Boletín de Estadística e Investigación Operativa
Vol. 30, No. 2, Junio 2014, pp. 105-124

Estadística

Statistical learning in materials engineering

Salvador Naya and Javier Tarrío-Saavedra

Escuela Politécnica Superior
Grupo de investigación MODES
Universidade da Coruña

✉ salva@udc.es, jtarrio@udc.es

Abstract

In this work we present different applications of statistical techniques such as modeling or supervised classification of engineering materials. Many of these techniques stack could be included within the Statistical Learning.

Statistical Learning refers to a set of tools for modeling and understanding complex datasets. It is a recently developed area in statistics and blends with parallel developments in computer science. With the explosion of "Big Data" problems, statistical learning has become a very hot field in many scientific areas as well as material engineering. The classification studies, analysis of variance and estimation of important materials characteristics are nowdays crucial in engineering. The proposed statistical learning algorithms have been performed using the R statistical software.

Keywords: Nonlinear regression, Supervised classification, Image segmentation, Thermal analysis.
AMS Subject classifications: 62P30, 62F99, 62H35

1. Introducción

El avance en nuevos materiales, como los nanomateriales o los smart materiales, y sus aplicaciones en la industria o a nivel biosanitario, es un tema actual que ocupa y preocupa a ingenieros e investigadores. Sin embargo, muchos de los modelos estadísticos que se utilizan en el área de ingeniería de los materiales están basados en leyes de la física, cuyas hipótesis para su aplicación son correctas sólo en ciertos supuestos en los que fueron planteadas, pero que no son aplicables a situaciones complejas como la modelización de procesos de degradación o el diseño y desarrollo de pruebas de vida aceleradas, de importancia creciente en la determinación de la vida útil de estos nuevos materiales. Desde hace décadas, cada vez que se descubre un nuevo material en los laboratorios, estalla

un entusiasmo generalizado, primero en la comunidad científica, después en los medios de comunicación y finalmente en la sociedad. Los investigadores intentan demostrar las ventajas y las posibles aplicaciones, generalmente revolucionarias, para lo que será preciso emplear técnicas estadísticas con el fin de avalar sus propiedades o estimar su vida útil mediante la aplicación de modelos acelerados.

Por otra parte, en los últimos años se ha producido un gran progreso en el desarrollo de nuevas metodologías estadísticas, como la estimación no paramétrica de curvas, el análisis de datos funcionales (FDA) o las técnicas de clasificación supervisada y no supervisada, que actualmente brillan por su ausencia en los estudios realizados en este contexto.

La ingeniería de los materiales, al igual que la estadística, es una ciencia emergente que permite posibilidades impensables hasta hace muy poco. Como ejemplo podría comentarse que los materiales nanoestructurados pueden mostrar propiedades muy diferentes a las que exhiben en una macroescala, posibilitando aplicaciones únicas. De este modo, sustancias opacas se vuelven transparentes (cobre), materiales inertes se transforman en catalizadores (platino), materiales estables se transforman en combustibles (aluminio), sólidos se vuelven líquidos a temperatura ambiente (oro), aislantes se vuelven conductores (silicona). Mucha de la fascinación que produce la nanotecnología proviene de estos peculiares fenómenos cuánticos y de superficie que la materia exhibe en nanoescala.

En estos momentos los nanomateriales y en especial algunos derivados del carbono, como los nanotubos o el grafeno, se han posicionado como las nuevas alternativas a otros materiales tradicionales, debido sobre todo a sus inigualables propiedades mecánicas, dando origen en la actualidad a muchas aplicaciones industriales. Así, los nanotubos de carbono son cien veces más fuertes que el acero y entre seis y diez veces más ligeros, sin por ello perder elasticidad. Es frecuente encontrarlos ya en la fabricación de determinados productos de uso cotidiano, como algunos artículos deportivos.

A nivel de la ingeniería la posibilidad de crear objetos con materiales compuestos mediante las nuevas y cada vez más extendidas impresoras 3D, que permiten la elaboración de todo tipo de instrumentos con casi cualquier tipo de material, está resultando toda una revolución en campos tan importantes como la medicina, donde ya existe la posibilidad de crear prótesis a medida en tiempo real a partir de distintos tipos de biomateriales, o sus aplicaciones al mundo de la gastronomía con la opción de crear alimentos con la forma y el material que se elija.

En este artículo se presentan una serie de aplicaciones del aprendizaje estadístico dentro de la ingeniería de materiales. Concretamente, 1) modelización de datos obtenidos del estudio de la degradación de materiales en laboratorio y la 2) clasificación supervisada de maderas industriales empleando técnicas FDA y de aprendizaje máquina multivariante.

Seguidamente, en la segunda sección del presente trabajo, se introduce de

forma general los aspectos más importantes de la modelización estadística de datos térmicos. En la tercera sección se presenta una propuesta de modelos cinéticos basados en mezclas de funciones logísticas, en la que se aborda el problema de la optimización de parámetros con algoritmos evolutivos, con ejemplos de su aplicación a casos concretos. En la cuarta sección se describen casos de aplicación de las técnicas de clasificación supervisada funcionales y multivariantes a distintas bases de datos, como es el caso de las curvas térmicas o termogramas y las imágenes obtenidas mediante microscopía electrónica de barrido. Finalmente, en la última sección se recogen las conclusiones más relevantes.

2. Modelización de datos de degradación por análisis térmico

Los estudios actuales en el campo de los materiales se apoyan en el manejo de sofisticados instrumentos de toma de medidas en laboratorio. Estos permiten obtener series de datos que relacionan algún tipo de variable dependiente (módulo elástico, flujo de calor, pérdida de masa, etc.) con variables como la composición del material, la temperatura a la que se somete o el tiempo del experimento. Los resultados obtenidos en el laboratorio proporcionan una gran cantidad de datos que requieren de una adecuada modelización a fin de interpretar correctamente la información obtenida.

