Facultad
de
Ciencias

ANÁLISIS DE PATRONES ELÉCTRICOS CLÍNICOS CON REDES COMPLEJAS E IDENTIFICACIÓN DE NODOS

(Clinical Electric Pattern Analysis with Complex Networks and Node Identification)

Trabajo de Fin de Grado
para acceder al
GRADO EN MATEMÁTICAS

Autora: Carlota García Fernández
Directora: Alicia Nieto
Co-Director: Paolo Bonifazi
Febrero 2025

Resumen

El objetivo de este proyecto es encontrar patrones que permitan automatizar el proceso de clasificación de áreas del cerebro relacionadas con la epilepsia. Los últimos estudios en medicina demuestran que la epilepsia es un fenómeno de red y que las diferentes áreas del cerebro pueden participar con distinto papel en el proceso de generación de crisis, de forma que pueden clasificarse en 3 grupos en función de si no están involucradas en los episodios epilépticos, si son la fuente de los episodios o si solo los promueven.

Generalmente, la identificación de nodos epileptogénicos (en su método clínico) se hace mirando la dinámica de las señales eléctricas al principio de las crisis, para poder identificar que zonas muestran actividad epiléptica en un primer momento. En este trabajo, proponemos la hipótesis de que los nodos epileptogénicos se pueden identificar también en la actividad basal (sin crisis) del cerebro. Por lo tanto, los datos utilizados serán dos horas de señales cerebrales eléctricas basales (sin actividad epiléptica) adquiridas en 11 pacientes con electrodos profundos intracraneales implantados en regiones potencialmente epileptogénicas.

A partir de estos datos se va a realizar un análisis de clasificación no supervisada de los nodos cerebrales basado en métricas de redes funcionales, utilizando diferentes técnicas y modelos estadísticos con los que se busca conseguir un algoritmo de asignación nodal concluyente.

El primer análisis explora varias bandas de frecuencia y duración de ventanas (desde fracciones hasta decenas de segundos) sobre la cual las redes funcionales son reconstruidas a partir de las series temporales. En cada ventana de tiempo, el grado de similitud entre grupos epilépticos conocidos y los grupos identificados con métricas de redes complejas indicará si la señal cerebral basal es una buena fuente de datos para este tipo de estudios sobre la epilepsia.

Para el segundo análisis, se calculan 55 métricas basadas en la media y la covarianza de diferentes métricas extraídas de las redes funcionales dinámicas y de nuevo se estudia la similitud entre grupos epilépticos conocidos y los grupos identificados con este nuevo procedimiento. Además, se compararán ambos análisis entre sí.

Las métricas y los patrones eléctricos más representativos de los nodos epileptogénicos son identificados y caracterizados para así poder sacar como conclusión de este trabajo una nueva estrategia e hipótesis sobre la identificación de regiones cerebrales epileptogenicas a partir de señales cerebrales interictales (basales y fácilmente adquiribles) sin necesitad de estudio de crisis epilépticas (ya que estos son episodios no siempre frecuentes).

Palabras clave: *Reconocimiento de Patrones, Aprendizaje no supervisado, Análisis Cluster, Métricas de centralidad, Redes Complejas, Asignación Nodal, Epilepsia, sEEG.*

Abstract

The main goal of this project is to find patterns that allows the automatization of the brain regions classification process in relation with epilepsy. The last medical studies show that epilepsy is a network problem and that the different brain regions could participate in a different way in the generation of an epilepsy crisis. In this way, these areas could be classified in 3 groups depending on whether they are not involved in the crisis, are the source of the episodes, or only promote them.

Generally, the identification of the epileptogenic nodes (in the clinic way) is done looking at the electric signal dynamic at the beginning of the crisis, to try to identify which zone shows epileptogenic activity first. In this project, we propose the hypothesis that epileptogenic nodes can also be identified in the brain's baseline activity (without seizures). So, the data that is going to be used is 2 hours of electric baseline cerebral signal (without seizures) for 11 patients with deep intracranial electrodes located in potentially epileptogenic zones.

