CARACTERIZACIÓN DE CURVAS OXIMÉTRICAS EN NIÑOS PREMATUROS MEDIANTE CLUSTERING DE DATOS FUNCIONALES

Papalardo, Cecilia [1]; Castro, Sebastián [1]
Chiappella, Lilian [2]; Criado, Alexandra [2]
Moreira, Isabel [3]; Scavone, Cristina [3]
ceciliap@iesta.edu.uy scastro@iesta.edu.uy

RESUMEN

La Displasia Broncopulmonar (DBP) es una enfermedad pulmonar crónica que ocurre con mayor frecuencia en recién nacidos pretérminos (RNPT) que requirieron oxigenoterapia y ventilación mecánica. La falta de oxígeno así como el aporte desmedido del mismo pueden ser factores desencadenantes de problemas en el sistema respiratorio y, como consecuencia, del desarrollo de la DBP. Actualmente, se utiliza el oxímetro de pulso (SpO_2) para monitorear en forma permanente y continua los requerimientos de oxígeno. En este trabajo se cuenta con oximetrías realizadas a un conjunto de niños prematuros, proporcionadas por el servicio de Neuropediatría del Centro Hospitalario Pereira Rossell (CHPR). El monitoreo de SpO_2 fue realizado a través de un oxímetro que permite obtener una gran cantidad de mediciones y almacenar los datos durante períodos de tiempo prolongados. Si bien se tiene para cada niño un conjunto finito de valores de SpO_2 puede pensarse que los mismos fueron generados por una función continua $x(t)$ que, en principio, podría ser evaluada en cada instante t si la herramienta de medición lo permitiera. De esta forma, los datos resultantes pueden ser vistos como un muestreo de las n curvas $x_1(t), ..., x_n(t)$, una para cada niño, lo que permite la utilización de las herramientas del *Análisis de Datos Funcionales*. El análisis de las oximetrías a través de este enfoque permite obtener una representación extendida y más suave de las curvas de SpO_2. En función de la dinámica que siguen las curvas oximétricas resulta de interés encontrar una tipología de recién nacidos prematuros y además la posible asociación de determinado comportamiento de las curvas con la presencia o no de broncodisplasia. En este trabajo se abordan estos problemas mediante técnicas de *clustering* o agrupamiento que permiten encontrar subgrupos homogéneos de curvas. A través de este análisis, se observan comportamientos diferenciados entre el grupo que presenta y el que no presenta la DBP que permiten entender mejor el comportamiento de las oximetrías en ambos grupos.

Palabras clave: *Broncodisplasia, oximetría, datos funcionales, clustering.*

[1]Departamento de Métodos Matemático-Cuantitativos. Área de Matemática/Instituto de Estadística.
[2]Escuela Universitaria de Tecnología Médica.
[3]Cátedra de Neuropediatría, Centro Hospitalario Pereira Rossell.

Índice

1. Introducción al problema

Un bebé recién nacido se considera pretérmino o prematuro (RNPT) si su nacimiento se produce antes de las 37 semanas de gestación. En un embarazo normal el parto tiene lugar entre las 38 a 40 semanas de edad gestacional, con lo cual si el nacimiento se produce antes es posible que los órganos del bebé no hayan madurado completamente y no estén preparados para la vida fuera del útero materno. Los niños prematuros suelen ser pequeños, con bajo peso y pueden, entre otras cosas, necesitar ayuda para respirar.

La *Displasia Broncopulmonar (DBP)* es una enfermedad pulmonar crónica que ocurre, con mayor frecuencia, en bebés prematuros con dificultad respiratoria aguda que requirieron aporte de oxígeno suplementario y ventilación mecánica. Esta patología, comenzó a ser tratada desde 1930 con aporte de oxígeno. Rápidamente se demostró que su utilización, al igual que sucede con otras drogas, puede tener efectos perjudiciales. Tanto la falta de oxígeno, como el aporte desmedido del mismo, pueden ser factores desencadenantes de problemas en el sistema respiratorio, y como consecuencia del desarrollo de la DBP.

Esta situación llevó a reconocer la necesidad de una adecuada monitorización del recién nacido. Se comenzó utilizando la medición de gases en sangre, desarrollándose con el tiempo, el oxímetro de pulso o saturómetro. La oximetría de pulso es un método continuo y no invasivo que permite medir en forma permanente el nivel de saturación parcial de oxígeno (SpO_2). En el mercado existen varios tipos de equipos cuyo principio de funcionamiento es básicamente el mismo, pero con marcada diferencia en la calidad de respuesta.

El monitoreo de SpO_2 realizado a través de algunos equipos permite obtener una gran cantidad de mediciones (cada 2 segundos) y almacenar los datos durante períodos de tiempo prolongados (hasta 24 horas). Además, es posible procesar los datos obtenidos con programas específicos que permitan un análisis detallado y graficar las señales en diferentes tiempos a través de una curva.