Uno de los instrumentos clásicos de un laboratorio, usado para medir el grado de degradación de un material con respecto al tiempo y/o a la temperatura, es el *analizador termogravimétrico*. Hoy en día, estos aparatos están compuestos, básicamente, por una termobalanza, un horno, un procesador de temperaturas, un circuito de gas de purga (normalmente aire o N_2) y una CPU o terminal con el software adecuado para almacenar, mostrar y procesar los datos obtenidos. Mediante este dispositivo, el personal de laboratorio puede programar la relación entre el tiempo y la temperatura, de hecho, los experimentos objeto de análisis en este artículo se han realizado a una velocidad de calentamiento constante (relación lineal entre tiempo y temperatura aplicada).

Este tipo de estudios se engloban dentro del denominado campo del *Análisis Térmico*. Así, éste puede definirse como el conjunto de técnicas mediante las cuales, una propiedad física o química de un material es medida en función de la temperatura o del tiempo (Turi, 1997). Por tanto, será esta herramienta la que se usará para estudiar propiedades de los materiales según avanzan y se suceden los procesos de degradación producidos por el paso del tiempo y acelerados por el incremento de temperatura.

2.1. Un ejemplo clásico: el oxalato de calcio

En la Figura 1, se presentan las curvas obtenidas por *calorimetría diferencial de barrido*, DSC (que expresan la diferencia de energía o flujo de calor entre

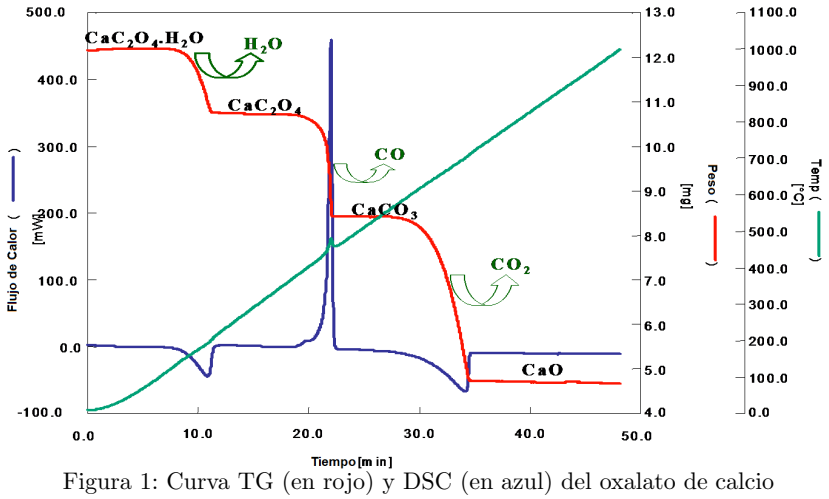

Figura 1: Curva TG (en rojo) y DSC (en azul) del oxalato de calcio

una muestra y la referencia), y las obtenidas mediante *termogravimetría*, TG, correspondientes al ensayo de una muestra de oxalato de calcio monohidrato en un *analizador simultáneo* STA. Se observa cómo cada uno de los escalones que presenta la gráfica TG se corresponde con una pérdida brusca de masa de la muestra al degradarse. Concretamente, en dicha Figura 1, puede verse cómo la muestra de oxalato de calcio se descompone en distintos *procesos de degradación* al ser sometida a un calentamiento lineal. Se aprecia que cada uno de los escalones de la curva TG está relacionado con un proceso o reacción de degradación en el material; por ejemplo, el primer escalón corresponde a la pérdida de agua, el segundo a una pérdida de CO, en el tercero lo que desaparece es el CO_2 y, al final del proceso, lo que ha quedado es óxido de calcio (CaO). Este ejemplo permite ilustrar la importancia que tiene encontrar cada uno de los valores en que ocurren los cambios en el proceso, y sus puntos críticos, para lo que el adecuado ajuste de estas funciones será primordial.

Entre las investigaciones en las que se apoya este artículo hay que destacar la propuesta de alternativas a los modelos clásicos tipo Arrhenius basada en un modelo paramétrico de mezcla de logísticas (Naya et al., 2003 y Naya, 2011). Estos modelos y algunas modificaciones posteriores, como las logísticas generalizadas, han sido aplicados con éxito a diferentes materiales, como las resinas epoxi (Cao et al. 2004 y López-Beceiro, 2011). El método de mezcla de logísticas ha sido aplicado para resolver problemas más concretos, como es el caso de la separación de varios procesos de degradación que aparecen solapados (e indistinguibles a simple vista) en el mismo rango de tiempos o temperaturas (Artiaga et al., 2005). Un estudio de sus propiedades estadísticas, con una propuesta de un contraste de hipótesis, que se basa en la comparación del ajuste paramétrico

con el suavizado no paramétrico, puede verse en Cao y Naya (2009).

Además, el empleo de este tipo de modelos ha permitido determinar características de interés en el estudio de nuevos materiales, entre las que están las siguientes: estudio de las propiedades de las resinas epoxi con nanoclays empleados en el recubrimiento de los depósitos de combustible de hidrógeno (Naya et al., 2009); aplicaciones al estudio de degradación del poliéster-poliuretano (Barbadillo et al., 2007); estabilidad térmica de nanocompuestos epoxi-humo de sílice (Tarrío-Saavedra et al., 2008 y 2011); de la estabilidad térmica, procesos de degradación, estimación de sus constituyentes principales y clasificación de especies de madera comercial (Tarrío-Saavedra et al., 2011a y Sebio-Puñal et al., 2012); propiedades específicas de la alumnita (López-Beceiro et al., 2011a); estudio térmico de la degradación de diferentes aceites comerciales (López-Beceiro et al., 2011b); propiedades térmicas de biocombustibles (Artiaga et al., 2011) o la caracterización de las propiedades viscoelásticas de nuevos nanocompuestos de matriz epoxídica (Tarrío-Saavedra et al., 2011) o el estudio del poliuretano modificado con nanotubos de carbono (Ríos et al., 2013).

2.2. Modelos cinéticos clásicos

La forma clásica de estudiar la descomposición térmica de una muestra se basa en suponer que la velocidad de pérdida de peso de las reacciones de descomposición térmica depende de la masa y de la temperatura, y su expresión general sería:

$$dm/dt = -f(m)k(T). \tag{2.1}$$

En la expresión (2.1) la función $f(m)$ se toma como m_t^n, lo que comúnmente se interpreta diciendo que la pérdida de peso de la muestra obedece a una *cinética de orden* n, donde m_t es la masa de la muestra en el tiempo t.