From this data, we are going to do an unsupervised classification analysis of the brain nodes based on the functional network's metrics, using different tecnics and statistics methods, looking for a conclusive nodal classification algorithm.

The first procedure explores different frequency bands and time windows duration (from fractions to dozens of seconds) on which functional networks are reconstructed from the time series. In each time window, the degree of similarity between known epileptogenic groups and the ones identified using complex network metric will indicate if the brain's baseline electric signal is a good data source for this type of epilepsy studies.

For the second procedure, we calculate 55 metrics based on the media and covariance of different metrics of dynamic functional networks and again we will study the similarity between the known epileptogenic groups and the ones identified in this new procedure. Also, we compare both procedures.

The most representative metrics and electrical patterns of epileptogenic nodes are identified and characterized to conclude this project with a new strategy and hypothesis about identification of epileptogenic brain regions from interictal electric signal (baseline and easy acquired), without the need of seizure studies, as these episodes are not always frequent.

Key Words: *Pattern Recognition, Unsupervised Learning, Clustering, Centrality metrics, Complex Networks, Node Asignation, Epilepsy, sEEG.*

Agradecimientos

A los dos tutores de este trabajo, Alicia y Paolo, por su entrega, dedicación y paciencia, además de por acompañarme y guiarme en este año de crecimiento y aprendizaje.

A Chema, mi tutor del grado, por confiar en que llegaría hasta aquí desde el principio y por acompañarme en todas las dificultades del camino.

Y a mi familia, por estar siempre ahí para mí.

Índice general

Capítulo 1: Introducción

1.1. Sobre la epilepsia

La epilepsia o condición de convulsiones recurrentes no provocadas es resultado de una amplia variedad de causas, siendo uno de los trastornos cerebrales graves más destacados del mundo afectando a unas 50 millones de personas en todo el mundo. Además, se estima que en un 30 % de los casos las convulsiones siguen estando mal controladas a pesar del tratamiento médico máximo [1].

Se conoce que la epilepsia se desarrolla dentro de circuitos del cerebro, donde se generan de forma localizada señales epilépticas que luego se propagan a otras áreas del cerebro y provocan las conocidas convulsiones. Estas convulsiones a nivel interno generan variaciones intensas de la señal eléctrica que el cerebro emite a lo largo del tiempo. Cada una de estas convulsiones perjudica a la salud del paciente, por lo que resulta interesante tratar no solo de eliminarlas, si no de que no sea necesario el estudio de las convulsiones de cada paciente para poder erradicar la enfermedad.

Hace años que esta enfermedad se considera un problema de red en el que el foco o focos se distribuyen a lo largo de las estructuras límbicas[1] [3, 4]. Es por esto por lo que son necesarias técnicas de monitorización de actividad cerebral completa con alta resolución, existen varias, pero las que se van a nombrar en este trabajo son las derivadas de la electroencefalografía (EEG). La EEG es una técnica que se utiliza para la monitorización no invasiva, a partir de electrodos que se colocan en el cuello cabelludo y registran la actividad eléctrica.

La resolución o calidad de los datos con esta técnica es baja debido a las interferencias con el cráneo entre la duramadre y el cuero cabello. Es por esto por lo que existen variaciones como la EEG intracraneal (iEEG), en la que los electrodos se colocan directamente en la parte expuesta del cerebro durante una cirugía [5]. Actualmente, en la fase de diagnóstico y tratamiento quirúrgico de la epilepsia se utiliza esta técnica.

La otra modalidad de monitorización que mencionaremos en el trabajo es la de la estereoelectroencefalografía (sEEG), lo que supone un punto medio entre la EEG y la iEEG, ya que los electrodos son introducidos en el cuero cabelludo por lo que tienen más resolución que cuando simplemente están en contacto con la superficie, pero no se necesita cirugía por lo que no es un método tan invasivo [5]. Los datos con los que se va a trabajar en este estudio han sido tomados con este método de sEEG.