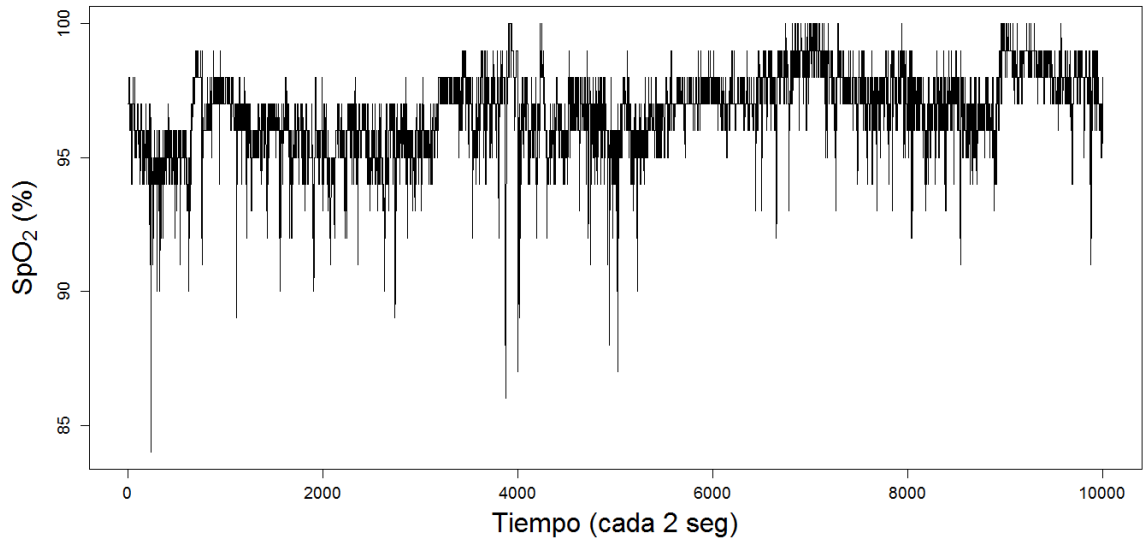

Figura 1: Ejemplo de una curva de SpO_2.

En función de la dinámica que siguen las curvas oximétricas resulta de interés encontrar una tipología de RNPT. Para ello, en este trabajo se utilizan técnicas de *clustering* o agrupamiento que permiten encontrar subgrupos homogéneos de curvas. El análisis de cluster es un proceso de aprendizaje no supervisado que realiza la división de un conjunto de datos en subgrupos, de forma que los objetos dentro de un grupo sean similares entre sí y diferentes a los objetos de otros grupos.

Además, interesa estudiar la posible asociación de determinado comportamiento de las curvas con la presencia o no de broncodisplasia. Por lo cual, se comparan los grupos obtenidos mediante las técnicas de clustering con los grupos que recibieron o no el diagnóstico de DBP. Como se observa en la Figura 2, las curvas de SpO_2 correspondientes a un niño que presenta DBP y uno que no padece la enfermedad pueden ser muy diferentes. En general, se espera que la curva oximétrica de un RNPT no pase más de un 10 % del tiempo de estudio por debajo de 90 %.

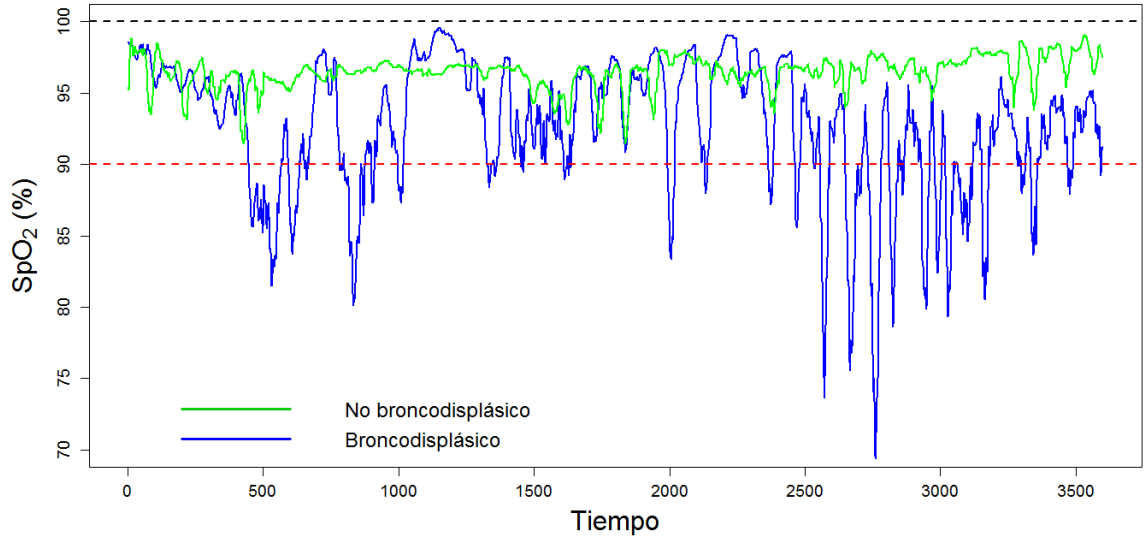

Figura 2: Comparación de la oximetría de pulso de dos niños.

