XII Congreso Galego de Estatística e Investigación de Operacións
Lugo, 22, 23 e 24 de outubro de 2015

CLASIFICACIÓN DE FONDOS MARINOS A PARTIR DE DATOS ACÚSTICOS

Javier Tarrío-Saavedra[1], Noela Sánchez-Carnero[2], Andrés Prieto Aneiros[3]

[1]Escola Politécnica Superior. Grupo MODES. Ferrol. Universidade da Coruña
[2]Grupo de Oceanografia Fisica. Universidade de Vigo
[3]Facultade de Informática. Grupo M2NICA. Universidade da Coruña.

RESUMEN

El objetivo de este estudio es la clasificación de fondos marinos en regiones costeras a partir de datos acústicos, en particular aquellos obtenidos mediante técnicas de sónar. Para ello se ha propuesto un enfoque alternativo de clasificación estadística supervisada y no supervisada aplicando técnicas de Análisis de Datos Funcionales (FDA). Se ha aplicado una versión FDA del algoritmo K-medias para identificar los tipos de fondos marinos en la ría de Cedeira. Además, se han aplicado métodos de clasificación supervisada FDA, en particular un modelo funcional lineal generalizado (GLM) y un modelo funcional aditivo generalizado (GSAM), para clasificar tres tipos distintos de suelo a partir de los datos obtenidos por sónar en el área de Cabo de Palos (Murcia). Los resultados indican que la introducción de técnicas FDA podría ser una alternativa al análisis tradicional multivariante para clasificar fondos marinos, sin tener que hacer una extracción selectiva de características relevantes de las curvas, y requiriendo por tanto un menor conocimiento previo de conceptos acústicos.

1. INTRODUCIÓN

Los datos acústicos, obtenidos mediante sónar, aparte de informar acerca de la batimetría y orografía del fondo, aportan información acerca de la textura del suelo (presencia o no de vegetación, tipo de vegetación, tipo de suelo: arenoso, limoso, pedregoso…). Es por esto último que han sido utilizados para identificar tipos de fondos marinos (Rodríguez-Pérez et al., 2014), tarea muy relacionada con el aprovechamiento de recursos (materias primas, flora, fauna) y el conocimiento del medio.

La clasificación de fondos marinos es una tarea compleja que en los últimos años se ha abordado desde el punto de vista de la estadística multivariante. Así, se suelen aplicar métodos de clasificación supervisada y no supervisada a una base de datos compuesta por las características relevantes de los pulsos acústicos obtenidos mediante sónar. La elección de un vector de características representativo de la textura del fondo no es en absoluto trivial. Hoy en día existen varias alternativas que combinan la elección de características con significación física a partir de los pulsos acústicos con métodos de reducción de dimensión como es el Análisis de Componentes Principales, PCA (Legendre et al, 2002; Preston y Kirlin, 2003). Sin embargo, la clasificación automática de fondos marinos es todavía un tema abierto. La aplicación de técnicas de clasificación de Análisis de Datos Funcionales, FDA, (Ramsay and Silverman, 2005) puede ser de gran utilidad en este campo al utilizar como variable regresora las propias curvas de pulsos acústicos, sin tener que partir de un conocimiento previo del problema físico para elegir el vector de variables independientes.

2. OBTENCIÓN DE DATOS

Se han obtenido medidas acústicas a través de 62 transectos realizados en la Ría de Cedeira (Figura 1), utilizando un single-beam echosounder (EA400P Simrad) que trabaja con un transdutor de 38/200

kHz, acoplado a un barco de 6.95 m de eslora. Las frecuencias de 200 kHz y 38 kHz han sido operadas con longitudes de pulso de 256 ms y 1024 ms. La velocidad del barco se ha mantenido entre 4 y 5 nudos (Rodríguez-Pérez et al., 2014). Con respecto a los datos tomados en Murcia, se ha empleado un transductor Simrad 38/200 Combi C, que combina dos transductores y un sensor de temperaturaen un único cuerpo.