Con estas técnicas, lo que se busca es monitorizar una crisis epiléptica, para poder detectar en que nodo o zona del cerebro se produce la crisis y posteriormente mediante cirugía eliminar (cauterizar o ablacir, en términos médicos) esa zona del cerebro, para que así no pueda volver a producirse.

1.2. Planteamiento y motivación

Un estudio reciente [6], de F. Bartolomei y coautores, analiza las redes neuronales relacionadas con la epilepsia, mostrando que existen 3 tipo de zonas en el cerebro de un paciente con epilepsia, aquellas que no están involucradas en las crisis epileptogénicas, denominadas como *Not involved networks* (NIZ), aquellas que propagan la crisis epiléptica *Propagation Network* (PZN) y aquellas en las que se encuentra la lesión denominadas *Epileptogenic zone Network* (EZN), como se muestra en el esquema de la Figura 1.1. Este estudio [6], además explica como la técnica de sEEG puede ayudar a la detección y análisis de señal eléctrica.

Como se ha explicado en la sección anterior, un estudio EEG (no invasivo) proporciona datos de baja resolución, por lo que en ocasiones no es capaz de proporcionar una identificación definitiva del foco epiléptico (que correspondería con un nodo Nivel 3 (EZN)), es por esto que se opta por insertar un electrodo intracraneal para una identificación con precisión del foco epiléptico.

En general, los médicos identifican las regiones que generan estas convulsiones epileptógeneas mediante la detección de los puntos de inicio de las convulsiones registradas por los electrodos intracraneales. Las convulsiones pueden ser espontáneas o inducidas por suspensión a través de medicación, pero destaca el que no ocurren con frecuencia, por lo que los pacientes sometidos a sEEG deben estar monitorizados durante al menos dos semanas

[1] Estructuras del cerebro interconectadas que median las emociones, el aprendizaje y la memoria [2].

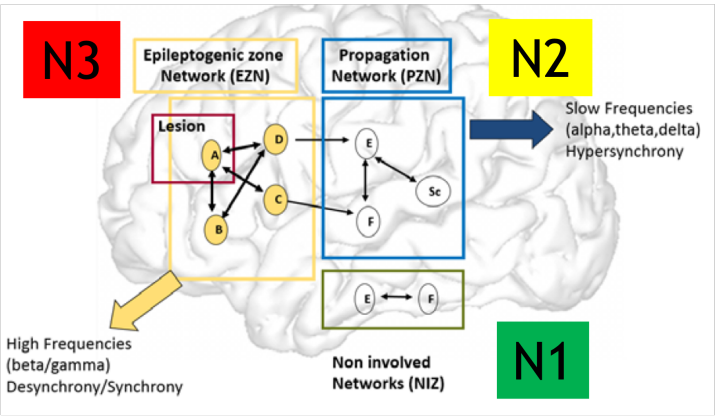

Figura 1.1: Esquema de la relación de las diferentes zonas del cerebro con las crisis epilépticas. Se muestran tres tipos de zonas, a las que asignaremos un nivel y un color. Nivel 1 (color verde) zonas no involucradas o NIZ, Nivel 2 (color amarillo) zonas de propagación o PZN y Nivel 3 (rojo) zonas lesionadas o EZN. Este dibujo ha sido obtenido de [6]

seguidas, ya que cada convulsión (aunque sea corta) es valiosa para que los médicos puedan observar si existe una identificación definitiva de las regiones que generan convulsiones epileptogéneas.

La originalidad de los datos usados en este TFG es el uso de rastros no epilépticos, que representan más del 99 % de los registros durante el periodo de monitorización de los pacientes con electrodos profundos. La hipótesis es que un cerebro epiléptico presenta, en su arquitectura de red funcional, anomalías (marcadores) que están presentes de manera basal, y que participan en el proceso que desencadena las crisis. Es como decir que en un cerebro epiléptico hay elementos o indicadores que están presentes independientemente de cuando se genera las crisis, y estos nos pueden indicar cuales son las regiones epileptogénicas.