A través de este agrupamiento se intenta brindar información adicional al momento de planificar los niveles apropiados de cuidados del recién nacido, como por ejemplo, al decidir si es necesario el aporte de oxígeno, la dosis y el tiempo que deberá suministrarse. Con esta información los especialistas pueden adoptar conductas terapéuticas diferenciadas para el tratamiento de los neonatos según el grupo en el que se encuentran.

2. Conjunto de datos a analizar

La información utilizada en este trabajo fue proporcionada por el servicio de Neuropediatría del Centro Hospitalario Pereira Rossell (CHPR). Este centro asistencial se encuentra ubicado en la ciudad de Montevideo, capital de la República Oriental del Uruguay. Los datos fueron obtenidos en el marco de un proyecto financiado por la Comisión Sectorial de Investigación Científica (CSIC) denominado *"Evaluación saturométrica y polisomnográfica de prematuros portadores de displasia broncopulmonar"* (Scavone *et al*,

2012). A partir de este proyecto, se realizaron entre el 22 de abril del 2009 y el 31 de diciembre del 2011 oximetrías de pulso prolongadas (durante por lo menos 12 hs.) a un conjunto de 207 niños que nacieron con peso de 1500 gramos o menor y/o con edad gestacional no mayor a 32 semanas.

La oximetría de pulso fue realizada en cada niño luego de transcurridas las 36 semanas de edad gestacional corregida (EGC) [4], abarcando varios períodos de ciclos sueño-vigilia. Se colocó el sensor en el pie del bebé. El oxímetro utilizado fue el Masimo modelo Rad - 5, que promedia cada 2 segundos, con memoria y software para descarga y análisis de datos.

Figura 3: Oxímetro de pulso Masimo Rad-5

Al visualizar los datos almacenados por el oxímetro se observó que algunas de las mediciones de SpO_2 correspondían a valores nulos. Considerando que estos registros pueden estar vinculados a movimientos bruscos del bebé durante el estudio, se decidió eliminar estos valores de cada una de las series. Adicionalmente, a causa de la alta irregularidad en las mediciones, se suavizaron los datos a través de promedios móviles. Para ilustrar este proceso, se presenta en la Figura 4 la curva suavizada correspondiente al ejemplo presentado al inicio en la Figura 1.

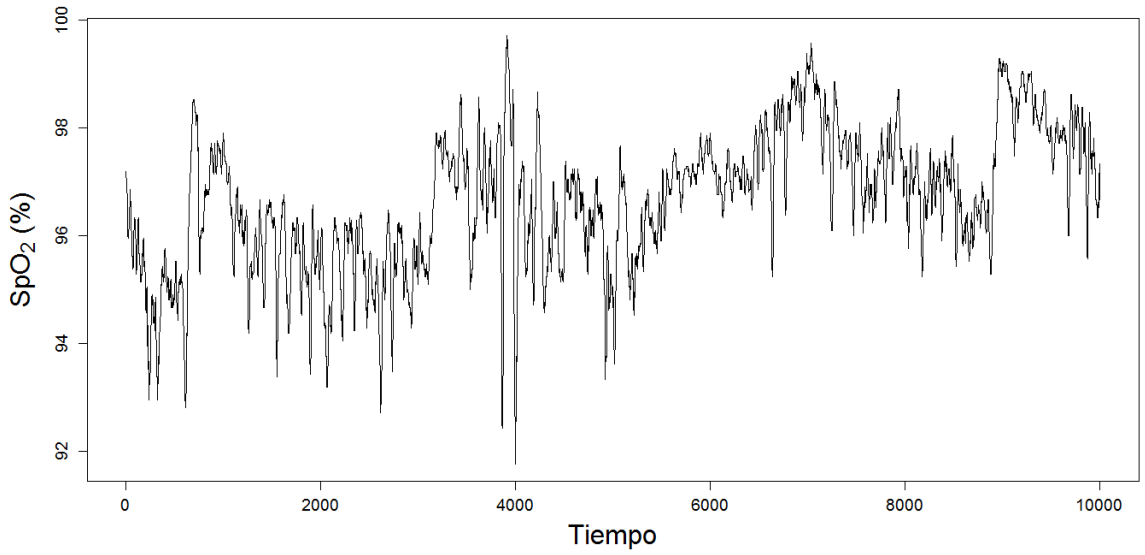

Figura 4: Curva suavizada mediante promedios móviles centrados de 21 observaciones.