Esta suposición del orden de reacción es aplicable a reacciones homogéneas y a algunas reacciones heterogéneas que se llevan a cabo en fase condensada. Sin embargo, algunos polímeros que se degradan siguiendo un mecanismo de escisión aleatoria de enlaces, no poseen un orden de reacción constante, de forma que sería erróneo el uso de esta expresión (2.1), dado que los parámetros estarían sobreestimados (véase Conesa (2000)).

En todo caso, la hipótesis fundamental de la relación (2.1) se basa en suponer que la pérdida de masa sigue la Ley de Arrhenius representada por la ecuación siguiente:

$$k(T) = A \exp\left(-\frac{E_a}{RT}\right). \tag{2.2}$$

La *ecuación de Arrhenius* (2.2) expresa la variación de la velocidad de reacción (k) con la temperatura (T), y parte de la suposición de que la velocidad de descomposición del proceso aumenta exponencialmente con la temperatura. La

constante A, llamada constante preexponencial, es independiente de la temperatura; E_a es la denominada energía de activación y R es la constante de los gases y proviene del papel que juega en la ley de los gases perfectos ($pV = nRT$). El nombre se debe a que, en principio, esta ecuación fue propuesta por el químico sueco Svante August Arrhenius para la velocidad de reacción química en los gases.

Para la determinación de los distintos parámetros existentes en la ecuación de Arrhenius (2.2), es necesario llevar a cabo algunas transformaciones, que consisten en expresar dicha ecuación en forma logarítmica:

$$\ln(k(T)) = \ln A - \frac{E_a}{R}\frac{1}{T}. \tag{2.3}$$

Esta ecuación (2.3) permite, representar $\ln(k(T))$ frente al inverso de la temperatura absoluta ($\frac{1}{T}$) y calcular los parámetros deseados. En la práctica, basta con calcular la recta de regresión de $\ln(k(T))$ frente al inverso de la temperatura absoluta ($\frac{1}{T}$) para, al menos, dos experimentos (hay que determinar dos constantes de la recta). La pendiente de dicha recta es la estimación del cociente $-\frac{E_a}{R}$, mientras que la ordenada en el origen es una estimación de $\ln A$.

Algunos autores critican este tipo de métodos basados en la suposición del modelo de Arrhenius y enumeran los siguientes inconvenientes: se manipulan muchos datos obtenidos en el equipo, por lo que el reiterado abuso de logaritmos puede esconder los verdaderos datos de la señal (Brown et al., 2000).

3. Modelos de mezcla de logísticas

3.1. Descripción y planteamiento del problema

Entre los modelos de tipo paramétrico que permiten un mejor ajuste de los datos TG están los que se basan en descomponer toda la curva TG en suma de diferentes funciones logísticas que llamaremos *modelo con mezcla de logísticas*. Este estudio puede verse como un modelo de regresión paramétrico del tipo:

$$y_i = m(t_i, \theta) + \varepsilon_i, \quad i = 1, 2, ...n,$$

donde la variable respuesta es y_i (la masa en las curvas TG) y la variable independiente será t_i (tiempo o temperatura), respectivamente; $m(t_i, \theta)$ es el modelo de ajuste, θ es el vector de parámetros del modelo, estimado por mínimos cuadrados y ε_i son los errores, supuestos con distribución normal de media 0 y varianza constante. La idea de esta modelización se basa en suponer que cada una de estas logísticas representa la cinética de descomposición de los distintos materiales simples de los que está compuesta la muestra inicial. La expresión matemática de este modelo es la siguiente:

$$m(t) = \sum_{i=1}^{k} w_i f(a_i + b_i t), \qquad f(t) = \frac{\exp t}{1 + \exp t}, \tag{3.1}$$

donde w_i son parámetros que representan pérdidas de peso en cada proceso; a_i son parámetros de localización y b_i representa la velocidad de pérdida de peso, para los k escalones que presente la función $m(t)$ en función del tiempo o la temperatura t.

La idea inicial parte de buscar una función que se ajuste lo mejor posible a los datos reales obtenidos para las curvas termogravimétricas. Las funciones candidatas a estimar estas curvas $(t, m(t))$, deben verificar que para valores grandes de $t\,(t \to \infty)$ las respuestas $(m(t))$ deben tender a 0, lo que obliga a que los parámetros b_i tienen que ser negativos, mientras que al inicio del proceso $(t = 0)$ la función debe tender al peso de la muestra en cada momento; por tanto, los valores w_i representan, aproximadamente, las pérdidas de masa en el $i-$ésimo proceso. Los distintos procesos estarían marcados por los "escalones" de cada experimento. Diferentes aplicaciones de este modelo pueden verse en: Naya et al. (2003), Cao et al. (2004), Naya et al. (2006), Naya y Cao (2009) y Naya (2011).

Una ampliación del método de logísticas, que permite una mejor interpretación físico-química de los parámetros consiste en considerar, en lugar de las logísticas simples, la versión de *logísticas generalizadas* que permiten incluir un parámetro adicional, lo que posibilita el ajuste en casos donde los procesos degradativos no sean simétricos.

Este modelo puede expresarse por la siguiente ecuación:

$$m(t) = \sum_{i=1}^{k} w_i g_i(t), \tag{3.2}$$

donde $g_i(t)$ son las funciones logísticas generalizadas de la forma

$$g_i(t) = \frac{c_i}{(1 + \tau_i \exp(-b_i(t - m_i)))^{\frac{1}{\tau_i}}}, \tag{3.3}$$

donde c_i representa el porcentaje de muestra involucrada en cada etapa de degradación, m_i la temperatura en la máxima razón de cambio, b_i está relacionado con la velocidad de cambio, τ_i es la medida para la asimetría y t es la temperatura o el tiempo. Entonces, aunque m_i, b_i y τ_i son parámetros de ajuste, en ausencia de procesos con gran solapamiento, m_i puede ser fácilmente identificado como la temperatura en el pico de la derivada (DTG). Por otra parte, τ_i está relacionado con un orden de reacción aparente, n, por $n = 1 + \tau$. (Véase López-Beceiro et al. (2011) y López-Beceiro (2011)).