Teniendo en cuenta esto y que la epilepsia se entiende como un fenómeno de red, nos planteamos estudiar la actividad basal del cerebro epiléptico utilizando la aproximación de red complejas, un campo matemático que proviene de la teoría de grafos que ha sido aplicado en las últimas décadas en diferentes sistemas biológicos, así como en otros muchos campos. ([3] [4])

Es por esto, que los médicos tienden a identificar los diferentes tipos de regiones a partir de los trazos del estudio sEEG, siguiendo las pautas del estudio [6], como regiones Nivel 1 (verde), Nivel 2 (amarillo) o Nivel 3 (rojo). En un estudio previo se pidió a los médicos que hicieran esta clasificación con 11 pacientes concretos para poder comparar sus observaciones con los resultados que se obtuviesen mediante una automatización de la clasificación utilizando algoritmos no supervisados. Estos datos servirán como referencia en este trabajo.

En resumen, este trabajo tiene el propósito de obtener información acerca de las crisis epileptogénicas, mediante la medición de métricas de red a partir de la correlación de las señales eléctricas del cerebro para discernir entre los tres tipos de nodos cerebrales, introduciendo la novedad de que la señal eléctrica del cerebro se ha recogido en dos horas en las que el paciente no sufre crisis epilépticas. Con esto, se desea confirmar la hipótesis de que en esta señal basal hay información sobre las crisis epilépticas.

Capítulo 2: Datos del problema

2.1. Variables del problema

Se tienen 11 pacientes, en cada uno de ellos se insertaron entre 3 y 8 electrodos para registrar actividad cerebral. Cada uno de estos electrodos tiene varios contactos eléctricos separados que registran información a diferente profundidad (separándose a una distancia de entorno a 3,5 mm) de forma que se obtienen entre 28 y 75 nodos de registro de señal eléctrica en el cerebro de cada paciente. Estos electrodos se colocan en el cerebro de los pacientes en diferente posición, se muestra un ejemplo de esta colocación en la Figura 2.1, donde se muestra una tomografía del cerebro del Paciente 3.

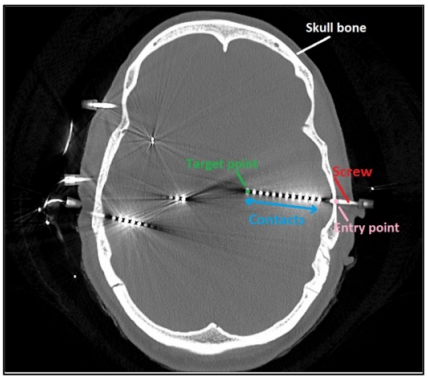

Figura 2.1: Ejemplo de la disposición de los nodos en el cerebro del Paciente 3 con 28 nodos. Se trata de una tomografía (CT) real, en la que destaca en contorno blanco el hueso del cráneo. En azul se ha destacado un conjunto de nodos, colocados por la introducción de un tornillo marcado en rojo a través del punto de entrada marcado en rosa. Esta tomografía ha sido obtenida junto con el conjunto de datos que ha proporcionado el codirector de este trabajo.

De cada uno de estos nodos de registro de señal se obtuvo una grabación eléctrica de 2 horas con 500 observaciones por segundo para 9 pacientes y 512 para los 2 restantes. Este número de observaciones por segundo se conoce como la *frecuencia de muestreo del electroencefalograma (f)* y se mide en hercios, donde 1 Hz corresponde a 1 observación por segundo. Para que la señal tomada por los sEEG no se distorsione, la *frecuencia en muestreo de sEEG* debe ser entorno al doble de la máxima frecuencia que interesa en el estudio, como se explicará en la sección 2.2, en este caso la frecuencia estudiada máxima será de 249Hz.

