NFDA method, the results were generally worse, although good probabilities of correct classification were obtained when the artificial TG curves were generated in scenarios of small variance and under independence. Note also that the K-NFDA method needed of a shorter computing time than the other methods.

In the case of the external validation, high probabilities of correct classification were obtained in all the scenarios considered. Even in the most unfavorable conditions, $\alpha=2, \beta=0.5$ or $\alpha=4, \beta=0$, and external validation with different sample size for each simulated species, LDA, NBC and SVM provided results over 80% of correct classification. In addition, the methodology proposed has resulted robust when the number of samples of each species in the training sample is varied.

Given all these results, the authors recommend the application of LDA to the parameters obtained by the logistic fit to perform this type of classification problems. This method has provided a general good performance in the initial application and in the simulations no matter the degree of variance or covariance of the data (in other words, no matter the heterogeneity of the real samples). Moreover, it

**Table 6**
Misclassification probabilities corresponding to the simulation study, applying an external validation to the raw TG curves and to the logistic regression parameters. Scenario of equal probabilities of generating a curve of each group. The probabilities are rounded using two significant figures.

| Scenarios | | Logistic parameters | | | | | Raw TG |
|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | LDA | NBC | $k$-NN | SVM | NN | K-NFDA |
| 0.05 | 0 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| 0.25 | 0 | 0.00 | 0.01 | 0.06 | 0.02 | 0.01 | 0.12 |
| 0.5 | 0 | 0.03 | 0.03 | 0.13 | 0.05 | 0.04 | 0.24 |
| 1 | 0 | 0.09 | 0.08 | 0.23 | 0.09 | 0.10 | 0.33 |
| 2 | 0 | 0.16 | 0.15 | 0.31 | 0.15 | 0.20 | 0.42 |
| 4 | 0 | 0.25 | 0.21 | 0.38 | 0.22 | 0.29 | 0.51 |
| 2 | 0.5 | 0.17 | 0.15 | 0.30 | 0.15 | 0.19 | 0.43 |
| 2 | 1 | 0.17 | 0.15 | 0.28 | 0.13 | 0.19 | 0.41 |
| 2 | 2 | 0.11 | 0.11 | 0.14 | 0.05 | 0.09 | 0.35 |

**Table 7**
Misclassification probabilities corresponding to the simulation study, applying an external validation to the raw TG curves and to the logistic regression parameters. Scenario of non-equal probabilities of generating a curve of each group. The probabilities are rounded using two significant figures.

| Scenarios | | Logistic parameters | | | | | Raw TG |
|---|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | LDA | NBC | $k$-NN | SVM | NN | K-NFDA |
| 0.05 | 0 | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 | 0.01 |
| 0.25 | 0 | 0.01 | 0.02 | 0.05 | 0.01 | 0.01 | 0.11 |
| 0.5 | 0 | 0.02 | 0.03 | 0.09 | 0.03 | 0.03 | 0.21 |
| 1 | 0 | 0.07 | 0.06 | 0.19 | 0.08 | 0.09 | 0.34 |
| 2 | 0 | 0.13 | 0.12 | 0.24 | 0.12 | 0.17 | 0.42 |
| 4 | 0 | 0.21 | 0.20 | 0.33 | 0.20 | 0.27 | 0.53 |
| 2 | 0.5 | 0.13 | 0.14 | 0.23 | 0.12 | 0.18 | 0.42 |
| 2 | 1 | 0.14 | 0.14 | 0.22 | 0.13 | 0.19 | 0.41 |
| 2 | 2 | 0.09 | 0.11 | 0.14 | 0.07 | 0.10 | 0.37 |

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

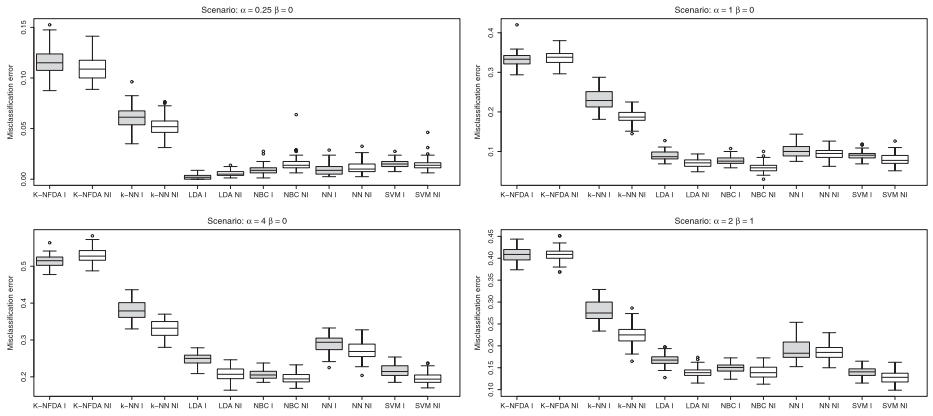


**Fig. 8.** Misclassification errors over 50 replicas for the external validation process. Boxplots in the case of equal probabilities of generating a curve of each group (denoted by 'I') and non-equal probabilities (denoted by 'NI') are shown.

needs of a lower computing time for its application than other approaches.

Finally, it is important to stress that this research was carried out with some specific wood species, but the methodologies used here could be applied with different species, or even with different materials.

## Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 (ERDF included) and Grant MTM2011-22392.

## References

[1] A. Guindeo Casasús, L. García Esteban, F. Peraza Sánchez, F. Arriaga Martitegui, Especies de maderas, Asociación de Investigación Técnica de las Industrias de la Madera y Corcho (AITIM), Madrid, 1997.
[2] I.R. Lewis, N.W. Daniel, N.C. Chaffin, P.R. Griffiths, Raman spectrometry and neural networks for the classification of wood types-1, Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy 50 (1994) 1943–1958.
[3] R. B. Miller, Structure of wood, in: Wood Handbook: Wood as an Engineering Material, Woodhead Publishing Limited, Madison, WI,U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, 1999.
[4] M. Khalid, E.L.Y. Lee, R. Yusof, M. Nadaraj, Design of an intelligent wood species recognition system, International Journal of Simulation: Systems, Science and Technology 9 (2008) 9–19.
[5] E.W.H. Hayek, P. Krenmayr, H. Lohninger, U. Jordis, W. Moche, F. Sauter, Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining, Analytical Chemistry 62 (1990) 2038–2043.
[6] J. Arno, G. Miller-Mead, J. Truini, The Art of Woodworking, Encyclopedia of Wood, Time-Life Books, Richmond, Virginia, 1993.
[7] R. Hoadley, Identifying Wood: Accurate Results with Simple Tools, Tautum press, Newtown, CT, 1990.
[8] R.B. Prime, H.E. Bair, P.K. Gallagher, A. Riga, Thermogravimetric analysis (TGA), In: in: J.D. Menczel, R.B. Prime (Eds.), Thermal Analysis of Polymers Fundamentals and Applications, John Wiley & Sons, San José, 2009.
[9] B.M. Lukasiak, R. Faria, S. Zomer, R.G. Brereton, J.C. Duncan, Pattern recognition for the analysis of polymeric materials, Analyst 131 (2006) 73–80.
[10] A.L. Pomerantsev, O.Y. Rodionova, Hard and soft methods for prediction of antioxidants' activity based on the DSC measurements, Chemometrics and Intelligent Laboratory Systems 79 (2005) 73–83.
[11] W. Miltyk, E. Antonowicz, L. Komsta, Recognition of tablet content by chemometric processing of differential scanning calorimetry curves—an acetaminophen example, Thermochimica Acta 507–508 (2010) 146–149.
[12] M. Ferriol, A. Gentilhomme, M. Cochez, N. Oget, J.L. Mieloszynski, Thermal degradation of poly(methyl methacrylate) (PMMA): modelling of DTG and TG curves, Polymer Degradation and Stability 79 (2003) 271–281.
[13] F. Ferraty, P. Vieu, Curves discrimination: a nonparametric functional approach, Computational Statistics & Data Analysis 44 (2003) 161–173.
[14] L. Bin, Y. Qingzhao, Classification of functional data: a segmentation approach, Computational Statistics & Data Analysis 52 (2008) 4790–4800.
[15] J.W. Lee, J.B. Lee, M. Park, S.H. Song, An extensive comparison of recent classification tools applied to microarray data, Computational Statistics & Data Analysis 48 (2005) 869–885.
[16] R. Artiaga, J. López-Beceiro, J. Tarrío-Saavedra, C. Gracia-Fernández, S. Naya, J.L. Mier, Estimating the reversing and non-reversing heat flow from standard DSC curves in the glass transition region, Journal of Chemometrics 25 (2011) 287–294, http://dx.doi.org/10.1002/cem.1347.
[17] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org 2011.
[18] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annual Eugenics 7 (1936) 179–188.
[19] T. Okada, S. Tomita, An optimal orthonormal system for discriminant analysis, Pattern Recognition 18 (1985) 139–144.
[20] K. Fukunaga, Introduction to Statistical Pattern Recognition, Academic Press Professional, Inc., San Diego, CA, 1990.
[21] A. Mallik, J. Tarrío-Saavedra, M. Francisco-Fernández, S. Naya, Classification of wood micrographs by image segmentation, Chemometrics and Intelligent Laboratory Systems 107 (2011) 351–362, http://dx.doi.org/10.1016/j.chemolab.2011.05.005.
[22] V. Piuri, F. Scotti, Design of an automatic wood types classification system by using fluorescence spectra, IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and reviews 40 (2010) 358–366.
[23] P. Camorani, M. Badiai, D. Francomacaro, M. Gamassi, P. Vincenzo, F. Scotti, M. Zanasi, A classification method for wood types using fluorescence spectra, In: Conference Record—IEEE Instrumentation and Measurement Technology Conference, 2008, pp. 1312–1315.
[24] R.D. Labati, M. Gamassi, V. Piuri, F. Scotti, A low-cost neural-based approach for wood types classification, In: 2009 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, CIMSA, 2009, pp. 199–203.
[25] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, Fourth edition Springer, New York, 2002, http://www.stats.ox.ac.uk/pub/MASS4.
[26] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, e1071: Misc Functions of the Department of Statistics (e1071), TU WienR package version 1.6, http://CRAN.R-project.org/package=e1071 2011.
[27] I. Kononenko, Machine learning for medical diagnosis: history, state of the art and perspective, Artificial Intelligence in Medicine 23 (2001) 89–109.
[28] L. Tóth, A. Kocsor, J. Csirik, On naive Bayes in speech recognition, International Journal of Applied Mathematics and Computer Science 2 (2005) 287–294.
[29] P. Gasson, R. Miller, D.J. Stekel, F. Whinder, K. Ziemińska, Wood identification of *Dalbergia nigra* (CITES Appendix I) using quantitative wood anatomy, principal components analysis and naïve Bayes classification, Annals of Botany 105 (2010) 45–56.
[30] E. Fix, J.L. Hodges, Discriminatory analysis, nonparametric discrimination: consistency properties, Technical Report 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.
[31] K. Hechenbichler, K. Schliep, Weighted k-Nearest-Neighbor Techniques and Ordinal Classification, In: Technical Report 399, SFB 386, Ludwing-Maximilians University Munich, 2004, http://epub.ub.uni-muenchen.de/1769/.

[32] K. Schliep, K. Hechenbichler, kknn: Weighted k-Nearest Neighbors, R package version 1.0-6, 2008.
[33] J. Tarrío-Saavedra, S. Naya, M. Francisco-Fernández, J. López-Beceiro, R. Artiaga, Functional nonparametric classification of wood species from thermal data, Journal of Thermal Analysis and Calorimetry 104 (2011) 87–100.
[34] V. Vapnik, Statistical Learning Theory, Wiley, 1998.
[35] A. Karatzoglou, D. Meyer, K. Hornik, Support vector machines in R, Journal of Statistical Software 15 (2006) 1–28.
[36] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of Machine Learning Research 2 (2001) 45–46.
[37] E. Chang, S. Tong, K. Goh, C.W. Chang, Support vector machine concept-dependent active learning for image retrieval, IEEE Transactions on Multimedia 2 (2005) 1–35.
[38] B.D. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, 1996.
[39] D. Michie, D.J. Spiegelhalter, C.C. Taylor, Machine learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence, Prentice Hall, 1994.
[40] R.P. Lippmann, Pattern classification using neural networks, IEEE Communications Magazine 27 (1989) 47–64.
[41] R. Jordan, F. Feeney, N. Nesbitt, J.A. Evertsen, Classification of wood species by neural network analysis of ultrasonic signals, Ultrasonics 36 (1998) 219–222.
[42] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer-Verlag, Berlin, 2006.
[43] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis, Springer-Verlag, New York, 2002.
[44] F. Ferraty, High-dimensional data: a fascinating statistical challenge, Journal of Multivariate Analysis 101 (2010) 305–306.
[45] J.O. Ramsay, H. Wickham, S. Graves, G. Hooker, FDA: Functional Data Analysis, R package version 2.2.7, 2011.
[46] M. Febrero-Bande, M. Oviedo de la Fuente, fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc), R package version 0.9.5, 2011.
[47] H. Yang, R. Yan, H. Chen, D.H. Lee, C. Zheng, Characteristics of hemicellulose, cellulose and lignin pyrolysis, Fuel 86 (2007) 1781–1788.
[48] R. Alén, E. Kuoppala, O. Pia, Formation of the main degradation compound groups from wood and its components during pyrolysis, Journal of Analytical and Applied Pyrolysis 36 (1996) 137–148.
[49] L. Gašparovič, Z. Koreňová, L. Jelemenský, Kinetic study of wood chips decomposition by tga, In: Proceedings 36th International Conference of SSCHE. Tatranské Matliare, vol. 178, 2009, pp. 1–14, World Scientific.
[50] A.F. Roberts, A review of kinetics data for the pyrolysis of wood and related substances, Combustion and Flame 14 (1970) 261–272.
[51] M.G. Grønli, G. Várhegyi, C. Blasi, Thermogravimetric analysis and devolatilization kinetics of wood, Industrial and Engineering Chemistry Research 41 (2002) 4201–4208.
[52] M. Müller-Hagedorn, H. Bockhorn, L. Krebs, U. Müller, A comparative kinetic study on the pyrolysis of three different wood species, Journal of Analytical and Applied Pyrolysis 68–69 (2003) 231–249.
[53] S. Wang, K. Wang, Q. Liu, Y. Gu, Z. Luo, K. Cen, T. Fransson, Comparison of the pyrolysis behavior of lignins from different tree species, Biotechnology Advances 27 (2009) 562–567.
[54] T. Sebio-Puñal, S. Naya, J. López-Beceiro, J. Tarrío-Saavedra, R. Artiaga, Thermogravimetric analysis of wood, holocellulose, and lignin from five wood species, Journal of Thermal Analysis and Calorimetry 1–5 (2012), http://dx.doi.org/10.1007/s10973-011-2133-1.
[55] D. Mohan, J.C.U. Pittman, P.H. Steele, Pyrolysis of wood/biomass for bio-oil: a critical review, Energy & Fuels 20 (2006) 848–889.
[56] J.A. Nelder, R. Mead, A simplex method for function minimization, The Computer Journal 7 (1965) 308–313.
[57] F. López-Granados, J.M. Peña Barragán, M. Jurado-Expósito, M. Francisco-Fernández, R. Cao, A. Alonso-Betanzos, O. Fontenla-Romero, Multispectral classification of grass weeds and wheat (*Triticum durum*) using linear and nonparametric functional discriminant analysis and neural networks, Weed Research 48 (2008) 28–37.
[58] R. Wehrens, Chemometrics with R. Multivariate Data Analysis in the Natural Sciences and Life Sciences, Springer-Verlag, New York, 2011.
[59] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, New York, 1993.
[60] T. Brandtberg, Individual tree-based species classification in high spatial resolution aerial images of forests using fuzzy sets, Fuzzy Sets Systems 132 (2002) 371–387.
[61] J.Y. Tou, Y.H. Tay, P.Y. Lau, Rotational invariant wood species recognition through wood species verification, In: in: N.T. Nguyen, H.P. Nguyen, A. Grzech (Eds.), Proceedings of the 2009 First Asian Conference on Intelligent Information and Database Systems (ACIIDS 2009), The Institute of Electrical and Electronics Engineers, Inc., Dong Hoi, 2009, pp. 115–120.

Please cite this article as: M. Francisco-Fernández, et al., A comprehensive classification of wood from thermogravimetric curves, Chemometrics and Intelligent Laboratory Systems (2012), doi:10.1016/j.chemolab.2012.07.003

# Apéndice C

# Functional nonparametric classification of wood species from thermal data

J Therm Anal Calorim (2011) 104:87–100
DOI 10.1007/s10973-010-1157-2

# Functional nonparametric classification of wood species from thermal data

**Javier Tarrío-Saavedra · Salvador Naya · Mario Francisco-Fernández · Jorge López-Beceiro · Ramón Artiaga**

ISBC XVI Conference Special Issue


**Abstract** In this study, thermogravimetric (TG) and differential scanning calorimetry (DSC) curves, obtained by means of a simultaneous TG/DSC analyzer, and statistical functional nonparametric methods are used to classify different wood species. The temperature ranges, where the highest probability of correct classification is reached, are also computed. As each observation is a curve, a nonparametric functional discriminant technique based on the Bayes rule and the Nadaraya–Watson regression estimator is used. It assigns a future observation to the highest probability predefined class (supervised classification). The smoothing parameter needed in this nonparametric method is selected according to the cross-validation technique. The method proposed is applied to a sample of 49 wood items (7 per wood class) and also to classify between hardwoods and softwoods. In all the cases, the samples have been successfully classified, obtaining better results with the TG curves. The results are compared with those obtained with other nonparametric methods based on boosting algorithm. A discussion about the relation of the obtained results with the referenced wood component degradation temperature ranks is presented.

**Keywords** Wood · Nonparametric classification · Functional data analysis · Thermal analysis

J. Tarrío-Saavedra (✉) · M. Francisco-Fernández
Departamento de Matemáticas. Facultad de Informática, Universidade da Coruña, Corunna, Spain
e-mail: jtarrio@udc.es

M. Francisco-Fernández
e-mail: mariofr@udc.es

S. Naya
Departamento de Matemáticas. Escuela Politécnica Superior, Universidade da Coruña, Corunna, Spain
e-mail: salva@udc.es

J. López-Beceiro · R. Artiaga
Departamento de Ingeniería Industrial II. Escuela Politécnica Superior, Universidade da Coruña, Corunna, Spain
e-mail: jlopezb@udc.es

R. Artiaga
e-mail: rartiaga@udc.es

## Introduction

The identification of wood is one of the most difficult tasks to perform related with the technology of this material, due to the wide variety of species and anatomical heterogeneity of its elements. Wood identification can often be made on the basis of readily visible characteristics such as color, odor, density, presence of pitch, or grain pattern. This analysis is typical in the furniture industries and the wood panel production. Often, the performed analysis has a non-uniform accuracy because of the operator. To achieve a correct classification is essential to use microscopy techniques, physical hardness tests, and chemical analysis [1–3]. Therefore, the implementation of quantitative models and automatic recognition methods of wood samples are justified and can be immediately useful. While there are various computational procedures to evaluate and rate the quality of a timber by inspecting its defects with the help of image processing techniques and spectral analysis [1, 4–8], these are not so generally used for species identification, despite several studies addressing this problem exist [2, 9–13].

A first step in a classification problem is to choose a discriminant feature from which it will be possible to classify. In the case of wood species classification, this discriminant feature could be the output of an experimental technique that really differentiates between them. In the

literature, wood samples are mainly classified based on the results of two techniques: image- and spectrum-based processing systems. A method of classification of 20 types of tropical timber from image processing, using extracting textural wood features, has been successfully tested in [9]. Those authors obtained a good classification proportion of 95%. On the other hand, in [2], the Fourier Transform Raman (FTR) spectroscopy and neural network technology have been coupled for spectral feature extraction and non-supervised classification. This represents the first time that both methodologies are combined. Later, neural networks and the FTR spectra for hardwoods and softwoods to differentiate temperate woods from tropical woods were also used [13]. Genetic algorithms and principal components analysis were used to classify 98 Raman spectra of temperate softwoods, hardwoods, and Brazilian and Honduran tropical woods [10]. Recently, in [12], an automatic wood type classification system based on the analysis of the fluorescence spectra, using nearest neighbor classifiers, linear and quadratic classifiers, and support vectors machines (SVMs) has been designed. However, it seems that the possibility of using thermal analysis as a source of data for statistical classification of wood species has not been sufficiently studied yet. In this article, the thermograms obtained by thermogravimetric analysis (TG) and differential scanning calorimetry (DSC) are used as a discriminant characteristic. These curves can be processed in a relatively simple way with functional analysis [14, 15, 16] and, as shown below, the shape of the TG curves is directly related to the wood composition. Therefore, TG analysis becomes an interesting option to discriminate between classes of timber.