[4]La EGC es la suma de la edad gestacional del bebé con la cantidad de semanas que transcurrieron luego de su nacimiento.

Dedido a que el tiempo de duración del estudio fue variable en cada bebé, se tomaron para cada serie los primeros 23017 valores que corresponde al número máximo disponible para todos los niños.

Por otro lado, utilizando el criterio diagnóstico para la DBP definido por el grupo de trabajo del Instituto Nacional de Salud de los Estados Unidos (NIH) (Tapia y González, 2008), se determinó que 64 de los niños estudiados eran portadores de la enfermedad. De esta manera, el conjunto de datos con el que se cuenta contiene información sobre el padecimiento de DBP, a través de una variable de tipo binaria (Si $= 1$, No $= 0$).

3. Metodología

Si bien se tiene para cada niño un conjunto finito de números, parece natural pensar que estos valores fueron generados por una función continua $x(t)$ que, en principio, podría ser evaluada en cada instante t si la herramienta de medición lo permitiera. De esta forma, los datos resultantes pueden ser vistos como un muestreo de las curvas $x_1(t), ..., x_n(t)$ ($n = 207$), lo cual permite la utilización de las herramientas del *Análisis de Datos Funcionales* (ADF).

3.1. Representación de la oximetría como dato funcional

Bajo el enfoque de ADF, el primer paso consiste en representar las funciones sobre un sistema de bases $\{\phi_1, \phi_2, ...\}$ a través de

$$x(t) = \sum_{k=1}^{\infty} c_k \phi_k(t)$$

y aproximar la misma mediante los primeros K términos

$$x(t) \approx \sum_{k=1}^{K} c_k \phi_k(t)$$

donde el parámetro K y los coeficientes c_k se determinan (estiman) a través de los datos disponibles (Ramsay y Silverman, 2005). El supuesto implícito en esta representación es que las funciones son relativamente *suaves* y que las mediciones fueron realizadas con algún nivel de ruido ajeno al objeto de estudio (Febrero Bande, 2008).

Existen distintos sistemas de bases como *Fourier*, *B-Splines*, etc. La primera suele ser utilizada para datos periódicos y la segunda para datos no periódicos. Debido a las características de los datos, en este trabajo se considera la utilización del sistema de bases *B-Splines*.

La determinación del parámetro K más adecuado es compleja, debido a que el mismo puede variar de individuo en individuo y además requiere de un conocimiento más profundo de cuán grande es el ruido que afecta las mediciones de SpO_2.

3.2. Clustering de datos funcionales

Se propone estudiar el agrupamiento de curvas oximétricas a través del análisis de cluster. Mediante este análisis se intenta separar un conjunto de datos en subgrupos homogéneos de forma que los individuos dentro de un grupo sean más similares entre sí que a los individuos de otros grupos.

Se utilizan dos técnicas de particionamiento (k-medias y PAM) y para medir la disimilaridad entre dos funciones $x_i(t)$ y $x_j(t)$ se utiliza la distancia L_2 que se calcula directamente sobre la representación funcional de las curvas como:

$$d(x_i, x_j) = \sqrt{\int_S (x_i(t) - x_j(t))^2 dt}$$

siendo S el dominio común de ambas funciones.

3.2.1. Algoritmo *k*-medias

Dado que la exploración de todas las particiones posibles de n individuos en k grupos es en general imposible, en la práctica se utilizan distintos algoritmos que son iterativos y analizan una cantidad limitada de particiones. Entre este tipos de técnicas k-medias es la más popular debido, entre otras cosas, a su eficiencia computacional.

3.2.2. Algoritmo *Partitioning Around Medoids* (PAM)

A diferencia de k-medias este algoritmo busca los k "individuos representativos" (o *medoids*) entre las observaciones del conjunto de datos. En general, PAM es más robusto que k-medias y requiere como argumento de entrada solamente la matriz de disimilaridades entre observaciones y no los datos originales. Como contrapartida es más intensivo computacionalmente (Izenman, 2008).

4. Resultados

Como se mencionó anteriormente, un primer paso en el análisis con el enfoque de ADF consiste en la representación de las oximetrías como datos funcionales. En las Figuras 5 y 6 se presentan ejemplos de ajuste de bases de Fourier y B-Splines con distinta cantidad de términos (K), para solo una de las 207 oximetrías. A simple vista no existen grandes diferencias entre los ajustes realizados y para ambos casos se puede observar que K es un parámetro de suavizado de los datos funcionales.