Cada contacto eléctrico coincide con un nodo de grabación que ha sido clasificado por los médicos, previamente a este estudio, como hemos comentado en la sección anterior, como verde (Nivel 1), amarillo (Nivel 2) o rojo (Nivel 3) siguiendo la metodología de clasificación de las zonas epileptogénicas del estudio [6].

De esta forma se tienen variables funcionales $X_1, X_2, ..., X_n$ con $n = 11$ que van a representar la información de cada paciente. Cada variable X_i con $i = 1, ..., n$ va a contener la información de k_i nodos, este número de nodos es diferente para cada paciente. Al existir más de un nodo por paciente, se trata de un problema de datos funcionales multivariantes. De forma que se definen las variables del problema:

$$X_i = \{X_{\{i,1\}}, \dots, X_{\{i,k_i\}}\}, \quad \text{con } i \in \{1, ..., 11\}. \tag{2.1}$$

Comentar que las $X_{\{i,j\}}$ son las variables funcionales unidimensionales y que los diferentes valores de k_i se muestran en la Tabla 2.1,

Estos datos funcionales que se observan a lo largo del tiempo no se pueden observar en un completo continuo por lo que ademas a cada variable le corresponde lo que hemos definido anteriormente como *frecuencia de muestreo del electroencefalograma* f_i (añadimos el subíndice i porque tendrá un valor para cada paciente). De forma que se añade la variable t_i que representa el número de datos que se obtienen a lo largo del tiempo (alrededor de 2 horas) que vendrá determinada por esta f_i de la siguiente forma: $t_i = f_i \cdot s_i$, siendo s_i los segundos totales

que dura la medición. De esta forma cada variable $X_{\{i,j\}}$, con $i = 1, ..., n$ y con $j = 1, ..., k_i$ en la práctica es $X_{\{i,j\}} = (X_{\{i,j,1\}}, ..., X_{\{i,j,t_i\}})$. La imagen de estas $X_{\{i,j,l\}}$, con $l \in \{1, ..., t_i\}$ se encuentra en los reales.

A continuación, en la Tabla 2.1 se muestran los valores de las variables mencionadas anteriormente de todos los pacientes.

i	1	2	3	4	5	6	7	8	9	10	11
k_i	57	66	28	64	56	35	71	35	55	34	75
f_i	500	500	500	500	500	500	500	500	500	512	512
t_i	3600000								3600166	3685945	3686238
s_i	7200 seg								7032 seg	7200 seg	7200 seg

Tabla 2.1: Datos de los pacientes a usar en el análisis.

Para los Pacientes 2 y 8, un estudio previo [7] nos indica que hay 3 nodos que deben ser eliminados en el estudio ya que la información que proporcionan es contradictoria y no representativa. Estos nodos son el 3, 7 y 13 para el Paciente 2 y el 9, 10 y 28 para el Paciente 8. Por lo tanto, en la práctica $k_2 = 63$ y $k_8 = 32$.

Para el caso concreto de estos 11 pacientes, se tiene como referencia una asignación de etiquetas previas para los k_i nodos; es decir para cada variable X_i se tiene una variable conocida $y_i = \{y_{\{i,1\}}, ..., y_{\{i,k_i\}}\}$, que contiene la asignación para cada nodo k_i de un nivel dentro del campo de la epilepsia 1 (NIZ), 2 (PZN) o 3 (EZN) de acuerdo a lo explicado en la introducción (Figura 1.1), que llamaremos vector de asignaciones esperadas. El objetivo será por tanto encontrar una asignación $Y_i = \{Y_{\{i,1\}}, ..., Y_{\{i,k_i\}}\}$ para cada variable X_i mediante métodos de aprendizaje no supervisado, en estos métodos solo se tienen en cuenta las variables X_i y una vez encontrada la asignación Y_i para comprobar si con el método se obtienen resultados concluyentes se compara con las variables conocidas y_i. Esta comparación se va a realizar teniendo en cuenta el porcentaje de asignaciones que coinciden en cada pareja de variables, creando la variable tasa de precisión ($accP$), o porcentaje de nodos con valores coincidentes entre Y_i e y_i. Entonces, para un paciente i, se denotará como $accP$ a lo siguiente:

$$accP = \frac{\sum_{j=1}^{k_i} cont_j}{k_i}, \quad \text{donde } cont_j = \begin{cases} 1, & \text{si } Y_{\{i,j\}} = y_{\{i,j\}} \\ 0, & \text{si } Y_{\{i,j\}} \neq y_{\{i,j\}} \end{cases} \text{ con } i \text{ fijo.} \tag{2.2}$$

Estos datos de referencia conocidos y_i han sido obtenidos a través de la experiencia médica del estudio a lo largo de los años mediante resonancias en medio de crisis epilépticas y cirugías, por tanto el objetivo de este estudio es obtener un algoritmo que nos proporcione esta información a cerca de los nodos mediante el estudio estadístico de la señal eléctrica en un periodo en el que no se dan crisis epilépticas, es decir, idealmente, identificar 3 grupos dentro de Y_i con $i \in \{1, ..., 11\}$ que coincidieran con los 3 grupos identificados por los médicos (que se encuentran en la variable y_i con $i \in \{1, ..., 11\}$).

2.2. Procesamiento de datos

Una vez se tiene la señal cerebral basal se necesitará realizar un preprocesado a los datos que se detallará a continuación.

En primer lugar, la técnica de sEEG recoge las oscilaciones de la actividad cerebral, que son muy variables debido a que el cerebro esta constantemente recibiendo estímulos sensoriales, también cuando estamos dormidos o no hacemos nada. No siempre se sabe descifrar de donde provienen estos estímulos ya que evocan a una continua actividad cerebral espontánea. Por lo que una vez recogida la actividad global se utilizan determinados estímulos sensoriales para acoplar las oscilaciones globales de forma coherente, dando lugar a ritmos inducidos dentro de la señal [8]. Es decir se filtra la señal global de X_i, de forma que se queden seleccionados solo los valores de la señal correspondientes a cada una de las siguientes actividades normales del cerebro que son: dormir, estar relajado sin ninguna atención, estar relajado con algo de atención, estar activo con algo de ansiedad y por último, estar concentrado. Cada uno de estos estados tiene un rango de frecuencias aproximado asociado [9].

Es por esto, que se van a filtrar los datos originales para separar la información correspondiente a estas 5 bandas de frecuencia estándar (en contexto neurológico y de neurociencia fundamental), de forma que se repetirán

Nombre de la banda	Rango de frecuencias	Estado cerebral
Delta δ	0.5-3 Hz	Dormido
Theta θ	3-7 Hz	Profundamente relajado
Alpha α	7-13 Hz	Relajado, atención pasiva
Beta β	13-30 Hz	Ansiedad dominante, Actividad, Atención al entorno
Gamma γ	>30 Hz	Concentrado

Tabla 2.2: Bandas de frecuencias asociadas a estados cerebrales, en las que se va a dividir la señal original a través de un filtrado de datos.

las metodologías con solo la señal de cada uno de estos 5 estados cerebrales cada vez, los valores concretos de frecuencias se muestran en la Tabla 2.2.

En la siguiente Figura 2.2 se muestra la representación de la señal eléctrica cerebral cruda, sin preprocesar, en cada nodo para el Paciente 6, X_6. Para que se pueda observar la señal de cada nodo, se va a mostrar una variable tipificada Z_6 que solo se utilizará para esta representación. Asumiendo que la media y la varianza teóricas de X_6 existen se obtendrá la variable Z_6 tipificada a partir de la media $\bar{x}$ y la varianza s^2 muestral de X_6 de forma que $Z_6 = (X_6 - \bar{x})/s^2$, creando así la nueva variable que se representa con media 0 y varianza 1. Destacar que la señal original tiene unidades de microvoltios.