In general, wood is defined as the set of xylem tissues forming the trunk, roots, and branches of woody plants, excluding the bark. The tubular cells size, shape and distribution, along with other anatomical elements such as wood radios, the presence of resin canals or vessels, pores, etc., in addition to the variable proportion of its chemical components, define the different wood species and their properties [1, 3, 13, 17–23]. Also, the different wood types can be generally divided in two broad categories: softwoods or conifers (gymnosperms) and hardwoods (dicot angiosperms) which can be subdivided into boreal, austral, and tropical hardwood types [1, 3]. Is it possible to observe these differences among species in the shape of the TG curves in a pyrolysis test? According to existing studies, the answer is yes [13, 17–23].

The wood degradation in an inert atmosphere is dominated by the degradation behavior of its three main components [17]. These are cellulose, lignin, and hemicellulose [13, 17–24]. Cellulose represents about 40 and 60% in the overall weight of dry wood (it accounts for 23–33% of the mass of softwoods), 23–33% of lignin in softwoods (16–25% in hardwoods), and 25–35% of hemicellulose (more in hardwoods than in softwoods) [3, 19, 25]. The three components decompose in temperatures ranging between 240–350 °C, 280–500 °C, and 200–260 °C, respectively [17, 19–23, 25]. As was reported in [13, 18], the TG curve describing the pyrolysis of wood nearly coincides with the sum of the degradation of its constituents. In many cases, no significant interaction between them has been concluded [21]. The proportion of each wood component varies depending on the species, to a greater or lesser extent [19, 20, 22, 23]. Therefore, it is expected that the TG curves are different depending on the type of wood to which they belong. While the effect of the wood structure appears to exist [22], this is much lower than that of the components [13, 18, 21]. Furthermore, differences in the pyrolysis of lignin and hemicellulose depending on whether these come from softwood or hardwood, or even of different species, were observed [20, 23]. These results suggest the use of discriminant characteristic TG curves.

Accordingly, the objectives of this study are

1. Evaluating the potential of functional nonparametric methods of discriminant analysis for the classification of hardwoods and softwoods and then for the classification of European oak, European chestnut, eucalyptus, scots, and insignis pine on the basis of TG and DSC data.
2. Comparing the accuracy performance of TG or DSC curve classification to discriminate between wood species or between major groups. The supervised kernel nonparametric classification and kernel nonparametric classification using the $k$-nearest method to select the bandwidth $h$ are used in [14]. In addition, two methods based on the boosting algorithm are also used to complete the study: using principal components analysis (PCA) and by the representation of functional data on a $b$-spline basis [15, 16, 26].
3. Finding the temperature range in TG and DSC curves where the highest probability of correct classification is reached.
4. Relating the results of classification analysis in each interval with the referenced cellulose, lignin, and hemicellulose degradation temperature ranks in a nitrogen atmosphere.

## Experimental

### Materials

Tests for five different hardwoods (European beech or *Fagus sylvatica*, European oak or *Quercus robur*, chestnut or *Castanea sativa*, *Eucalyptus globulus*, and jatobá or

*Hymenaea courbaril*) and two softwoods (scots pine—*Pinus silvestris* and insignis pine—*Pinus radiata*) are carried out. Seven samples per each one of the above mentioned species, obtained from wood of different trees are tested. The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the time of experimentation. The samples are not dried to avoid disturb as much as possible their structure and composition, and to test the automatic classification method with a minimal sample preparation, under the worse conditions.

Measurement methods

The test is performed on a SDT 2960 TA Instruments thermo balance. This apparatus provides both TG and DSC curves used in the classification analysis. A heating ramp of 20 °C $\text{min}^{-1}$ is applied in the range from 20 to 600 °C at a rate of 50 mL $\text{min}^{-1}$ of $N_2$ [27]. The nitrogen is purged for 10 min, before starting the heating program for establishing an inert environment. The used heating rate is chosen to obtain a proper balance between time test and resolution [27]. It aims to assess the discriminatory power of the resulting curves, using the minimum experimental time. The sample mass chosen is between 6 and 8 mg. Alumina crucibles are used. In particular, TG and DSC measurements are affected by some experimental parameters such as heating rate, amount of mass, type of atmosphere, or sample geometry [27]. Therefore, all these parameters are remained constant to obtain a better classification.

**Classification techniques**

Nonparametric functional techniques based on kernel methods [14] and two nonparametric methods based in the boosting algorithm are applied to construct a classification rule to discriminate between hardwoods and softwoods, and between the different species: European beech, European oak, European chestnut, eucalyptus, jatobá, scot, and insignis pine, based on a sample of 49 TG and DSC curves. A DSC or TG 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 (K-NPFDA), shown in (1), is applied. Given a new TG or DSC 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:

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

where the observed TG/DSC 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 $\{0, 1, \ldots, G\}$, the different classes. The parameter $h$ is the bandwidth or smoothing parameter and $\|\cdot\|$ denotes the following distance between curves:

$$d(X_i, X_j) = \int_a^b \left(X_i(t) - X_j(t)\right)^2 dt, \tag{2}$$

where $[a, b]$ is one of the 1,280 temperature intervals studied.

After a careful examination of the TG/DSC curves, further processing of the data has been found useful for standardizing the curves [28]. Denoting by $f(x)$ a curve, a linear transformation, $\tilde{f}(x) = \alpha f(x) + \beta$ with

$$\alpha = \frac{\sqrt{b-a}}{\sqrt{\int_a^b \left(f(t) - \frac{1}{b-a}\int_a^b f(s)ds\right)^2 dt}}$$

and

$$\beta = 1 - \frac{\int_a^b f(t)dt}{\sqrt{b-a}\sqrt{\int_a^b \left(f(t) - \frac{1}{b-a}\int_a^b f(s)ds\right)^2 dt}},$$

is done to achieve

$$\frac{1}{b-a}\int_a^b \tilde{f}(t)dt = 0$$

and

$$\frac{1}{b-a}\int_a^b \left(\tilde{f}(t) - \frac{1}{b-a}\int_a^b \tilde{f}(s)ds\right)^2 dt = 1$$

This transformation should act on the mean and variance to improve the discriminant power of the curves.

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 misclassifying a future observation and it is selected according to the cross-validation method [29]. This method consists in minimizing the cross-validation function:

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

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

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

It can be useful and efficient to replace the $h$ parameter, a real number, by an integer parameter $k$ from a finite subset.

A way to achieve this is to consider a $k$ Nearest Neighbors (kNN) version of the kernel estimator [14]. In this study, it is named KNN-NPFDA method. The number of neighbors and the bandwidth is selected using the cross-validation method.

When the kNN estimator is used, the parameter $h$ is replaced by $h_k$, which is the bandwidth allowing us to take into account $k$ terms in the weighted average [14]:

$$\hat{p}_k^{(g)}(x) = \frac{\sum_{\{i:Y_i=g\}}^{n} K\left(h_k^{-1} \cdot d(x, X_i)\right)}{\sum_{i=1}^{n} K\left(h_k^{-1} \cdot d(x, X_i)\right)},$$

where $h_k$ is a bandwidth such that

$$\#\{i : d(x, X_i) < h_k\} = k,$$

with # the cardinal of the set.

Two additional nonparametric methods based on the boosting algorithm, the B method and the B-PCA method, are implemented to be compared with the kernel methods. They are specially designed to perform a nonparametric supervised classification for functional data. The boosting algorithm used is the Adaboost algorithm for classification [26].

In the B method, the boosting algorithm estimates the optimal number of basis and the optimum depth of the tree partition using Functional Data Object for obtaining the best possible estimation.

In the B-PCA method, the Adaboost algorithm is applied to a set of data using Principal Component Analysis. The optimal number of Principal Components and the optimum depth for one or more classifiers are estimated.

The free statistical software R [30] is employed to implement the nonparametric functional methods used in this article. Mainly, the R packages `fda` and `fda.usc` are used to perform the classification applying nonparametric functional analysis. Adaboost algorithm is a modified algorithm of the function `adabag.M1` of `adabag` R package that fits adaboost algorithm with classification trees as individual classifiers.

## Results and discussion

In this section, the methods previously presented are applied to the TG and DSC curves to classify between different species and main groups. First, a descriptive analysis of the data is shown. It is important to note that each method is validated through cross-validation, which is the technique widely used for the validation of an empirical model. It works by leaving out one TG/DSC curve; then a model is trained with the remaining thermograms and, finally, the developed model is used for the classification of the left out TG/DSC curve. This is repeated until all the curves have been left out once. As the data set available contain 49 samples, 48 samples are used for training and 1 sample for testing. This process is repeated 49 times, and the percentage (measured as per one) of correct classification are calculated.

### Descriptive analysis of the TG curves

Figure 1 shows the 49 TG curves obtained (7 per class). As can be observed, each curve represents a functional data. A particular wood specie is highlighted in each panel. The last panel (in row 4, column 2) shows all the TG curves. Different trends are observed for almost all types of wood. Even so, the variability in each class prevents from discerning clearly in some intervals. Apart from this, some species such as oak and beech tend to overlap. This is correlated to their similar densities, hardness and mechanical properties. In addition, they belong to the broader group of boreal hardwood [1].

In Fig. 2a, the means of the TG curves for each class of wood, as defined by Fraiman and Muniz [31], are plotted. They can report on the possible degree of overlapping among the degradation trends for the different species and in what intervals this happens. In fact, it appears that differences between species are starting to take place from 200 °C onwards, coinciding with the beginning of the hemicellulose degradation [22, 25]. These differences become maxima in the range of temperatures where the cited maximum decomposition rate of the cellulose and lignin occurs [13].

As the storage period is long enough (over a year) and the storage conditions of all wood samples are the same, through the TG curves is possible to measure the water absorption capacity of each timber. In fact, it is observed that the height of the first step is slightly different for some species, being able to build two groups: oak, beech, insignis pine and, on the other hand, chestnut, eucalyptus, and jatobá. It can be observed that the existing residue in the range of 400–600 °C is different depending on the species (chestnut and jatobá > eucalyptus > oak, scots pine, beech and insignis pine).

Figure 3a shows the variability in each class. By working with functional data, the variance of the TG curves for a class is not a value, but a curve. The greatest variability occurs in the range of temperatures where the maximum decomposition rate of the cellulose and lignin is produced, according to Yang et al [13]. Therefore, the use of classification methods should work worse at these temperatures (320–370 °C, see Fig. 3a), but this depends on the magnitude of the variability among species in this interval, which is also higher (see mean differences in Fig. 2a). This variability is expected given the heterogeneity of wood [1, 3].

Figure 4 shows the DSC curves obtained for every sample tested, in the same form as in Fig. 1. There is a

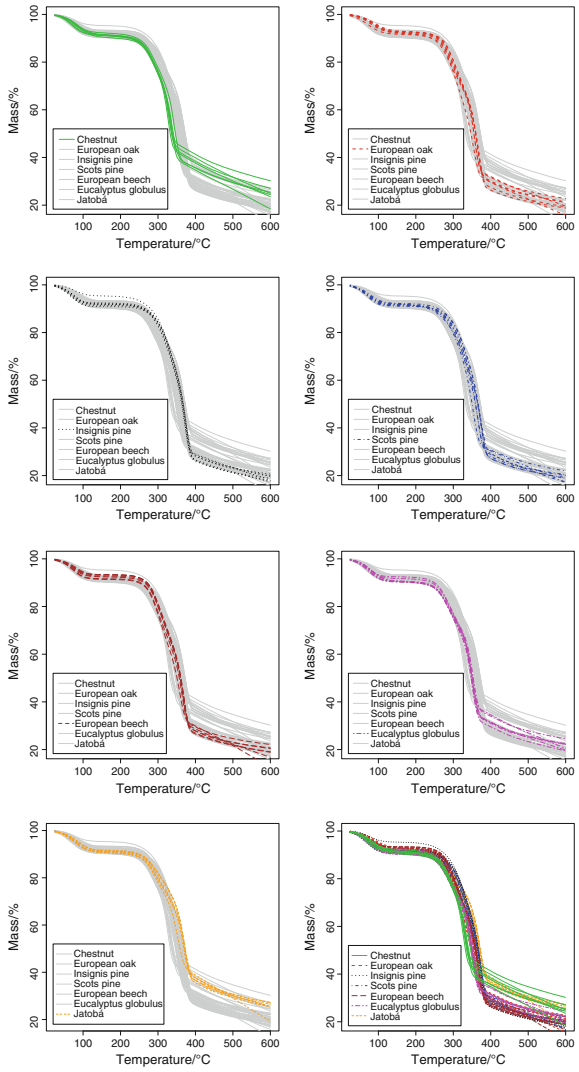


**Fig. 1** Original TG curves (7 per class), where each particular wood specie is *highlighted* in the corresponding *panel*. In the last *panel* (row 4, column 2) all the *curves* are presented

wide variability in each class, higher than in the TG curves. Therefore, it is expected that the DSC curves discriminate worse than TG between classes of wood. According to the literature, the endo and exo DSC events of hemicellulose, cellulose, and lignin overlap in the 220–520 °C range [13, 22, 25]. The differences among species can be set

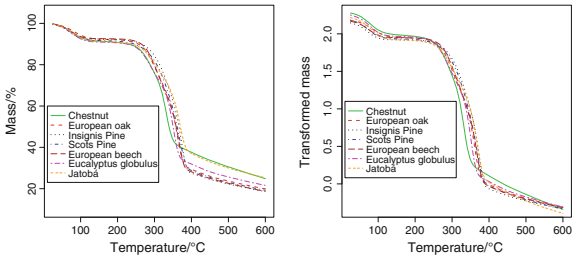


**Fig. 2** **a** Functional means of the original TG curves for each specie (*left panel*). **b** Functional means of the location-scale transformed TG curves (*right panel*)

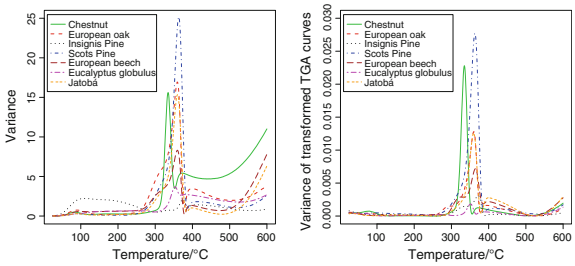


**Fig. 3** **a** Functional variances of the original TG curves for each specie (*left panel*). **b** Functional variances of the location-scale transformed TG curves (*right panel*)

according to the displacement and magnitude of these three previous peaks (see Fig. 5), which may be due to different weight percentage of hemicellulose, cellulose and lignin, interactions, or differences in the structure [19, 20, 22, 23]. There are also differences in the peak corresponding to the water loss (see Fig. 5). The classification methods implemented may determine whether these differences are reliable.

By implementing a location-scale transformation, a reduction of the variance in each class is sought. Figures 6 and 7 show the TG and DSC data after this processing step. A reduction in the dispersion of the curves in each class is observed (Fig. 3). There is also an increasing distance between curves from different classes in certain temperature ranges.

## Result of the data transformation

We apply nonparametric kernel methods for functional data (K-NPFDA and KNN-NPFDA) and methods based on boosting algorithm (B and B-PCA). It is observed that the transformation of the data [28] significantly helps to distinguish among the different groups (hardwoods and softwoods) and also among species. For example, if the TG curves are analyzed by the K-NPFDA method, classifying samples from seven kinds of wood, then a probability of correct classification equal to 0.79 is obtained in the best of possible intervals (180–330 °C); on the other hand, if we use the transformed data, then a probability of 0.88 is obtained in the range 192.5–292.5 °C This is repeated for all models and data analysis; therefore, the results using the transformed data are shown, see Table 1.

## TG curve classification

In Table 2 the probabilities of correct classification and the temperature ranges for which they are maxima are shown. They are computed in two settings, classifying among the seven different species and in the more general case of classifying into three different groups. This is the result of evaluating the probabilities at 1,280 intervals of eight different sizes, from 50 to 400 °C. The four methods of classification described above are calculated.

The results in Table 2 can be grouped in two blocks, those techniques based on a kernel and the Nadaraya–Watson estimator and those based on boosting algorithm. In the case of classification in three groups or general classes

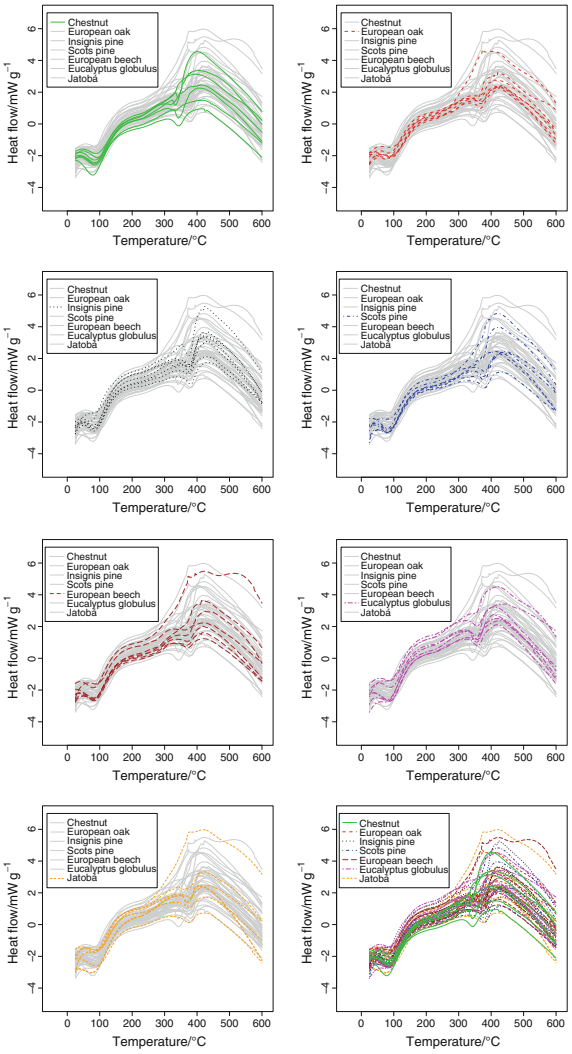


**Fig. 4** Original DSC curves (7 per class), where each particular wood specie is highlighted in the corresponding *panel*. In the last *panel* (row 4, column 2) all the *curves* are presented

of wood, the K-NPFDA (with $0.0005 < h < 0.001$) and KNN-NPFDA techniques, with an optimal number of neighbors equal to 1, are the best methods with an estimated probability of success of almost equal to 1 (0.94). It is interesting to note that the four methods practically coincide in the optimal range (192.5–292.5 °C and 210–310 °C), which in turn coincides with the region of hemicellulose degradation reported by several authors [22, 25]. The

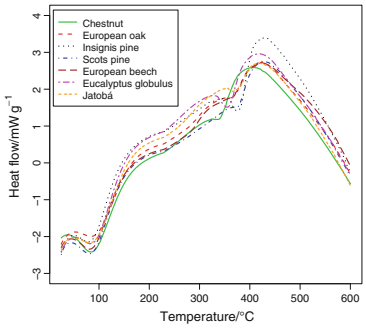


**Fig. 5** Functional means of DSC curves

hardwoods tend to have a higher content of hemicellulose [3, 19, 25] and their weight proportion influences the total degradation of wood. Classification between hardwoods and conifers has been successfully achieved. Also, the correct classification among austral, boreal and tropical hardwoods is obtained. In fact, there is very small confusion between the groups. In Table 3, the classification matrices in the optimal intervals presented in Table 2 using 3 different classes (boreal hardwoods, softwoods and other hard woods) are shown. It can be observed that using the K-NPFDA method the probability of classifying a boreal as a tropical or austral hardwood is only 0.05 and the probabilities of correct classification in each group are very high in this case (0.95 for boreal hardwoods, 0.86 for softwoods and 1 for other hardwoods). This result is similar to that obtained by other techniques using image- and spectrum-based processing systems [2, 9–13].