De aquí en adelante se trabaja solo con el sistema de bases B-Splines con el mayor valor de K planteado en el ejemplo, considerando que de esta forma será posible captar buena parte de la información de los datos originales sin caer en un excesivo sobreajuste. La utilización de otros valores de K, así como métodos automáticos para su determinación podrían constituir futuras aproximaciones de este problema.

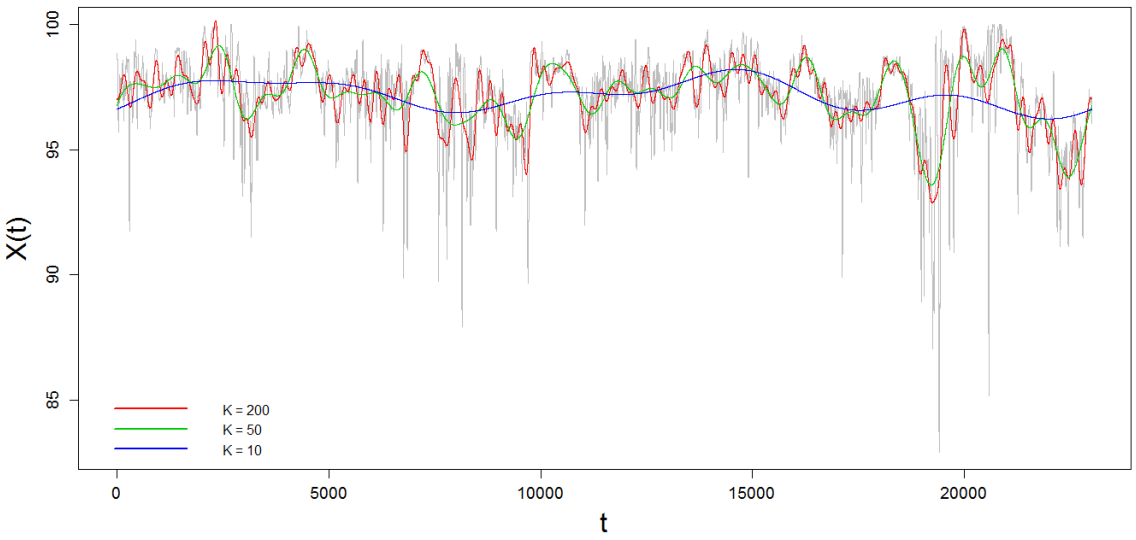

Figura 5: Ajuste con funciones bases de Fourier para distintos valores de K.

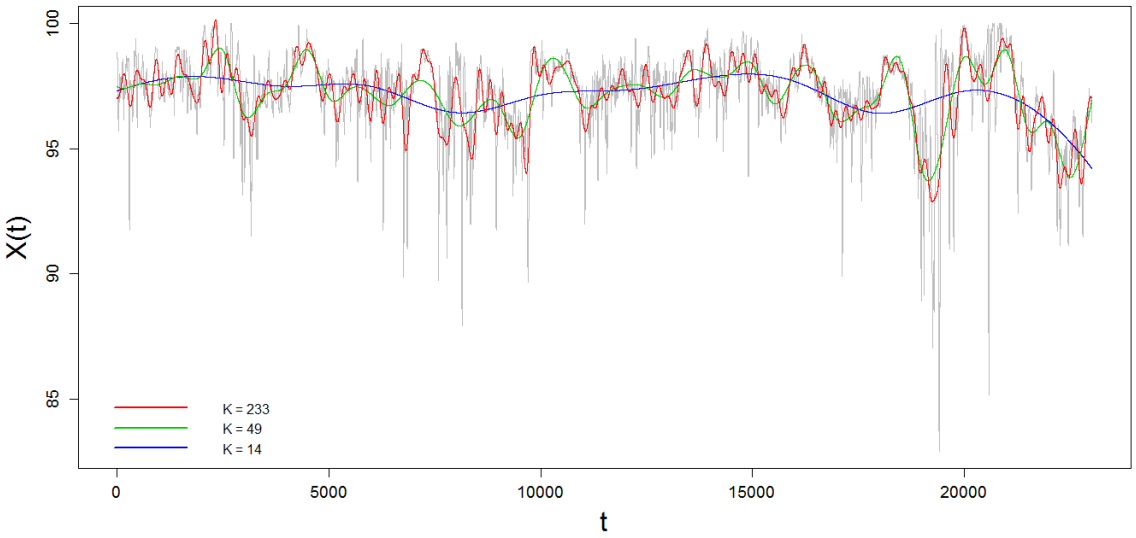

Figura 6: Ajuste con funciones bases B-Splines para distintos valores de K.