El método de mezcla de logísticas o su extensión a las logísticas generalizadas permite el ajuste de las curvas TG, basado en la composición de tantas

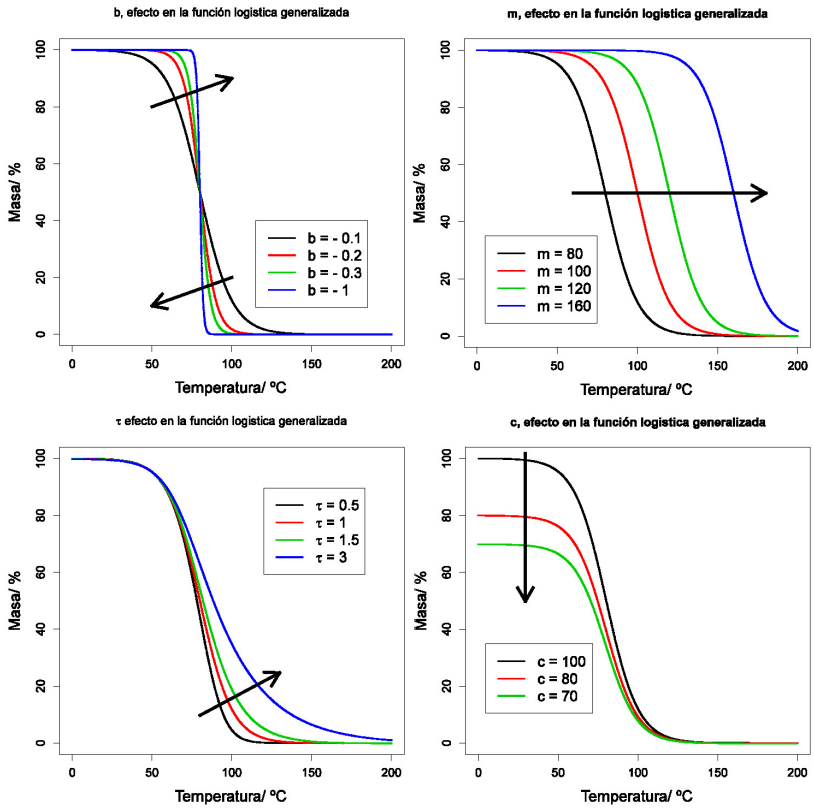

Figura 2: Efecto de los parámetros en el modelo de logísticas generalizadas

funciones logísticas como procesos distintos sufra la muestra, y supone un modelo alternativo a los existentes para la explicación de la cinética.

Un aspecto esencial en el ajuste de este tipo de modelos es la elección del número total de componentes a usar. El número de componentes logísticas está relacionado con el número de procesos de diferentes que caracterizan la degradación de un material. En los casos que se presentan a continuación, dicho número se estima a partir del conocimiento del material en particular (por ejemplo, se conoce que la madera está constituída principalmente por hemicelulosa, celulosa, lignina y agua) y con la ayuda que representa la aplicación del test basado en la estimación polinómica local lineal propuesto por Cao y Naya (2009). El previo conocimento estimativo del número de procesos de degradación permite reducir problemas tales como el que distintas combinaciones de parámetros lleven al mismo ajuste.

3.2. Optimización de los parámetros

Uno de los aspectos con mayor dificultad en este tipo de modelización paramétrica es encontrar los parámetros óptimos del modelo propuesto, para lo que se precisa emplear los métodos de optimización adecuados. En este apartado se presenta un *algoritmo memético* para optimización del ajuste de este modelo no lineal de suma de funciones logísticas generalizadas que se ha implementado recientemente para dar solución a este problema (Ríos-Fachal et al., 2014).

Partiendo de la idea de medir la bondad del ajuste mediante el método de mínimos cuadrados, cuya expresión más habitual es

$$\text{mín} \sum_{i=1}^{n} \left((y_i - m(t_i; \theta_i) \right)^2 , \tag{3.4}$$

donde y_i hace referencia a los valores de la muestra simulada, o, en el caso real, el valor de la masa y $m(t_i; \theta_i)$ representa los valores calculados con los diferentes modelos cinéticos para un conjunto de parámetros θ_i. Por ejemplo, supuesta una relación logística para la cinética de descomposición $Y(t)$ con el tiempo t.

Una forma práctica de conseguir algoritmos más eficaces y adaptados a problemas concretos consiste en hibridar los algoritmos evolutivos con otras técnicas. Los algoritmos evolutivos son una rama de la inteligencia artificial que engloba una serie de métodos de optimización basados en las premisas de la evolución biológica (diversas poblaciones se cruzan y compiten evolucionando a cada vez mejores soluciones). Se han aplicado con éxito principalmente en problemas no lineales con una gran variedad de soluciones posibles para los parámetros. Además, los algoritmos evolutivos obtenidos mediante la hibridación con técnicas de búsqueda local son denominados algoritmos meméticos. Un procedimiento usual de mejora es aplicar el método de búsqueda local a los nuevos miembros de la población, para explorar las mejores regiones de búsqueda obtenidas durante el muestreo global del algoritmo evolutivo, o bien para utilizar las soluciones como

valor de los parámetros iniciales del siguiente algoritmo a aplicar.

Se propone una nueva metodología basada en la utilización de algoritmos que no precisen de un conjunto inicial de parámetros, como el *Diferential Evolution* (DE) o el *Covarianza Matrix* (CMA-ES), que parten de una región factible de valores iniciales, por lo que al dar un intervalo amplio se evitan problemas de convergencia. Luego, una vez aplicado este primer método, que en algunos casos ya consigue un óptimo global del problema, se utilizará la solución propuesta con un método más preciso y que converge con mayor probabilidad al óptimo global, el *Simulated Anneling* (SA) para encontrar el óptimo de los parámetros partiendo de esa solución inicial y alcanzar la solución óptima.

Actualmente están disponibles muchas implementaciones de estos algoritmos en R, así el DE está en la librería *DEoptim*; el algoritmo CMAES está implementado en R por una parte en *cmaes*; mientras que el SA está disponible en la librería *GenSA*.

