ANÁLISIS DE DATOS FUNCIONALES APLICADO EN ELECTROENCEFALOGRAMAS: AGRUPAMIENTO POR K-MEDIAS FUNCIONAL.

Alexis Enrique Carrillo Ramírez
Olga Cecilia Garatejo Escobar

Fundación Universitaria Los Libertadores
Departamento de Ciencias Básicas
Especialización en Estadística Aplicada

Bogotá D.C.
2016

ANÁLISIS DE DATOS FUNCIONALES APLICADO EN ELECTROENCEFALOGRAMAS: AGRUPAMIENTO POR K-MEDIAS FUNCIONAL.

Alexis Enrique Carrillo Ramírez
Olga Cecilia Garatejo Escobar

Asesor:
Wilmer Pineda Ríos

Fundación Universitaria Los Libertadores
Departamento de Ciencias Básicas
Especialización en Estadística Aplicada

Bogotá D.C.
2016

Nota de Aceptación

_____

_____

_____

_____

_____

_____

_____

Firma del presidente del jurado

_____

Firma del Jurado

_____

Firma del Jurado

Bogotá, D.C Julio del 2016

Las Directivas de la Universidad de
Los Libertadores, los jurados calificadores y el cuerpo
Docente no son responsables por los
Criterios e ideas expuestas En el presente documento.
Estos corresponden únicamente a los autores

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

GLOSARIO

Complejos k: Corresponden a ondas lentas bifásicas, caracterizadas por una descarga lenta, negativa, de amplitud elevada y de una deflexión positiva en el Electroencefalograma.

Electroencefalograma: Gráfico en el que se registra la actividad del cerebro y es obtenido por un electroencefalógrafo.

Hipnograma: Es la forma gráfica en la cual se observa el registro de diferentes tipos de actividades en relación con el ciclo sueño – vigilia a lo largo del tiempo.

Neurociencias: Conjunto de disciplinas científicas que estudian la estructura, la función, el desarrollo de la bioquímica, la farmacología, y la patología del sistema nervioso y de cómo sus diferentes elementos interactúan, dando lugar a las bases biológicas de la conducta.

REM: (Rapid Eyes Movement) Es una fase del ciclo del sueño, la cual se caracteriza principalmente por movimientos oculares rápidos y por registros electroencefalográficos de ondas de alta frecuencia y baja amplitud. Esta fase está relacionada con la experiencia de soñar y con procesos de consolidación de memoria y aprendizaje.

Sinapsis: Proceso de comunicación electroquímica de las células del sistema nervioso, en el cual el botón terminal de la neurona presináptica recibe un impulso eléctrico que activa una serie de reacciones químicas en la membrana, la cual libera a la brecha sináptica un neurotransmisor que entra en contacto con la membrana de la neurona post sináptica. Este contacto hace que la neurona receptora realice en su membrana un intercambio de iones, alterando su potencial eléctrico que se propaga por la membrana hasta la región terminal.

SWS: (Slow Wave Sleep) Es una fase del ciclo del sueño caracterizada por ondas de gran amplitud y una transición de altas frecuencias al inicio del ciclo del sueño, llegando a bajas frecuencias en estadios profundos del sueño.

Vigila: Fase del ciclo del sueño en que la persona se encuentra despierta, vigilante, interactúa de manera efectiva y consciente a los estímulos ambientales y se encuentra orientada en tiempo y espacio.

Resumen

El análisis de datos funcionales se basa en el estudio de la función que describe la variabilidad de un conjunto de datos en un espacio de n muestras, y dentro de sus modelos se encuentra el análisis de conglomerados por k-medias funcional. Dado que la actividad cerebral responde a una función de onda de la carga eléctrica de las neuronas sobre el tiempo, observamos la oportunidad de aplicar el análisis de datos funcionales a este tipo de registros. El objetivo de este proyecto es describir la aplicabilidad del análisis de conglomerados por k-medias funcional para clasificación de la actividad cerebral en ratas Norvegicus Wistar. Se realizó la conversión de los registros en funciones de onda en bases de Fourier, las cuales fueron procesadas con análisis de componentes principales funcionales, algoritmo de k-medias funcional (k=6) y un análisis de correspondencias 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; elementos a tener en cuenta para la clasificación de señales. El análisis de datos funcionales es aplicable a la clasificación de registros de electroencefalograma, dado que toma un dato que es n-dimensional y permite manejarlo como un único valor (una función de onda) y así ser procesado con diferentes técnicas de minería de datos.

Palabras clave: Electroencefalografía (EEG), datos funcionales, series de Fourier, Aprendizaje automático, K –Medias funcional, Análisis de componentes principales funcionales.

Abstract

Functional data analysis is the study of the function that describes the variability of a data set in a space of n samples, and the clustering functional k-means is one of their techniques. As the brain activity can be described as a wave function of the electric charge of the neurons over time, hence we see the opportunity to apply it functional data analysis. The objective of this project is to describe the applicability of cluster analysis by functional k-means as an unsupervised model for classification of brain activity in rats (*Norvegicus Wistar*). We transform the records to wave functions with a Fourier basis, then we run functional principal components analysis, functional k-means (k = 6) and a correspondence analysis between the clusters and the phases of activity in the hypnogram. The functional K-means clustering is consistent to the attributes of frequency and regularity of the waves, elements to be considered for classifying signals. The functional data analysis is applicable to the classification of electroencephalographic records. It allows to take n-dimensional values and process them as a unique feature over several machine learning algorithms.