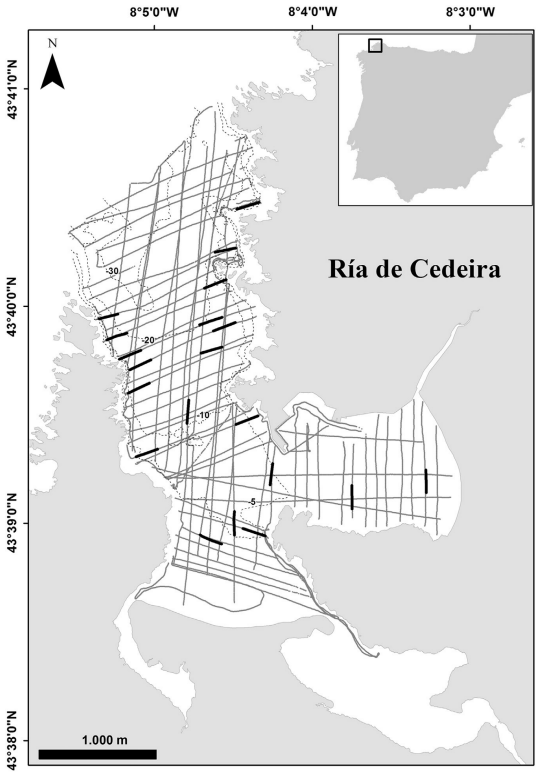

Figura 1. Transectos realizados en la Ría de Cedeira (Rodríguez-Pérez et al., 2014).

Una vez obtenidos los datos, han de hacerse realizarse una seria de transformaciones con el objeto de que el dato transformado aporte sólo información acerca de la textura del fondo. La corrección de profundidad se practica para que dos ecos que han sido obtenidos en un mismo tipo de fondo, con una misma inclinación, aunque a diferentes profundidades, sean prácticamente iguales. Se define una profundidad de referencia y, de acuerdo con el método de Pouliquen (Rodríguez-Pérez et al., 2014), los ecos son también escalados en tiempo y potencia.

3. TÉCNICAS DE CLASIFICACIÓN FDA

El Análisis de Datos Funcionales (FDA) es una nueva rama de la estadística que analiza la información contenida en curvas, super_cies u otra estructura variando sobre un continuo: tiempo, espacio, longitud de onda, etc. El FDA está relacionado con el análisis de datos en forma de funciones (Ferraty y Vieu, 2006). X se dice variable funcional si toma valores en un espacio funcional normado o seminormado, E, y una base de datos funcional $\{X_1,\ldots,X_n\}$ es la observación de n variables funcionales $X_1,\ldots,X_n$ idénticamente distribuidas como X (Tarrío-Saavedra et al., 2011).

Por otro lado, se define clasificación como un conjunto de técnicas de Análisis Multivariante y FDA diseñadas para agrupar o clasificar los elementos de una muestra, definidos por un conjunto de características (variables). Puede ser supervisada y no supervisada.

La clasificación no supervisada es la agrupación de individuos de los que no se conoce su clase (ni el número de grupos diferentes) según sus características. Existen diferentes métodos de clasificación no supervisada como el clúster jerárquico, K-medias, etc. En el caso funcional, existen diversas alternativas entre las que se encuentra una versión K-medias funcional (Febrero-Bande y Oviedo de la Fuente, 2012). Se elige el nº de grupos (K) y se establecen los centros (funciones) al azar. Se asignan los individuos a los

grupos y se vuelven a calcular sus centros (funciones). Se repiten los pasos anteriores hasta que se repite la clasificación, se llega a un máximo número de iteraciones o el desplazamiento de centros es menor que una determinada tolerancia.

La clasificación supervisada tiene por objeto asignar una clase a un individuo usando un modelo de clasificación previamente construido a partir de una muestra de entrenamiento (compuesta por individuos de los que se sabe su clase). Cada individuo está definido por un vector de características o mismo una función. Se define muestra de entrenamiento como un conjunto de individuos, caracterizados por un vector de variables de los que se conoce su clase. El modelo de clasificación o función discriminante se estima a partir de estos datos. Finalmente, la muestra de test es un grupo de individuos de los que no sabemos su clase. El objetivo de la clasificación supervisada es la determinación de las funciones discriminantes a partir de las variables originales, que permiten decidir en qué clase debe estar cada elemento, utilizando como criterio de asignación la proximidad de cada elemento a las distintas clases.