El objetivo es implementar un proceso de optimización utilizando la menor cantidad posible de hipótesis. De hecho, la optimización se lleva a cabo con una sola limitación relacionada con el modelo de regresión de mezcla de logísticas: la suma de todos los parámetros c_i tiene que ser igual a la masa inicial de la muestra. En concreto, se propone la siguiente secuencia: se aplica el algoritmo de DE para comenzar, ya que no necesita la asignación de una única solución inicial y se indica una amplia región de posibles soluciones. Esta región de valores posibles de los parámetros ha sido elegida suponiendo que no se sabe prácticamente nada de los procesos individuales de la degradación que se superponen en cada curva de TG y entonces, se valida el procedimiento en el peor de los casos. La función objetivo a minimizar es suma de errores de predicción. La solución final se obtiene después de 120.000 iteraciones en los casos complejos, y después de 10.000 iteraciones en las situaciones más simples (Ríos-Fachal et al., 2014).

Finalmente, se realizaron distintos estudios de simulación para comparar los algoritmos propuestos usando distintos escenarios con varios procesos solapados. La solución final obtenida por el DE se utiliza como solución inicial en estos métodos. La ventaja es que se obtuvo una buena solución inicial por el DE sin saber nada acerca de la importancia, la ubicación y la forma de los procedimientos individuales que componen cada curva de TG. Las soluciones obtenidas por separado por SA y CMA-ES se comparan con un valor de referencia; si la diferencia es menor que una cierta tolerancia (Tol) esa solución es elegida como óptima. La mejor solución se obtuvo mediante el algoritmo memético que combina dos algoritmos evolutivos, DE y SA en el caso más complejo de cuatro procesos solapados dos a dos.

3.3. Aplicación al estudio de maderas industriales

En López-Beceiro et al. (2011a) se empleó un nuevo modelo de regresión compuesto por una suma de componentes logísticas generalizadas para ajustar

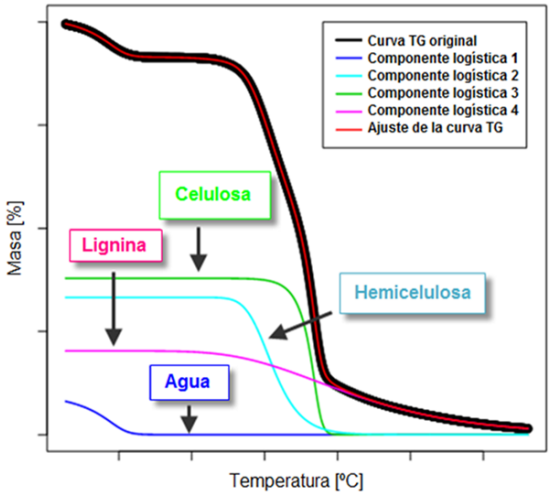

Figura 3: Descomposición de la degradación de una muestra de madera en sus componentes básicas, representadas como logísticas

curvas TG y sus derivadas, de manera que se pudieran estudiar separadamente los procesos de degradación de muestras de madera comercial. Como puede verse en la Figura 3, cada uno de los procesos se corresponde con la degradación de los cuatro componentes de la madera, el agua, la lignina, la celulosa y la hemicelulosa. Es importante observar que el camino de degradación de cada una de estas sustancias se corresponde con una determinada curva logística; de ahí la importancia de la correcta modelización y optimización.

Posteriormente, este tipo de modelos de mezcla de logísticas generalizadas ha sido utilizado en el trabajo de Francisco-Fernández et al. (2012) para seleccionar las características representativas de las curvas TG obtenidas a partir de muestras de madera y su posterior clasificación supervisada.

4. Clasificación de materiales en ingeniería

La clasificación automática de materiales o sustancias químicas a partir de los datos obtenidos por técnicas analíticas es una labor fundamental en ingeniería, con aplicaciones industriales inmediatas. Entre las más conocidas se encuentran el control de calidad en materias primas y semiproductos, la ingeniería inversa y la prevención del fraude comercial. Precisamente en relación al control de calidad y la lucha contra el fraude, se sitúa la necesidad de obtener modelos de aprendizaje estadístico que permitan la correcta clasificación de la madera de uso comercial.

La identificación de la madera, además de necesaria, es una de las tareas

más difíciles de realizar dentro de la tecnología de este material debido a su alta heterogeneidad estructural, mecánica y química. Su resolución puede considerarse una piedra de toque para la evaluación de procedimientos de clasificación.

4.1. Clasificación de madera a partir de datos térmicos

Hasta la fecha, las muestras de madera se habían clasificado automáticamente aplicando sistemas basados en el procesamiento de imágenes o de espectros. Sin embargo, en los últimos años se han desarrollado nuevas metodologías de clasificación supervisada (estimación de la clase correspondiente a una muestra desconocida entre un conjunto finito de clases posibles) aplicadas a bases de datos compuestas por curvas térmicas: termogravimétricas (Tarrío-Saavedra et al., 2010 y Francisco-Fernández et al., 2012), calorimétricas (Tarrío-Saavedra et al., 2010) y calorimétricas obtenidas a altas presiones (Tarrío-Saavedra et al., 2013). Estas nuevas propuestas representan alternativas factibles que combinan la aplicación de técnicas relativamente novedosas, o más o menos sofisticadas, como son el FDA, el aprendizaje máquina (máquinas de vector soporte, redes neuronales, etc.), la reducción de dimensión mediante análisis de componentes principales (PCA) y mínimos cuadrados parciales (PLS), y por otro lado el uso de nuevas bases de datos de carácter térmico, que aportan información acerca del modo en que un material se degrada. Obviamente, estas metodologías son de aplicación en una gran gama de sustancias y materiales.

La Figura 4 muestra los métodos de clasificación, tanto FDA como multivariantes, aplicados a datos térmicos en Tarrío-Saavedra et al. (2010, 2013) y Francisco-Fernández et al. (2012). Para evaluar los resultados obtenidos por cada método se suele utilizar como medida la proporción de clasificación correcta o incorrecta, obtenida mediante procedimientos estándar de doble validación cruzada o validación externa. Se utilizó el procedimiento de *validación cruzada leave-one-out* (LOO), que puede verse en Tarrío-Saavedra et al. (2010, 2013) y Francisco-Fernández et al. (2012) con 49 curvas experimentales TG o DSC a presión.