Key words: Electroencephalography, Functional data analysis, Fourier series, Machine learning, functional K-Means, Functional principal component analysis.

1. Introducción

En el campo de las neurociencias, la electroencefalografía (EEG) es el método más utilizado para medir la actividad cerebral en diferentes especies de animales y su análisis numérico se conoce como Electroencefalografía Cuantitativa (QEEG). En este proyecto se van a utilizar modelos estadísticos del análisis de datos funcionales para clasificar las ondas en los registros de actividad cerebral en ratas Norvegicus Wistar.

El interés por este tipo de investigación nace debido a que los investigadores de neurociencias normalmente tienen formación de base en campos como medicina, biología o psicología, y pueden llegar a tener conocimientos matemáticos insuficientes al momento de comprender y aplicar los modelos de análisis de las señales electroencefalografías. Otra debilidad puede ser el hecho de tener poca familiaridad o dificultades de acceso a software especializado como MATLAB o R, el uso de estos es necesario para los investigadores dado que permiten identificar registros de evidencia estadísticamente significativa para demostrar efectos experimentales, o tener la capacidad de hacer análisis descriptivos de las ondas.

2. Formulación o Pregunta Problema

¿Cuál es la aplicabilidad del análisis de conglomerados por k-medias funcional sobre las señales de la actividad cerebral en la rata Norvegicus Wistar?

3. Justificación

En neurociencias, al medir las señales emitidas por la actividad cerebral, se encuentra como principal dificultad un gran volumen de información. A pesar 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 tanto se recurre al análisis de datos funcionales.

En el análisis de datos funcionales, al igual que en el multivariado, los registros se pueden procesar de tal forma que los métodos estadísticos como análisis de correspondencias y conglomerados son de total validez y efectividad. En este caso se estudia la aplicabilidad del método de conglomerados por K-medias funcional para la clasificación de la actividad cerebral de la rata norvegicus wistar durante las 24 horas del registro del EEG.

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 usando las librerías fda y fda.usc, las cuales permiten manipular adecuadamente los registros entregados por la electroencefalografía y de esta manera solventar tales dificultades.

4. Objetivo General

Describir la aplicabilidad del análisis de conglomerados por k-medias funcional sobre las señales de la actividad cerebral en la rata Norvegicus Wistar.

5. Objetivos específicos

- Segmentar los registros en fragmentos de 2 segundos (800 datos).
- Seleccionar los registros con valores entre -350mv y 350 mv.
- Convertir los registros en datos funcionales, por medio de las bases de Fourier.
- Aplicar el algoritmo k-medias funcional a la base de datos por medio del software R usando las librerías fda y fda.usc.
- Analizar la correspondencia entre la asignación del algoritmo de k-medias con las fases del registro del hipnograma.

6. Marco de Referencia

6.1 Análisis de Datos Funcionales

Las señales eléctricas que producen las células del cerebro al comunicarse pueden ser registradas por medio de un Electroencefalograma (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, 2006).

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

Los resultados obtenidos a partir del EEG, se presentan como observaciones discretas de las variables funcionales en un conjunto finito de instantes de tiempo. Para un correcto análisis de estas variables, primero se obtiene la forma funcional la cual debe cumplir 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):

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

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 los registros del EEG. Las funciones varían en amplitud y frecuencia; por tanto la aproximación conveniente es en series de Fourier. Teniendo en cuenta que existen otras aproximaciones como wavelets y B-Spline.

6.1.1 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:

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

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.

6.1.2 Estadísticos Descriptivos en Análisis Funcional de datos.

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: $Cov\big(f(t_1), f(t_2)\big) = \frac{1}{n-1}\sum_{j=1}^{n}(f_j(t_1) - \bar{f}(t_1))'(f_j(t_2) - \bar{f}(t_2))$
- Correlación: $Cor\big(f(t_1), f(t_2)\big) = \frac{Cov\big(f(t_1), f(t_2)\big)}{\sqrt{s(f(t_1))s(f(t_2))}}$

Por tanto, los estadísticos descriptivos del análisis multivariado, aplican igualmente para datos funcionales. El análisis descriptivo de los datos funcionales obtenidos a partir de los registros del EEG se realiza por medio del algoritmo de k-medias funcional y se complementa con el análisis de componentes principales. A continuación se expone la teoría necesaria a desarrollar.

6.1.3 Análisis de Conglomerados para datos funcionales

El análisis de datos funcionales para este proyecto 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 mostrara la teoría referente a componentes principales para funcionales.

6.1.3.1 Componentes principales para datos funcionales.

El objetivo del análisis de componentes principales es considerar la máxima información dentro de una combinación lineal de auto-funciones, obteniendo una base de menor dimensión. Se busca que la primera componente de dicha base contenga la mayor proporción posible de la variabilidad original, para la 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 auto-valores y auto-funciones de la función covarianza $Cov\big(f(t_1), f(t_2)\big)$.