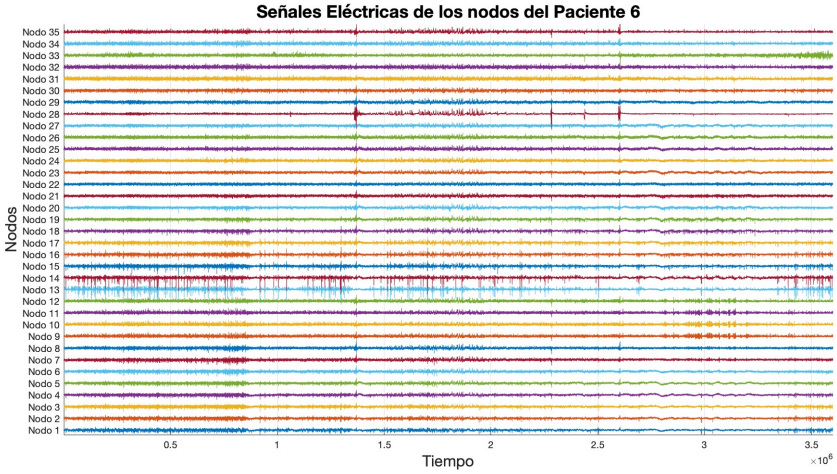

Figura 2.2: Señal eléctrica tipificada de la forma $Z_6 = (X_6 - \bar{x})/s^2$ de los 35 nodos para el ejemplo del Paciente 6.

A simple vista vemos que la señal de ciertos nodos se comporta de manera similar, lo que puede indicar correlación entre los nodos. Sin embargo, observamos que la señal varía a lo largo de toda la línea temporal, y que hay zonas en las que la señal es similar en varios nodos y en otras menos. Vemos por ejemplo, los primeros 11 nodos hasta los 800000 datos o los nodos 13 y 14 en el tiempo final, cuya señal parece ensancharse en el mismo momento para todos.

Esto nos lleva a plantear la cuestión de si a lo largo de las dos horas de medición hay intervalos de tiempo que aportan información sobre las crisis epilépticas y otros que no, es por esto por lo que se plantea la opción de dividir la señal en ventanas de tiempo. A raíz de esta división optamos por dos tipos de estudio, por un lado, estudiar cada ventana de tiempo como caso separado y analizar si hay ventanas de tiempo en las que las coincidencias entre Y_i e y_i sean significativas u optar por un segundo análisis, en el que basándonos en estas ventanas de tiempo en su totalidad, estudiamos la red compleja completa obteniendo un resultados único a partir de la variabilidad media de todas las ventanas.

De esta forma, se partirá de la variable $X_{\{i,j\}} = (X_{\{i,j,1\}}, ..., X_{\{i,j,t_i\}})$, con $i = 1, ..., n$, con $j = 1, ..., k_i$, y con t_i como se definió en la sección anterior (número de datos a lo largo del tiempo) $t_i = f_i \cdot s_i$, que se dividirá en ventanas de tiempo y se obtendrá un valor de asignaciones para cada ventana de tiempo.

Por tanto, para el primer procedimiento se obtendrá un vector de asignaciones no solo por cada paciente sino también por cada ventana de tiempo. Entonces cada paciente i tendrá un Y_i formado por un número v_f de $Y_{\{i,j\}}$, de

la siguiente manera: $Y_{\{i,j\}} = (Y_{\{i,j\}_1}, ..., Y_{\{i,j\}_{v_f}})$ donde v_f es el número de ventanas de tiempo para cada división del número total de datos. Este valor v_f variará en función del tipo de preprocesamiento (de ahí el subíndice f) como explicaremos a continuación, y recordemos que j es el identificador del nodo en cuestión.

