UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2017, Vol. 10, No. 1, pp. 129–144

Análisis multivariado de datos funcionales aplicado a curvas de encefalogramas

Multivariate functional data applied to encephalogram curves

Alexis Carrillo Ramirez[a]
alexiscarrillor@gmail.com

Olga Garatejo Escobar[b]
olga311003@gmail.com

Wilmer Pineda-Ríos[c]
wilmerpineda@usantotomas.edu.co

Resumen

Los desarrollos tecnológicos han hecho posible que los investigadores de muchas áreas dispongan de grandes volúmenes de información para un mismo individuo. Usualmente, estos datos pueden ser representados a través de curvas o en general de funciones. De ahí surge un nuevo campo de estudio en estadística denominado Análisis de Datos Funcionales (ADF). En el ADF la unidad básica de información es la función completa, más que un conjunto de valores (Ramsay & Dalzell 1991). Los métodos estadísticos usuales han sido adaptados a esta situación, en particular se ha desarrollado el análisis de conglomerados funcional por el método de k-medias. Dado que la actividad cerebral responde a una función de onda de la carga eléctrica de las neuronas sobre el tiempo, surge la oportunidad de aplicar el ADF a este tipo de registros. El objetivo de este trabajo es describir la aplicabilidad del análisis de conglomerados funcional por el método de k-medias para clasificar la actividad cerebral en ratas Norvegicus Wistar. Se realizó la conversión de los registros en funciones de onda, usando bases de Fourier, las que fueron analizadas de acuerdo con la metodología desarrollada en (Yamamoto 2012) y un análisis de correspondencias simples entre los conglomerados y las fases de actividad registradas manualmente en el hipnograma. Los conglomerados obtenidos hacen una categorización no supervisada consistente, especialmente respecto a los atributos de frecuencia y regularidad de las ondas.

***Palabras clave*:** electroencefalografía (EEG), datos funcionales, series de Fourier, aprendizaje automático, k-medias funcional, análisis de componentes principales funcionales..

[a]Especialista en Estadística Aplicada, Fundación Universitaria los Libertadores
[b]Especialista en Estadística Aplicada, Fundación Universitaria los Libertadores
[c]Docente de Tiempo Completo. Universidad Santo Tomás, Bogotá

Abstract

Technological developments have made it possible for researchers in many areas to have large volumes of information for the same individual. Usually these data can be represented through curves or in general functions. From this arises a new field of study in statistics called Functional Data Analysis (FDA). In the FDA the basic unit of information is the complete function, rather than a set of values (Ramsay & Dalzell 1991). The usual statistical methods have been adapted to this situation, in particular the analysis of functional conglomerates by the k-means method has been developed. Since the brain activity responds to a wave function of the neuronal charge over time, the opportunity arises to apply the FDA to this type of record. The objective of this work is to describe the applicability of the functional cluster analysis by the k-means method to classify brain activity in Norvegicus Wistar rats. The conversion of the registers into wave functions was carried out using Fourier bases, which were analyzed according to the methodology developed in (Yamamoto 2012) and a simple correspondence analysis between the clusters and the phases of activity manually recorded in the hypnogram. The obtained conglomerates make a consistent unsupervised categorization, especially with respect to the attributes of frequency and regularity of the waves.

***Keywords*:** electroencephalography, functional data analysis, Fourier series, machine learning, functional k-means, functional principal component analysis..

1. Introducción

A estudiar las señales emitidas por la actividad cerebral, el análisis de datos en neurociencias, suele ser complejo debido al gran volumen de información (Moser et al. 2009). A pesar de que el registro tiene características discretas, en realidad son funciones de onda por su naturaleza continua dado que depende del tiempo. Es decir, los datos no tienen una estructura escalar por cada unidad muestral, sino para cada una de ellas (electrodos) se cuenta con N respuestas a través del tiempo; por tanto la unidad básica de información es una función. Para el estudio de estas señales el análisis estadístico multivariado es insuficiente, dado que al procesar los datos, cada uno de estos serán funciones, por ello se recurre al ADF (Ramsay & Silverman 2005).

El ADF multivariado y, en particular, los análisis de componentes principales funcional (ACPF) y de conglomerados funcional (ACF), surgen como una alternativas de análisis de datos como los mencionados. En este trabajo se hace una aplicación del método de conglomerados funcional por k-medias a la clasificación de la actividad cerebral de la rata Norvegicus Wistar durante las 24 horas del registro del EEG. En análisis anteriores de este conjunto de datos, se utilizaban espectrogramas, que toman un segmento de la curva y, mediante un histograma se promedian los puntos del EEG. La desventaja de este tratamiento de la información es que los datos son estudiados como puntos y no como curvas.

Para procesar los registros obtenidos por el EEG, frecuentemente se recurre a paquetes como EEGLAB en MATLAB. El inconveniente con estos programas radica en que al ser código cerrado, los investigadores dependen de los desarrolladores para hacer modificaciones o ajustar los análisis a las condiciones particulares de su proyecto; además de los costos que conllevan la adquisición o actualización de los mismos. En vista de lo anterior, se identifica la oportunidad de poder realizar éstos análisis por medio del software `R` (R Core Team 2013) usando las librerías `fda` y `fda.usc`, los cuales permiten manipular adecuadamente los registros entregados por la electroencefalografía y de esta manera solventar tales dificultades.