Así, sean $\{f_1, f_2, \cdots, f_n,\}$ observaciones como se definieron en la sección (6.1.2) y sus correspondientes estadísticos media y covarianza:

$$\overline{f(t)} = \frac{1}{n}\sum_{j=1}^{n} f_j(t)$$

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

Se asume que cada f_j con $j = 1,2, \cdots n$ tiene una expansión en series de Fourier como en la sección (6.1.1):

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

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

$$\sum_{j=1}^{n}\Big(f_j(t_1) - \bar{f}(t_1)\Big)' = A'\varphi(t_1)' \quad \text{Y} \quad \sum_{j=1}^{n}\Big(f_j(t_2) - \bar{f}(t_2)\Big) = A\varphi(t_2) \quad \textbf{(2)}$$

Sustituyendo **(2)** en la función covarianza $Cov\big(f(t_1), f(t_2)\big)$, se tiene:

$$Cov\big(f(t_1), f(t_2)\big) = \tfrac{1}{n-1}\varphi(t_1)'A'A\varphi(t_2) \qquad \textbf{(3)}$$

Ahora, los auto-valores y auto-funciones de la función de covarianzas se encuentran solucionando la siguiente integral con $t_1, t_2 \in [0, T]$ y cada auto-función con expansión en base $\varphi(t)'$, $f_j(t) = b_j\varphi(t)'$ se plantea:

$$\int_0^T Cov\big(f(t_1), f(t_2)\big) f_j(t_2) dt_2 = \lambda_j f(t_1)$$

Reemplazando **(3),** en la anterior integral se tiene:

$$\int_0^T \frac{1}{n-1} \varphi(t_1)' A' A \varphi(t_2) b_j \varphi(t_2)' dt_2 = \lambda_j b_j \varphi(t_1)'$$

$$\frac{1}{n-1} \varphi(t_1)' A' A \int_0^T \varphi(t_2)\varphi(t_2)' dt_2 \, b_j = \lambda_j b_j \varphi(t_1)'$$

Sea $W = \int_0^T \varphi(t_2)\varphi(t_2)' dt_2$, simplificando $\varphi(t_1)'$ se llega a un problema de auto-valores multivariado o por matrices:

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

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:

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

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

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

Los auto-valores λ_j y las auto-funciones $b_i = u_j W^{-\frac{1}{2}}$, 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.

6.1.3.2 Conglomerados por el Método de K-medias

Para el análisis de objetos funcionales como los descritos en la **definición 1**, tal que $f(t) \in L^2$, se usara el algoritmo de componente principal funcional k-medias (Yamamoto M. , 2012).

En primer lugar se deben definir los siguientes espacios:

- Sea $V = \{v_l\}$ con $(l = 1, \cdots, r)$, $v_l \in \ell = L^2(T)$, $r < \infty$. Son las funciones que conforman la base ortonormal del subespacio de proyección.
- Sea P_v el operador proyección ortogonal definido como:

$$P_v: \ell \to \ell_v$$

  Es decir, el operador P_v va del espacio de los datos funcionales ℓ sobre el subespacio ℓ_v, el cual 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 $y \in \ell$, $s \in T$

- Función objetivo (4)

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

Según Yamamoto (2012), la función objetivo (4), se puede escribir como:

(5)

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

(6)

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

El algoritmo k-medias componentes principales funcional (KCPF), se reduce a minimizar la función objetivo (4), 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 $U = \{u_{ik}\}$ 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 (5), fijando U sobre V.

Paso 3: Se minimiza el segundo término en la ecuación (6), 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 el k-medias es sensible a óptimos locales.

Dado que el objetivo de este proyecto es aplicar este análisis a datos funcionales, a continuación se explicara el origen de los datos a ser analizados, los cuales son mediciones de las señales emitidas por las neuronas en el cerebro de la rata Norvegicus Wistar.

6.1.4 Análisis de Correspondencias

El análisis de correspondencias es una técnica descriptiva para representar tablas de contingencia, es decir, tablas donde se recoge las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos. Puede interpretar como una manera de representar las variables en un espacio de dimensión menor, análoga a componentes principales o también como un procedimiento objetivo de asignar valores numéricos a variables cualitativas

.

6.2 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 que se aplican en la base del cráneo sin necesidad de procedimiento quirúrgico; o 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 Estereo Electroencefalograma (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.

7. 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. Estos datos corresponden a un registro de 24 horas de una rata del 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.

Para el desarrollo del código, se utilizó R-Studio, el cual 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

Procedimiento de análisis.

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.

8. Resultados

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 6.1.1 cada función tiene una aproximación en series de Fourier como sigue:

$$f_j(t) \approx \frac{a_0}{2} + \sum_{i=1}^{800} (a_i cos \frac{2\pi it}{800} + b_i sen \frac{2\pi it}{800})$$