Los mejores resultados se han obtenido aplicando un método de clasificación FDA basado en el estimador no paramétrico de Nadaraya-Waton a las curvas TG en el intervalo de degradación de la hemicelulosa (90 % de clasificación correcta). La Figura 5, muestra las curvas TG y DSC transformadas linealmente (Tarrío-Saavedra et al., 2010) para mejorar los resultados de clasificación. En dicho gráfico aparece indicado el intervalo de temperaturas donde se obtiene una mayor proporción de clasificación correcta.

Como ya se ha comentado, para llevar a cabo la clasificación de curvas TG (y en general para cualquier tipo de espectro compuesto por una cantidad ingente de datos, mayor que el número de curvas disponibles) aplicando métodos multivariantes, es necesario aplicar métodos de discretización o reducción de dimensión. En Francisco-Fernández et al. (2012) se aplicó el método PCA; de

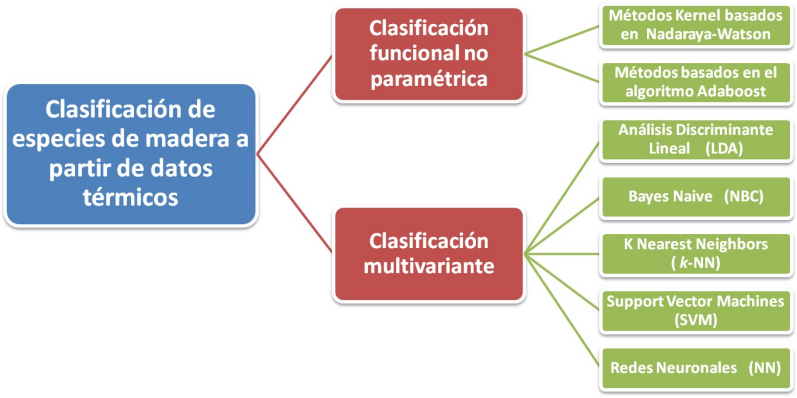

Figura 4: Métodos de clasificación supervisada FDA y multivariantes aplicados a datos térmicos correspondientes a muestras de madera

hecho, se optó por seleccionar las componentes de la curva que expliquen la mayor parte de la variabilidad de los datos (99 % en este caso). Otra de las posibles alternativas consiste en ajustar un modelo logístico generalizado a las curvas TG, usando posteriormente los parámetros de este modelo como características representativas de las curvas. El modelo ajustado consiste en una combinación de 4 funciones logísticas generalizadas relacionadas con los principales constituyentes de la madera: celulosa, hemicelulosa, lignina y agua.

En Tarrío-Saavedra et al. (2013), además de las alternativas anteriormente expuestas, se discretizaron curvas DSC obtenidas a altas presiones (para acelerar el proceso de oxidación de la madera), se aplicó la técnica PLS. Esta técnica proporciona los mismos resultados que el PCA, pero utilizando menos componentes.

La madera, como sucede con otros tipos de materiales (nanocompuestos, por ejemplo), es un material de alta heterogeneidad; de hecho, es difícil obtener una muestra real totalmente representativa de este material, que estime correctamente su variabilidad. Por ello, en Francisco-Fernández et al. (2012) se propone el uso generalizado de estudios de simulación estadística. En particular, los autores proponen generar nuevas curvas TG articiales que imiten a las experimentales mediante el uso de los parámetros obtenidos a partir del ajuste del modelo logístico generalizado a las curvas TG reales. Esto permite comparar los diferentes procedimientos de clasificación y establecer conclusiones acerca de la clasificación de materiales (madera) en muy diferentes escenarios, ahorrando tiempo de experimentación. Cada curva simulada se obtiene a partir de la distribución multinormal que tiene por media el vector medio de los parámetros para cada

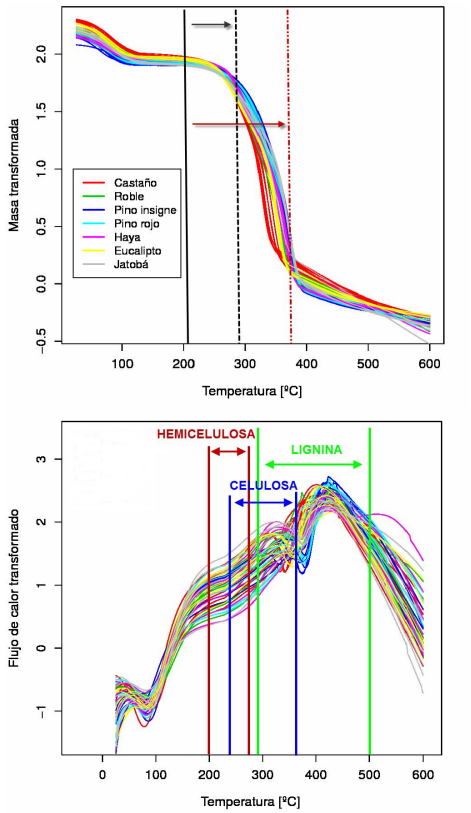

Figura 5: Curvas TG y DSC transformadas correspondientes a 7 especies de madera distintas

especie y matriz de varianzas covarianzas la muestral multiplicada por ciertos parámetros que controlan la variabilidad y la dependencia. Incluso en los peores escenarios, definidos por alta varianza y dependencia entre parámetros, se han obtenido porcentajes de clasificación correcta superiores siempre al 70 %.

4.2. Clasificación de materiales mediante segmentación de imágenes de microscopio SEM

Como ya se ha comentado, las técnicas de obtención y procesado de imágenes aportan una valiosa fuente de datos para la *clasificación supervisada* de todo tipo de materiales. De hecho, en Mallik et al. (2011), se muestra una nueva metodología para la clasificación de maderas comerciales a partir de la segmentación de micrografías obtenidas por un microscopio electrónico de barrido a 1500 aumentos. En este artículo se muestra paso a paso, la metodología propuesta: preparación de muestras, toma de micrografías a 1500 aumentos, mejora del contraste de las imágenes, segmentación de las mismas, extracción de vectores de características representativas relacionadas con la geometría y distribución de las traqueidas de la madera y, finalmente, aplicación de los modelos de clasificación multivariante (LDA, clasificación cuadrática, regresión logística, SVM y redes neuronales).