The overall probabilities of correct classification when one wants to discriminate among the seven existing species of wood is also very high. Especially interesting are the results obtained using the methods K-NPFDA and B (7 elements in the basis and depth of tree equal to 3). In the first case, a probability of correct classification of 0.88 for the interval 192.5–292.5 °C is obtained (see Table 2). This could be due to the different hemicellulose content and differences in hemicellulose degradation depending on the species, but we must do more experiments to prove it. The optimal interval (217.5–417.5 °C) obtained by the method B includes the degradation processes of the hemicellulose, cellulose and lignin getting a slightly higher probability of correct classification. In fact, according to Müller [20], the differences in wood species are mainly due to the different thermochemical behavior of lignin degradation and that of the first step of the hemicellulose degradation.

Table 4 shows the classification matrices in the optimal intervals presented in Table 2 using seven different classes. It can be observed that the results obtained by Methods B and K-NPFDA are complementary. The probabilities of correct classification of each kind of wood are relatively high in both cases. The K-NPFDA method only fails to discriminate the scots pine class ($P = 0.58$). Instead, a probability of correct classification equal to 1 of the scots pine class is obtained by the B method. Among other things, it can be due to the much larger optimal temperature range, encompassing much of the lignin degradation which provides more information to differentiate the species. On the other hand, the K-NPFDA method classifies slightly better oak and jatobá woods than the B method (see Table 4).

In general, it appears that the cited range of decomposition of the hemicellulose and, to a lesser extent, of lignin and cellulose is the temperature range where more differences between species exist.

Apart from using leave-one-out cross-validation, the prediction power of the K-NPFDA method is measured. For this, the whole set of 49 curves is divided into two groups: a training sample of 42 curves, and a test sample of seven curves (one for each class of timber). Our aim is try to classify correctly the test sample using the training sample. This problem is more common in industry. Table 5 shows the classification matrix obtained when classifying among the seven considered species. In this table, the results in the temperature interval of 207–307 °C are shown. This is the optimal interval using cross-validation with the training sample. It can be observed that the 100% of the test sample is successfully classified using the K-NPFDA method. Same results for the case of classifying among the three main groups are shown in Table 6 (in the interval 192.5–292.5 °C), obtaining the same success.

In conclusion, it is proved that using TG curves as discriminant characteristic is possible to classify different species of wood.

## DSC curves classification

Tables 7, 8, and 9 show similar results to those presented in Tables 2, 3, and 4, respectively, but using the DSC curves obtained by a simultaneous SDT. It can be observed that DSC curves have less discriminating power than TG curves. Nevertheless, very good results are obtained when we try to distinguish among boreal hardwoods, softwoods, and tropical or austral hardwoods (Tables 7 and 8). The kernel nonparametric functional methods (K-NPFDA and KNN-NPFDA) work better than the others based on the boosting algorithm. A good classification probability equal to 0.80 is obtained in the temperature range of 322.5–485 °C. This

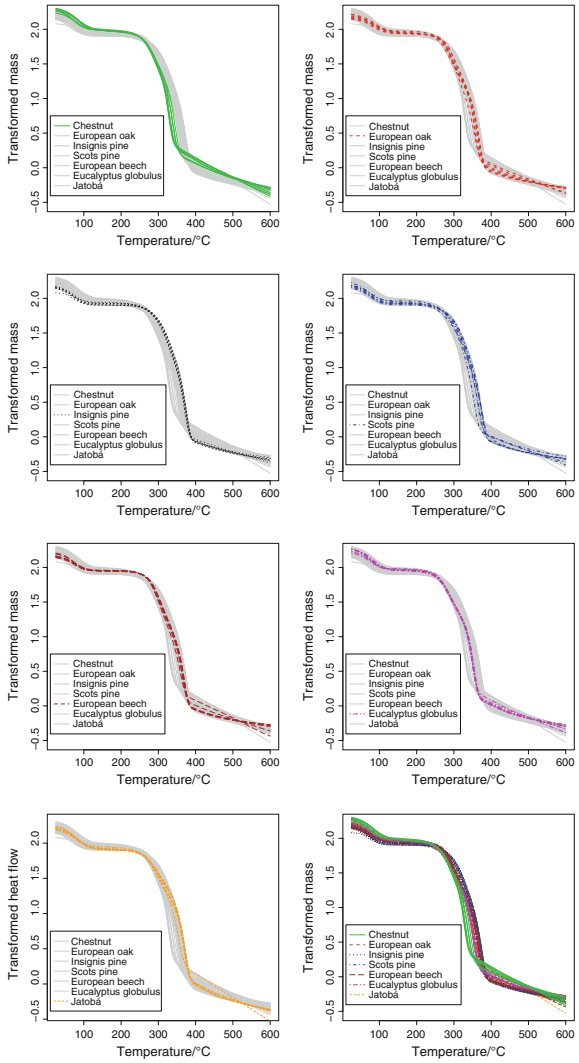


**Fig. 6** Location-scale transformed TG curves

interval corresponds to the region of maximum degradation rate of cellulose and lignin, reported by several authors [13, 22, 25]. Therefore, the DSC curves have a higher classification power in this area; there are more differences between species. These differences may be due to the nature and weight percentage of lignin [11].

When we want to classify among the seven species, the DSC results are worse than the TG ones. The best

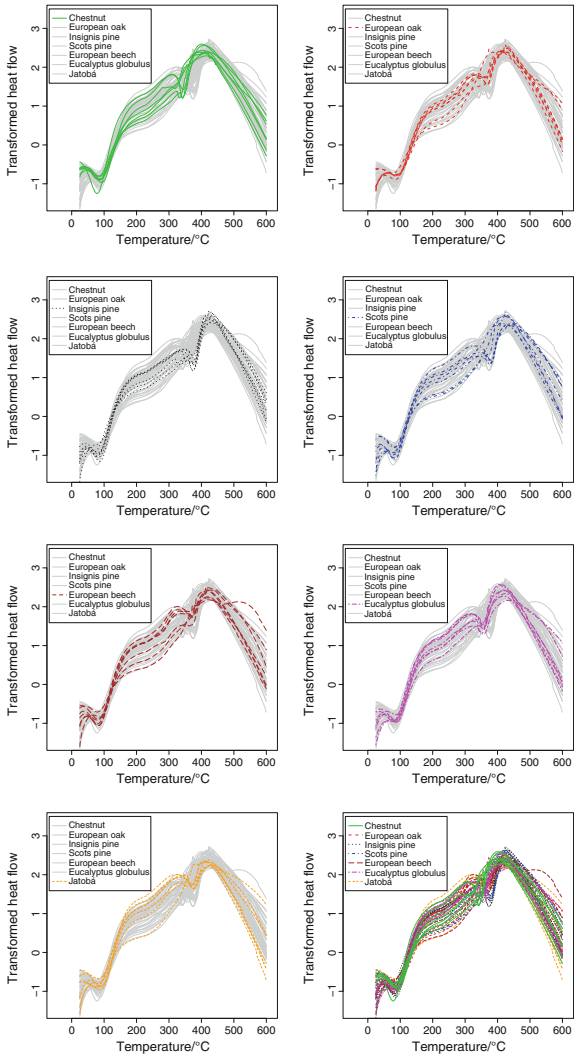


**Fig. 7** Location-scale transformed DSC curves

methods are again K-NPFDA and KNN-NPFDA with a maximum probability of correct classification equal to 0.60 for the interval 322.5–347.5 °C Thus, using the K-NPFDA method, good classification results are only obtained in the case of eucalyptus, chestnut and jatobá (Table 9) but the two kinds of pine are often misclassified and the beech curves are often classified as oak curves ($P = 0.29$). In fact, both pairs of species have very similar mechanical properties, hardness and density [1].

**Table 1** Correct classification probabilities using original and transformed data, with 3 or 7 classes

| Data | Original data | | Transformed data | |
|---|---|---|---|---|
| | Number of classes | Corr. class. prob. | Number of classes | Corr. class. prob. |
| TG | 3 | 0.94 | 3 | 0.94 |
| TG | 7 | 0.79 | 7 | 0.88 |
| DSC | 3 | 0.42 | 3 | 0.79 |
| DSC | 7 | 0.24 | 7 | 0.57 |

**Table 2** Correct classification probabilities and optimal intervals obtained by each classification method

| Methods | 3 Groups classification | | 7 Groups classification | |
|---|---|---|---|---|
| | Optimal interval/°C | Corr. class. prob. | Optimal interval/°C | Corr. class. prob. |
| K-NPFDA | 192.5–292.5 | 0.94 | 192.5–292.5 | 0.88 |
| KNN-NPFDA | 192.5–292.5 | 0.94 | 182.5–282.5 | 0.88 |
| B | 210.0–310.0 | 0.90 | 217.5–417.5 | 0.90 |
| B-PCA | 210.0–310.0 | 0.90 | 195.0–345.0 | 0.83 |

The TG data were tested with 3 (boreal hardwoods, softwoods, other hardwoods) and seven classes

**Table 3** Classification matrices using 3 different classes (boreal hardwoods, softwoods, and other hard woods) obtained by different nonparametric classification methods, using TG data

| Methods | Estimated | Theoretical | | |
|---|---|---|---|---|
| | | Boreal hardwoods | Softwoods | Other hardwoods |
| K-NPFDA | Boreal hardwoods | 0.95 | 0.07 | 0.00 |
| | Softwoods | 0.00 | 0.86 | 0.00 |
| | Other hardwoods | 0.05 | 0.07 | 1.00 |
| B | Boreal hardwoods | 0.97 | 0.00 | 0.23 |
| | Softwoods | 0.00 | 1.00 | 0.06 |
| | Other hardwoods | 0.03 | 0.00 | 0.71 |

**Table 4** Classification matrices using seven different classes (chestnut, oak, insignis pine, scots pine, beech, eucalyptus, and jatobá) obtained by two different nonparametric classification methods, using TG data

| Methods | Estimated | Actual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Chesn. | Oak | Insig. P. | Scots P. | Beech | Eucal. | Jat. |
| K-NPFDA | Chestnut | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Oak | 0.00 | 0.72 | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 1.00 | 0.14 | 0.00 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 0.58 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.14 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.14 | 0.00 | 0.00 | 0.00 | 1.00 | 0.14 |
| | Jatobá | 0.00 | 0.00 | 0.00 | 0.14 | 0.00 | 0.00 | 0.86 |
| B | Chestnut | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Oak | 0.00 | 0.67 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.17 |
| | Jatobá | 0.00 | 0.33 | 0.00 | 0.00 | 0.00 | 0.00 | 0.83 |

**Table 5** Classification matrix using seven different classes (chestnut, oak, insignis pine, scots pine, beech, eucalyptus, and jatobá) obtained by K-NPFDA using a training sample with 42 TG curves

| 207–307 °C, K-NPFDA | Estimated | Actual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Chesn. | Oak | Insig. P. | Scots P. | Beech | Eucal. | Jat. |
| New sample $P = 1$ | Chestnut | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Oak | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.00 | 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 | 1.00 |

Probabilities of correct classification of a new sample consisting of seven curves (one per class)

**Table 6** Classification matrix using three different classes (boreal hardwoods, softwoods and tropical, and austral hardwoods) obtained by K-NPFDA using a training sample with 42 TG curves

| 192.5–292.5 °C, K-NPFDA | Estimated | Actual | | |
|---|---|---|---|---|
| | | Boreal hardwoods | Softwoods | Other hardwoods |
| New sample $P = 1$ | Boreal hardwoods | 1.00 | 0.00 | 0.00 |
| | Softwoods | 0.00 | 1.00 | 0.00 |
| | Other hardwoods | 0.00 | 0.00 | 1.00 |

Probabilities of correct classification of a new sample consisting of three curves

**Table 7** Correct classification probabilities and optimal intervals obtained by each classification method

| Methods | 3 Groups classification | | 7 Groups classification | |
|---|---|---|---|---|
| | Optimal interval/°C | Corr. class. prob. | Optimal interval/°C | Corr. class. prob. |
| K-NPFDA | 322.5–485.0 | 0.80 | 322.5–472.5 | 0.57 |
| KNN-NPFDA | 322.5–485.0 | 0.80 | 322.5–347.5 | 0.60 |
| B | 330.0–575.5 | 0.60 | 325.0–375.0 | 0.46 |
| B-PCA | 330.0–575.5 | 0.67 | 325.0–375.0 | 0.38 |

The DSC data were tested using three (boreal hardwoods, softwoods, other hardwoods) and seven classes

**Table 8** Classification matrix using three different classes (boreal hardwoods, softwoods, and other hard woods) obtained by different non-parametric classification methods, using DSC curves

| Methods | Estimated | Theoretical | | |
|---|---|---|---|---|
| | | Boreal hardwoods | Softwoods | Other hardwoods |
| K-NPFDA | Boreal hardwoods | 0.90 | 0.29 | 0.21 |
| | Softwoods | 0.05 | 0.71 | 0.00 |
| | Other hardwoods | 0.05 | 0.07 | 0.79 |

In Table 10 we measure the prediction power of a new sample taking a training sample of 42 items. The new sample consists of seven curves, six of which have been classified successfully ($P = 0.86$). A higher misclassification is obtained precisely in the most heterogeneous group (other hardwoods).

**Table 9** Classification matrix using 7 different classes (chestnut, oak, insignis pine, scots pine, beech, eucalyptus, and jatobá) obtained by K-NPFDA method, using DSC data

| Methods | Estimated | Actual | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Chesn. | Oak | Insig. P. | Scots P. | Beech | Eucal. | Jat. |
| K-NPFDA | Chestnut | 0.86 | 0.14 | 0.00 | 0.13 | 0.00 | 0.00 | 0.14 |
| | Oak | 0.00 | 0.57 | 0.14 | 0.00 | 0.14 | 0.14 | 0.14 |
| | Insignis P. | 0.00 | 0.00 | 0.29 | 0.29 | 0.14 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.00 | 0.43 | 0.29 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.29 | 0.14 | 0.29 | 0.43 | 0.00 | 0.00 |
| | Eucalyptus | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.86 | 0.00 |
| | Jatobá | 0.14 | 0.00 | 0.00 | 0.00 | 0.29 | 0.00 | 0.72 |

**Table 10** Classification matrix using three different classes (boreal hardwoods, softwoods and tropical, and austral hardwoods) obtained by K-NPFDA using a training sample with 42 DSC curves

| 322.5–485.5 °C K-NPFDA | Estimated | Actual | | |
|---|---|---|---|---|
| | | Boreal hardwoods | Softwoods | Other hardwoods |
| New sample $P = 0.86$ | Boreal hardwoods | 1.00 | 0.00 | 0.50 |
| | Softwoods | 0.00 | 1.00 | 0.00 |
| | Other hardwoods | 0.00 | 0.00 | 0.50 |

Probabilities of correct classification of a new sample consisting of three curves

## Conclusions

Classifying different species of wood using the TG curves as discriminant characteristic has been proved possible (percentage of correct classification = 90%). Also, the classification between hardwoods, softwoods, and tropical or austral hardwoods have been successfully carried out using these curves (percentage of correct classification = 94%). The results are comparable to those obtained from image-based processing systems and spectrum-based processing systems. It was observed that the temperature ranges corresponding to the higher probabilities of correct classification basically match with those reported for the single components decomposition (mainly hemicellulose).

The DSC curves obtained by a simultaneous SDT have less discriminant power than that of the TG curves. Nevertheless, using these curves, very good results are obtained when we try to distinguish among boreal hardwoods, softwoods, and tropical or austral hardwoods (percentage of correct classification = 80%) and among certain types of hardwoods (chestnut, jatobá, and eucalyptus). Moreover, the referenced temperature range corresponding to the maximum rate of decomposition of lignin and cellulose is the range where more differences among species were found, using DSC curves.

In general, K-NPFDA and KNN-NPFDA methods, based on the nonparametric Nadaraya–Watson functional regression estimator, have provided probabilities of correct classification superior to the others based on the boosting algorithm, and with a shorter computing time.

**Acknowledgements** This study was supported by the Ministry of Education and Science MTM2008-00166 (ERDF included), and by Xunta de Galicia, under Grant No.PGIDIT07PXIB105259PR. The authors wish to express special thanks to Manuel Febrero Bande for his valuable comments.

## References

1. Guindeo Casasús A, García Esteban L, Peraza Sánchez F, Arriaga Martitegui F. Especies de maderas. Madrid: Asociación de investigación técnica de las industrias de la madera ycorcho (AITIM); 1997.
2. Lewis IR, Daniel NW, Chaffin NC, Griffiths PR. Raman spectrometry and neural networks for the classification of wood types-1. Spectrochim Acta A-Mole Biomole Spectrosc. 1994;50: 1943–58.
3. Miller RB. Structure of wood. In: Wood handbook: Wood as an engineering material. Madison, WI: Woodhead Publishing Limited, Department of Agriculture, Forest Service, Forest Products Laboratory; 1999.
4. Cavalin P, Oliveira LS, Koerich AL, Britto AS. Wood defect detection using grayscale images and an optimized feature set. In: Proceedings IEEE Ind. Electron (IECON). Singapore: World Scientific; 2006. pp. 3408–12.
5. Fuentealba C, Simon C, Choffel D, Chawentier P, Massons D. Wood products identification by internal characteristics readings. In: Proceedings IEEE-ICIT. Hammamet: IEEE; 2004.
6. Gu IY, Andersson H, Vicen R. Automatic classification of wood defects using support vector machines. In: Bole L, Kulikowski

JL, Wojciechoswski K, editors. ICCVG 2008, Lecture Notes in Computer Science. Berlin Heilderberg: Springer-Verlag; 2009. pp. 356–67.
7. Lampinen J, Smolander S, Korhonen M. Wood surface inspection system based on generic visual features. In: Fogelman F, Gallinari SP, editors. Industrial applications of neural networks. Paris: World Scientic; 1995. pp. 35–42.
8. Watanabe K, Hart JF, Mansfield SD, Avramidis S. Near-infrared technology applications for quality control in wood processing. In: Ridley-Ellis DJ, Moore JR, editors. Proceedings of the final conference of COST Action E53, quality control for wood & wood products. Edinburgh, UK: Forest Products Research Institute/Centre for Timber Engineering, Edinburgh Napier University; 2010.
9. Khalid M, Lee ELY, Yusof R, Nadaraj M. Design of an intelligent wood species recognition system. Int J Simul Syst Sci Technol. 2008;9:9–19.
10. Lavine BK, Davidson CE, Moores AJ, Griffiths PR. Raman spectroscopy and genetic algorithms for the classification of wood types. Appl Spectrosc. 2001;55:960–66.
11. Nuopponen MH, Birch GM, Sykes RJ, Lee SJ, Stewart DJ. Estimation of wood density and chemical composition by means of diffuse reflectance mid-infrared fourier transform (DRIFT-MIR) spectroscopy. J Agric Food Chem. 2006;54:34–40.
12. Piuri V, Scotti F. Design of an automatic wood types classification system by using fluorescence spectra. IEEE Trans Syst Man Cybern C-Appl Rev. 2010;40:358–66.
13. Yang H, Lewis IR, Griffiths PR. Raman spectrometry and neural networks for the classification of wood types. 2. Kohonen self-organizing maps. Spectrochim Acta A-Mole Biomol Spectrosc. 1999;55:2783–91.
14. Ferraty F, Vieu P (2006) Nonparametric functional data analysis. Berlin: Springer-Verlag.
15. Ramsay JO, Silverman BW. Functional data analysis 2nd ed. New York, Springer-Verlag; 2005.
16. Ramsay JO, Silverman BW. Applied functional data analysis. New York: Springer-Verlag; 2002.
17. Alén R, Kuoppala E, Pia O. Formation of the main degradation compound groups from wood and its components during pyrolysis. J Anal Appl Pyrolysis. 1996;36:137–48.
18. Gašparovič L, Koreňová Z, Jelemenský L. Kinetic study of wood chips decomposition by TGA. Chem Pap. 2009;64:174–81.
19. Grønli MG, Várhegyi G, Blasi C. Thermogravimetric analysis and devolatilization kinetics of wood. Ind Eng Chem Res. 2002; 41:4201–08.
20. Müller-Hagedorn M, Bockhorn H, Krebs L, Müller U. A comparative kinetic study on the pyrolysis of three different wood species. J Anal Appl Pyrolysis. 2003;68–69:231–49.
21. Raveendran K, Ganesh A, Khilar KC. Pyrolysis characteristics of biomass and biomass components. Fuel. 1996;75:987–98.
22. Roberts AF. A review of kinetics data for the pyrolysis of wood and related substances. Combust Flame. 1970;14:261–72.
23. Wang S, Wang K, Liu Q, Gu Y, Luo Z, Cen K, Fransson T. Comparison of the pyrolysis behavior of lignins from different tree species. Biotechnol Adv. 2009;27:562–7.
24. Korošec RC, Lavrič B, Rep G, Pohleven F, Bukovec P. Thermogravimetry as a possible tool for determining modification degree of thermally treated Norway spruce wood. J Therm Anal Calorim. 2009;98:189–95.
25. Mohan D, Pittman JCU, Steele PH. Pyrolysis of wood/biomass for bio-oil: a critical review. Energy Fuel 2006;20:848–89.
26. Bühlmannand P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. Stat Sci. 2007;22:477–505.
27. Prime RB, Bair HE, Gallagher PK, Riga A. Thermogravimetric analysis (TGA). In: Menczel JD, Prime RB, editors. Thermal analysis of polymers Fundamentals and applications. San José, CA: Wiley; 2009. pp. 7–240.
28. López-Granados F, Peña Barragán JM, Jurado-Expósito M, Francisco-Fernández M, Cao R, Alonso-Betanzos A, Fontenla-Romero O. Multispectral classification of grass weeds and wheat (*Triticum durum*) using linear and nonparametric functional discriminant analysis and neural networks. Weed Res. 2008;48: 28–37.
29. Naya S, Cao R, Artiaga R, García A. New method for material classification from TGA data by nonparametric regression. Mater Sci Forum. 2006;514–516:1452–6.
30. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2008. http://www.R-project.org. Accesed 1 Sep 2010.
31. Fraiman R, Muniz G. Trimmed means for functional data. Test. 2001;10:419–40.