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

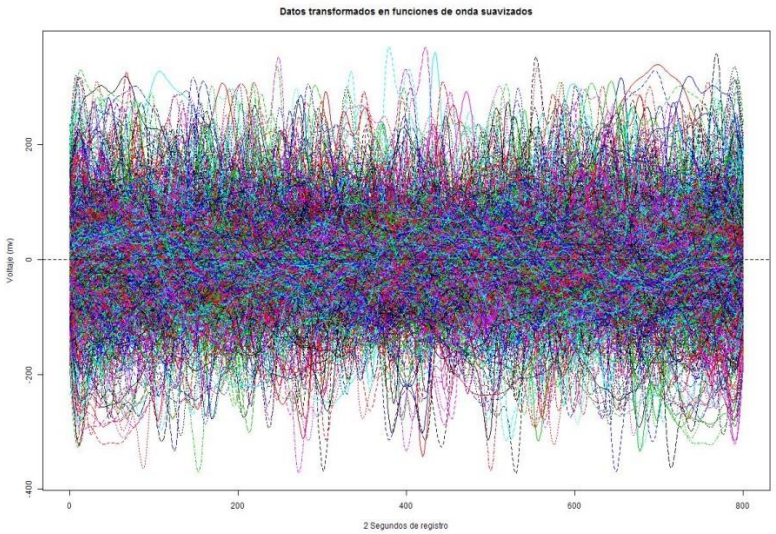
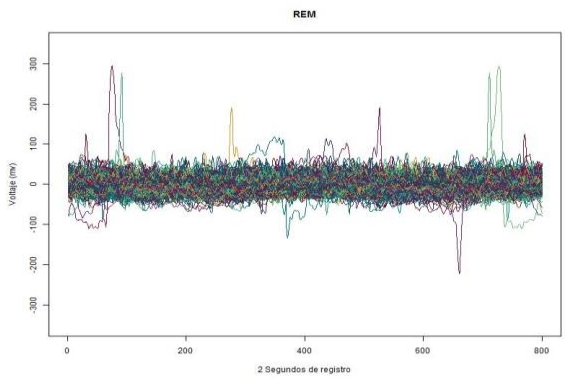
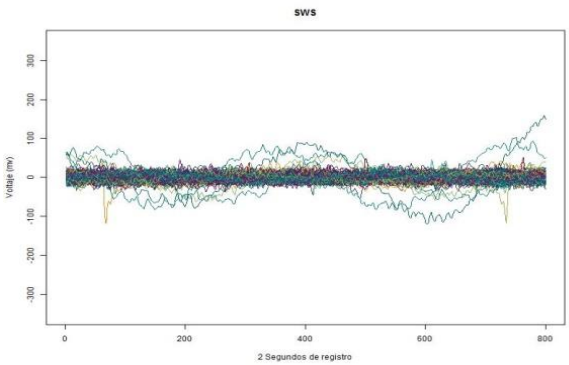


Figura 1: Datos obtenidos del registro, transformados en funciones de onda suavizados por medio de bases de Fourier.

Cada una de estas ondas tiene su correspondiente clasificación según el registro del hipnograma, como se observan en la en la figura 2.

a) 	b) 
c)	d)

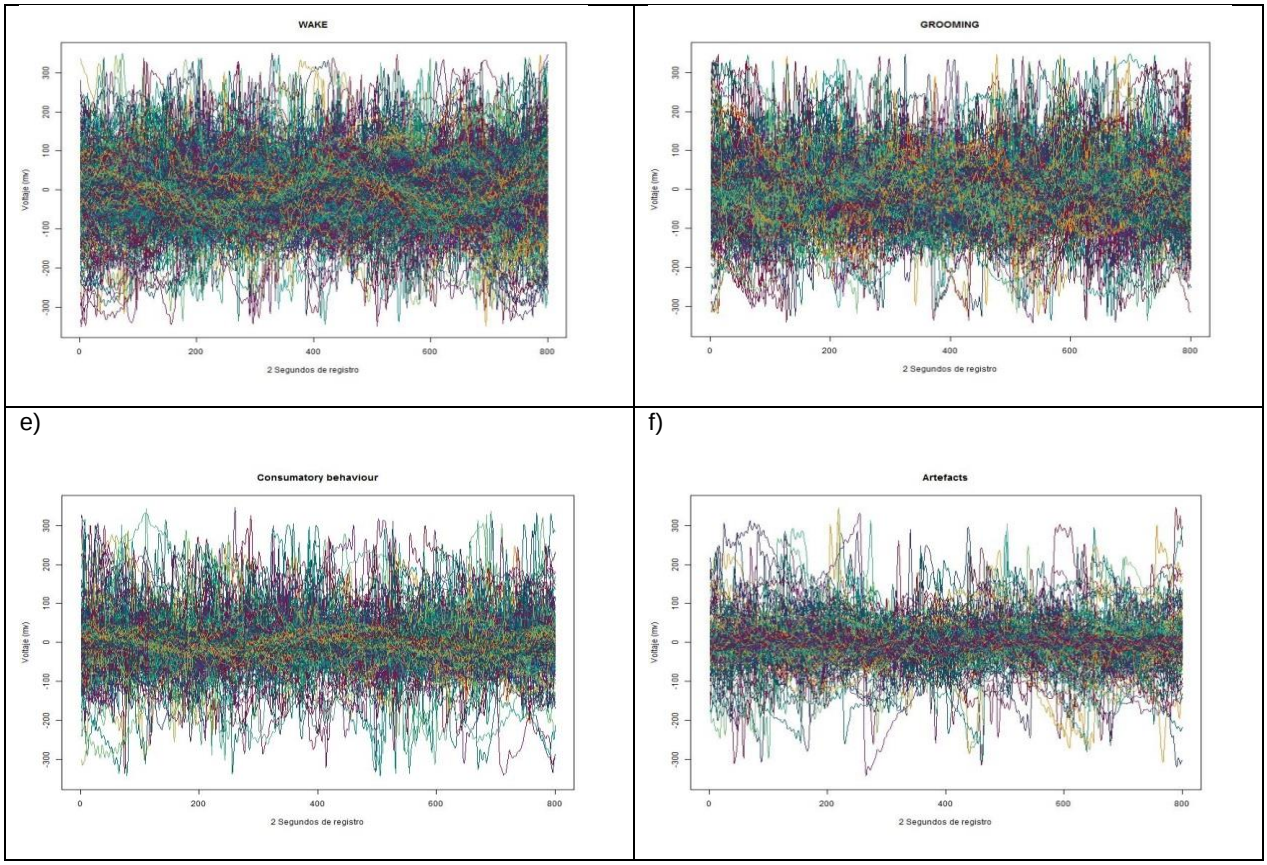
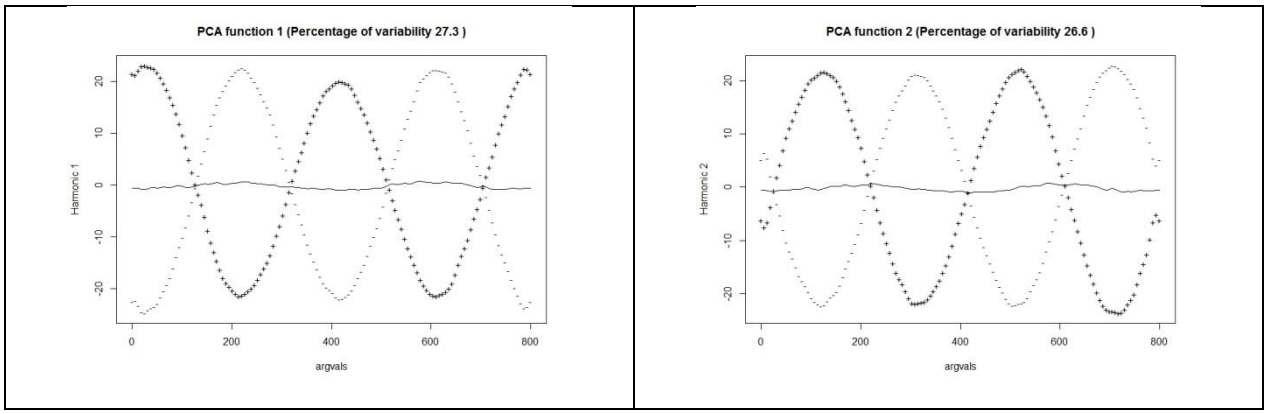