En un problema clásico de clasificación supervisada, se predice la clase *Y* a partir de un vector de características *X* (Wehrens, 2011). En la clasificación FDA, *X* es una función, con lo que el objetivo es predecir *Y* a partir de una varaible funcional *X*. En particular, se estima la probabilidad a posteriori de pertenencia a cada grupo (regla de Bayes):

$$p_g(X) = P(Y = g \vee ¿\chi = X) = E(1_{Y=g} \vee ¿\chi = X)\ \hat{Y} = arg\,max\ \hat{p}_g(X)$$

La clase de cada individuo se puede estimar utilizando clasificadores logísticos o no paramétricos mediante el paquete R fda.usc (Febrero-Bande y Oviedo De la Fuente, 2012): clasificadores k vecinos más próximos, kernel y logística a través de modelos aditivos generalizados y modelos aditivos generalizados.

En este trabajo se aplican los modelos de clasificación FDA lineal generalizados y aditivos generalizados. En el modelo lineal funcional generalizado (GLM) la respuesta escalar y se estima con variables funcionales $\{X_q(t)\}_{q=1}^{Q}$ y también no funcionales $Z = \{Z_j\}_{j=1}^{J}$,

$$y_i = g^{-1}\left(\alpha + Z_i\beta + \sum_{q=1}^{Q} \left\langle X_i^Q(t), \beta_q(t) \right\rangle\right) + \varepsilon_i$$

donde $g()$ es el link inverso y ε_i los errores de media cero y varianza σ^2 .

En estos modelos se requiere representar $X(t)$ y $\beta(t)$ en una base adecuada. Ramsay and Silverman (2005) propusieron su representación en función de una base B-spline, Fourier o Wavelets. Por otro lado, Cardot et al. (1999) introdujeron la regresión con Componentes Principales funcionales (FPC), mientras que Preda et al. (2007) propusieron la regresión con Mínimos Cuadrados Parciales funcionales (FPLS).

El modelo lineal aditivo funcional espectral generalizado, GSAM (Febrero-Bande y González-Manteiga , 2013), tiene la forma

$$y_i = g^{-1}\left(\alpha + \sum_{j=1}^{J} f_j\left(Z_j^J\right) + \sum_{q=1}^{Q} s_q\left(X_i^Q(t)\right)\right) + \varepsilon_i$$

donde $f()$ y $s()$ son funciones suaves.

Como no se puede probar un modelo con la misma muestra que se usa para entrenarlo, para asegurar que un método de clasificación es válido para la población, no sólo para la muestra, es necesario aplicar modelos de validación como el método de validación cruzada. En el caso del presente trabajo, se ha empleado el procedimiento 10-fold cross-validation (la muestra es grande) para obtener una estimación de la proporción de clasificación incorrecta así como su variabilidad. La muestra se divide en 10 bloques de

forma aleatoria, se entrena el modelo de clasificación con nueve de ellos y se deja fuera el restante. Una vez estimado el modelo, se predicen las clases de los individuos del bloque que quedó fuera y se obtiene una proporción de clasificación incorrecta. Esto se repite 10 veces, hasta haber sacado fuera una vez cada uno de los bloques que componen la muestra.

4. RESULTADOS

En primer lugar se ha aplicado un método de clasificación no supervisada K-medias FDA para evaluar la posibilidad de identificar los distintos fondos marinos de la Ría de Cedeira a partir de curvas acústicas. Presenta la ventaja de ser un procedimiento más automático que los actualmente utilizados, no precisando un conocimiento profundo de la naturaleza física de los datos. En la Figura 2 se muestran cada una de las curvas acústicas obtenidas para cada muestra de terreno. Cada curva está compuesta de dos ecos relacionados con la incidencia de la onda sonora en el fondo marino y su posterior rebote, respectivamente. La Figura 2 muestra el grupo asignado a cada curva, de cuatro posibles (que debería relacionarse con dos tipos de roca, arena gruesa y arena fina). Se han calculado los centros representativos, en este caso también curvas acústicas, las medias funcionales.