# Apéndice D

# Classification of wood micrographs by image segmentation

Chemometrics and Intelligent Laboratory Systems 107 (2011) 351–362

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Classification of wood micrographs by image segmentation

Abhirup Mallik [a], Javier Tarrío-Saavedra [b], Mario Francisco-Fernández [c,*], Salvador Naya [b]

[a] Indian Institute of Technology, Department of Mathematics, Kharagpur 721302, India
[b] University of A Coruña, Higher Polytechnic School. Campus de Esteiro, Ferrol 15403, Spain
[c] University of A Coruña, Faculty of Computer Science, Campus de Elviña, s/n, A Coruña 15071, Spain

ARTICLE INFO

*Article history:*
Received 10 February 2011
Received in revised form 14 April 2011
Accepted 8 May 2011
Available online 14 May 2011

*Keywords:*
Wood
Image segmentation
Supervised classification
Machine learning
Scanning electron microscopy

ABSTRACT

The principal aim of this study is to classify wood species using scanning electron microscopy (SEM) micrographs obtained with 1500× magnification and processed by image segmentation. The results show that it is possible to observe differences among species in the wood texture at this magnification. The micrographs have been processed in a simple way using segmentation and object recognition to identify the cross-section tracheids belonging to earlywood of 7 different timber species: *Fagus sylvatica*, *Castanea sativa*, *Juglans regia*, *Eucalyptus globulus*, *Hymenaea courbaril*, *Pinus silvestris* and *Pinus radiata*. We have analyzed the shape, number and distribution of the tracheids using 5 features: circularity, rectangularity, number of tracheids, distance between tracheids and average area. The extracted features are classified using different statistical methods: Linear Discriminant Analysis (LDA), Quadratic classification, Logistic regression, *K* Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machines (SVM) and Neural Networks. A comparative study using gray level co-occurrence based features is also presented, with the improvement of using the segmentation method. Moreover, some additional results showing the possibility of using fractal analysis in this framework complete the research.



## 1. Introduction

Whether a piece of furniture made from an unfamiliar wood has to be restored or you are debating the authenticity of a particular board with a local lumberyard, a knack for identifying a piece of lumber is a useful skill. This skill becomes very important if we transfer these cases to the industry. Depending on the specie to which it belongs, a timber has a certain physical–chemical properties determining its industrial applications. Thus, given its properties, there are woods suitable for the development of flats, others for the manufacture of structural elements in a building, some for cabinet or furniture of different qualities, etc. Uses or applications are not always interchangeable. In addition to the own properties of the wood, such as chemical resistance, tensile, bending, compression strength, hardness, elastic modulus, density, porosity, odor, etc., another important factor determining the application or not of a particular specie of wood is the price. Therefore, identifying types of wood becomes crucial to check a possible fraud as some timber traders tend to mix different types of wood to increase their profit margin [1]. It is also the case when the cheaper timber (similar at first glance) is directly sold as it were the more expensive one. Often the identification of the actual building material is the most daunting task for a furniture researcher, a construction foreman or even for an archeologist [2]. In fact, the identification of wood is one of the most difficult tasks to perform related with the technology of this material, due to the wide variety of species and anatomical heterogeneity of its elements. Wood identification can often be made on the basis of readily visible characteristics such as color, odor (usually caused by oils in the heartwood), density, presence of pitch, grain pattern, texture (depending on the size and distribution of its cell), type of transition from earlywood to latewood, presence of resin canals, or pores. The principal standard tool for macroscopic viewing of wood is 10× hand lens. Very often the identification can be performed using these lenses and some identification keys from books [3–5]. This analysis is typical in the furniture industries and the wood panel production. However, many woods are impossible to tell apart without using a microscope. Sometimes a great deal of knowledge and laboratory equipment is needed to identify species [3–8]. But even having the best equipment, many times the performed analysis has a non-uniform accuracy due to the operator. Correctly identifying an unfamiliar wood sample out of thousands of possibilities requires close observation, and a thorough knowledge of wood and its properties. Training a skilled worker takes years, with the corresponding cost, and the industry trend for automatization has meant to dispense with the traditional trades. Accordingly, these workers are increasingly scarce. Therefore, the implementation of statistical models and automatic recognition methods of wood samples are justified and can be immediately useful. While there are various computational procedures to evaluate and rate the quality of a timber inspecting its defects by image processing techniques and spectral analysis [6,9–13], these are not so generally

* Corresponding author. Tel.: +34 981167000x1222; fax: +34 981167160.
*E-mail address:* mariofr@udc.es (M. Francisco-Fernández).

0169-7439/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2011.05.005

used for species identification, although there are also several works addressing this problem [1,7,14–17]. A first step in a classification problem is to choose a discriminant feature from which it will be possible to classify. In the case of wood species classification, this discriminant feature could be the output of an experimental technique that really differentiates between them. In the literature, wood samples are mainly classified based on the results of two techniques: image-based and spectrum-based processing systems. In [7], the Fourier Transform Raman (FTR) spectroscopy and Neural Network technology have been coupled for spectral feature extraction and non-supervised classification. This represents the first time that both methodologies are combined. Later, Neural Networks and the FTR spectra for hardwoods and softwoods to differentiate temperate woods from tropical woods were also used [17]. Genetic algorithms and principal component analysis were used to classify 98 Raman spectra of temperate softwoods, hardwoods and Brazilian and Honduran tropical woods [14]. Recently, in [16], an automatic wood type classification system based on the analysis of the fluorescence spectra, using Nearest Neighbor classifiers, Linear and Quadratic classifiers, and Support Vectors Machines (SVM) is designed. Another alternative is to classify attending to the thermograms obtained by TGA (thermogravimetric analysis) [18]. These curves can be processed in a relatively simple way with functional analysis methods [19–21] and their shape is directly related to the wood composition. On the other hand, a method of classification of 20 types of tropical timber from image processing, using extracting textural wood features from wood images obtained with 10× magnification has been successfully tested in [1]. They obtained a good classification proportion of 95% using Neural Networks and test samples of 10 items. In [22], different wood species using high spatial resolution infrared color aerial photographs (taken from the tree crowns) are classified. Nine different features of each image object are estimated, and transformed using principal component analysis (PCA). The accuracy, using the supervised grade of membership (GoM) model with cross-validation was 67%. In [23], macroscopic images belonged to six different tropical wood species, taken from cross-sections, are classified obtaining a success of 80% with test samples of 60 items. They apply a rotational invariant method using the gray level co-occurrence matrices (GLCM) as the features, an energy value representing the similarity between the test sample and the template. However, it seems that the possibility of using micrographs with a bigger magnification as a source of data for statistical classification of wood species has not been sufficiently studied yet. Is it possible to observe these differences among species in the wood texture at a magnification of 1500×, attending to the shape, number, and distribution of the tracheids? This work pretends to answer this question. Generally, wood is defined as the set of xylem tissues forming the trunk, roots and branches of woody plants, excluding the bark. The tubular cells size, shape and distribution, along with other anatomical elements such as wood radios, the presence of resin canals or vessels, etc., in addition to the variable proportion of its chemical components, define the different wood species and their properties [6,8,17,24–30]. Also, the different wood types can be generally divided in two broad categories: softwoods or conifers (gymnosperms) and hardwoods (dicot angiosperms), which can be subdivided in boreal, austral and tropical hardwood [6,8]. The difference between softwood and hardwood, and also among species, is readily apparent when viewed under microscope magnification. For both hardwoods and softwoods, the side of the board to be examined is often discussed. In fact, wood has three distinct surfaces: the cross-section (end-grain), the tangential surface, and the radial surface. Moreover, trees growing in temperate areas with a winter season display rings. Growth rings consist of two separate layers where the first, called earlywood, is laid down at the beginning of the growing season and the second one, latewood, is formed toward the end. Earlywood is more porous than latewood [3]. There are many differences in wood texture between these two zones [3,8,31]. Therefore, in this study, only the earlywood region is studied, defining earlywood as wood zones where double cell wall of the tracheids is smaller than the lumen [31]. In the present paper, some features corresponding to micrographs obtained by scanning electron microscopy (SEM) with 1500× magnification are used as discriminant characteristics. These images, taken in cross-sections, can be processed in a relatively simple way using segmentation and object recognition to identify the elemental cells and analyze their shape, number and distribution. The extracted features are classified using Linear Discriminant Analysis (LDA), Quadratic classification, Logistic regression, *K* Nearest Neighbors (KNN), Naïve Bayes, Support Vector Machines (SVM) and Neural Networks. A comparative study using GLCM based features is also presented, showing the effectiveness of the segmentation method using these data. Moreover, some additional results showing the possibility of using fractal analysis in this context complete the research.

Accordingly, the objectives of this study are:

1. Checking that the correct classification of timber species (and also between softwoods and hardwoods) is feasible, using all the features obtained from the segmentation of micrographs taken with an electron microscope at 1500× magnification, where the tracheids structure can be observed.
2. Showing the advantages of image segmentation for this particular case of wood micrographs classification problem.
3. Evaluating the potential of supervised classification methods, such as LDA, Quadratic classification, Logistic regression, KNN, Naïve Bayes, SVM and Neural Networks for distinguishing among European chestnut, European beech, eucalyptus, jatobá, walnut, and Scots and insignis pine on the basis of some features obtained by image segmentation; and additionally, for distinguishing between hardwoods and softwoods.

The content of the paper is as follows. In Section 2, the materials and experimental techniques used to obtain the micrographs are described. In Section 3, the image enhancement, the segmentation process and the classification methods used in the field of statistical analysis of multivariate data are explained. In Section 4, these methods are applied and the results are obtained and analyzed. Finally, Section 5 collects the main conclusions.

## 2. Experimental

Tests for 5 different hardwoods (European beech or *Fagus sylvatica*, chestnut or *Castanea sativa*, common walnut or *Juglans regia*, *Eucalyptus globulus* and jatobá or *Hymenaea courbaril*) and 2 softwoods (Scots pine — *Pinus silvestris* and insignis pine — *Pinus radiata*) are carried out. At least ten samples per each one of the above mentioned species, obtained from commercial wood of different trees are tested (see Table 1). The aim of this sampling process is to obtain a compromise between capturing the existing variability and minimizing the time of experimentation. All samples have been cut using a razor blade so that the cross-section of wood can be studied. This is a common technique used when wood samples are classified by visual

**Table 1**
Number of samples per wood class.

| Type of wood | No. of samples |
|---|---|
| Eucalyptus | 15 |
| European beech | 11 |
| Scots pine | 16 |
| European chestnut | 15 |
| Jatobá | 18 |
| Insignis pine | 15 |
| Walnut | 11 |

inspection using 10× hand lens [3–5]. This method may very slightly affect the structure of the wood, but, in this work, we have wanted to conduct a timber classification based on standard methods for collecting wood samples, minimizing the preparation time, and thus adjusting to the industry requirements. However, this is a destructive method (although required samples are very small) and it is not applicable in case of artworks (or it must be applied very carefully). Each sample has approximately a prismatic shape with dimensions 0.5×0.5×0.1 cm. The samples were dehydrated in a series of graded concentrations of ethanol and critical-point dried using liquid $CO_2$. The dried samples were then affixed on aluminum stubs with adhesives, covered with gold (BAL-TEC SCD004 sputter coater) and examined in a scanning electron microscope (Jeol JSM-6400), using a magnification equal to 1500×.

## 3. Image treatment methodology and statistical methods

### 3.1. Image enhancement

The acquired images (micrographs) are enhanced for better accuracy of the results. The image is cropped for the important part of the picture. Sometimes there can be too much noise in the picture due to higher resolution in the microscope. Image noise is the random variation of color and brightness information in the image. This can occur from several sources including electrical sensor noise, photographic grain noise and channel errors. Noise can be of many types, like amplifier noise (Gaussian noise), salt and pepper noise, shot noise, quantization noise, etc. Image noise arising from a noisy sensor or channel transmission errors usually appears as discrete isolated pixel variations that are not spatially correlated. Pixels that are in error often appear visually to be markedly different from their neighbors [32]. So the noise is reduced by applying a median filter by a 3-by-3 neighborhood to the gray scale image. Median filter is a nonlinear digital noise filtering technique, developed by Tukey [33]. In one-dimensional form, the median filter consists of a sliding window encompassing an odd number of pixels. The center pixel in the window is replaced by the median of the pixels in the window [32]. The intensity values are corrected so that 1% of the pixels are saturated at low and high intensities. By this method the contrast of the picture is enhanced. This is further achieved by histogram equalization of the image. Many times the luminance histogram of the image is skewed towards either brighter or darker side. Histogram equalization is a process for which the histogram of the enhanced image is forced to be uniform. The method is useful in images with backgrounds and foregrounds that are both bright or both dark [34–36]. Fig. 1 shows the effect of enhancement for a Scots pine micrograph.

### 3.2. Segmentation

Image segmentation refers to the process of partitioning a digital image into multiple segments (sets of pixels). The goal is to analyze a part of the image only over the whole image, to infer more accurately. We have chosen here to concentrate on the shapes and distribution of tracheids in the pictures, to segment the image in such a way to get only those parts of the micrographs. There are many methods of segmenting an image. In the present paper, thresholding and edge detection segmentation methods have been used.

#### 3.2.1. Thresholding

The structure of the wood sample is analyzed by identifying the tracheids and their properties and patterns. Here, thresholding is used to get connected objects. As here the micrographs can be characterized as containing *holes* (tracheids) with reasonably uniform brightness against a background of different brightness, luminance can be used as a distinguished feature to segment the tracheids from the rest of the image. We have set an optimal value of the threshold depending on the brightness, and the tracheids are marked if their pixel values are less than that. The value of the threshold can be also automatically calculated [37]. The image is then inverted so that the important parts (tracheids) can be seen as white pixels. There can be a lot of noise coming in by the process of inversion. So, the small objects are removed from the image and then the closed boundaries are identified and filled with white pixels to get complete tracheids.

#### 3.2.2. Edge detection

Edge detection is an important way of image segmentation. Here, we identify the points in an image where the brightness changes abruptly. This is usually detected using difference of the pixels with their surrounding pixels. There are two major classes of differential edge detection: first order and second order derivative. For first order, some form of spatial first order differentiation is performed, and the resulting edge gradient is compared to a threshold value. An edge is judged present if the gradient exceeds the threshold. For the second-order derivative class of differential edge detection, an edge is judged present if there is a significant spatial change in the polarity of the second derivative.

A Sobel edge detector (a well-known first order edge detection method) has been used to find the edges of the tracheids. This operation is necessary because uneven darkness in the tracheids. Therefore, many times thresholding operation misses some parts of them. Those parts need to be filled by detecting their boundaries using edge detection. The detected edges are dilated to get connected boundaries, then we fill the closed areas. The small objects are rejected to reduce noise. We use logical OR to sum two binary images.

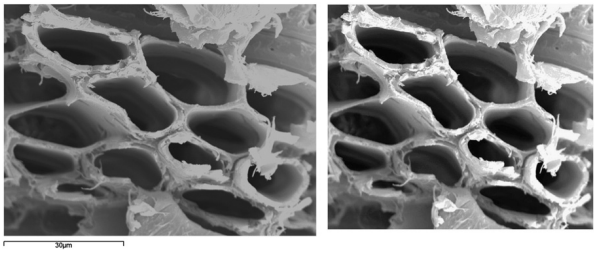


**Fig. 1.** Effect of enhancement for a Scots pine micrograph.

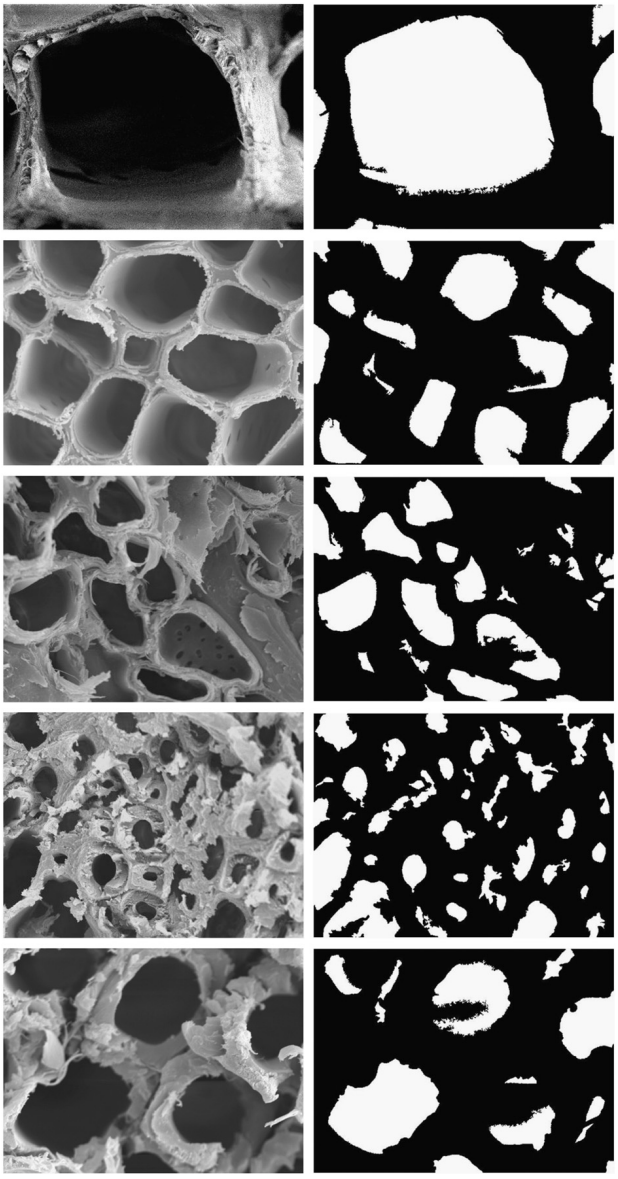

**Fig. 2.** Segmentation process for hardwood species. Row 1: eucalyptus. Row 2: beech. Row 3: chestnut. Row 4: jatobá. Row 5: walnut.

Now, this image can be used for extracting features from the tracheids. Figs. 2 and 3 show the segmentation result for several micrographs of hardwood and softwood species, respectively (each row in these figures corresponds to each one of the considered species in the present paper).