Figura 2: Funciones de Onda según la fase del hipnograma a) ciclo de sueño de movimientos oculares rápidos REM; b) ciclo de sueño de movimientos oculares lentos SWS; c) estado de vigila (despierta) de la Rata; d) conductas de acicalamiento; e) conducta consumatoria (comida o bebida); f) registros artefactos (errores producidos por agentes externos).

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, mientras que la figura 4 muestra el gráfico de sedimentación de las proporciones de varianza.



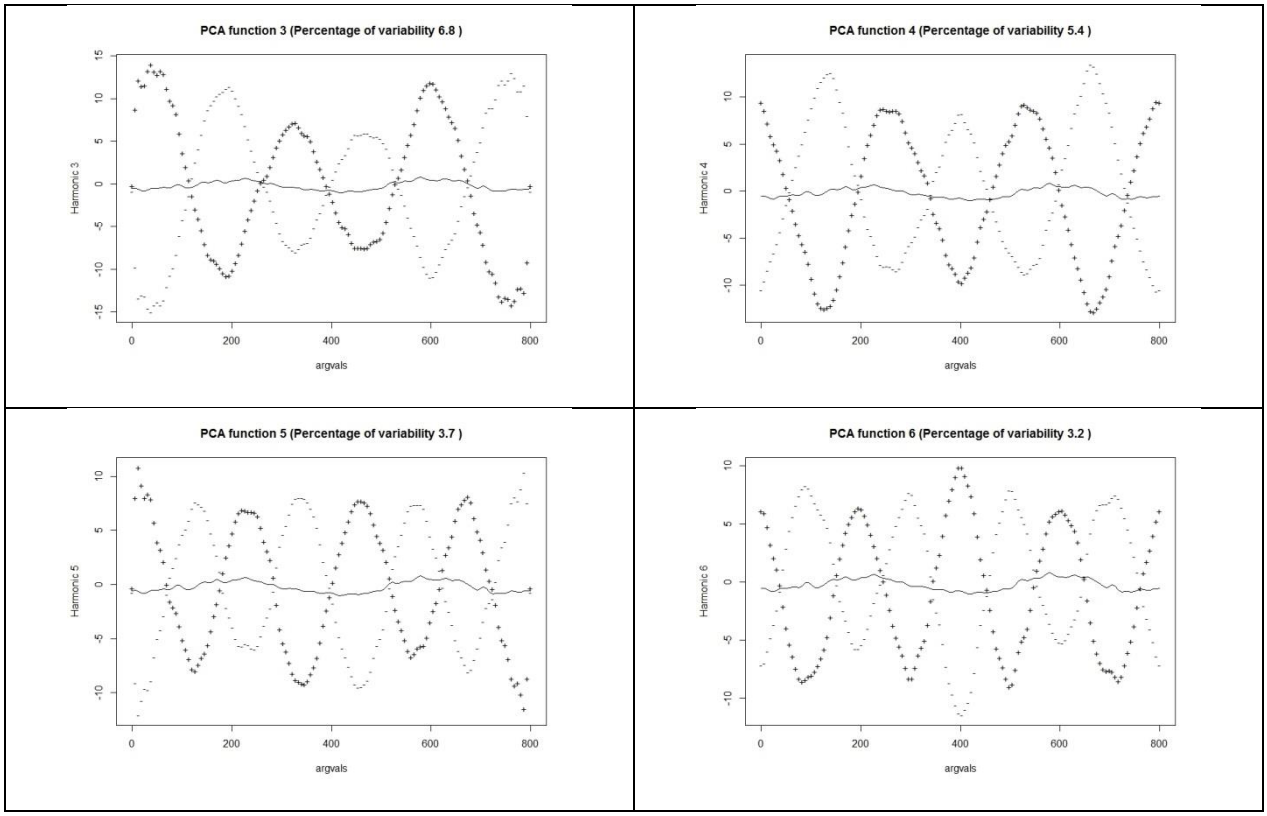


Figura 3: El porcentaje de variabilidad en el análisis de componentes principales.