Luego se estudia el agrupamiento de curvas oximétricas a través del análisis de cluster. Para ello se utilizan dos técnicas de particionamiento (k-medias y PAM) y una medida de disimilaridad (L_2) calculada directamente sobre la representación funcional de las curvas[5]. Los grupos obtenidos se cruzan con los de broncodisplásicos y no broncodisplásicos y, por último, los resultados se comparan con los del análisis de cluster bajo el enfoque multivariado clásico.

[5]Estas distancias, al igual que el algoritmo de k-medias, se aplicaron utilizando la biblioteca `fda.usc` del entorno `R` versión 2.15.2 (R Core Team, 2012).

4.1. Algoritmo *k*-medias

Los resultados de la aplicación de este algoritmo sobre el conjunto de las representaciones de las oximetrías en las bases B-Splines, se muestran en las Figuras 7, 8 y 9 y en el Cuadro 1.

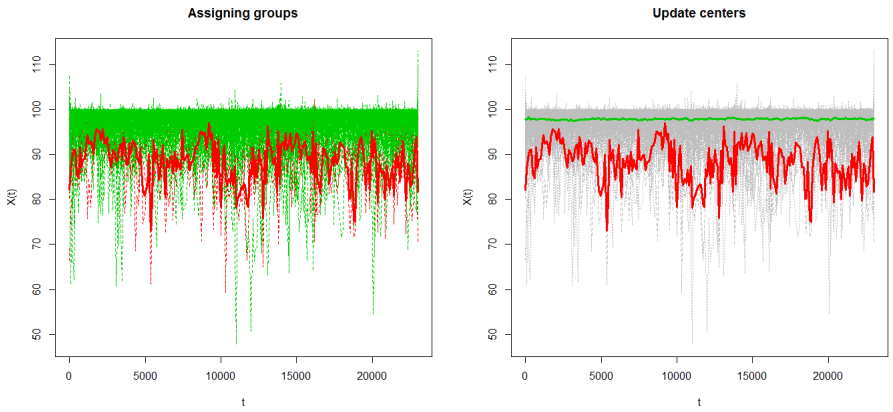

Figura 7: Gráficos de asignación de grupos y centros en k-medias para 2 grupos.

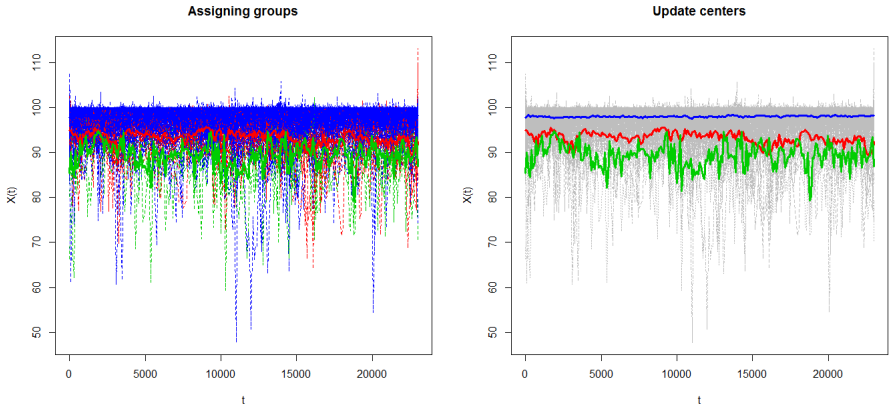

Figura 8: Gráficos de asignación de grupos y centros en k-medias para 3 grupos.

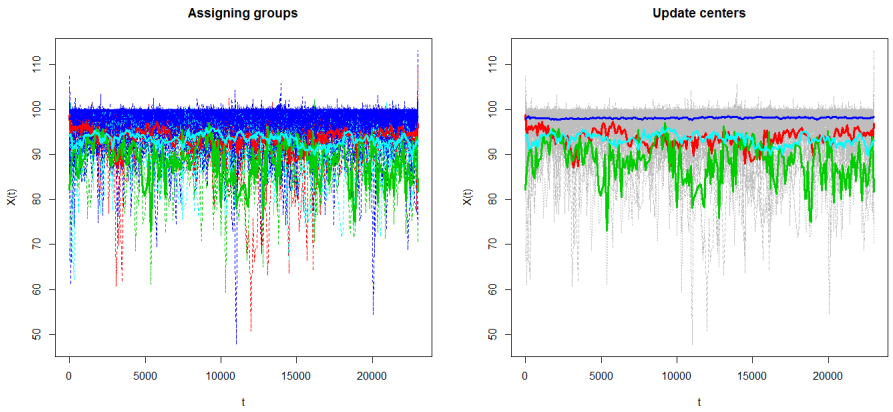

Figura 9: Gráficos de asignación de grupos y centros en k-medias para 4 grupos.