2. Marco de Referencia

2.1. Electroencefalografía

Las neuronas, a través de una serie de reacciones químicas, incorporan o liberan iones (en su mayoría de sodio, potasio o calcio), produciendo cambios en las cargas eléctricas que se propagan a través de su membrana y se transmiten a otras neuronas en un sistema de comunicación electro-químico llamado sinapsis. De esta forma las neuronas codifican y transfieren la información que procesan. Los cambios pueden ser registrados como señales eléctricas, midiendo las diferencias de voltaje de un punto específico del cráneo en relación a un punto neutro del cuerpo. Por lo tanto, la actividad bioeléctrica cerebral puede captarse sobre el cuero cabelludo, en la base del cráneo, en cerebro expuesto, o en localizaciones cerebrales profundas. Para capturar la señal se utilizan diferentes tipos de electrodos, como: los superficiales que se aplican sobre el cuero cabelludo; los basales pueden aplicarse en la base del cráneo sin necesidad de procedimiento quirúrgico; los quirúrgicos, en cuya aplicación es necesaria la cirugía y pueden ser corticales o intracerebrales.

El registro de la actividad bioeléctrica cerebral recibe distintos nombres según la forma de captación. Se conoce como electroencefalograma (EEG) cuando se utilizan electrodos de superficie o basales, electrocorticograma (ECoG) si se utilizan electrodos quirúrgicos en la superficie de la corteza y Estéreo EEG (E-EEG) cuando se utilizan electrodos quirúrgicos de aplicación profunda (Doris, 2009). Junto con el registro de la actividad cerebral, también se puede registrar el nivel de actividad del sujeto evaluado. Normalmente se hacen observaciones del nivel de actividad, las cuales se pueden dividir en dos grandes categorías: sueño o vigilia. Dependiendo del proceso de investigación que se esté desarrollando, se pueden utilizar más clases de comportamiento. Su representación gráfica se conoce como hipnograma.

2.2. Análisis de datos funcionales

Las señales eléctricas que producen las células del cerebro al comunicarse pueden ser registradas por medio de un EEG. Cada registro depende del tiempo por ende

cada unidad de información en un tiempo determinado es una función. A continuación se presentan las definiciones necesarias para el posterior análisis descriptivo de datos que en adelante serán representados por funciones.

Una variable aleatoria toma valores en un espacio de funciones, como un espacio infinito dimensional. Así, una observación $f(t)$ de la variable aleatoria se denomina dato funcional en un instante t (Ferraty & Vieu 2006).

Definición 1. Un dato funcional $f(t)$, $t \in T \subset \mathbb{R}$, se representa como un conjunto finito de pares (t_i, x_i), $t_i \in T$, $i = 1, 2, \cdots, N$, donde N representa la cantidad de puntos de la variable funcional de interés.

Para un correcto análisis de las variables funcionales, es necesaria la siguiente definición:

Definición 2. Sea $L^2(T)$, con $T = [a, b] \subset \mathbb{R}$, el espacio de las funciones cuadrado integrable (Espacio de Hilbert):

$$L^2(T) = \left\{ f : \mathbb{R} \to \mathbb{R} | \int_a^b f(t)^2 dt < \infty \right\}$$

con producto interno

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt$$

A partir de la definición 2, se encuentra un conjunto de funciones que permiten aproximar las curvas por medio de suavizacion. Las funciones varían en amplitud y frecuencia, por lo tanto, la aproximación conveniente es en series de Fourier. Teniendo en cuenta que existen otras aproximaciones como wavelets y B-Splines.

2.2.1. Representación en series de Fourier:

Para modelar los datos experimentales como datos funcionales se aproxima a una función $f(t)$ por medio de la combinación lineal de funciones. La mejor representación para el estudio de la frecuencia y amplitud por dato funcional es en series de Fourier. Así, para un conjunto de datos discretos determinados en el tiempo se aproxima al dato funcional $f(t)$ de acuerdo con la siguiente expresión:

$$f(t) \approx \frac{a_0}{2} + \sum_{i=1}^{N} \left\{ a_i \cos\left(\frac{2\pi it}{N}\right) + b_i \sin\left(\frac{2\pi it}{N}\right) \right\}$$

Donde, a_0, a_i y b_i constantes con $i = 1, \cdots, N$. Una vez se obtienen los registros representados como funciones $f(t)$, es posible realizar el respectivo análisis estadístico de los objetos funcionales, como medidas de tendencia central, de dispersión o conglomerados por k-medias funcional entre otros.