### 3.3. Dilation

Dilation is a morphological operation. Morphological image processing is a type of processing in which the spatial form or structure of objects within an image is modified. With dilation, an object grows uniformly in spatial extent. We needed this because we are losing some of the boundary part of the tracheids in the time of thresholding as the boundary can have higher brightness than the inner part. So to restore the objects to the approximate original shape, dilation is applied.

### 3.4. Features

Now from this image, we identify and label the objects. The following features are extracted from the images:

1. Number of tracheids ($N$) detected in the image. This is a very important feature, as there are clearly different number of tracheids per image for different types of wood.
2. Average circularity of the tracheids. The circularity of the tracheids is measured as:

$$C = \frac{4\pi \times A}{P^2},$$

where $P$ denotes the object perimeter, and $A$ is the area of a tracheid, approximately measured as the number of pixels in the object. For a perfectly round tracheid this should be 1, and less than 1 for any other kind of shape.

3. Average rectangularity of the tracheids.
The rectangularity of the tracheids is measured as:

$$R = \frac{A}{A_r},$$

where $A_r$ is the area of the surrounding rectangle. This index will be 1 for a perfect rectangle and less than 1 for other shape.
4. Average area per tracheid.
This gives the measure of the size of the tracheids. It is a distinctive feature for the different kinds of wood micrographs.
5. Average distance between the tracheids.
This is measured by the perimeter of the polygon formed by joining the centers of the tracheids in the image. This gives the information about the distribution and spacing between the tracheids in the micrograph. This is also a very distinctive feature for the images.

The objects are chosen on the basis of these features. As very small, or objects with very small circularity are unlikely to be tracheids, we reject them by setting some thresholds. Fig. 4 presents an example of segmentation and some of the features extracted for a Scots pine micrograph.

### 3.5. Classification

A classification task usually involves training and testing data consisting of some observed instances. Each instance in the training set contains one target value or class label and several attributes or features. The goal of classification methods is to produce a model to predict class labels of data instances in the testing set for which only the attributes are known. In a first stage, each method compared in our study is validated through leave-one-out cross-validation. This is a technique widely used for the validation of an empirical model. It works by leaving out one instance (the testing sample); then a model is trained with the remaining samples and, finally, the developed model is used for the classification of the left out instance. This is repeated until all the micrographs have been left out once. As the data set available contains 101 samples, 100 samples are used for training

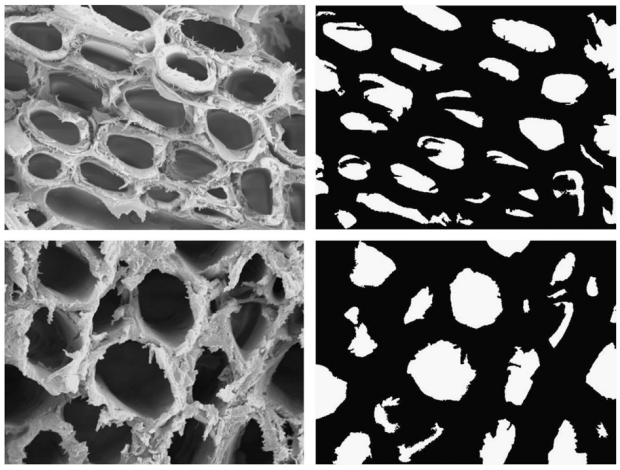

**Fig. 3.** Segmentation process for softwood species. Row 1: Scots pine. Row 2: insignis pine.

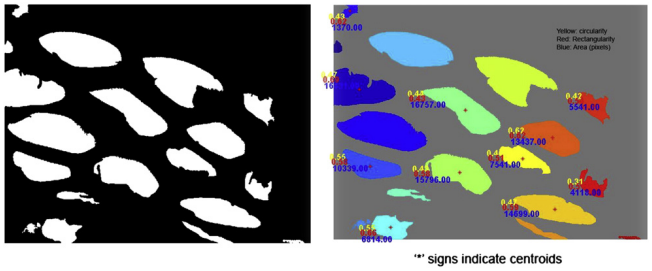


**Fig. 4.** Segmentation and some extracted features for a Scots pine micrograph.

and 1 sample for testing. This process is repeated 101 times, and the percentages (measured as per one) of correct classification are calculated. Being more prone to lead to overfitting, i.e. minimum perturbation in the training set, leave-one-out cross-validation has a higher probability of resulting in mispredictions on external samples than 10-fold cross-validation; however, the probabilities of wrong assignment inside the training samples are lower. This is an important fact to consider since compare directly the results obtained by both methods can be misleading. To complete the study, an external validation process is also carried out. There are many supervised classification methods performed to work with multivariate data, such as traditional Linear Discriminant Analysis (LDA), Quadratic classification and Logistic regression; besides advanced statistical methods belonging to Machine Learning as Support Vector Machines (SVM), *K* Nearest Neighbors (KNN), Naïve Bayes, Classification Trees, Neural Networks or classification methods based on Adaboost algorithm. In this study, we have applied LDA, Quadratic, Logistic, Naïve Bayes, KNN, SVM and Neural Networks classifiers on the feature set, previously scaled. A comparative study is shown in the following section.

The most common classification model used in this study is the LDA, where a linear classifier is built among the different classes assuming normal densities and equal covariance matrices in the input data. For more information see [38–40].

If accepting the normality of the observations and the hypothesis of equal variances were not admissible, the procedure for solving the problem is to classify the observation in the group with maximum posterior probability. This is the case of Quadratic classification that use a second-order mapping of the input [38,39].

The Logistic model is applied to a wide range of situations where the explicative variables do not have a multivariate normal distribution [39–41]. Considering only two classes ($C_1$ and $C_2$), the Logistic regression equation is the following:

$$\log\frac{p}{1-p} = \alpha + \beta' x,$$

or

$$p = \frac{\exp(\alpha + \beta' x)}{1 + \exp(\alpha + \beta' x)},$$

where $p = P(Y = C_1|x)$ is the posterior probability of $Y$ equal to $C_1$, $\log\left(\frac{p}{1-p}\right)$ is the logit transformation, $x$ is the p-dimensional vector of predictor variables, $\beta$ is a vector of $p$ parameters and $\frac{p}{1-p}$ the odds ratio. The logit model can be generalized to more than two populations, i.e. for qualitative response with more than two possible levels. If we suppose $G$ populations, then, defining $p$ as the probability that the observation $i$ belongs to the class $g$, it is possible to write:

$$p_{ig} = \frac{\exp\left(\beta_{0g} + \beta'_{1g} x_i\right)}{1 + \sum_{j=1}^{G-1} \exp\left(-\beta_{0j} - \beta'_{1j} x_i\right)}, \quad j = 1, \ldots, G-1$$

Therefore, we can say that the posterior probabilities, $p_{ig}$, satisfy a multivariate logistic distribution. The comparison between two categories is made in the usual way

$$\frac{p_{ig}}{p_{ij}} = \frac{\exp\left(\beta_{0g} + \beta'_{1g} x_i\right)}{\exp\left(\beta_{0j} + \beta'_{1j} x_i\right)}.$$

The Naïve Bayes classifier technique is based on the Bayes' theorem and is particularly suited when the dimensionality of the inputs is high. Given a set of variables, $X = \{x_1, x_2, \ldots, x_d\}$, we want to construct the posterior probability for the event $C_j$ among a set of possible outcomes $C = \{c_1, c_2, \ldots, c_d\}$. Using the Bayes rule [40,42]:

$$p\left(C_j|x_1, x_2, \ldots, x_d\right) \propto p\left(x_1, x_2, \ldots, x_d|C_j\right) p\left(C_j\right),$$

where $p(C_j|x_1, x_2, \ldots, x_d)$ is the posterior probability of class membership, i.e., the probability that $X$ belongs to $C_j$. Since Naïve Bayes assumes that the conditional probabilities of the independent variables are statistically independent, we can decompose the likelihood as a product of terms:

$$p\left(X|C_j\right) \propto \prod_{k=1}^{d} p\left(x_k|C_j\right)$$

and rewrite the posterior probabilities as:

$$p\left(C_j|X\right) \propto p\left(C_j\right) \prod_{k=1}^{d} p\left(x_k|C_j\right)$$

Using the Bayes' rule, we label a new case $X$ with a class level $C_j$ that achieves the highest posterior probability. Although the assumption that the predictor variables are independent is not always accurate, Naïve Bayes reduces a high-dimensional density estimation task to a one-dimensional kernel density estimation. Furthermore, the assumption does not seem to greatly affect the posterior probabilities. Naïve Bayes can be modeled in several different ways, including normal, log-normal, gamma and Poisson density functions. The Naïve

Bayes classifier used in the present study assumes independence of the predictor variables, and a Gaussian distribution (given the target class) of metric predictors, i.e.

$$p\left(x_k|C_j\right) = \frac{1}{\sigma_{kj}\sqrt{2\pi}}\exp\left(\frac{-\left(x-\mu_{kj}\right)^2}{2\sigma_{kj}}\right),$$

where $\mu_{kj}$ and $\sigma_{kj}$ are the corresponding means and standard deviations, respectively.

*K* Nearest Neighbors is a simple nonparametric classification procedure that has been successful used with non-normal populations. It performs as follows [39,42]:

1. Defining a measure of distance between points, normally Mahalanobis distance.
2. Calculating the distances from the test sample $x_0$ to the others points.
3. Selecting the $k$ nearest sample points to the one we intend to classify. Calculating the proportion of these $k$ points belonging to each of the populations. Classifying the point $x_0$ in the population corresponding to a higher points frequency from the $k$ points. In this study, the $k$ value has been selected minimizing the cross-validation misclassification error.

Support Vector Machine (SVM) is a classifier method developed by Vapnik and coworkers performing classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. They found that if a wide margin between the regions of distribution of sample points of different kinds exists, the mathematical model obtained as an optimal hyperplane will exhibit very good prediction ability, even if the dimension of the feature space was very high and the equation of this optimal hyperplane had to be expressed by many adjustable parameters [42–44]. In Fig. 5, the optimal hyperplane denotes the unique hyperplane having largest distances with the sample points of different classes. The sample points located on the border of the margin are called support vectors. It can be seen that the position of the optimal hyperplane is only decided by the support vectors.

The principle of SVM is much different from the others commonly used methods. The most important task in SVM is not dimension reduction, but dimension elevation using a kernel function to map the sample points of the input space into a feature space with higher dimensionality by nonlinear transformation. In these high dimensional feature spaces, nonlinear separable sample points in the input space can become linearly separable with wide margin in the feature space, and a linear separation algorithm can be used to make a mathematical model with good prediction ability [39,40,42–44].

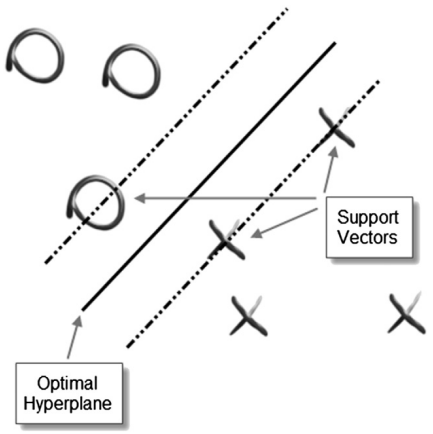


**Fig. 5.** Optimal hyperplane and large margin in SVM.

For constructing an optimal hyperplane, SVM method uses an iterative training algorithm, which is used to minimize an error function. According to the form of the error function, the SVM classification models can belong to two different groups: C-SVM and nu-SVM classification. In this study, the first one is used. So, the error function to be minimized in training is the following:

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i,$$

subject to

$$y_i\left(w^T\phi(x_i) + b\right) \geq 1-\xi_i,\ \ \xi_i \geq 0,\ \ i = 1,\ldots,N,$$

where $C$ is the capacity constant, $w$ is the vector of coefficients, $b$ a constant and $\xi_i$ are parameters for handling non-separable data (inputs). The index $i$ labels the $N$ training cases. Note that $y_i$ are the class labels and $x_i$ the independent variables. The kernel $\phi$ is used to transform data from the input to the feature space. It should be noted that the larger the $C$, the more the error is penalized. Thus, $C$ should be chosen with care to avoid over-fitting. Different kernel types are used in the present study.

Additionally, a Neural Network method has been implemented. In particular, a single hidden layer perceptrons (feed-forward Neural Networks) classification method is used. For further information see [39,40,42,45,46].

Before discussing the results, it is necessary to define what parameters were defined and the value thereof when we use the KNN, SVM and Neural Networks methods. Using KNN method, we obtained the optimal correct classification probability for 7 classes with $k=2$ neighbors. In the case of SVM, 4 different kernels (linear, Gaussian, polynomial and sigmoid) were tested using the C-SVM algorithm. The optimal result was obtained using a polynomial kernel $((gamma \times u'v + coef)^{degree})$ with parameters $C=36$ (testing from 1 to 100), $gamma=0.03$ (testing from 0.01 to 1), $degree=3$ (testing from 3 to 5) and $coef=14$ (testing from 1 to 30). Finally, regarding Neural Networks, a single-hidden-layer neural network with 2 units, skip-layer connections, initial random weights on $[-0.4, 0.4]$ and parameter of weight decay equal to 0.001. These optimal parameters were obtained using leave-one-out cross-validation. For classifying between 2 classes, the parameters obtained are the following: KNN ($k=2$), SVM ($C=12$, $gamma=0.01$, $degree=5$, $coef=4$), Neural Networks (hidden-layer size = 1, weight decay = 0.01 and initial random weights on $[-0.66, 0.66]$).

In the present study, the following R (free software) libraries are used: `nnet` (for Logistic regression and Neural Networks), `e1071` (for Naïve Bayes and SVM) and MASS (for LDA and Quadratic classification) [47]. For applying the KNN classifier, the Matlab platform is used [48].

Fig. 6 shows the flowchart of the classification process.

## 4. Results and discussion

In this section, the methods previously presented are applied to the features matrix to classify between different species and main groups (hardwoods and softwoods). First, the procedure for obtaining and scaling the data and, additionally, a brief descriptive analysis is shown.

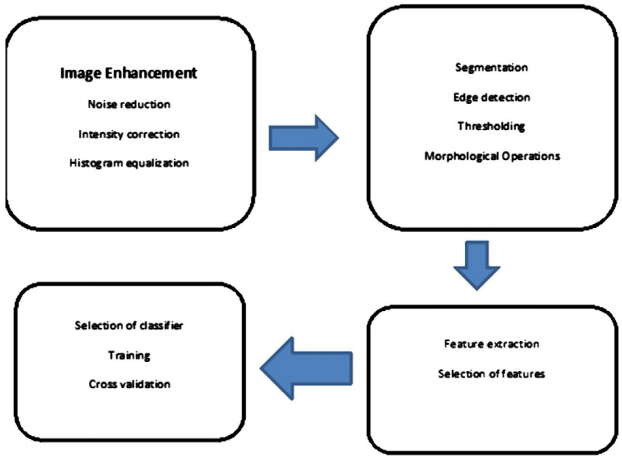


**Fig. 6.** Flowchart of the classification process.

*4.1. Obtaining, scaling and descriptive analysis of features*

The image analysis has been done on Matlab platform, using the image processing toolbox. We have used 7 different classes of wood as our test sample. Figs. 2 and 3 show some micrographs obtained using a scanning electron microscope and the resulting image after applying image enhancement, segmentation and dilation processes, as described in Section 3 of this study.

The features extraction, described in Section 3.4 is performed to the processed images. We have used a total of 101 micrographs, with variable number of samples for each group. So our feature matrix is of size $101 \times 5$. In Table 1, the number of samples per class is shown.

Data preparation is a very important step in Machine Learning. In fact, the scaling of data is a common practice in multivariate analysis, especially when techniques aimed at the prediction are used [44]. In this study, each column of matrix features is centered (the mean of each column is subtracted from it) and then divided by its standard deviation. Using this standardization, the features variabilities are matched [39].

After obtaining the standardized features, we have used the metric multidimensional scaling (MDS) technique to obtain preliminary information about the data [39,42]. The MDS method can be applied to many multivariate data, provided that the calculation of the distances or similarities be doable. After calculating the distance matrix corresponding to the data, the aim of this first descriptive analysis is to obtain information about the structure of the data, i.e., try to find out what elements present similar properties, if there are distinct groups, outliers, etc. The information provided by the distance matrix can be approximated by using two principal coordinates (corresponding to the two largest eigenvalues in the similarities matrix) [39]. Fig. 7 shows all the obtained samples on the basis of these two principal coordinates.

Using two dimensions has been appropriate because we achieve to represent a high percentage of data variability. In this case, the explained variability proportion using the two first eigenvalues is 84% [39]. It can be observed that the samples belonging to the same wood specie form groups. Some groups of species appear particularly distinct with respect to the others, such as insignis pine, jatobá, European chestnut, and especially eucalyptus. Eucalyptus samples displayed no overlap with samples of other kind of wood. The differences between eucalyptus and the other species are much more pronounced than any other. However, there are different groups or types showing some overlap. These are, above all, European beech, Scots pine and walnut. In particular, the overlapping between beech and Scots pine samples may hinder the classification of new samples belonging to both timber species. The MDS analysis shows that the chosen features characterize, in greater or lesser extent, the wood species.

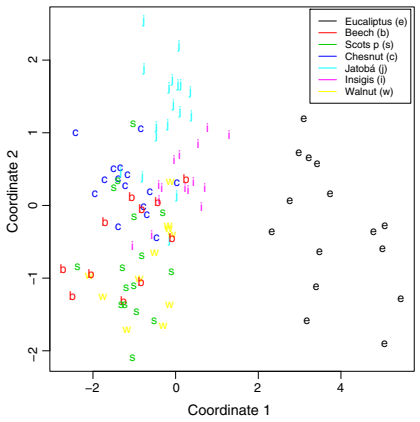


**Fig. 7.** Samples from different species of wood, using MDS with two principal coordinates.

*4.2. Supervised classification results*

We apply the different classification methods defined in Section 3.5. Traditional methods estimating the population parameters by statistical inference are used: LDA, Quadratic classification and Logistic regression. Moreover, other methods related to Machine Learning are applied: SVM, Naïve Bayes, KNN and Neural Networks. These last classifiers focused on the accuracy of prediction rather than on the interpretation of the models generating them. The correct classification probabilities obtained by the methods previously mentioned are presented in Table 2. The leave-one-out cross-validation process (described in Section 3.5) has been used to obtain these results and to compare the classifiers. In the next subsection, an external validation test completes the comparison between the different approaches. The probabilities are computed in two settings, classifying among the 7 different species and in the more general case of classifying into 2 different groups: hardwoods and softwoods.

Usually, we get better results when using Machine Learning methods, i.e., SVM classifiers, KNN and Neural Networks. High probabilities of good classification between 0.87 (SVM) and 0.89 (Neural Networks) are obtained using these methods for classifying between softwoods and hardwoods (Table 2). We also obtain good results when we want to classify among 7 different classes: the correct classification probabilities are between 0.78 (KNN) and 0.81 (SVM). Note that the results obtained when classifying between 2 classes are slightly better than those obtained in the case of 7 species. As pointed out previously, classifiers related to Machine Learning usually seem to work better than traditional methods, especially when classifying in 2 classes. For example, the LDA method is the worst method when 7 classes are used (0.72), but the difference with respect to the three Machine Learning techniques is even larger in the case of 2 classes (0.75 for LDA). Although the probabilities obtained by LDA are not low in absolute terms, better results are obtained by other nonlinear methods such as Logistic (0.80 with 7 classes) or Quadratic classification.

In general, the results in Table 2 show a relatively high probability of correct classification, especially if we take into account the heterogeneity of the wood and the results obtained in other studies [22,23]. Nevertheless, since the probability of misclassification is larger than zero, an interesting question is: what are the species tending to get confused using these features and methods? The answer is in the confusion matrices shown in Tables 3 and 4.

Table 3 shows the confusion matrices corresponding to the SVM, Neural Networks and KNN methods applied to classify between hardwoods and softwoods. It can be observed that the hardwood samples generally are better predicted than softwoods (see SVM and Neural Networks methods). There is an exception, KNN classifier predicts correctly all the softwoods (the posterior probability is equal to 1). These little confusions may be due to the beech and Scots pine overlapping that can be observed below in Table 4.