DBP	2 grupos		3 grupos			4 grupos			
No (143)	0	143	1	0	142	0	0	141	2
Si (64)	3	61	12	5	47	7	3	42	12
Total	3	204	13	5	189	7	3	183	14

Cuadro 1: Resumen de k-medias considerando 2, 3 y 4 grupos, y comparación con grupos de DBP (los colores se corresponden con los de las curvas en las respectivas figuras).

Se observa que en todas las configuraciones se obtiene al menos un grupo formado solo por niños broncodisplásicos, aunque son muy pocos en relación al total de niños que presentan la enfermedad. Además, teniendo en cuenta las curvas de los centros, estos grupos se corresponden con niveles promedios de SpO_2 más bajos e irregulares. En los restantes grupos formados se encuentran niños con y sin DBP.

4.2. Algoritmo *Partitioning Around Medoids* (PAM)

En el Cuadro 2 y en la Figura 10 se presentan los resultados de aplicar el algoritmo. En comparación con k-medias, para 2 grupos los clusters que se obtienen en PAM son más balanceados. Por otro lado, aunque no hay un grupo formado únicamente por broncodisplásicos, éstos predominan en uno de ellos y son minoría en el otro.

DBP	2 grupos		3 grupos			4 grupos			
No (143)	129	14	98	44	1	91	42	0	10
Si (64)	33	31	20	29	15	17	21	10	16
Total	162	45	118	73	16	108	63	10	26

Cuadro 2: Resumen de PAM considerando 2, 3 y 4 grupos, y comparación con grupos de DBP (los colores se corresponden con los de las curvas en las respectivas figuras).

Es interesante destacar que al ser en PAM los representantes de cada cluster individuos del conjunto de datos, las curvas de los mismos reflejan en mejor medida la irregularidad presente en las observaciones.

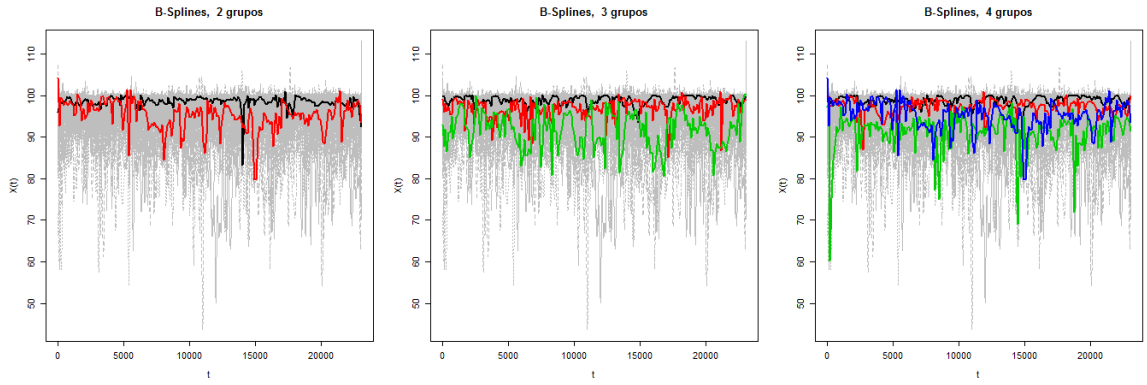

Figura 10: Resultados de clustering mediante algoritmo PAM para 2, 3 y 4 grupos.

4.3. Comparación con el enfoque multivariado

En esta sección se comentan los resultados que se obtienen realizando análisis de cluster sobre una representación multivariada de los datos de oximetrías. Para ello se considera cada observación como un vector multivariado de dimensión 23017 y se aplica sobre estos datos las versiones clásicas de k-medias y PAM. Los resultados se resumen en los Cuadros 3 y 4, y en la Figura 11. Se observa que en la representación multivariada las curvas son "más rugosas" que en el enfoque de datos funcionales, y que esto se traslada, en cierto modo, tanto a los centros de k-medias como a los de PAM. De esta manera, se dificulta la distinción entre las principales características de los representantes de los grupos obtenidos. Por otro lado, la asignación de grupos que se obtiene no parece ser muy diferente, en cuanto a las características principales de los clusters, que la resultante para los datos funcionales.

DBP	2 grupos		3 grupos			4 grupos			
No (143)	4	139	40	103	0	16	57	0	70
Si (64)	25	39	28	22	14	19	17	14	14
Total	29	178	68	125	14	35	74	14	84

Cuadro 3: Resumen de k-medias sobre datos multivariados, considerando 2, 3 y 4 grupos, y comparación con grupos de DBP.