2.2.2. Estadísticos descriptivos en datos funcionales

Sea el conjunto de datos funcionales $f_1,\ f_2\ ,\cdots,f_n$, definidos en $t \in [a,b]$ es un intervalo de tiempo. Las funciones descriptivas están dadas por las expresiones. (Ramsay, 2005)

Media: $\overline{f(t)} = \frac{1}{n}\sum_{i=1}^{n} f_i(t)$
Varianza: $s(t) = \frac{1}{n-1}\sum_{j=1}^{n}(f_j(t) - \overline{f(t)})^2$
Desviación estándar: $\sigma(t) = \sqrt{s(t)}$
Covarianza:
$Cov(f(t_1), f(t_2)) = \frac{1}{n-1}\sum_{j=1}^{n}(f_j(t_1) - \overline{f(t_1)})'(f_j(t_2) - \overline{f(t_2)})$
Correlación: $Cor(f(t_1), f(t_2)) = \frac{Cov(f(t_1), f(t_2))}{\sqrt{s(f(t_1))s(f(t_2))}}$

3. Análisis de conglomerados para datos funcionales

El ADF para este trabajo se realizará por medio de conglomerados aplicando el algoritmo de k-medias funcional. En general, el análisis de conglomerados clasifica toda muestra de datos con mínima variabilidad en grupos, de tal forma que entre grupos sean lo más variable posible, así los datos quedan clasificados en categorías. Para hallar conglomerados óptimos el algoritmo de k-medias hace uso de componentes principales. A continuación, se mostrará la teoría referente a componentes principales para funcionales.

3.1. Componentes principales para datos funcionales

El objetivo del análisis de componentes principales es considerar la máxima información dentro de una combinación lineal de autofunciones, obteniendo de tal forma que se tenga una base de menor dimensión. Se busca que la primera componente de dicha base contenga la mayor proporción posible de la variabilidad original, luego para el segunda componente se busca que contengan la máxima variabilidad restante y así sucesivamente para los otros componentes. El problema de Análisis de componentes principales es hallar los autovalores y autofunciones de la función covarianza $Cov(f(t_1), f(t_2))$.
Así, sean $\{f_1, f_2, \cdots, f_n,\}$ observaciones con sus correspondientes estadísticos media y covarianza:

$$\overline{f(t)} = \frac{1}{n}\sum_{i=1}^{n} f_i(t)$$

$$Cov(f(t_1), f(t_2)) = \frac{1}{n-1}\sum_{j=1}^{n}(f_j(t_1) - \overline{f(t_1)})'(f_j(t_2) - \overline{f(t_2)})$$

Se asume que cada f_j con $j = 1, 2, \cdots, n$ tiene una expansión en series de Fourier como en la sección 2.2.1:

$$f_j(t) = a_j \varphi(t) \tag{1}$$

Sea la matriz A, cuyas filas son los elementos a_j y $\varphi(t) \in L^2$ son las funciones de la base de Fourier. Así los factores de la función $Cov(f(t_1), f(t_2))$, se pueden escribir como:

$$\sum_{j=1}^{n} (f_j(t_1) - \overline{f(t_1)})' = \varphi(t_1)' A' \tag{2}$$

$$\sum_{j=1}^{n} (f_j(t_2) - \overline{f(t_2)}) = A\varphi(t_2) \tag{3}$$

Sustituyendo (2) y (3) en la función covarianza $Cov(f(t_1), f(t_2))$, se tiene:

$$Cov(f(t_1), f(t_2)) = \frac{1}{n-1} \varphi(t_1)' A' A \varphi(t_2) \tag{4}$$

Ahora, los autovalores y autofunciones de la función de covarianzas se encuentran solucionando la siguiente integral con $t_1, t_2 \in [a, b]$ y cada autofunción con expansión en base $\varphi(t)'$, $f_j(t) = b_j \varphi(t)'$ se plantea:

$$\int_a^b Cov(f(t_1), f(t_2)) f_j(t_2) dt_2 = \lambda_j f(t_1)$$

Reemplazando (4), en la anterior integral se tiene:

$$\int_a^b \frac{1}{n-1} \varphi(t_1)' A' A \varphi(t_2) b_j \varphi(t_2)' dt_2 = \Lambda_j b_j \varphi(t_1)'$$

$$\frac{1}{n-1} \varphi(t_1)' A' A \int_a^b \varphi(t_2) \varphi(t_2)' dt_2 b_j = \lambda_j b_j \varphi(t_1)'$$

Si se define la matriz W como sigue:

$$W = \int_0^T \varphi(t_2) \varphi(t_2)' dt_2$$

y se simplifica $\varphi(t_1)'$, se llega a un problema de autovalores multivariado o por matrices:

$$\frac{1}{n-1}A'AWb_j = \lambda_j b_j$$

Para la solución se tiene en cuenta, $W = W^{\frac{1}{2}}W^{\frac{1}{2}}$ y se multiplica a ambos lados, por $W^{\frac{1}{2}}$ así:

$$\frac{1}{n-1}W^{\frac{1}{2}}A'AW^{\frac{1}{2}}W^{\frac{1}{2}}b_j = \lambda_j W^{\frac{1}{2}}b_j$$

Si $u_j = W^{\frac{1}{2}}b_j$, entonces:

$$\frac{1}{n-1}W^{\frac{1}{2}}A'AW^{\frac{1}{2}}u_j = \lambda_j u_j$$

Los autovalores λ_j y las autofunciones $b_i = u_j W^{\frac{-1}{2}}$, ahora la solución se reduce a encontrar la matriz $W^{\frac{-1}{2}}$. (Julien Jacques, 2013)

A partir del análisis de los componentes principales, se logra la reducción dimensional que posteriormente permitirá agrupar los datos funcionales en conglomerados. Para este estudio se considera el método de k-medias funcional para obtener los conglomerados.