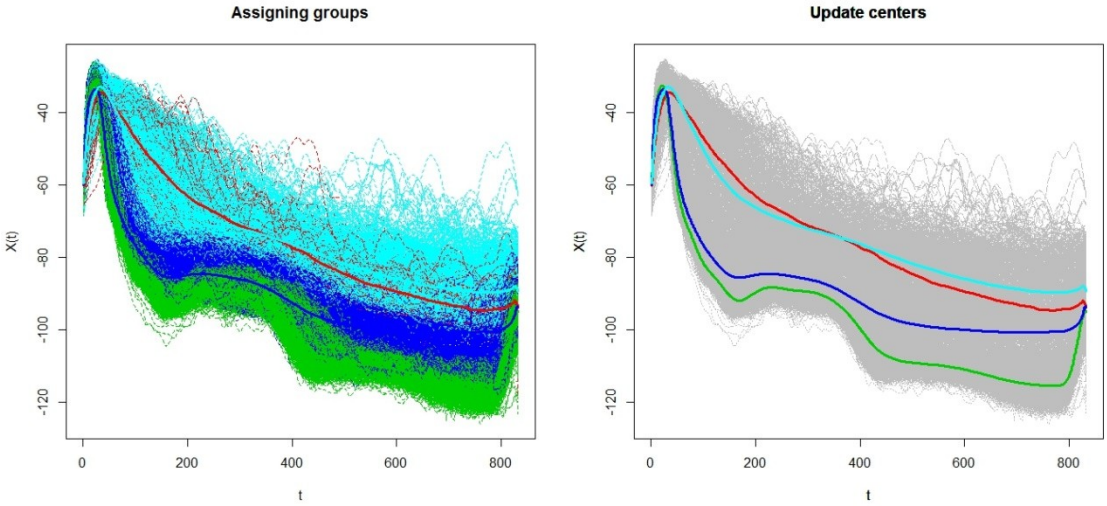

Figura 2. Panel izquierdo: ecos de cada muestra y grupo asignado (cuatro en total). Panel derecho: medias funcionales de cada grupo (centros).

En la Figura 3 se observa el mapa de la Ría de Cedeira con los transectos realizados por el barco. Se ha marcado el grupo al que pertenece cada muestra a partir de su curva de ecos utilizando el método K-medias funcional y la distancia L_1. Si se definen tres clases, se distinguen correctamente dos tipos de roca (en rojo y negro) y un tipo de arena si se comparan los resultados con otros estudios previos (Rodríguez-Pérez et al., 2014). Si se clasifican las curvas acústicas en cuatro clases, se distinguen tres tipos de roca y uno de arena. Esto es debido a que existen más diferencias desde un punto de vista acústico entre los grupos de rocas que entre los tipos de arena.

En el caso de Cabo de Palos (Murcia) se han evaluado métodos de clasificación supervisada FDA para identificar el tipo de fondos marino partiendo de un número definido de clases posibles: arena (CP01), arena con vegetación (CP02), rocas (CP03). De la primera, segunda y tercera clase se han obtenido 9295 curvas, 11235 y 4967, respectivamente, en total 25497 curvas evaluadas en 260 puntos/tiempos. Primeramente se comprobó que los datos acústicos son una característica discriminante del tipo de fondo. Para ello se realizó un contraste ANOVA funcional por el método de proyecciones aleatorias (Cuesta-Albertos y Febrero-Bande, 2010 ; Tarrío-Saavedra et al., 2011), resultando que la media de cada grupo son significativamente diferentes (ver también Figura 4).