DBP	2 grupos		3 grupos			4 grupos			
No (143)	118	25	73	70	0	65	70	0	8
Si (64)	25	39	35	14	15	25	14	11	14
Total	143	64	108	84	15	90	84	11	22

Cuadro 4: Resumen de PAM sobre datos multivariados considerando 2, 3 y 4 grupos, y comparación con grupos de DBP.

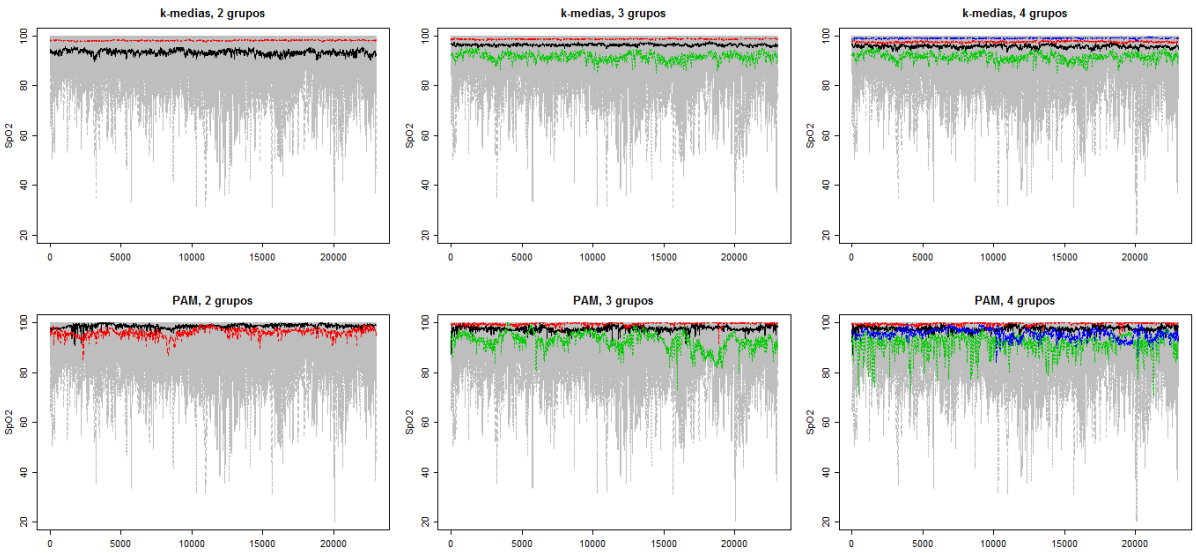

Figura 11: Resultados de clustering multivariado utilizando k-medias y PAM, para 2, 3 y 4 grupos.

5. Comentarios finales

El análisis de las oximetrías a través del enfoque de datos funcionales permite obtener una representación extendida y más *suave* de las curvas de SpO_2, más allá de los datos disponibles. Luego, con esta representación es posible utilizar los desarrollos de las técnicas de clustering para datos funcionales. En ambos casos se observó ciertos comportamientos diferenciados entre los grupos de broncodisplásicos y no broncodisplásicos que permiten entender mejor la dinámica de las curvas en ambos grupos. Si bien la representación multivariada de los datos es posible, la misma no permite visualizar con igual claridad algunos de los resultados mencionados anteriormente.

En un futuro sería interesante profundizar en parte de los análisis realizados (suavizado inicial de los datos, determinación del parámetro K en ADF, otras técnicas de clustering y medidas de disimilaridad, etc.), así como explorar nuevas alternativas como modelos de clasificación supervisada para la presencia o no de DBP en función de las curvas oximétricas y otras covariables.

La posibilidad de encontrar una tipología de RNPT según sus curvas oximétricas permitiría acercarse a la determinación sistemática de la necesidad de aporte suplementario de oxígeno. El conocimiento de requerimiento de oxígeno en estos niños, permite realizar una utilización racional de la oxigenoterapia, ajustando el aporte a las necesidades individuales. Este aspecto es fundamental para asegurar un adecuado crecimiento y desarrollo de los bebés prematuros.

6. Bibliografía

Febrero Bande, M. (2008). A present overview on functional data analysis. *Boletín de Estadística e Investigación Operativa*, **24**, pp. 7–14.

Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques : Regression, Classification, and Manifold Learning*. Springer Texts in Statistics. Springer New York.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org/`

Ramsay, J. y Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer, 2ª ed..

Scavone, C.; Lorenzo, D.; Moreira, I.; Chiappella, L.; Criado, A. y Sastre, L. (2012). Evaluación saturométrica y polisomnográfica de prematuros con y sin displasia broncopulmonar. *Archivos de Pediatría del Uruguay*, **B3(3)**, pp. 170–175.

Tapia, J. L. y González, Á. (2008). *Neonatología*. Mediterraneo, 3ª ed..