3.2. Conglomerados por el método k-medias

Para el análisis de objetos funcionales como los descritos en la definición 1, tal que $f(t) \in L^2$, se usará el algoritmo de componentes principales funcionales por k-medias (Yamamoto M, 2012).

En primer lugar se deben definir los siguientes argumentos:

- Sea $V = v_l$ con $(l = 1, \cdots, r)$, $v_l \in L^2, r < \infty$ las funciones que conforman la base ortonormal del subespacio de proyección.
- Sea P_v el operador proyección ortogonal definido como: $P_v : L^2 \to L^2_v$, es decir, el operador P_v va del espacio de los datos funcionales L^2 sobre el subespacio L^2_v, que es expandido por V.
- Sea $U = u_{ik}$ con $(i = 1, \cdots, n; k = 1, \cdots, q)$, donde u_{ik} es 1 si pertenece al conglomerado k y cero si pertenece a otro.
- Sea n_k el número de datos asignados al conglomerado k.
- Los centroides de cada conglomerado son:

$$\overline{x_k} = \frac{1}{n_k}\sum_{i=1}^{n} u_{ik}x_i$$

- Sea B_C el operador integral definido como:

$$B_C y(s) = \sum_{k=1}^{q} \frac{n_k}{n} \langle \overline{x_k}, y \rangle \overline{x_k}(s)$$

Con $y \in L^2$, $s \in T$

- Función objetivo:

$$g(U,V) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{q} u_{ik} \|x_i - p_v \overline{x_k}\|^2 \tag{5}$$

Según Yamamoto (2012), la función objetivo (5), se puede escribir como:

$$g(U,V) = \frac{1}{n} \sum_{i=1}^{n} \|x_i\|^2 - \sum_{l=1}^{r} \langle v_l, B_c v_l \rangle \tag{6}$$

$$g(U,V) = \frac{1}{n} \sum_{i=1}^{n} \|x_i - p_v x_i\|^2 + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{q} u_{ik} \|p_v x_i - p_v \overline{x_k}\|^2 \tag{7}$$

El algoritmo k-medias vía componentes principales funcionales (KCPF), se reduce a minimizar la función objetivo (5), respecto a U y V simultáneamente, en particular si en (4) $n = q$, entonces el algoritmo de (KCPF) se convierte en el método usual de análisis de componentes principales (ACP). Así, para minimizar la función g(U,V), se siguen los siguientes pasos:

- Paso 1: Se inicia definiendo a $U = u_i k$ con $(i = 1, \cdots, n; k = 1, \cdots, q)$, teniendo en cuenta los parámetros antes descritos.
- Paso 2: Se minimiza el segundo término en la ecuación (6), fijando U sobre V.
- Paso 3: Se minimiza el segundo término en la ecuación (7), Fijando V sobre U.
- Paso 4: Se va al paso 2 hasta que los centroides $\overline{x_k}$ queden fijos.

Sin embargo, no se garantiza que converja en un mínimo global, ya que en particular el algoritmo (KCPF) es sensible a óptimos locales. Dado que el objetivo de este trabajo es aplicar este análisis a datos funcionales. A continuación, se explicará el origen de los datos a ser analizados, que son mediciones de las señales emitidas por las neuronas en el cerebro de la rata Norvegicus Wistar.

4. Metodología

La base de datos fue facilitada por integrantes del Semillero Neurociencia y Comportamiento de la Universidad de los Andes, dirigido por el profesor Fernando

Cárdenas. Los datos corresponden a un registro de 24 horas de una rata de laboratorio de la especie Norvegicus Wistar. Los valores registrados corresponden a los siguientes canales:

- Hipnograma: Registro del estado de actividad del sujeto experimental.
- Electromiograma EMG.
- Registro de actividad cerebral en la zona parietal, electrodo 1 (P1), cuya unidad de medida son voltios.
- Registros de actividad cerebral, lóbulo frontal, electrodo 3 (F3), medida en voltios.

Los datos describen el valor de la diferencia de carga del electrodo de registro respecto a un electrodo de referencia. La frecuencia de registro es de 400 datos por segundo. En primera instancia se seleccionó el canal F3 para el procesamiento, puesto que en el lóbulo frontal se encuentran las áreas corticales asociadas a la actividad motora, por lo cual se observa con mayor claridad el sueño paradójico y la actividad cerebral en vigilia. Una vez cargados los datos se procede a segmentarlos cada dos segundos. Esto se logró convirtiendo el vector en una matriz de datos orientado por columnas y con un límite de 800 filas. Posteriormente, se seleccionaron los segmentos cuyos valores estuvieran dentro del rango -350 a 350 mv, ya que valores por fuera de éstos indicaban una anomalía en el registro por variables extrañas y ajenas a la actividad cerebral. La base de datos organizada y filtrada se convirtió en un objeto tipo dato funcional con una base de Fourier, que finalmente fueron procesados con el algoritmo de k-medias funcional para seis conglomerados. El valor de $k = 6$ se planteó con el fin de poder comparar la relación entre los conglomerados y las categorías del registro de hipnograma.