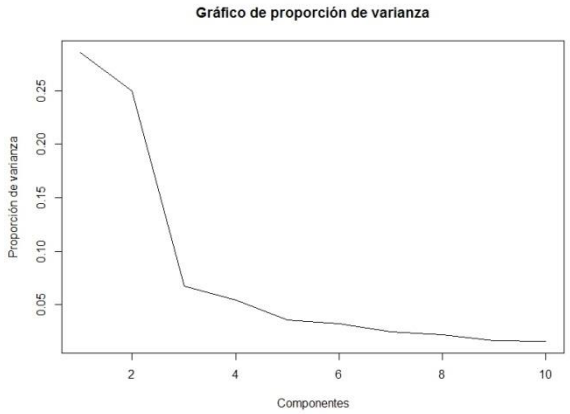


Figura 4: Gráfico de sedimentación de las proporciones de varianza.

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

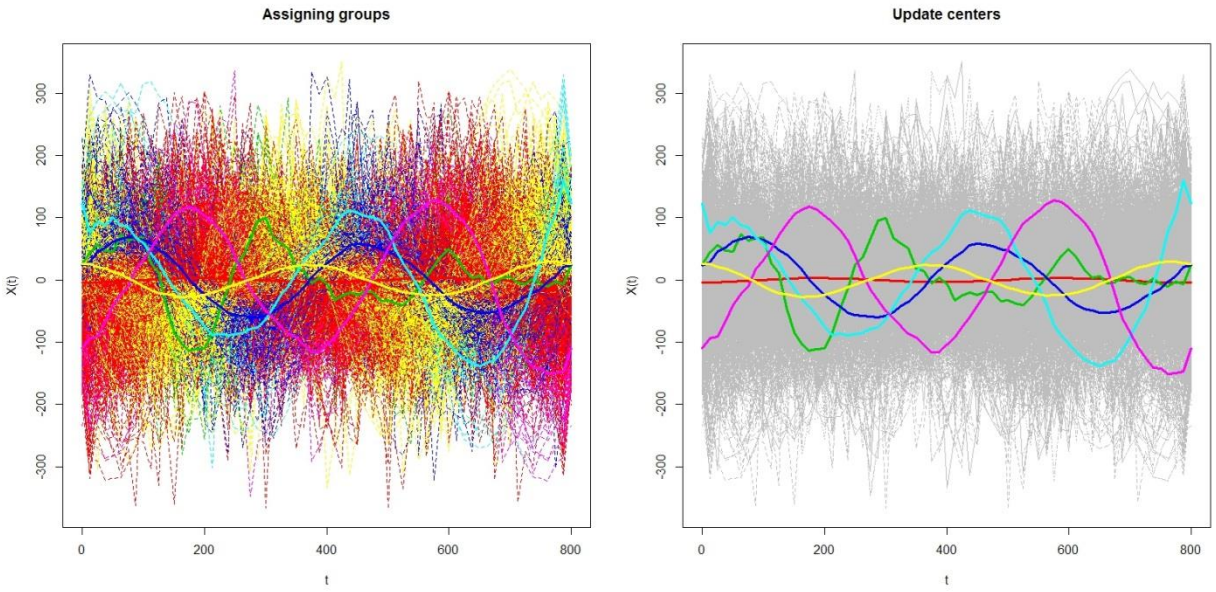
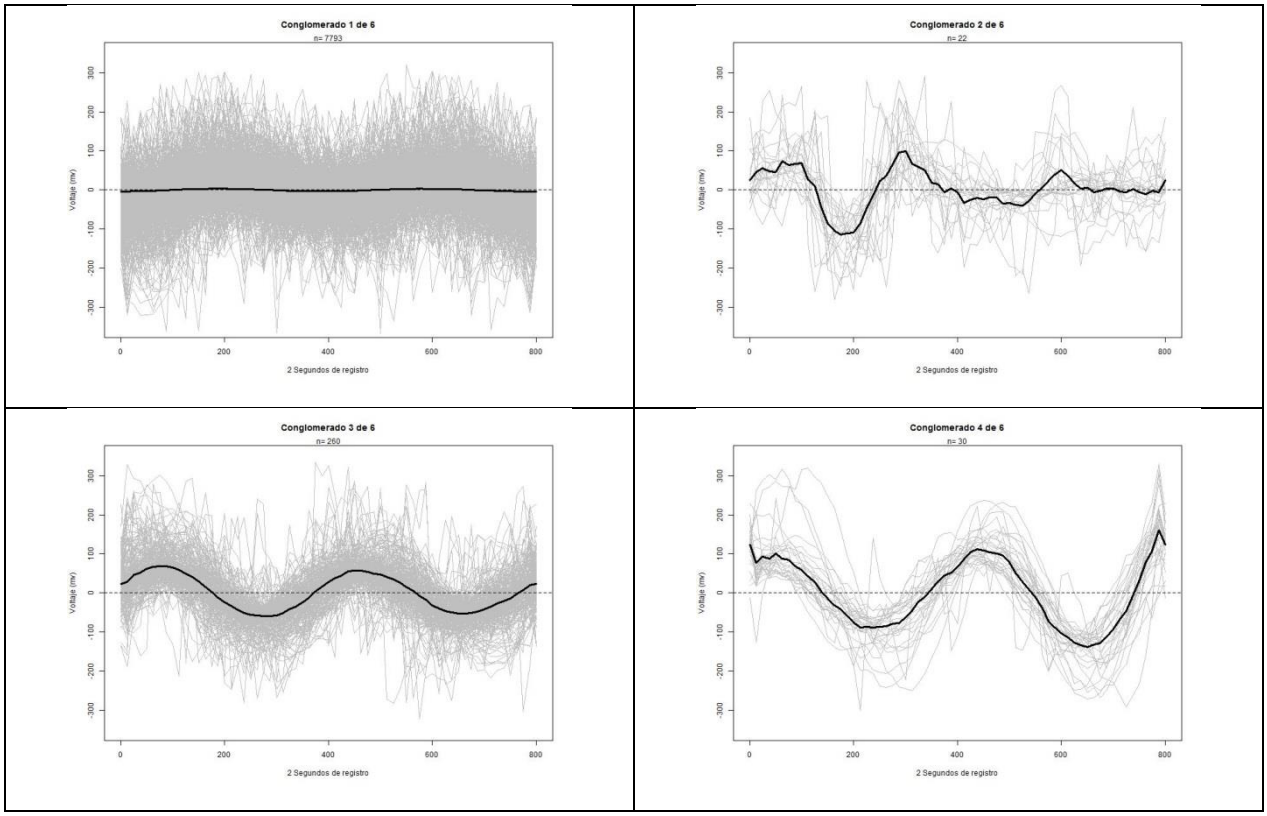