Table 4 shows the confusion matrices corresponding to the SVM, Logistic regression, KNN and Neural Networks methods applied to classify between 7 different types of wood. The first two methods presented in Table 4 are representative of all others (except KNN), in addition, they produce the best results. As the descriptive analysis had pointed, the posterior probability of eucalyptus is 1, i.e., all eucalyptus samples have been predicted correctly (see leave-one-out cross-validation in Section 3.5). Moreover, there is no confusion with any other wood species. Other distinct groups that have been observed using the MDS technique present high posterior probabilities (correct classification). This is the case of insignis pine (0.93), jatobá (0.89) and chestnut (0.80), using SVM or Logistic regression. The use of Neural Networks improves the insignis pine prediction (1.00), but the result with chestnut is worse (0.60). Moreover, the SVM method has been able to correctly classify walnut (0.82). This is a very good result, because according to the descriptive analysis previously shown, this specie presented an overlapping with other wood species. In general, beech and Scots pine are the most difficult species to predict. The beech samples can be confused with the Scots pine ones and Scots pine samples can be predicted as walnut or beech (see Table 3). The best results in this case correspond to the use of the KNN method: this technique is able to correctly predict all the samples of beech wood,

**Table 2**
Prediction probabilities obtained by each classification method and leave-one-out cross-validation, using features based in segmentation process. The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

| Methods | 7 groups classification | 2 groups classification |
|---|---|---|
| | Prediction | Prediction |
| LDA | 0.72 | 0.75 |
| Quadratic classification | 0.77 | 0.78 |
| Logistic regression | 0.80 | 0.75 |
| Naïve Bayes | 0.79 | 0.75 |
| KNN | 0.78 | 0.88 |
| SVM | 0.81 | 0.87 |
| Neural Networks | 0.80 | 0.89 |

**Table 3**
Probabilities of correct classification, using leave-one-out cross-validation, in 2 different classes (hardwoods and softwoods) obtained by SVM, Neural Networks and KNN. The probabilities are rounded using two significant figures.

| Methods | Actual | Estimated | |
|---|---|---|---|
| | | Hardwoods | Softwoods |
| SVM | Hardwoods | 0.90 | 0.10 |
| | Softwoods | 0.23 | 0.77 |
| Neural Networks | Hardwoods | 0.93 | 0.07 |
| | Softwoods | 0.20 | 0.80 |
| KNN | Hardwoods | 0.83 | 0.17 |
| | Softwoods | 0.00 | 1.00 |

**Table 4**
Probabilities of correct classification, using leave-one-out cross-validation, in 7 different classes (chestnut, walnut, insignis pine, Scots pine, beech, eucalyptus and jatobá) obtained by SVM, Logistic regression, KNN and Neural Networks. The probabilities are rounded using two significant figures.

| Methods | Actual | Estimated | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Eucal. | Beech | Scots P. | Chesn. | Jat. | Insig. P. | Walnut |
| SVM | Eucalyptus | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.64 | 0.18 | 0.09 | 0.00 | 0.00 | 0.09 |
| | Scots P. | 0.00 | 0.06 | 0.63 | 0.13 | 0.00 | 0.06 | 0.13 |
| | Chesnut | 0.00 | 0.07 | 0.07 | 0.80 | 0.00 | 0.00 | 0.07 |
| | Jatobá | 0.00 | 0.11 | 0.00 | 0.00 | 0.89 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 0.07 | 0.00 | 0.00 | 0.93 | 0.00 |
| | Walnut | 0.00 | 0.00 | 0.09 | 0.18 | 0.00 | 0.00 | 0.73 |
| Logistic R. | Eucalyptus | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.45 | 0.45 | 0.00 | 0.09 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.06 | 0.63 | 0.13 | 0.00 | 0.06 | 0.13 |
| | Chesnut | 0.00 | 0.07 | 0.07 | 0.80 | 0.00 | 0.00 | 0.07 |
| | Jatobá | 0.00 | 0.06 | 0.06 | 0.00 | 0.89 | 0.00 | 0.00 |
| | Insignis P. | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.93 | 0.00 |
| | Walnut | 0.00 | 0.00 | 0.18 | 0.00 | 0.00 | 0.00 | 0.82 |
| KNN | Eucalyptus | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.38 | 0.63 | 0.13 | 0.00 | 0.06 | 0.13 |
| | Chesnut | 0.00 | 0.13 | 0.00 | 0.87 | 0.00 | 0.00 | 0.00 |
| | Jatobá | 0.00 | 0.06 | 0.06 | 0.11 | 0.78 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.13 | 0.07 | 0.00 | 0.00 | 0.80 | 0.00 |
| | Walnut | 0.00 | 0.09 | 0.36 | 0.09 | 0.09 | 0.00 | 0.36 |
| Neural Networks | Eucalyptus | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | Beech | 0.00 | 0.64 | 0.18 | 0.18 | 0.00 | 0.00 | 0.00 |
| | Scots P. | 0.00 | 0.00 | 0.69 | 0.13 | 0.00 | 0.06 | 0.13 |
| | Chesnut | 0.00 | 0.13 | 0.07 | 0.60 | 0.07 | 0.00 | 0.13 |
| | Jatobá | 0.00 | 0.00 | 0.06 | 0.06 | 0.89 | 0.00 | 0.00 |
| | Insignis P. | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| | Walnut | 0.00 | 0.09 | 0.09 | 0.09 | 0.00 | 0.00 | 0.73 |

but at the cost of worse walnut predictions. Therefore, Table 4 shows three methods giving complementary information.

The results in Tables 2–4 indicate that, overall, classifying between hardwoods and softwoods and, on other hand, between 7 different species, has been possible. Therefore, the existence of differences between wood species from the earlywood tracheids structure (defined using the 5 proposed features obtained at 1500× magnification) has been proven. On the other hand, the SVM method (particularly suitable for small samples [44], as in this case), Neural Networks and KNN have shown the best prediction behavior.

*4.2.1. External validation*

The validation scheme used in the previous section could be not enough to guarantee a proper generalizability of the outcomes and to get an accurate idea of the model performances on an actually external set. This is because Machine Learning classification methods, such as KNN, Neural Networks or SVM require to fix the values of some adjustable parameters (learning coefficients and number of hidden neurons, value of $k$ in KNN, learning parameters in SVM) which are chosen based on the minimum error in leave-one-out cross-validation, and the same cross-validation approach has been used to get an estimate of the predictive ability of the models. Therefore, the cross-validation samples may be not entirely external for the above mentioned models themselves.

Given these arguments, a new test set, completely external, is included to properly evaluate the predictive ability of the models. Considering the available number of samples, the procedure is the following: in the hardwood and softwood (2 classes) classification case, an external test set (consisting of 30 samples) based on the Kennard–Stone intelligent criterion [49] is selected. The adjustable model parameters are obtained using the training set and the 10-fold cross-validation error minimization procedure. On the other hand, for the 7 class problem, taking into account the reduced number of samples per class, a first set of samples (one per class) is randomly extracted as external validation set. The remaining samples are used as training set, and models choosing the parameters according to a 10-fold cross-validation procedure are built. This global procedure is repeated 100 times to ensure that, with high probability, each sample is included in the test set at least once.

Table 5 shows the results obtained in this framework. Compared with the previous leave-one-out cross-validation results (Table 2), and when classifying between 2 classes, a better performance for all the methods (except SVM) is observed. Considerable high probabilities of good classification (0.93) are obtained by KNN and Neural Networks. It is important to note that the third best method using external validation is now the Quadratic classification procedure, instead of SVM.

When the samples are classified according to the 7 different wood species, the results obtained by LDA, Logistic regression and Naïve Bayes methods using external validation (Table 5) are very similar to those obtained when using the leave-one-out cross-validation criterion (Table 2). Note that the result produced by the Quadratic classification approach improved slightly in this setting, giving a probability of correct classification equal to 0.80. Other Machine Learning methods, such as Neural Networks and KNN produced slightly worse results and, to a lesser extent, SVM (0.76). In short, the best correct classification percentages obtained by leave-one-out cross-validation are similar to those obtained using the external validation test.

*4.2.2. Additional studies*

The study presented in the previous sections can be completed in different directions. In this section, we focus on two possible extensions of our research; on one hand, using standard GLCM features and, on the other hand, applying fractal analysis techniques. These two approaches are related to the acquisition of new attributes or variables from which to classify.

Firstly, we apply the same statistical methods to the standard GLCM features of the micrographs. The corresponding results are shown in Table 6, using leave-one-out cross-validation (similar results are obtained with the external validation test described in Section 4.2.1). They are clearly worse than those obtained with the segmentation study (Table 2), proving the advantage of using the approach presented in this paper.

On the other hand, a possible extension of the present work could be using fractal analysis procedures in our research. The concept of fractal was first introduced by Mandelbrot [50]. Fractal refers to entities, especially sets of pixels, which display a degree of self-similarity at different scales. It has been successfully used in [51] to study the capability to accurately describe the surface roughness of bread crumb. In the same context, in [52], the relationship between different features and panelists' perception of bread crumb is analyzed. In both papers, the study is carried out computing the fractal dimension (FD) of digital images of bread crumb using different methods: fractional Brownian motion (FBM) method, frequency domain method, box-counting (BC) method, morphological fractal method, mass fractal method and random walks method. Fractal analysis has been also applied in medical image applications. For example, in [53], the FD computed using the FBM method is used to classify breast ultrasound images. Recently, in [54], some applications of fractal theory in wood science are proposed.

Following the lines in the previous papers, we obtained the FD of our 101 images, using the FBM and BC approaches. A detailed description of these two methods can be found in [51,53,55] and [51,56], respectively. The means and standard deviations of the FD's for each one of the 7 wood species are given in Table 7. As in [51], we compute the correlation coefficients between the 5 original texture features and the fractal dimensions determined with the FBM and BC methods. Unlike in [51], no strong correlations are observed in our case. Therefore, the FD vectors are directly used as two extra features,

**Table 5**
Prediction probabilities by each classification method and an external validation test, using features based in segmentation process. The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

| Methods | 7 groups classification | 2 groups classification |
|---|---|---|
| | Prediction | Prediction |
| LDA | 0.73 | 0.80 |
| Quadratic classification | 0.80 | 0.90 |
| Logistic regression | 0.77 | 0.83 |
| Naïve Bayes | 0.79 | 0.87 |
| KNN | 0.73 | 0.93 |
| SVM | 0.76 | 0.87 |
| Neural Networks | 0.73 | 0.93 |

**Table 6**
Prediction probabilities obtained by each classification method and leave-one-out cross-validation, using GLCM as the features. The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

| Methods | 7 groups classification | 2 groups classification |
|---|---|---|
| | Prediction | Prediction |
| LDA | 0.41 | 0.67 |
| Quadratic classification | 0.37 | 0.59 |
| Logistic regression | 0.37 | 0.69 |
| Naïve Bayes | 0.36 | 0.53 |
| KNN | 0.39 | 0.78 |
| SVM | 0.37 | 0.69 |
| Neural Networks | 0.47 | 0.73 |

**Table 7**
Means and standard deviations of FD computed with the FBM and BC methods for the 7 wood species.

| Species | FBM | | BC | |
|---|---|---|---|---|
| | Mean | Standard deviation | Mean | Standard deviation |
| Eucalyptus | 2.986 | 0.0019 | 1.769 | 0.0171 |
| Beech | 2.988 | 0.0007 | 1.840 | 0.0164 |
| Scots P. | 2.986 | 0.0010 | 1.841 | 0.0110 |
| Chesnut | 2.985 | 0.0018 | 1.839 | 0.0125 |
| Jatobá | 2.985 | 0.0030 | 1.831 | 0.0233 |
| Insignis P. | 2.986 | 0.0014 | 1.841 | 0.0106 |
| Walnut | 2.983 | 0.0019 | 1.803 | 0.0354 |

jointly with the 5 original ones, in the classification procedures. Table 8 shows the percentages of correct classification obtained when the FD's are included as new features, using the external validation framework described in Section 4.2.1. The results presented in Table 8 correspond to three scenarios, depending on the features considered in the classification approaches: the 5 original features jointly with the FD computed by the FBM method (denoted by FBM in the table), the 5 original features jointly with the FD computed by the BC method (denoted by BC in the table), and the 5 original features jointly with the FD computed by the FBM and BC methods (denoted by FBM + BC in the table). As in previous tables, the results are divided in 7 and 2 classes.

The results indicate that, in general, including the FD computed with the FBM method improves the classification process when the aim is classifying in 7 classes. Adding this fractal feature (FBM), the best classification probability is 0.83, corresponding to the Naïve Bayes method. Moreover, the results obtained by Logistic regression (0.81), SVM (0.81), KNN (0.77) and Neural Networks (0.78) are clearly better than those obtained without including in the feature set the FD by the FBM method (see Table 5). This improvement is mainly due to the existence of a minor confusion in predicting classes corresponding to beech and Scots pine samples. Nevertheless, when we want to classify between hardwoods and softwoods, the percentages of correct classification, having added the feature related to the fractal dimension by FBM (and selecting again a new test sample by Kennard Stone method) does not improve the previous results (see Table 5). The only exception is the correct classification probability improvement obtained by SVM (0.90). On the other hand, the percentages of correct classification for the statistical classification techniques, adding the fractal dimension computed by the BC method to the previous 5 extracted features (columns BC in Table 8) show that, generally, including only the BC dimension feature does not improve the classification process. Finally, the last two columns in Table 8 present the percentages of correct classification for each statistical classification technique, adding the two proposed fractal dimensions to the original 5 feature data set. It is observed that, in this setting, the best classification probability (among 7 classes) is 0.83, corresponding to the Naïve Bayes method. Moreover, a maximum correct classification probability equal to 0.93 by KNN and Neural Networks methods is obtained when the aim is classifying between hardwoods and softwoods. In general, compared with the case of just using the 5 original texture features, if the two additional fractal dimensions are used jointly with them, the resulting probabilities of correct classification for the 7 class case are improved, while similar probabilities of correct classification are obtained for classifying into 2 groups.

**Table 8**
Prediction probabilities obtained by each classification method an external validation test, using the 5 original features jointly with the FD computed by the FBM method (FBM), the 5 original features jointly with the FD computed by the BC method(BC), and the 5 original features jointly with the FD computed by the FBM and BC methods (FBM + BC). The feature data set was tested with 2 (hardwoods and softwoods) and 7 classes.

| Methods | FBM | | BC | | FBM + BC | |
|---|---|---|---|---|---|---|
| | 7 Class. | 2 Class. | 7 Class. | 2 Class. | 7 Class. | 2 Class. |
| LDA | 0.77 | 0.77 | 0.75 | 0.80 | 0.75 | 0.90 |
| Quadratic classification | 0.75 | 0.87 | 0.76 | 0.87 | 0.74 | 0.87 |
| Logistic regression | 0.81 | 0.77 | 0.71 | 0.87 | 0.76 | 0.90 |
| Bayes Naïve | 0.83 | 0.80 | 0.78 | 0.87 | 0.83 | 0.87 |
| KNN | 0.77 | 0.87 | 0.75 | 0.80 | 0.75 | 0.93 |
| SVM | 0.81 | 0.90 | 0.72 | 0.87 | 0.79 | 0.90 |
| Neural Networks | 0.78 | 0.80 | 0.69 | 0.90 | 0.80 | 0.93 |

## 5. Conclusions

In the present paper, micrographs obtained by SEM with 1500× magnification, taken in cross-sections, have been processed in a simple way using segmentation and object recognition to identify tracheids corresponding to the earlywood of 7 different timbers. Segmentation process has allowed to extract 5 features related to the shape, number, area and distribution of this type of cells.

Classifying wood species using these 5 features extracted from SEM micrographs with 1500× magnification has been possible. High correct classification probabilities have been obtained when we want to discern between hardwoods and softwoods (0.89 using leave-one-out cross-validation, and 0.93 using an external validation test) and among 7 different wood species (0.81 using one-leave-out cross-validation, and 0.80 using an external validation test), taking into account the heterogeneity of wood. The results show that it is possible to observe differences among species in the wood texture at this magnification (1500×).

Observing the shape, number, area and distribution of the earlywood tracheids, eucalyptus is very different to the others species. On the other hand, beech can be confused with Scots pine using these 5 features.

Fractal analysis has been successfully used in this work. We obtained the FD of our 101 images using the FBM and BC approaches. Adding to the original 5 features these fractal dimensions, the best classification probability (among 7 classes) is 0.83, corresponding to the Naïve Bayes method. Moreover, a maximum correct classification probability equal to 0.93 by KNN and Neural Networks methods, when we want to classify between hardwoods and softwoods, is obtained. If the two fractal dimensions are added to the 5 texture features considered in this research, a clear improvement in the resulting probabilities of correct classification for the 7 class case is obtained, while similar results are obtained if the aim is classifying into 2 groups. A more complete study including fractal analysis would be of great interest in this context, for example, determining the FD using some other methods and using these vectors as new features. This will be analyzed in a future research.

Classifiers related to Machine Learning seem to work better than traditional methods, especially when we want to classify between 2 classes. The drawback when SVM or Neural Networks classifiers are used is the relatively high computing time.

A comparative study using gray level co-occurrence (GLCM) based features is also presented, showing the effectiveness of the image segmentation method, at least with this data set. In fact, the correct classification probabilities obtained using segmentation are much higher.

## Acknowledgments

This research has been partially supported by the Spanish Ministry of Science and Innovation, Grant MTM2008-00166 (ERDF included). The authors thank two anonymous referees for constructive comments that improved the presentation of this article.