Para el desarrollo del código, se utilizó `R-Studio`, que es un entorno de desarrollo integrado (IDE, por sus siglas en Inglés) para `R`. Para la ejecución de las pruebas estadísticas pertinentes se utilizaron los paquetes `fda` y `fda.usc`. El software funcionó sobre el sistema operativo `Ubuntu 14.04 LTS`, la versión de `R` para el desarrollo del análisis es la 3.2.1, la versión de `R-Studio` es la 0.98.1091. La versión del paquete `fda` es la 2.4.4 y del paquete `fda.usc` es la 1.2.1

5. Resultados y discusión

Con el software `R` haciendo uso de las librerías, `fda` y `fda.usc`, se procesaron los datos obtenidos en el (EEG), convirtiéndolos en funciones por medio de bases de Fourier. Cada función representa 800 datos transcurridos por 2 segundos, el registro total se realiza durante 24 horas. En la figura 1 se observan las funciones de onda. Por la sección 2.2.1 con cada función tiene una aproximación en series de Fourier como sigue:

$$f_j(t) \approx \frac{a_0}{2} + \sum_{i=1}^{800} \left\{ a_i \cos\left(\frac{2\pi i t}{800}\right) + b_i \sin\left(\frac{2\pi i t}{800}\right) \right\}, \tag{8}$$

donde a_0, a_i y b_i constantes con $i = 1, \cdots, 800$ y $f_j(t) \in L^2$ con $j = 1, \cdots, \approx 43200$.

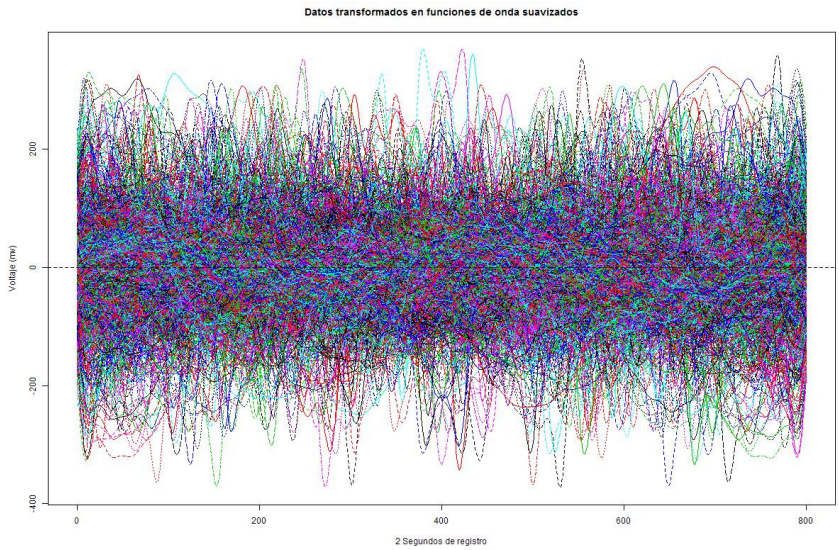

Figura 1: *Datos obtenidos del registro, transformados en funciones de onda suavizados por medio de bases de Fourier. Fuente: elaboración propia.*

Cada una de estas ondas tiene su correspondiente clasificación, según el registro del hipnograma, en fases estas son: ciclo de sueño de movimientos oculares rápidos REM (*Rapid Eyes Movement*), ciclo de sueño de movimientos oculares lentos SWS (*Slow Wave Sleep*), estado de vigila (despierta) de la rata, conductas de acicalamiento, conducta consumatoria (comida o bebida), registros artefactos (errores producidos por agentes externos) como se observa en la en la figura 2. En una fase exploratoria descriptiva de las funciones de onda, se realizó un análisis de componentes principales, la figura 3 señala la proporción de varianza.

La ejecución del algoritmo (KCPF) sobre las funciones de onda con 6 centroides arrojó los siguientes resultados. En la figura 4 se observan las funciones de onda representativas y la figura 5 muestra las ondas correspondientes a cada conglomerado con su función de onda centroide.

Una vez asignados los centroides según las componentes principales, se procedió a contrastar la clasificación del algoritmo de K-medias funcional y la fase de actividad del hipnograma.