Figura 5: Resultado del algoritmo de K-medias en R clasificando ondas según 6 centroides funcionales. Así la función roja corresponde al centroide 1, la función verde, azul, azul marino, violeta y amarilla corresponde a los centroides 2, 3, 4, 5 y 6 respectivamente.



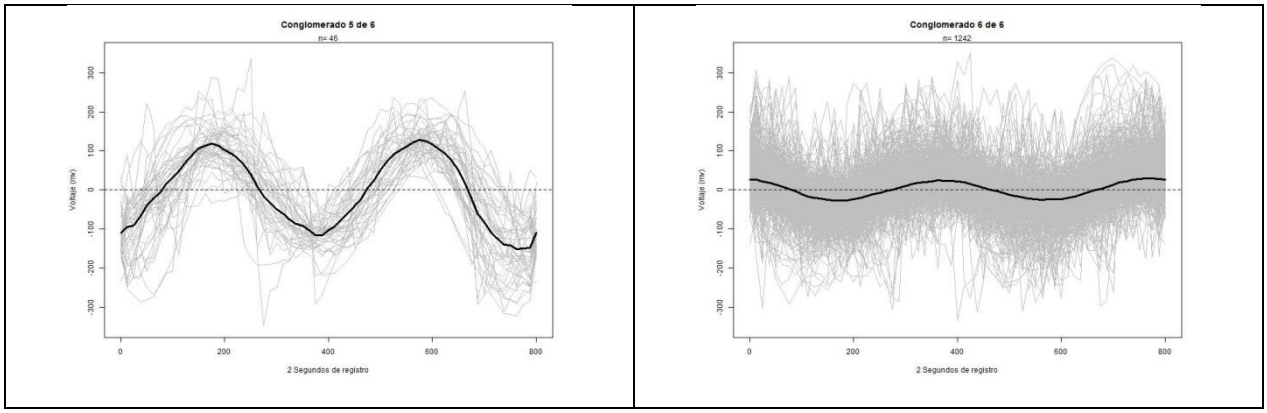


Figura 6: Conglomerados por k-medias funcional para cada centroide.

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. Inicialmente con una tabla de contingencia (Tabla 1) y posteriormente con un análisis de correspondencias (Figura 7).

		Fase_Hyp						
		REM	**SWS**	**WAKE**	**GROOMING**	**CONSUMATORY BEHAVIOUR**	**ARTEFACTS**	
Conglomerados	**Co_1**	0,1595	0,0520	0,1899	0,3681	0,0414	0,0187	0,83
	Co_2	0,0000	0,0000	0,0012	0,0006	0,0004	0,0001	0,00
	Co_3	0,0004	0,0001	0,0169	0,0057	0,0023	0,0021	0,03
	Co_4	0,0000	0,0000	0,0020	0,0009	0,0001	0,0002	0,00
	Co_5	0,0000	0,0000	0,0029	0,0015	0,0003	0,0002	0,00
	Co_6	0,0151	0,0006	0,0641	0,0280	0,0175	0,0069	0,13
		0,18	0,05	0,28	0,40	0,06	0,03	1,00