El siguiente paso es la aplicación de los modelos de clasificación supervisada FDA mencionados. La Tabla 1 muestra las proporciones de clasificación incorrecta. *Y* es la respuesta cualitativa, clase. *X*, los ecos, variable funcional. Se evalúan los siguientes modelos de clasificación funcional GLM y GSAM mediante un procedimiento 10 fold cross-validation:

a) GLM1, donde X y β se representan suavizados con dos bases B-spline, de 15 y 7 elementos, respectivamente.
b) GLM2: donde X y β se representan suavizados con dos bases B-spline de 7 elementos.
c) GLM3: X se representa mediante una base de 10 FPC.
d) GSAM1: X se representa por una base de 10 FPC.
e) GSAM2: X se representa por una base de 10 FPC.

La Tabla 2 muestra que, utilizando el modelo GSAM2, se clasifican correctamente el 98.7% de los ecos, prácticamente la totalidad de los ecos. Por tanto, a la vista de las altas proporciones de identificación correcta, se demuestra que la aplicación de técnicas FDA de clasificación supervisada podría ser de gran utilidad para la clasificación de fondos marinos.

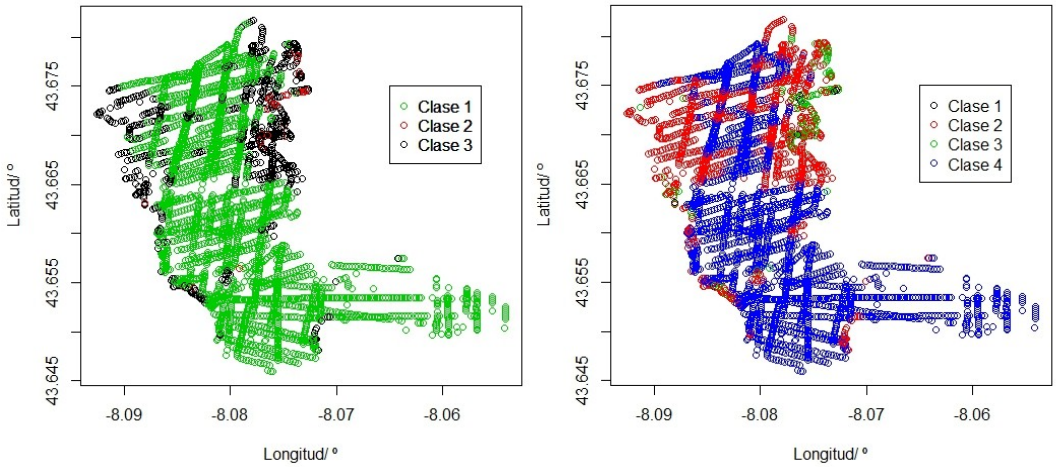

Figura 3. Mapa de la Ría de Cedeira con los transectos realizados por el barco. Se ha marcado el grupo al que pertenece cada muestra a partir de su curva de ecos utilizando el método K-medias funcional. Panel izquierdo: fijando 3 clases o grupos. Panel derecho: fijando 4 grupos diferentes.