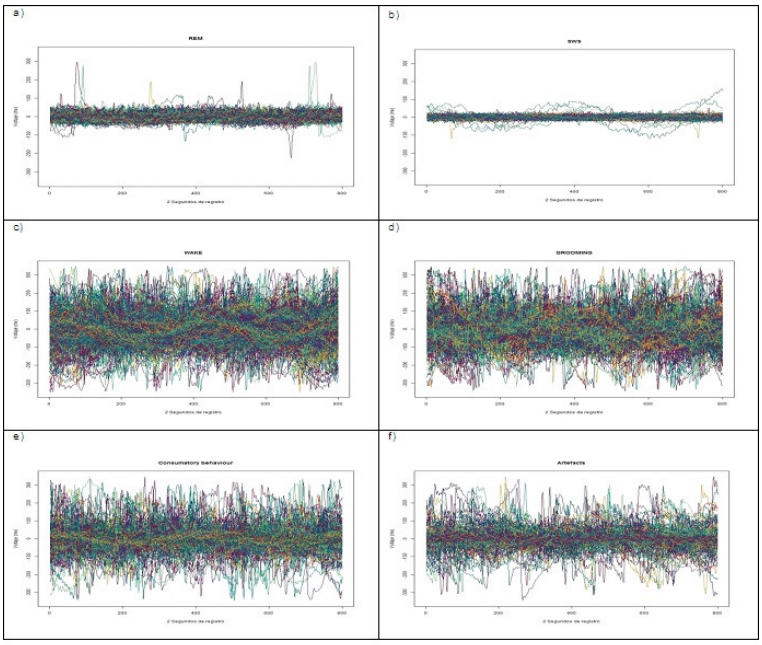

Figura 2: *Funciones de onda según la fase del hipnograma (a) ciclo de sueño de movimientos oculares rápidos REM (Rapid Eyes Movement), (b) ciclo de sueño de movimientos oculares lentos SWS (Slow Wave Sleep), (c) estado de vigila (despierta) de la Rata, (d) conductas de acicalamiento, e) conducta consumatoria (comida o bebida), (f) registros artefactos (errores producidos por agentes externos). Fuente: elaboración propia.*

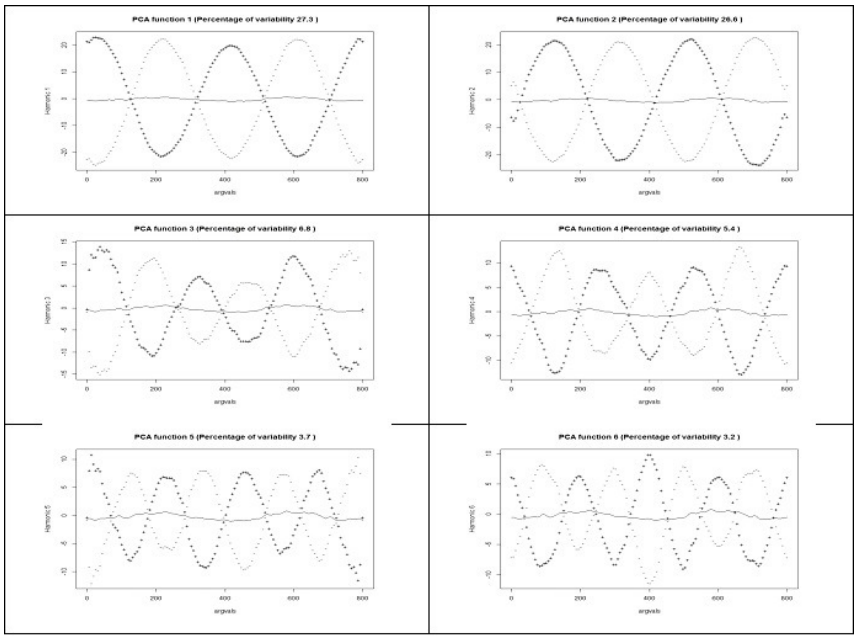

Figura 3: *El porcentaje de variabilidad en el análisis de componentes principales. Fuente: elaboración propia.*

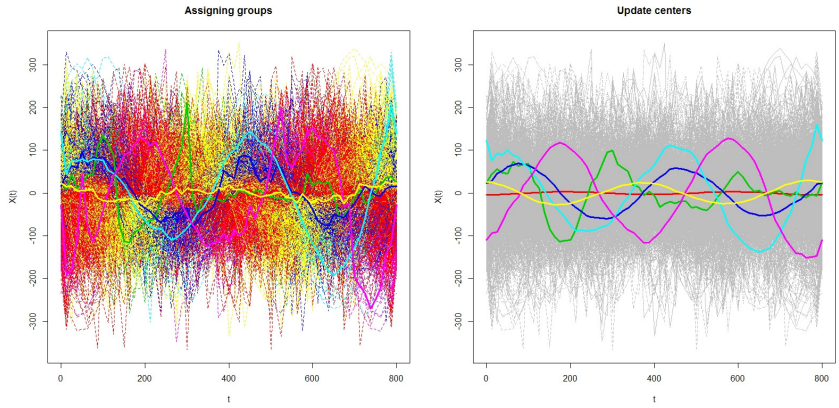

Figura 4: *Resultado del algoritmo de K-medias en R clasificando ondas, según 6 centroides funcionales. Así, la función roja corresponde al centroide 1, la función verde, azul, azul marino, violeta y amarilla corresponde a los centroides 2, 3, 4, 5 y 6 respectivamente. Fuente: elaboración propia.*

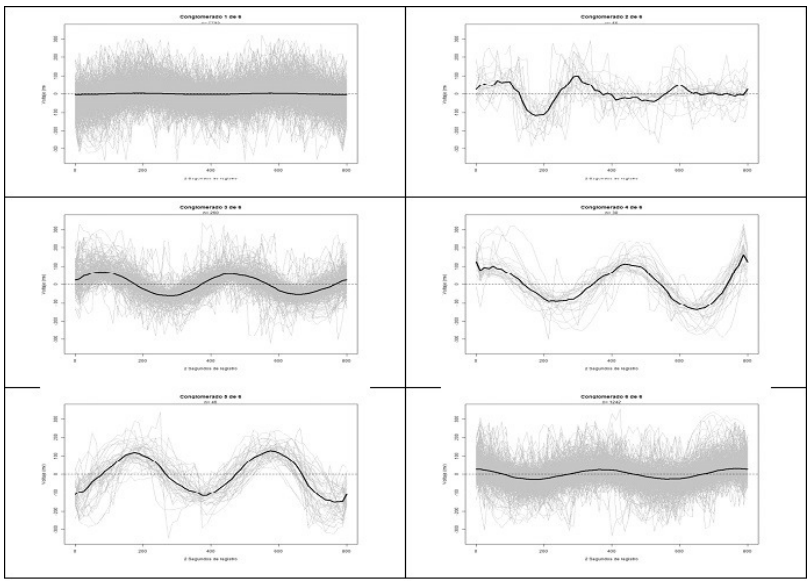

Figura 5: *Conglomerados por k-medias funcional para cada centroide. Fuente: elaboración propia.*