En la Figura 6 se observa el resultado de la aplicación de procesos de segmentación de imágenes a las micrografías de una de las muestras de madera industrial. De forma muy esquemática, se puede definir este proceso como el conjunto de técnicas diseñadas para reducir una imagen a, en este caso, únicamente dos tonos de píxeles, de forma que se puedan observar objetos conexos (las traqueidas).

Después de la extracción de características relevantes (5 en este caso de la madera) y la aplicación de los métodos de clasificación, se consiguieron porcentajes de clasificación correcta en torno al 80 %, aplicando para ello procesos de validación cruzada leave-one-out y procesos de validación externa. Es interesante destacar que una muy buena alternativa para la extracción de caracterísicas representativas de una imagen, en este caso micrografía SEM, fue la estimación de la dimensión fractal (Mallik et al., 2011).

5. Conclusiones

En este artículo se expuso la aplicación de técnicas de modelización y clasificación en el campo de la ingeniería de los materiales, concretamente se abordó el caso del modelado óptimo de curvas de degradación térmica. Se presentó un método basado en la mezcla de logísticas generalizadas. Para la optimización se propone el empleo de una técnica híbrida que permite aprovechar las características de los algoritmos evolutivos al combinarlos con algoritmos clásicos. Estos métodos propuestos se aplicaron a casos reales como la modelización de maderas

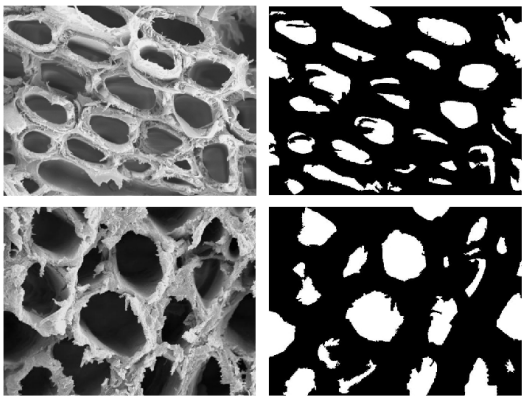

Figura 6: Proceso de segmentación de imágenes. A la izquierda, imágenes de pino insigne (arriba) y pino rojo (abajo) después de la mejora del contraste; a la derecha, micrografías después del proceso de segmentación

o a nanocompuesto, lo que además permiten el estudio de diferentes propiedades ingenieriles de los materiales.

También se presentaron algunas aplicaciones de clasificación supervisada de maderas industriales, usando diferentes técnicas estadísticas como la clasificación de datos funcionales o la segmentación de imágenes.

La progresiva aparición de nuevos materiales y su futura aplicación en distintos campos, como es el caso de los smart materiales o los biomateriales de uso en medicina, generará nuevas necesidades y soluciones para cuya consecución la aplicación de herramientas estadísticas es fundamental. Se concluye que, dado que los problemas de modelización de datos de laboratorio generan interesantes y complejos retos para ser investigados, divulgados e implantados, la propuesta de nuevos modelos estadísticos adaptativos como los aquí abordados, será sin duda una de las líneas emergentes que verán su expansión en un futuro inmediato.

6. Agradecimientos

Los autores de este trabajo agradecen las ayudas recibidas del proyecto del Ministerio de Ciencia e Innovación MTM2011-22392 (fondos FEDER incluídos).

Referencias

[1] Artiaga, R., Varela, A., Mier, J. L., García, A., Losada, R. y Naya, S. (2003). Study of a curing reaction of an epoxy resin. *Mater. Sci.* **426-432**, 2163-2168.

[2] Artiaga, R., Varela, A., Mier J.L., García, A., Losada, R. y Naya. S. (2003). Microstructural changes in a Ni-Based Super-Alloy induced by Thermal Treatment. *Mat. Sci.* **426-432**, 749-754.

[3] Artiaga, R., Cao, R., Naya, S., González-Martín, B., Mier, J.L. y García, A. (2005). *Separation of overlapping proceses from TGA data and verification by EGA.* J. of ASTM International. New York, (USA).

[4] Artiaga, R., López-Beceiro, J., Tarrío-Saavedra, J. Gracia-Fernández, C. Naya, S. Mier, J. L. (2011). Estimating the reversing and non-reversing heat flow from standard DSC curves in the glass transition region. *J. Chemom.* **25,** (6), 287-294.

[5] Artiaga R., López-Beceiro J., Tarrío-Saavedra J., Mier J., Naya S., Gracia C. (2011). Oxidation Stability of Soy and Palm Based Biodiesels Evaluated by Pressure Differential Scanning Calorimetry. *ASTM Special Technical Publication.* **1477**, 29-41.

[6] Barbadillo, F., Fuentes, A. Naya, S,. Cao, R., Mier, J. L. y Artiaga, R. (2007). Evaluating the logistic mixture model on real and simulated TG curves. *J. Therm. Anal. Calorim.* **87**, 1, 223-227.

[7] Brown. M.E. Maciejewski. M. Vyazovkin S. Nomen R. Chao. L. Malek J. y Mitsuhashi T. (2000). *Computational aspects of kinetic analysis. Thermochim. Acta.* **355**, 125-143.

[8] Cao, R., Naya S., Artiaga, R., García, A. y Varela, A. (2004). Logistic approach to polymer degradation in dynamic TGA. *Polym. Degrad. Stab.* **85**, 667-674.

[9] Cao, R. y Naya, S. (2009). Nonlinear regression checking via local polynomial smoothing with applications to thermogravimetric analysis. *J. Chemom.* **23, 6**; 275-282.

[10] Conesa, J. A. (2000). *Curso Básico de Análisis Térmico.* Editorial Club Universitario. Madrid.