## References


[1] M. Khalid, E.L.Y. Lee, R. Yusof, M. Nadaraj, Design of an intelligent wood species recognition system, Int. J. Simul. Syst. Sci. Technol. 9 (2008) 9–19.
[2] E.W.H. Hayek, P. Krenmayr, H. Lohninger, U. Jordis, W. Moche, F. Sauter, Identification of archaeological and recent wood tar pitches using gas chromatography/mass spectrometry and pattern recognition, Anal. Chem. 62 (1990) 2038–2043.
[3] J. Arno, G. Miller-Mead, A. Poynter, J. Truini, The art of woodworking, Encyclopedia of Wood, Time-Life Books, Richmond, Virginia, 1993.
[4] S.A. Leavengood, Identifying common northwest wood species : a woodworker's guide, in: Oregon State University. Extension Service, 1998.
[5] R. Hoadley, Identifying Wood: Accurate Results with Simple Tools, Taunton Press, Newtown, CT, 1990.
[6] A. Guindeo Casasús, L. García Esteban, F. Peraza Sánchez, F. Arriaga Martitegui, Especies de Maderas, Asociación de investigación técnica de las industrias de la madera y corcho (AITIM), Madrid, 1997.
[7] I.R. Lewis, N.W. Daniel, N.C. Chaffin, P.R. Griffiths, Raman spectrometry and neural networks for the classification of wood types—1, Spectroc. Acta Pt. A—Molec. Biomolec. Spectr. 50 (1994) 1943–1958.
[8] R.B. Miller, Structure of wood, in: Wood Handbook: Wood as an Engineering Material, Woodhead Publishing Limited, Madison, WI,U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, 1999.
[9] P. Cavalin, L.S. Oliveira, A.L. Koerich, A.S. Britto, Wood defect detection using grayscale images and an optimized feature set, Proc. IEEE Ind. Electron (IECON), World Scientific, 2006.
[10] C. Fuentealba, C. Simon, D. Choffel, P. Chawentier, D. Massons, Wood products identification by internal characteristics readings, IEEE International Conference on Industrial Technology (ICIT), Institute of Electric and Electronics Engineers (IEEE), 2004, pp. 763–768.
[11] I.Y. Gu, H. Andersson, R. Vicen, Wood defect classification based on image analysis and support vector machines, Wood Sci. Technol. 44 (2009) 693–704.
[12] J. Lampinen, S. Smolander, M. Korhonen, Wood surface inspection systemb based on generic visual features, in: F. Fogelman, S.P. Gallinari (Eds.), Industrial Applications of Neural Networks, World Scientific, Paris, 1995, p. 3542.
[13] K. Watanabe, J.F. Hart, S.D. Mansfield, S. Avramidis, Near-infrared technology applications for quality control in wood processing, The Future of Quality Control for Wood & Wood Products, The final conference of COST action E53, 2010, http://cte.napier.ac.uk/e53, Accessed 25 May 2010.
[14] B.K. Lavine, C.E. Davidson, A.J. Moores, P.R. Griffiths, Raman spectroscopy and genetic algorithms for the classification of wood types, Appl. Spectrosc. 55 (2001) 960–966.
[15] M.H. Nuopponen, G.M. Birch, R.J. Sykes, S.J. Lee, D.J. Stewart, Estimation of wood density and chemical composition by means of diffuse reflectance mid-infrared fourier transform (DRIFT-MIR) spectroscopy, J. Agric. Food Chem. 54 (2006) 34–40.
[16] V. Piuri, F. Scotti, Design of an automatic wood types classification system by using fluorescence spectra, IEEE Trans. Syst. Man Cybern. Part C—Appl. Rev. 40 (2010) 358–366.
[17] H. Yang, I.R. Lewis, P.R. Griffiths, Raman spectrometry and neural networks for the classification of wood types. 2. Kohonen self-organizing maps, Spectroc. Acta Pt. A—Molec. Biomolec. Spectr. 55 (1999) 2783–2791.
[18] J. Tarrío-Saavedra, S. Naya, M. Francisco-Fernández, J. López-Beceiro, R. Artiaga, Functional nonparametric classification of wood species from thermal data, J. Therm. Anal. Calorim. 104 (2011) 87–100.
[19] F. Ferraty, P. Vieu, Nonparametric Functional Data Analysis, Springer-Verlag, Berlin, 2006.
[20] J.O. Ramsay, B.W. Silverman, Functional Data Analysis, Springer-Verlag, New York, 2005.
[21] J.O. Ramsay, B.W. Silverman, Applied Functional Data Analysis, Springer-Verlag, New York, 2002.
[22] T. Brandtberg, Individual tree-based species classification in high spatial resolution aerial images of forests using fuzzy sets, Fuzzy Sets Syst. 132 (2002) 371–387.
[23] J.Y. Tou, Y.H. Tay, P.Y. Lau, Rotational invariant wood species recognition throughwood species verification, in: N.T. Nguyen, H.P. Nguyen, A. Grzech (Eds.), Proceedings of the 2009 First Asian Conference on Intelligent Information and Database Systems (ACIIDS), The Institute of Electrical and Electronics Engineers, Inc., Dong Hoi, 2009, pp. 115–120.
[24] R. Alén, E. Kuoppala, O. Pia, Formation of the main degradation compound groups from wood and its components during pyrolysis, J. Anal. Appl. Pyrolysis 36 (1996) 137–148.
[25] L. Gašparovič, Z. Koreňová, L. Jelemenský, Kinetic study of wood chips decomposition by TGA, Proceedings 36th International Conference of SSCHE. Tatranské Matliare, vol. 178, World Scientific, 2009, pp. 1–14.
[26] M.G. Grønli, G. Várhegyi, C. Blasi, Thermogravimetric analysis and devolatilization kinetics of wood, Ind. Eng. Chem. Res. 41 (2002) 4201–4208.
[27] M. Müller-Hagedorn, H. Bockhorn, L. Krebs, U. Müller, A comparative kinetic study on the pyrolysis of three different wood species, J. Anal. Appl. Pyrolysis 68–69 (2003) 231–249.
[28] K. Raveendran, A. Ganesh, K.C. Khilar, Pyrolysis characteristics of biomass and biomass components, Fuel 75 (1996) 987–998.
[29] A.F. Roberts, A review of kinetics data for the pyrolysis of wood and related substances, Combust. Flame 14 (1970) 261–272.
[30] S. Wang, K. Wang, Q. Liu, Y. Gu, Z. Luo, K. Cen, T. Fransson, Comparison of the pyrolysis behavior of lignins from different tree species, Biotechnol. Adv. 27 (2009) 562–567.
[31] F.H. Schweingruber, Wood Structure and Environment, Springer-Verlag, Berlin, 2007.
[32] T.S. Huang, G.J. Yang, G.Y. Tang, A fast two-dimensional median filtering algorithm, IEEE Trans. Acoust., Speech, Signal Process. 27 (1979) 13–18.
[33] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley, Reading, MA, 1971.
[34] H.C. Andrews, A.G. Tescher, R.P. Kruger, Image processing by digital computer, IEEE Spectr. 9 (1972) 20–32.
[35] E.L. Hall, R.P. Kruger, S.J. Dwyer, D. Hall, R.W. Mclaren, G.S. Lodwick, A survey of preprocessing and feature extraction techniques for radiographic images, IEEE Trans. Comput. 20 (1972) 1032–1044.
[36] E.L. Hall, Almost uniform distribution for compute image enhancement, IEEE Trans. Comput. 23 (1974) 207–208.
[37] L.G. Shapiro, G.C. Stockman, Computer Vision, Prentice Hall, New Jersey, 2001.
[38] G.A. Marcoulides, S.L. Hershberger, Multivariate Statistical Methods. A First Course, Lawrence-Erlbaum Associates, Mahwah, New Jersey, 1997.
[39] D. Peña, Análisis de Datos Multivariantes, McGraw-Hill, Madrid, 2002.
[40] J. Hernández Orallo, M.J. Ramírez Quintana, C. Ferri Ramirez, Introducción a la Minería de Datos, Pearson-Prentice Hall, Madrid, 2004.
[41] G.L. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley-Interscience, Hoboken, New Jersey, 2004.
[42] T. Hill, P. Lewicki, STATISTICS Methods and Applications, Springer, Tulsa, OK, 2007.
[43] V. Vapnik, Stat. Learning Theory, Wiley, 1998.
[44] N. Chen, W. Lu, J. Vang, G. Li, Support Vector Machines in Chemistry, World Scientific Publishing, Singapore, 2004.
[45] W.N. Venables, M.J. Ripley, B.D. Quintana, C. Ferri Ramirez, Modern Applied Statistics with S, Springer, 2002.
[46] B.D. Ripley, Neural network and related methods for classification (with discussion), J. R. Stat. Soc. Ser. B—Stat. Methodol. 56 (1994) 409–456.
[47] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2008, http://www.R-project.org.
[48] MATLAB, Version 7.10.0 (R2010a), The MathWorks Inc, Natick, Massachusetts, 2010.
[49] R.W. Kennard, L.A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148.
[50] B.B. Mandelbrot, The Fractal Geometry of Nature, W. H. Freeman and Company, New York, 1983.
[51] U. Gonzales-Barron, F. Butler, Fractal texture analysis of bread crumb digital images, Eur. Food Res. Technol. 226 (2008) 721–729.
[52] U. Gonzales-Barron, F. Butler, Prediction of panellists' perception of bread crumb appearance using fractal and visual textural features, Eur. Food Res. Technol. 226 (2008) 779–785.
[53] D.R. Chen, R.F. Chang, C.J. Chen, M.F. Ho, S.J. Kuo, S.T. Chen, S.J. Hung, W.K. Moon, Classification of breast ultrasound images using fractal feature, Clin. Imaging 29 (2005) 235–245.
[54] W. Fenghu, Z. Mengying, S. Jianping, Z. Xiaodong, The application of fractal theory in wood science research, Adv. Mater. Res. 113–114 (2010) 801–806.
[55] J.C. Russ, The Image Processing Handbook, CRC Press, Boca Raton, 2002.
[56] S. Buczkowski, S. Kyriacos, F. Nekka, L. Cartilier, The modified box-counting method: analysis of some characteristic parameters, Pattern Recognit. 31 (1998) 411–418.

# Bibliografía

*Cuando bebas agua, recuerda la fuente.*

Proverbio chino


Alén, R., Kuoppala, E. y Pia, O. Formation of the main degradation compound groups from wood and its components during pyrolysis. *J. Anal. Appl. Pyrolysis*, vol. 36, páginas 137–148, 1996.

Andrews, H. C., Tescher, A. G. y Kruger, R. P. Image processing by digital computer. *IEEE Spectr.*, vol. 9, páginas 20–32, 1972.

Antoniadis, A. y Sapatinas, T. Estimation and inference in functional mixed effects models. *Comput. Stat. Data Anal.*, vol. 51, páginas 4793–4813, 2007.

Arno, J., Miller-Mead, G., Poynter, A. y Truini, J. *The Art of Woodworking. Encyclopedia of Wood*. Time-Life Books, Richmond, Virginia, 1993.

Artiaga, R., López-Beceiro, J., Tarrío-Saavedra, J., Gracia-Fernández, C., Naya, S. y Mier, J. L. Estimating the reversing and non-reversing heat flow from standard dsc curves in the glass transition region. *J. Chemometr.*, vol. 25, páginas 287–294, 2011. DOI: 10.1002/cem.1347.

ASTM International E2009-08. *American standard test method for oxidation onset temperature of hydrocarbons by differential scanning calorimetry*. ASTM International, West Conshohocken, PA, 2008.

del Barrio, E., Cuesta-Albertos, J., Fraiman, R. y Matran, C. The random projection method for goodness of fit for functional data. *Comput. Stat. Data Anal.*, vol. 51, páginas 4814–4831, 2007.

BIN, L. y QINGZHAO, Y. Classification of functional data: A segmentation approach. *Comput. Stat. Data Anal.*, vol. 52, páginas 4790–4800, 2008.

BOOKSTEIN, F. L. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge, United Kingdom, 1991.

BRANDTBERG, T. Individual tree-based species classification in high spatial resolution aerial images of forests using fuzzy sets. *Fuzzy Sets Syst.*, vol. 132, páginas 371–387, 2002.

BRUMBACK, B. A. y RICE, J. A. Smoothing spline models for the analysis of nested and crossed samples of curves. *J. Amer. Stat. Assoc.*, vol. 93, páginas 961–994, 1998.

BUCZKOWSKI, S., KYRIACOS, S., NEKKA, F. y CARTILIER, L. The modified box-counting method: Analysis of some characteristic parameters. *Pattern Recognit.*, vol. 31, páginas 411–418, 1998.

BÜHLMANNAND, P. y HOTHORN, T. Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, vol. 22, páginas 477–505, 2007.

CAMORANI, P., BADIAI, M., FRANCOMACARO, D., GAMASSI, M., VINCENZO, P., SCOTTI, F. y ZANASI, M. A classification method for wood types using fluorescence spectra. En *Conference Record - IEEE Instrumentation and Measurement Technology Conference*, páginas 1312–1315. 2008.

CARDOT, H., CRAMBES, C., KNEIP, A. y SARDA, P. Smoothing spline estimators in functional linear regression with errors-in-variables. *Comput. Stat. Data Anal.*, vol. 51, páginas 4832–4848, 2007.

CARDOT, H., FERRATY, F. y SARDA, P. Functional linear model. *Stat. Probab. Lett.*, vol. 45, páginas 11–22, 1999.

CAVALIN, P., OLIVEIRA, L. S., KOERICH, A. L. y BRITTO, A. S. Wood defect detection using grayscale images and an optimized feature set. En *Proc. IEEE Ind. Electron (IECON)*. World Scientific, 2006.

CHANG, E., TONG, S., GOH, K. y CHANG, C. W. Support vector machine concept-dependent active learning for image retrieval. *IEEE Trans. Multimedia*, vol. 2, páginas 1–35, 2005.

CHEN, D. R., CHANG, R. F., CHEN, C. J., HO, M. F., KUO, S. J., CHEN, S. T., HUNG, S. J. y MOON, W. K. Classification of breast ultrasond images using fractal feature. *Clin. Imaging*, vol. 29, páginas 235–245, 2005.

CHEN, N., LU, W., VANG, J. y LI, G. *Support Vector Machines in Chemistry*. World Scientific Publishing, Singapore, 2004.

CUESTA-ALBERTOS, J. A. y FEBRERO-BANDE, M. A simple multiway anova for functional data. *Test*, vol. 19, páginas 537–557, 2010.

CUEVAS, A., FEBRERO, M. y FRAIMAN, R. Linear functional regression: the case of fixed design and functional response. *Can. J. Stat.*, vol. 30, páginas 285–300, 2002.

CUEVAS, A., FEBRERO, M. y FRAIMAN, R. An anova test for functional data. *Comput. Stat. Data Anal.*, vol. 47, páginas 111–122, 2004.

CUEVAS, A., FEBRERO, M. y FRAIMAN, R. On the use of the bootstrap for estimating functions with functional data. *Comput. Stat. Data Anal.*, vol. 51, páginas 1063–1074, 2006a.

CUEVAS, A., FEBRERO, M. y FRAIMAN, R. Robust estimation and classification for functional data via projection-based depth notions. *Comput. Stat.*, vol. 22, páginas 481–496, 2006b.

DABO-NIANG, S. y RHOMARI, N. Kernel regression estimate in a banach space. *J. Stat. Plan. Infer.*, vol. 131, páginas 1421–1434, 2009.

DIMITRIADOU, E., HORNIK, K., LEISCH, F., MEYER, D. y WEINGESSEL, A. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien*, 2011. R package version 1.6. http://CRAN.R-project.org/package=e1071.

EFRON, B. y TIBSHIRANI, R. J. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.

FAN, J., LIN y K, S. Functional anova models for proportional hazards regression. *J. Amer. Stat. Assoc.*, vol. 93, páginas 1007–1021, 1998.

FEBRERO-BANDE, M., DE LA FUENTE, O. y M. *fda.usc: Functional Data Analysis and Utilities for Statistical Computing (fda.usc)*, 2011. R package version 0.9.5.

FENGHU, W., MENGYING, Z., JIANPING, S. y XIAODONG, Z. The application of fractal theory in wood science research. *Adv. Mater. Res.*, vol. 113-114, páginas 801–806, 2010.

FERRATY, F. High-dimensional data: a fascinating statistical challenge. *J. Multi. Anal*, vol. 101, páginas 305–306, 2010.

FERRATY, F. En *Recent advances in functional data analysis and related topics* (editado por F. Ferraty). Springer-Verlag, Berlin Heilderberg, 2011.

FERRATY, F. y ROMAIN, Y. *The oxford handbook of functional data analysis*. Oxford University Press, 2010.

FERRATY, F. y VIEU, P. The functional nonparametric model and application to spectrometric data. *Comput. Stat.*, vol. 17, páginas 545–564, 2002.

FERRATY, F. y VIEU, P. Curves discrimination: a nonparametric functional approach. *Comput. Stat.*, vol. 44, páginas 161–173, 2003.

FERRATY, F. y VIEU, P. *Nonparametric Functional Data Analysis*. Springer-Verlag, Berlin, 2006.

FERRATY, F., VIEU, P. y VIGUIER-PLA, S. Factor based comparison of groups of curves. *Comput. Stat. Data Anal.*, vol. 51, páginas 4903–4910, 2007.

FERRIOL, M., GENTILHOMME, A., COCHEZ, M., OGET, N. y MIELOSZYNSKI, J. L. Thermal degradation of poly(methyl methacrylate) (PMMA): modelling of DTG and TG curves. *Polym. Degrad. Stabil.*, vol. 79, páginas 271–281, 2003.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annual Eugenics*, vol. 7, páginas 179–188, 1936.

FIX, E. y HODGES, J. L. Discriminatory analysis, nonparametric discrimination: Consistency properties. Informe Técnico 4, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

FRAIMAN, R. y MUNIZ, G. Trimmed means for functional data. *Test*, vol. 10, páginas 419–440, 2001.

FRANCISCO-FERNÁNDEZ, M., TARRÍO-SAAVEDRA, J., MALLIK, A. y NAYA, S. A comprehensive classification of wood from thermogravimetric curves. *Chemometrics Intell. Lab. Syst.*, 2012. Doi:10.1016/j.chemolab.2012.07.003.

FUENTEALBA, C., SIMON, C., CHOFFEL, D., CHAWENTIER, P. y MASSONS, D. Wood products identification by internal characteristics readings. En *IEEE International Conference on Industrial Technology (ICIT)*, páginas 763–768. Institute of Electric and Electronics Engineers (IEEE), 2004.

FUKUNAGA, K. *Introduction to Statistical Pattern Recognition*. Academic Press Professional, Inc., San Diego, CA, 1990.

GASSON, P., MILLER, R., STEKEL, D. J., WHINDER, F. y ZIEMIŃSKA, K. Wood identification of Dalbergia nigra (CITES Appendix I) using quantitative wood anatomy, principal components analysis and naï¿½ve bayes classification. *Ann. Bot.*, vol. 105(1), páginas 45–56, 2010.

GAŠPAROVIČ, L., KOREŇOVÁ, Z. y JELEMENSKÝ, L. Kinetic study of wood chips decomposition by tga. En *Procedings 36th International Conference of SSCHE. Tatranské Matliare*, vol. 178, páginas 1–14. World Scientific, 2009.

GONZALES-BARRON, U. y BUTLER, F. Fractal texture analysis of bread crumb digital images. *Eur. Food Res. Technol.*, vol. 226, páginas 721–729, 2008a.

GONZALES-BARRON, U. y BUTLER, F. Prediction of panellists'perception of bread crumb appearance using fractal and visual textural features. *Eur. Food Res. Technol.*, vol. 226, páginas 779–785, 2008b.

GONZÁLEZ MANTEIGA, W. y VIEU, P. Statistics for functional data. *Comput. Stat. Data Anal.*, vol. 51, páginas 4788–4792, 2007.

GRACIA-FERNÁNDEZ, C., DAVIS, P., GÓMEZ-BARREIRO, S., LÓPEZ-BECEIRO, J., TARRÍO-SAAVEDRA, J. y ARTIAGA, R. A vitrification and curing study by simultaneous tmdsc-photocalorimetry. *J. Therm. Anal. Calorim.*, vol. 102, páginas 1057–1062, 2010a.

GRACIA-FERNÁNDEZ, C., GÓMEZ-BARREIRO, S., LÓPEZ-BECEIRO, J., TARRÍO-SAAVEDRA, J., NAYA, S. y ARTIAGA, R. Comparative study of the dynamic glass transition temperature by dma and tmdsc. *Polym. Test*, vol. 29, páginas 1002–1006, 2010b.

GRACIA-FERNÁNDEZ, C., TARRÍO-SAAVEDRA, J., LÓPEZ-BECEIRO, J., GÓMEZ-BARREIRO, S., NAYA, S. y ARTIAGA, R. Temperature modulation in pdsc for monitoring the curing under pressure. *J. Therm. Anal. Calorim.*, vol. 106, páginas 101–107, 2011.

GRØNLI, M. G., VÁRHEGYI, G. y BLASI, C. Thermogravimetric analysis and devolatilization kinetics of wood. *Ind. Eng. Chem. Res.*, vol. 41, páginas 4201–4208, 2002.

GU, I. Y., ANDERSSON, H. y VICEN, R. Wood defect classification based on image analysis and support vector machines. *Wood Sci. Technol.*, vol. 44, páginas 693–704, 2009.

GUINDEO CASASÚS, A., GARCÍA ESTEBAN, L., PERAZA SÁNCHEZ, F. y ARRIAGA MARTITEGUI, F. *Especies de Maderas*. Asociación de investigación técnica de las industrias de la madera y corcho (AITIM), Madrid, 1997.

HALL, E. L. Almost uniform distribution for compute image enhancement. *IEEE Trans. Comput.*, vol. 23, páginas 207–208, 1974.

HALL, E. L., KRUGER, R. P., DWYER, S. J., HALL, D., MCLAREN, R. W. y LODWICK, G. S. A survey of preprocessing and feature extraction techniques for radiographic images. *IEEE Trans. Comput.*, vol. 20, páginas 1032–1044, 1972.

HARSCH, M., KARGER-KOCSIS, J. y HOLST, M. Influence of fillers and additives on the cure kinetics of an epoxy/anhydride resin. *Eur. Polym. J.*, vol. 43, páginas 1168–1178, 2007.

HAYEK, E. W. H., KRENMAYR, P., LOHNINGER, H., JORDIS, U., MOCHE, W. y SAUTER, F. Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *Anal. Chem.*, vol. 62, páginas 2038–2043, 1990a.

HAYEK, E. W. H., KRENMAYR, P., LOHNINGER, H., JORDIS, U., MOCHE, W. y SAUTER, F. Identification of archaeological and recent wood tar pitches using gas chromatography/mass spectrometry and pattern recognition. *Anal. Chem.*, vol. 62, páginas 2038–2043, 1990b.