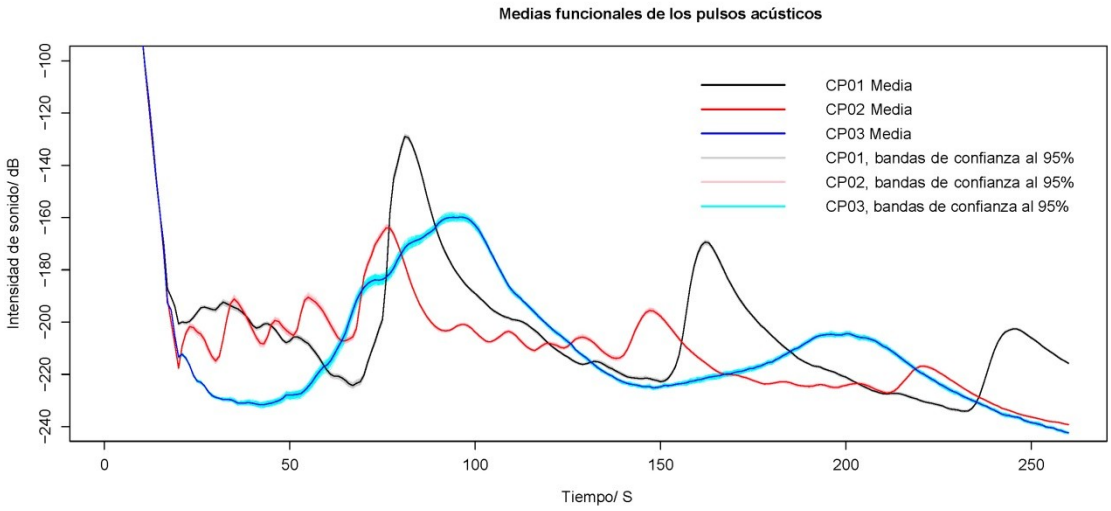

Figura 4. Medias funcionales y bandas de confianza al 95% para las curvas acústicas correspondientes a las tres clases de suelo de Cabo de Cabo de Palos (Murcia).

Tabla 1. Proporciones de clasificación incorrecta (media y desviación típica) con 10-fold cross-validation.

	Proporción de clasificación incorrecta	
Métodos	Media	Desviación estándar
GSAM1	0.050	0.003
GLM1	0.110	0.006
GLM2	0.151	0.007
GSAM2	0.013	0.002
GLM3	0.082	0.007

5. CONCLUSIONES

En este trabajo se han aplicado técnicas FDA de clasificación supervisada y no supervisada para la clasificación de tipos de fondo marino a partir de los datos obtenidos mediante una sonda de sónar. Se ha conseguido clasificar correctamente (entre fondo rocoso, arenoso y arenoso con vegetación) más del 98% de las muestras, utilizando para ello un modelo de clasificación supervisada funcional GSAM. Por tanto, el uso de técnicas FDA podría ser de gran utilidad en este campo.

Agradecimientos

Los autores agradecen la ayuda de los proyectos Simulación numérica de problemas hidroacústicos de alta frecuencia en medios mariños – SIMNUMAR (EM2013/052) y MTM2013-41383P (fondos FEDER incluidos).

REFERENCIAS

Cardot H., Ferraty F., Sarda P. (1999). Functional Linear Model. Statistics and Probability Letters, 45(1), 11-22.

Cuesta-Albertos J.A., Febrero-Bande M. (2010). A simple multiway anova for functional data. Test. 19, 537-57.

Febrero-Bande M., Oviedo de la Fuente M. (2012). Statistical computing in functional data analysis: The R package fda.usc. Journal of Statistical Software, 51(4):128.

Febrero-Bande M., González-Manteiga W. (2013). Generalized additive models for functional data. TEST, 22(2):278292.

Ferraty F., Vieu P. (2006). Nonparametric functional data analysis. Springer Series in Statistics, New York.

Francisco M., Tarrío J., Mallik A., Naya S.(2012). A comprehensive classification of wood from thermogravimetric curves. Chemometrics and Intelligent Laboratory Systems. 118, 159-172.

Legendre P., Ellingsen K. E., Bjørnbom E., Casgrain P. (2002). Acoustic seabed classification: improved statistical method. Canadian Journal of Fisheries and Aquatic Sciences, 59, 1085-1089.

Preda C., Saporta G., Lévéder C. (2007). PLS classification of functional data. Computational Statistics, 22, 223-235.

Preston J. M., Kirlin T. L. (2003). Comment on “Acoustic seabed classification: improved statistical method”, Canadian Journal of Fisheries and Aquatic Sciences, 60, 1299-1300

Ramsay J., Silverman B.W. (2005). Functional Data Analysis. Springer.

Rodríguez-Pérez, D., Sánchez-Carnero, N., Freire, J. (2014). A pulse-length correction to improve energy-based seabed classification in coastal areas. Continental Shelf Research, 77, 113.

Tarrío, J., Naya, S., López, J., Artiaga, R. (2011). Application of functional ANOVA to the study of thermal stability of micronano silica epoxy composites. Chemometrics and Intelligent Laboratory Systems. 105, 114-124.

Wehrens, R. (2011). Chemometrics with R. Multivariate Data Analysis in the Natural Sciences and Life Sciences. Berlin Heidelberg: Springer.