[11] Francisco-Fernández, M., Tarrío-Saavedra, J., Mallik A., Naya, S. (2012). A comprehensive classification of wood from thermogravimetric curves. *Chemometrics Intell. Lab. Syst.* **118**, 159-172.

[12] López Beceiro, J. (2011). *Modelización de la transición vítrea con relajación entálpica a partir de datos térmicos.* Tesis Doctoral. Universidade da Coruña.

[13] López-Beceiro, J., Artiaga, R,. Gracia-Fernández, C., Tarrío- Saavedra, J., Naya, S. y Mier J. L. (2011). Comparison of olive, corn, soybean and sunflower oils by PDSC. *J. Therm. Anal. Calorim.* **104,** (1), 169-175.

[14] López-Beceiro, J., Pascual-Cosp, J., Artiaga, R., Tarrío- Saavedra, J. y Naya S. (2011). Thermal characterization of ammonium alum. *J. Therm. Anal. Calorim.* **104,** (1), 127-130.

[15] López-Beceiro, J., Gracia-Fernández, C. y Artiaga, R. (2013). A kinetic model that fits nicely isothermal and non-isothermal bulk crystallizations of polymers from the melt. *Eur. Polym. J.*, **49**, 8, 2233-2246.

[16] Mallik, A., Tarrío-Saavedra, J., Francisco-Fernández, M. y Naya, S. (2011). Classification of wood micrographs by image segmentation. *Chemometrics Intell. Lab. Syst.* ,**107**, 351-362.

[17] Naya, S., Cao, R. y Artiaga, R. (2003). Local polynomial estimation of TGA derivatives using logistic regression for pilot bandwidth selection. *Thermochim. Acta.* **6**, 319-322.

[18] Naya, S., Cao, R., Artiaga, R. y García A. (2006). New method for polymer classification by nonparametric regression with functional data. *Mater. Sci.* **512**, 1094-1098.

[19] Naya, S., Cao, R., López-de-Ullibarri, I., Artiaga, R., Barbadillo, F. y García, A. (2006). Logistic mixture model vs Arrhenius for kinetic study of degradation of materials by dynamic thermogravimetric analysis. *J. Chemom.* **20**, 158-163.

[20] Naya, S, Martínez-Vilariño, S. y Artiaga. R. (2009). Effects of thermal cycling on permeability and thermal properties of nanoclay-epoxy composites. *DYNA*. **84**, 2. 151-156.

[21] Naya, S. (2011). *Modelización de curvas en Análisis Térmico.* Editorial Academia Española. Saarbrücken. (Alemania).

[22] R Development Core Team. (2013). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org.

[23] Rios-Fachal, M., Gracia-Fernández, C., López-Beceiro, J., Gómez-Barreiro, S., Tarrío-Saavedra, J., Ponton, A., y Artiaga, R. (2013). Effect of nanotubes on the thermal stability of polystyrene. *J. Therm. Anal. Calorim.* **113**, 2, 481-487.

[24] Ríos-Fachal, M., Tarrío-Saavedra, J., López-Beceiro, J, Naya, S. y Artiaga R. (2014). Optimizing fitting parameters in thermogravimetry. *J. Therm. Anal. Calorim.* DOI: 10.1007/s10973-013-3623-0.

[25] Tarrío-Saavedra, J., López-Beceiro, J. Naya S. y Artiaga R. (2008). Effect of silica content on fumed silica/epoxy composites. *Polym. Degrad. Stab.* **93**, 12, 2133-2137.

[26] Tarrío-Saavedra, J., Lopez-Beceiro, J., Naya, S. y Artiaga, R. (2010). Influential factors on the oxidation stability of biodiesel: statistical study. *DYNA.* **85-4**, 341-350.

[27] Tarrío-Saavedra, J., Lopez-Beceiro, J., Naya, S., Gracia, C. y Artiaga, R. (2011). Controversial effects of fumed silica on the curing and thermomechanical properties of epoxy composites. *Express Polym Lett.* **48**, 4-6, 382-395.

[28] Tarrío-Saavedra, J., Francisco-Fernández, M., Naya, S., López-Beceiro, J. y Artiaga, R. (2010). Functional nonparametric classification of wood species from thermal data. *J. Therm. Anal. Calorim.* **104**, 87-100.

[29] Tarrío-Saavedra, J., Francisco-Fernández, M., Naya, S., López-Beceiro, Gracia-Fernández, C. y Artiaga, R. (2013). Wood identification using pressure DSC data. *J. Chemom.* **427**, 475-487.

[30] Turi, A. (1997). *Thermal characterization of polymeric materials.* Academic Press, San Diego, CA. (USA).

Acerca de los autores

Salvador Naya Fernández es profesor titular de Estadística en la Universidad de A Coruña, contando con la acreditación de catedrático del área por la ANECA desde el 2012. Imparte docencia en los grados de ingeniería industrial y naval en la Escuela Politécnica Superior de Ferrol y en diferentes másteres, como el máster interuniversitario de Técnicas Estadísticas o el doble máster internacional de la Universidad de A Coruña con la Paris Diderot en Materiales Complejos. Cuenta con varias publicaciones en estadística aplicada a la ingeniería y libros de divulgación estadística. Es miembro electo del ISI y ha ocupado distintos cargos de gestión en la universidad. En la actualidad es el director de la cátedra Jorge Juan, órgano que depende de la Universidad de A Coruña y del Ministerio de Defensa.

Javier Tarrío Saavedra es profesor de Estadística en la Universidad de A Coruña, impartiendo materias de control estadístico de calidad y fatiga termomecánica, entre otras. Cuenta con la titulación en Ingeniería Industrial y con el doctorado en Estadística e Investigación Operativa por la UDC. Imparte

docencia en materias del área de ingeniería industrial y naval y en diferentes másteres como el máster interuniversitario de Técnicas Estadísticas o el doble máster internacional de esta universidad gallega con la Paris Diderot en Materiales Complejos. Sus actividades como investigador se centran en la aplicación de técnicas estadísticas a la caracterización térmica y reológica de materiales complejos y combustibles, aprendizaje estadístico y bibliometría. Cuenta con varias publicaciones en estadística aplicada a estas áreas.