HECHENBICHLER, K. y SCHLIEP, K. Weighted k-nearest-neighbor techniques and ordinal classification. Informe Técnico 399, SFB 386, Ludwing-Maximilians University Munich, 2004. http://epub.ub.uni-muenchen.de/1769/.

HERNÁNDEZ ORALLO, J., RAMÍREZ QUINTANA, M. J. y FERRI RAMIREZ, C. *Introducción a la Minería de Datos*. Pearson-Prentice Hall, Madrid, 2004.

HILL, T. y LEWICKI, P. *STATISTICS Methods and Applications*. Springer, Tulsa, OK, 2007.

HOADLEY, R. *Identifying Wood: Accurate Results with Simple Tools*. Taunton Press, Newtown, CT, 1990.

HUANG, J., KOOPERBERG, C., STONE, C. J. y TRUONG, Y. K. Functional anova models for proportional hazards regression. *Ann. Stat.*, vol. 28, páginas 961–999, 2000.

HUANG, T. S., YANG, G. J. y TANG, G. Y. A fast two-dimensional median filtering algorithm. *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, páginas 13–18, 1979.

JORDAN, R., FEENEY, F., NESBITT, N. y EVERTSEN, J. A. Classification of wood species by neural network analysis of ultrasonic signals. *Ultrasonics*, vol. 36(1-5), páginas 219–222, 1998.

KARATZOGLOU, A., MEYER, D. y HORNIK, K. Support vector machines in R. *J. Stat. Softw.*, vol. 15, páginas 1–28, 2006.

KAWAGUCHI, A., YONEMOTO, K., TANIZAKI, Y., KIYOHARA, Y., YANAGAWA, T. y TRUONG, Y. K. Application of functional anova models for hazard regression to the Hisayama data. *Stat. Med.*, vol. 27, páginas 3515–3527, 2008.

KENNARD, R. W. y STONE, L. A. Computer aided design of experiments. *Technometrics*, vol. 11, páginas 137–148, 1969.

KHALID, M., LEE, E. L. Y., YUSOF, R. y NADARAJ, M. Design of an intelligent wood species recognition system. *Int. J. Simul. Syst., Sci. Technol.*, vol. 9, páginas 9–19, 2008.

KNEIP, A. y GASSER, T. Convergence and consistency results for self-modeling nonlinear regression. *Ann. Stat.*, vol. 16, páginas 82–112, 1988.

KNEIP, A., LI, X., MACGIBBON, K. B. y RAMSAY, J. O. Curve registration by local regression. *Can. J. Stat.*, vol. 28, páginas 19–29, 2000.

KONONENKO, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.*, vol. 23, páginas 89–109, 2001.

LABATI, R. D., GAMASSI, M., PIURI, V. y SCOTTI, F. A low-cost neural-based approach for wood types classification. En *2009 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications, CIMSA 2009*, páginas 199–203. 2009.

LAMPINEN, J., SMOLANDER, S. y KORHONEN, M. Wood surface inspection systemb based on generic visual features. En *Industrial Applications of Neural Networks* (editado por F. Fogelman y S. P. Gallinari), página 3542. World Scientific, Paris, 1995.

LAVINE, B. K., DAVIDSON, C. E., MOORES, A. J. y GRIFFITHS, P. R. Raman spectroscopy and genetic algorithms for the classification of wood types. *Appl. Spectrosc.*, vol. 55, páginas 960–966, 2001.

LEAVENGOOD, S. A. Identifying common northwest wood species : a woodworker's guide. En *Oregon State University. Extension Service*. 1998.

LEE, A. y LICHTENHAN, J. D. Thermal and viscoelastic property of epoxy-clay and hybrid inorganic-organic epoxy nanocomposites. *J. Appl. Polym. Sci.*, vol. 73, páginas 1993–2001, 1999.

LEE, J. W., LEE, J. B., PARK, M. y SONG, S. H. An extensive comparison of recent classification tools applied to microarray data. *Comput. Stat. Data Anal.*, vol. 48, páginas 869–885, 2005.

LEWIS, I. R., DANIEL, N. W., CHAFFIN, N. C. y GRIFFITHS, P. R. Raman spectrometry and neural networks for the classification of wood types-1. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.*, vol. 50, páginas 1943–1958, 1994.

LIPPMANN, R. P. Pattern classification using neural networks. *IEEE Commun. Mag.*, vol. 27, páginas 47–64, 1989.

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. y COHEN, K. L. Robust principal component analysis for functional data (with discussion). *Test*, vol. 8, páginas 1–74, 1999.

LÓPEZ-BECEIRO, J., GRACIA-FERNÁNDEZ, C., TARRÍO-SAAVEDRA, J., GÓMEZ-BARREIRO, S. y ARTIAGA, R. Study of gypsum by pdsc. *J. Therm. Anal. Calorim.*, vol. 109, páginas 1177–1183, 2012.

LÓPEZ-BECEIRO, J., PASCUAL-COSP, J., ARTIAGA, R., TARRÍO-SAAVEDRA, J. y NAYA, S. Thermal characterization of ammonium alum. *J. Therm. Anal. Calorim.*, vol. 104, páginas 127–130, 2011a.

LÓPEZ-BECEIRO, R., J.AND ARTIAGA, GRACIA-FERNÁNDEZ, C., NAYA, S., TARRÍO-SAAVEDRA, J. y MIER-BUENHOMBRE, J. Comparison of olive, corn, soybean and sunflower oils by pdsc. *J. Therm. Anal. Calorim.*, vol. 104, páginas 169–175, 2011b.

LÓPEZ-GRANADOS, F., PEÑA BARRAGÁN, J. M., M., J.-E., FRANCISCO-FERNÁNDEZ, M., CAO, R., ALONSO-BETANZOS, A. y FONTENLA-ROMERO, O. Multispectral classification of grass weeds and wheat (*Triticum durum*) using linear and nonparametric functional discriminant analysis and neural networks. *Weed Research*, vol. 48, páginas 28–37, 2008.

LÓPEZ-PINTADO, S. y ROMO, R. Depth-based inference for functional data. *Comput. Stat. Data Anal.*, vol. 51, páginas 4957–4968, 2007.

LUKASIAK, B. M., FARIA, R., ZOMER, S., BRERETON, R. G. y DUNCAN, J. C. Pattern recognition for the analysis of polymeric materials. *Analyst*, vol. 131(1), páginas 73–80, 2006.

MALLIK, A., TARRÍO-SAAVEDRA, J., FRANCISCO-FERNÁNDEZ, M. y NAYA, S. Classification of wood micrographs by image segmentation. *Chemometrics Intell. Lab. Syst.*, vol. 107, páginas 351–362, 2011. DOI: 10.1016/j.chemolab.2011.05.005.

MANDELBROT, B. B. *The Fractal Geometry of Nature*. W. H. Freeman and Company, New York, 1983.

MARCOULIDES, G. A. y HERSHBERGER, S. L. *Multivariate Statistical Methods. A First Course*. Lawrence-Earlbaum Associates, Mahwah, New Jersey, 1997.

MATLAB. *Version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts, 2010.

MCLACHLAN, G. L. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley-Interscience, Hoboken, New Jersey, 2004.

MEHTA, S., MIRABELLA, F. M., RUFENER, K. y BAFNA, A. Thermoplastic olefin/clay nanocomposites: Morphology and mechanical properties. *J. Appl. Polym. Sci.*, vol. 92, páginas 928–936, 2004.

MICHIE, D., SPIEGELHALTER, D. J. y TAYLOR, C. C. *Machine learning, Neural and Statistical Classification (Ellis Horwood Series in Artificial Intelligence)*. Prentice Hall, 1994.

MILLER, R. B. Structure of wood. En *Wood Handbook: Wood as an Engineering Material*. Woodhead Publishing Limited, Madison, WI,U.S. Department of Agriculture, Forest Service, Forest Products Laboratory, 1999.

MILLER, R. G. *Simultaneous Statistical Inference*. Springer-Verlag, New York, USA, 1991.

MILTYK, W., ANTONOWICZ, E. y KOMSTA, L. Recognition of tablet content by chemometric processing of differential scanning calorimetry curves - an acetaminophen example. *Thermochim. Acta*, vol. 507-508, páginas 146–149, 2010.

MOHAMMAD, A. y SIMON, G. P. Rubber-clay nanocomposites. En *Polymer nanocomposites* (editado por M. Yiu-Wing y Y. Zhong-Zhen). Woodhead Publishing Limited, 2006.

MOHAN, D., PITTMAN, J. C. U. y STEELE, P. H. Pyrolysis of wood/biomass for bio-oil: a critical review. *Energ. Fuel.*, vol. 20, páginas 848–889, 2006.

MÜLLER-HAGEDORN, M., BOCKHORN, H., KREBS, L. y MÜLLER, U. A comparative kinetic study on the pyrolysis of three different wood species. *J. Anal. Appl. Pyrolysis*, vol. 68–69, páginas 231–249, 2003.

NAYA, S. *Nuevas aplicaciones de la estimación paramétrica y no paramétrica de curvas al análisis térmico*. Tesis Doctoral, University of A Coruña, Spain, 2003.

NAYA, S., CAO, R., ARTIAGA, R. y GARCÍA, A. New method for material classification from TGA data by nonparametric regression. *Mater. Sci. Forum*, vol. 514-516, páginas 1452–1456, 2006.

NELDER, J. A. y MEAD, R. A simplex method for function minimization. *Comput. J.*, vol. 7, páginas 308–313, 1965.

NERINI, D. y GHATTAS, B. Classifying densities using functional regression trees: application in oceanology. *Comput. Stat. Data Anal.*, vol. 51, páginas 4984–4993, 2007.

NUOPPONEN, M. H., BIRCH, G. M., SYKES, R. J., LEE, S. J. y STEWART, D. J. Estimation of wood density and chemical composition by means of diffuse reflectance mid-infrared fourier transform (DRIFT-MIR) spectroscopy. *J. Agric. Food Chem.*, vol. 54, páginas 34–40, 2006.

OKADA, T. y TOMITA, S. An optimal orthonormal system for discriminant analysis. *Pattern Recognit.*, vol. 18, páginas 139–144, 1985.

PEÑA, D. *Análisis de Datos Multivariantes*. McGraw-Hill, Madrid, 2002.

PIURI, V. y SCOTTI, F. Design of an automatic wood types classification system by using fluorescence spectra. *IEEE Trans. Syst. Man Cybern. Part C-Appl. Rev.*, vol. 40, páginas 358–366, 2010.

POMERANTSEV, A. L. y RODIONOVA, O. Y. Hard and soft methods for prediction of antioxidantsáctivity based on the dsc measurements. *Chemometrics Intell. Lab. Syst.*, vol. 79(1-2), páginas 73–83, 2005.

PREGHENELLA, M., PEGORETTI, A. y MIGLIARESI, C. Thermo-mechanical characterization of fumed silica-epoxy nanocomposites. *Polymer*, vol. 46, páginas 12065–12072, 2005.

PRIME, R. B. Thermosets. En *Thermal characterization of polymeric materials (second edition)* (editado por E. Turi), páginas 1380–1766. Academic Press, San Diego, 1997.

PRIME, R. B., BAIR, H. E., GALLAGHER, P. K. y RIGA, A. Thermogravimetric analysis (TGA). En *Thermal Analysis of Polymers Fundamentals and Applications* (editado por J. D. Menczel y R. B. Prime). John Wiley & Sons, San José, 2009.

R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. http://www.R-project.org.

RAMSAY, J. O. Estimating smooth monotone functions. *J. Royal Stat. Soc. Ser. B*, vol. 60, páginas 365–375, 1998.

RAMSAY, J. O., HOOKER, G. y S., G. *Functional data analysis with R and Matlab*. Springer, New York, 2009.

RAMSAY, J. O. y LI, X. Curve registration. *J. Royal Stat. Soc. Ser. B*, vol. 60, páginas 351–363, 1998.

RAMSAY, J. O. y SILVERMAN, B. W. *Applied Functional Data Analysis*. Springer-Verlag, New York, 2002.

RAMSAY, J. O. y SILVERMAN, B. W. *Functional Data Analysis*. Springer-Verlag, New York, 2005.

RAMSAY, J. O., WICKHAM, H., GRAVES, S. y HOOKER, G. *fda: Functional Data Analysis*, 2011. R package version 2.2.7.

RAVEENDRAN, K., GANESH, A. y KHILAR, K. C. Pyrolysis characteristics of biomass and biomass components. *Fuel*, vol. 75, páginas 987–998, 1996.

RIPLEY, B. D. Neural network and related methods for classification (with discussion). *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, vol. 56, páginas 409–456, 1994.

ROBERTS, A. F. A review of kinetics data for the pyrolysis of wood and related substances. *Combust. Flame*, vol. 14, páginas 261–272, 1970.

RUSS, J. C. *The Image Processing Handbook*. CRC Press, Boca Raton, 2002.

SALKOE, H. y CHIBA, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions, ASSP-26*, vol. 1, páginas 43–49, 1978.

SCHADLER, L. S. Polymer-based and polymer-filled nanocomposites. En *Nanocomposite Science and Technology* (editado por P. M. Ajayan, L. S. Schadler y P. V. Braun), páginas 77–135. Wiley-VCH, Weinheim, 2003.

SCHLIEP, K. y HECHENBICHLER, K. *kknn: Weighted k-Nearest Neighbors*, 2008. R package version 1.0-6.

SCHWEINGRUBER, F. H. *Wood Structure and Environment*. Springer-Verlag, Berlin, 2007.

SEBIO-PUÑAL, T., NAYA, S., LÓPEZ-BECEIRO, J., TARRÍO-SAAVEDRA, J. y ARTIAGA, R. Thermogravimetric analysis of wood, holocellulose, and lignin from five wood species. *J. Therm. Anal. Calorim.*, vol. 109, páginas 1163–1167, 2012.

SHAO-YUN, F., XI-QIAO, F., LAUKE, B. y YIU-WING, M. Effects of particle size, particle/matrix interface adhesion and particle loading on mechanical properties of particulate-polymer composites. *Compos. Pt. B-Eng.*, vol. 39, páginas 933–961, 2008.

SHAPIRO, L. G. y STOCKMAN, G. C. *Computer Vision*. Prentice Hall, New Jersey, 2001.

SHEN, Q. y FARAWAY, J. J. An $F$ test for linear models with functional responses. *Stat. Sin.*, vol. 14, páginas 1239–1257, 2004.

SILVERMAN, B. W. Incorporating parametric effects into functional principal components analysis. *J. Royal Stat. Soc. Ser. B*, vol. 57, páginas 673–689, 1995.

STONE, C. J., HANSEN, M. H., KOOPERBERG, C. y TRUONG, Y. K. Polynomial splines and their tensor products in extend linear modeling. *Ann. Stat.*, vol. 25, páginas 1371–1470, 1997.

TARRÍO-SAAVEDRA, J., GRACIA-FERNÁNDEZ, C., LÓPEZ-BECEIRO, J., NAYA, S. y ARTIAGA, R. Temperature modulation in pdsc for monitoring the curing under pressure. *J. Therm. Anal. Calorim.*, 2012. Doi:10.1007/s10973-012-2568-z.

TARRIO-SAAVEDRA, J., LÓPEZ-BECEIRO, J., NAYA, S. y ARTIAGA, R. Effect of silica content on thermal stability of fumed silica/epoxy composites. *Polym. Degrad. Stabil.*, vol. 93, páginas 2133–2137, 2008.

TARRÍO-SAAVEDRA, J., LÓPEZ-BECEIRO, J., NAYA, S., GRACIA, C. y ARTIAGA, R. Controversial effects of fumed silica on the curing and thermomechanical properties of epoxy composites. *Express Polym. Lett.*, vol. 4, páginas 382–395, 2010a.

TARRÍO-SAAVEDRA, J., LÓPEZ-BECEIRO, J., NAYA, S., GRACIA-FERNÁNDEZ, C., MIER-BUENHOMBRE, J. y ARTIAGA, R. Factores influyentes en la estabilidad a la oxidación del biodiesel. estudio estadístico. *Dyna*, vol. 85, páginas 341–350, 2010b.

TARRÍO-SAAVEDRA, J., NAYA, S., FRANCISCO-FERNÁNDEZ, M., ARTIAGA, R. y LÓPEZ-BECEIRO, J. Application of functional anova to the study of thermal stability of micro-nano silica epoxy composites. *Chemometrics Intell. Lab. Syst.*, vol. 105, páginas 114–124, 2011.

TARRÍO-SAAVEDRA, J., NAYA, S., FRANCISCO-FERNÁNDEZ, M., LÓPEZ-BECEIRO, J. y ARTIAGA, R. Functional nonparametric classification of wood species from thermal data. *J. Therm. Anal. Calorim.*, vol. 104, páginas 87–100, 2011.

TARRÍO-SAAVEDRA, J., NAYA, S., LÓPEZ-BECEIRO, J., GRACIA-FERNÁNDEZ, C. y ARTIAGA, R. Biodiesel. En *Thermooxidative properties of biodiesels and other biological fuels* (editado por G. Montero y M. Stoycheva), páginas 47–62. INTECH, Rijeka, 2011.

TONG, S. y KOLLER, D. Support vector machine active learning with applications to text classification. *J. Mach. Learn. Res.*, vol. 2, páginas 45–46, 2001.

TÓTH, L., KOCSOR, A. y CSIRIK, J. On naive Bayes in speech recognition. *Int. J. Appl. Math. Comput. Sci.*, vol. 2, páginas 287–294, 2005.

TOU, J. Y., TAY, Y. H. y LAU, P. Y. Rotational invariant wood species recognition throughwood species verification. En *Proceedings of the 2009 First Asian Conference on Intelligent Information and Database Systems*

*(ACIIDS 2009)* (editado por N. T. Nguyen, H. P. Nguyen y A. Grzech), páginas 115–120. Dong Hoi, The Institute of Electrical and Electronics Engineers, Inc., 2009.

TSIATIS, A. A. y DAVIDIAN, M. Joint modeling of longitudinal and time-to-event data: an overview. *Stat. Sin.*, vol. 14, páginas 793–818, 2004.

TUKEY, J. W. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1971.

VALDERRAMA, M. J. An overview to modelling functional data. *Comput. Stat. Data Anal.*, vol. 22, páginas 331–334, 2007.

VAPNIK, V. *Statistical Learning Theory*. Wiley, 1998.

VENABLES, W. N. y RIPLEY, B. D. *Modern Applied Statistics with S*. Springer, New York, fourth edición, 2002. Http://www.stats.ox.ac.uk/pub/MASS4.

VENABLES, W. N., RIPLEY, M. J., B. D. QUINTANA y FERRI RAMIREZ, C. *Modern Applied Statistics with S*. Springer, 2002.

WANG, S., WANG, K., LIU, Q., GU, Y., LUO, Z., CEN, K. y FRANSSON, T. Comparison of the pyrolysis behavior of lignins from different tree species. *Biotechnol. Adv.*, vol. 27, páginas 562–567, 2009.

WATANABE, K., HART, J. F., MANSFIELD, S. D. y AVRAMIDIS, S. Near-infrared technology applications for quality control in wood processing. En *The Future of Quality Control for Wood & Wood Products*. The final conference of COST action E53, 2010. http://cte.napier.ac.uk/e53. Accessed 25 May 2010.

WEHRENS, R. *Chemometrics with R. Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer-Verlag, New York, 2011.

YANG, H., LEWIS, I. R. y GRIFFITHS, P. R. Raman spectrometry and neural networks for the classification of wood types. 2. Kohonen self-organizing maps. *Spectroc. Acta Pt. A-Molec. Biomolec. Spectr.*, vol. 55, páginas 2783–2791, 1999.

YOUSEFI, A., LAFLEUR, P. G. y GAUVIN, R. Kinetic studies of thermoset cure reactions: a review. *Polym. Compos.*, vol. 18, páginas 157–168, 1997.

ZHANG, H., ZHANG, Z., FRIEDRICH, K. y EGER, C. Property improvements of in situ epoxy nanocomposites with reduced interparticle distance at high nanosilica content. *Acta Mater.*, vol. 54, páginas 1833–1842, 2006.

*Et je l'ai, mais ça ne me suffit pas encore*
*Je ne suis jamais content.*

*Je voudrais pas crever*
*Boris Vian*