Tabla 1: *Tabla de contingencia para las proporciones entre las fases del hipnograma y la asignación a los conglomerados. Fuente: elaboración propia.*

	REM	**SWS**	**WAKE**	**GROOMING**	**CONS. BEHAV**	**ARTEFACTS**	**TOTAL**
CO_1	0,1598	0,052	0,1899	0,3681	0,0414	0,0187	0,8299
CO_2	0	0	0,0012	0,0006	0,0004	0,0001	0,0023
CO_3	0,0004	0,0001	0,0169	0,0057	0,0023	0,0021	0,0275
CO_4	0	0	0,002	0,0009	0,0001	0,0002	0,0032
CO_5	0	0	0,0029	0,0015	0,0003	0,0002	0,0049
CO_6	0,0151	0,0006	0,0641	0,028	0,0175	0,0069	0,1322
TOTAL	0,1753	0,0527	0,277	0,4048	0,062	0,0282	1

En la figura 6 y la tabla 1, se observa el análisis de correspondencias con una variabilidad explicada del 95.2 % para la dimensión uno sobre el eje horizontal y una variabilidad explicada del 4.3 % sobre el eje vertical. Un ejercicio de interpretación más detallado lleva a pensar que el conglomerado uno (CO_1) abarca el 83 % del registro de las actividades cerebrales, teniendo mayor proximidad en ondas rápidas REM, acicalamiento y vigilia WAKE con unas proporciones del 15 %, 36 %, y 18 % respectivamente. Menor proximidad en el comportamiento de ondas lentas SWS, conducta consumatoria y artefactos con 5,2 %, 4,14 % y 1,8 % respectivamente. Así, en el componente se describe el 83 % de la variabilidad explicada de la actividad cerebral durante las 24 horas. Las componentes dos, cuatro y cinco contienen información acerca de la actividad cerebral de solo cuatro comportamientos con una proporción de la variabilidad explicada del 0.23 %, 0.32 % y 0.42 % respectivamente con mayor proximidad con los comportamientos vigilia WAKE y acicalamiento (*Grooming*) y menor proximidad con conducta consumatoria y artefactos.

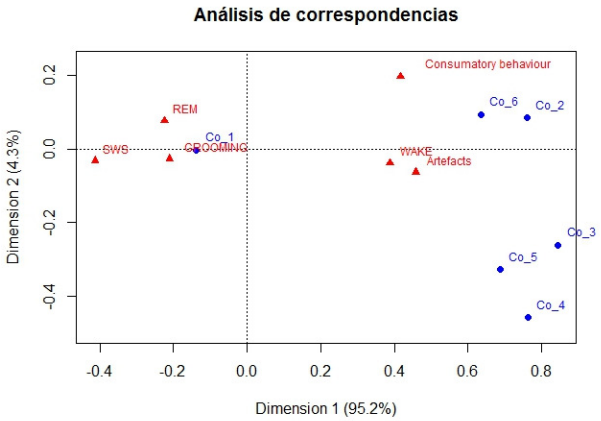

Figura 6: *Análisis de correspondencias para los conglomerados y las fases del hipnograma. Fuente: elaboración propia.*

El conglomerado seis (CO_6) también da información acerca de todas las actividades cerebrales, pero con una variabilidad explicada del 13 % de la siguiente forma: mayor proximidad con vigilia (WAKE), acicalamiento Grooming y conducta consumatoria con una variabilidad del 6.41 %, 2.8 % y 1.75 % respectivamente y menor proximidad con artefactos, ondas rápidas REM y ondas lentas SWS con 0.69 %, 1.5 % y 0.06 % de variabilidad explicada. El conglomerado 3 (CO_3) también da información acerca de todas las actividades cerebrales, pero con una variabilidad explicada del 3 % de la siguiente forma: mayor proximidad con vigilia WAKE, acicalamiento Grooming y conducta consumatoria con una variabilidad del 1.69 %, 0.57 % y 0.23 % respectivamente y menor proximidad con artefactos, ondas rápidas REM y ondas lentas SWS con 0.21 %, 0.04 % y 0.01 % de variabilidad explicada.