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

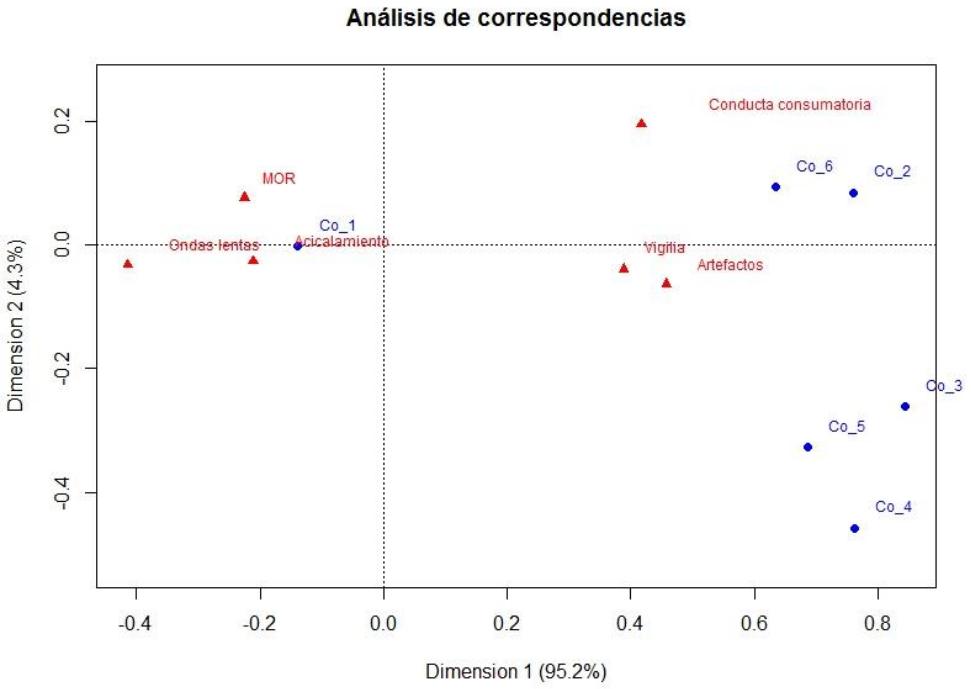


Figura 7: Análisis de correspondencias para los conglomerados y las fases del hipnograma.

En la figura 7 y la tabla de contingencias correspondiente a la tabla 1, se observa el análisis de correspondencias con una variabilidad explicada del 95.2% para la dimensión 1 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 1 (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 2, 4 y 5 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.

El conglomerado 6 (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.

9. Discusión

En el análisis de componentes principales podemos identificar que el componente 1 se relaciona con la frecuencia de las ondas, siendo el extremo negativo las ondas con más baja frecuencia, lo cual 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 componente 2 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, podemos interpretar los conglomerados del k-medias funcional, 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 del k-medias para la agrupación de los conglomerados que se observa en la figura 5. Se hace notar mayor aglomeración de funciones en los componentes 1, 3 y 6. Lo cual es coherente con lo encontrado con el análisis de correspondencias.

Observamos que el componente 1 tiene mayor cercanía con las fases de sueño y relajación (acicalamiento), ante lo cual podemos pensar en un conglomerado de relajación. Los conglomerados 2 y 6 se encuentran en el extremo positivo del componente 1, siendo éstos grupos de actividad, relacionados con fases de vigilia, conducta consumatoria, e incluso señales clasificadas como artefactos.

Por otro lado, los conglomerados 3, 4 y 5, 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.

10. Conclusiones

Comparar un sistema de clasificación manual, como es el hipnograma frente a un modelo de análisis de datos no supervisado como el algoritmo 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 análisis de datos funcionales tiene una gran aplicabilidad para el análisis de señales como lo son el registro de la actividad cerebral.

Se resalta el potencial del análisis de datos funcionales en el sentido que la conversión de onda permite tomar una serie de n datos y tomarla como un único objeto el cual 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*.

Para futuras investigaciones, se sugiere aplicar algoritmos de clasificación para evaluar su potencial aplicación en la clasificación automática de señales, basados en datos funcionales.

Referencias

Acharya, R. (2005). Non-Linear analysis of EEG signals at various sleep stages. *Computer methods and programs in Biomedicine* , 37-45.

Ancoli, I., Chesson, A., & Quan, S. (2007). *The Aasm manual for the Scoring of Sleep and Associated.*

Anderson, M. (2008). Effects of sleep loss on sleep architecture in wistar rats: Gender-specific rebound. *Progress in Neuro-Psychopharmacology & Biological psychiatry*, 975-983.

Doris, M. (2009). Sleep classification according to AASM and rechtschaffen & kales. 139-149.

Ferraty, F. y. (2006). Nonparametric functional data analysis. *Springer-Verlag*.

Giles, G. R. (2009). Functiona data analysis with R and Matlabl. *Springer dordrecht heidelberg*, 99-115.

Julien Jacques, C. P. (2013). Funtional data clustering:a survey. *Springer-Verlag Berlin Heidelberg*.

Ramsay, J. a. (2005). Funtional Data Analysis. *Springer*.

T, T. (2012). Linear tranformations and the k-means clustering algorithm. *The american statistician*.

Yamamoto. (2012). Clustering of functional data in a low-dimensional subspace. *Springer-Verlag*, 219-225.

Yamamoto, M. T. (2014). Functional factorial K-means analysis . *Computational statistics & data analysis*, 133-148.