En el análisis de correspondencias se puede identificar que el eje 1 se relaciona con la frecuencia de las ondas, siendo el extremo negativo las ondas con más baja frecuencia, lo que es característico de las fases acicalamiento, SWS y REM; mientras que las fases de frecuencias altas (conducta consumatoria, vigilia y artefactos) se encuentran en el extremo positivo. Por su parte, el eje dos responde a la regularidad de la onda; es decir, que durante ese momento de registro no se presenten alteraciones de la señal, acercándose a la forma de una onda ideal. En el extremo positivo se encuentran las fases en las que existe mayor probabilidad de la regularidad de la señal, como lo son la fase REM y la conducta consumatoria. El lado negativo del componente dos indica registros irregulares en la forma de la onda. Por ejemplo, en el sueño de ondas lentas se presentan pequeños cambios en las ondas que caracterizan sus fases, como lo son los complejos k, los husos de sueño o el cambio alternado de frecuencias que indican transiciones en las fases del ciclo de sueño. De la misma forma, la probabilidad de tener señales irregulares en vigilia es alta, puesto que la actividad muscular adiciona ruido; además que las ondas

clasificadas como artefactos, son consideradas errores de medida, siendo también irregulares en extremo.

Una vez identificadas las características de los componentes, se logra interpretar la técnica (KCPF), respecto a las fases registradas en el hipnograma y a los componentes como tal. El análisis de correspondencias nos permite interpretar la tabla de contingencia entre las fases registradas y los conglomerados. Los resultados del algoritmo k-medias para la agrupación de los conglomerados que se observa (figura 5). Se hace notar mayor aglomeración de funciones en los componentes uno, tres y seis. Lo cual es coherente con lo encontrado con el análisis de correspondencias.

También se puede observar que la componente uno tiene mayor cercanía con las fases de sueño y relajación (acicalamiento), ante lo que se puede pensar en un conglomerado de relajación. Los conglomerados dos y seis se encuentran en el extremo positivo del componente uno, siendo estos grupos de actividad, relacionados con fases de vigilia, conducta consumatoria, e incluso señales clasificadas como artefactos.

Por otro lado, los conglomerados tres, cuatro y cinco, aunque corresponden a una proporción muy pequeña, se pueden considerar como el grupo de señales irregulares con alta frecuencia, que se caracterizan por ser de transición de fase, alteraciones por ruido o artefactos.

6. Conclusiones

Al comparar un sistema de clasificación manual, como el hipnograma frente a un modelo de análisis de datos no supervisado como el método de clasificación por k-medias, en datos funcionales, se encontró que los 6 conglomerados guardan una relativa consistencia con las fases del hipnograma del EEG. Por tanto, se puede concluir que el ADF permite describir el comportamiento de las señales asociadas con la actividad cerebral y, en particular, las ondas de sueño.

Se resalta el potencial del ADF en el sentido que la conversión de onda permite tomar una serie de n datos y verla como un único objeto que puede ser procesado en labores de agrupamiento, clasificación o asociación. Una ventaja importante es el aumento en la eficiencia de los algoritmos, ya que en lugar de seleccionar un solo factor (amplitud o frecuencia) se está analizando la señal per se.

El trabajo ofrece una alternativa a las técnicas actuales de procesamiento de señales, abriendo una linea de investigación donde se toma la señal como una funcion y no una serie de puntos. En trabajos futuros se pretende evaluar otros algoritmos de clasificación para datos funcionales y, de esta manera, evaluar la eficiencia computacional de los mismos en el contexto de mineria de datos, además de evaluar su potencial aplicación en la clasificación automática de señales.

Recibido: 27 de julio de 2016
Aceptado: 21 de octubre de 2016

Referencias

Acharya, R., Faust, O., Kannathal, N., Chua, T. & Laxminarayan, S. (2005), 'Non-linear analysis of eeg signals at various sleep stages', *Computer methods and programs in biomedicine* **80**(1), 37–45.

Andersen, M., Antunes, I., Silva, A., Alvarenga, T., Baracat, E. & Tufik, S. (2008), 'Effects of sleep loss on sleep architecture in wistar rats: gender-specific rebound sleep', *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **32**(4), 975–983.

Ferraty, F. & Vieu, P. (2006), *Nonparametric functional data analysis: theory and practice*, Springer Science & Business Media.

Iber, C. (2007), *The AASM manual for the scoring of sleep and associated events: rules, terminology and technical specifications*, American Academy of Sleep Medicine.

Jacques, J. & Preda, C. (2014), 'Functional data clustering: a survey', *Advances in Data Analysis and Classification* **8**(3), 231–255.

Moser, D., Anderer, P., Gruber, G., Parapatics, S., Loretz, E., Boeck, M., Kloesch, G., Heller, E., Schmidt, A., Danker-Hopfe, H. et al. (2009), 'Sleep classification according to aasm and rechtschaffen & kales: effects on sleep scoring parameters', *Sleep* **32**(2), 139–149.

Peña, D. (2002), *Análisis de datos multivariantes*, Vol. 24, McGraw-Hill Madrid.

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

Ramsay, J. O. & Dalzell, C. (1991), 'Some tools for functional data analysis', *Journal of the Royal Statistical Society* **3**, 539–572.

Ramsay, J. O., Hooker, G. & Graves, S. (2009), *Functional data analysis with R and MATLAB*, Springer Science & Business Media.

Ramsay, J. & Silverman, B. (2005), 'Functional data analysis'.

Yamamoto, M. (2012), 'Clustering of functional data in a low-dimensional subspace', *Advances in Data Analysis and Classification* **6**(3), 219–247.

Yamamoto, M. & Terada, Y. (2014), 'Functional factorial k-means analysis', *Computational Statistics & Data Analysis* **79**, 133–148.