En cambio para el segundo procedimiento se obtendrá un solo vector de asignaciones Y_i por paciente al estudiar el conjunto de todas las ventanas de tiempo en global.

El análisis de los resultados por tanto consistirá en estudiar la precisión entre la asignación Y_i, para el primer procedimiento en cada ventana de tiempo, es decir en cada $Y_{\{i,j\}_s}$ con $s \in \{1, ..., v_f\}$ con la asignación de referencia y_i, y el número de ventanas de tiempo en el que esta precisión se da por concluyente. Y en obtener este porcentaje de precisión ($accP$ (2.2)) de la asignación única que obtenemos para el segundo análisis.

Para poder dividir los datos de una banda de frecuencia en concreto en ventanas de tiempo se debe tener en cuenta el concepto de *Ciclos de frecuencia por ventana de tiempo en cada banda de frecuencia*. Un ciclo de frecuencia corresponde a un período completo de la oscilación (llamando oscilación a la señal variante, de forma que el periodo empieza cuando la señal corta en un punto como puede ser el eje x y termina cuando vuelve a pasar por ese punto) dentro de la señal eléctrica.

Esta división se debe realizar de tal forma que cada ventana de tiempo contenga oscilaciones completas para obtener datos concluyentes acerca de la señal. Para conocer el tamaño que debe tener esta ventana de tiempo en función de los ciclos que se deseen incluir en ella se utiliza el valor de frecuencia lp (la más baja que hay en la banda de frecuencia que se ha establecido), medida en herzios, que será la mínima en el intervalo, para asegurarnos que la ventana siempre sea lo suficientemente larga para incluir los ciclos que se deseen. De esta forma, si c es el número de ciclos que debe contener la ventana, se usa lp, la frecuencia más baja dentro del rango de frecuencia que se desea filtrar, para establecer la duración de la ventana d_f (medida en segundos) de la siguiente forma:

$$d_f = \frac{c}{lp}.$$

Esta d_f es la duración de la ventana en segundos (que depende de la banda de frecuencia y el numero de ciclos por eso se añade el subindice f).

Con este planteamiento, se va a seguir el siguiente esquema de preprocesamiento de datos.

- **Paso 1:** División de los datos en ventanas de tiempo.
- **Paso 2:** Aislamiento de la señal X_i en un rango de frecuencia concreto.
- **Paso 3:** Eliminación del ruido de la señal (que se observa en los 50 y 150 Hz).

El primer paso, por tanto, es dividir estos datos $X_{\{i,j\}}$ en ventanas de tiempo de un tamaño concreto, para ello se van a escoger 3 valores para el número de ciclos por ventana de tiempo para el estudio que son 5, 10 y 15. Una vez escogido esto, se va a dividir cada $X_{\{i,j\}}$ en ventanas de tiempo buscando este número de ciclos en cada una. Se encuentran muchas diferencias entre dividir la señal teniendo 5 ciclos por ventana o 15 ciclos por ventana, por lo que puede ser que para algunos casos sea favorable un número de ciclos u otro.

Para esta división, se utilizará la variable de paso p_f, que se calcula por la siguiente Ecuación (2.3):

$$p_f = \frac{c}{lp} \cdot f_i = d_f \cdot f_i. \tag{2.3}$$

Esta variable paso, determinará el número de puntos en cada ventana de tiempo y en consecuencia el número de ventanas en la que se dividen los datos de la siguiente forma: $v_f = t_i/p_f$, tanto v_f como p_f dependerán tanto de la variable t_i, como de la banda de frecuencia con la que se desea trabajar.

Para aclarar la definición de estas variables, se va a presentar un caso práctico, por ejemplo para el Paciente 1. Supongamos que queremos trabajar en la banda de frecuencia de $> 30Hz$ con 15 ciclos completos por ventana ($lp = 30$ y $c = 15$). El número de puntos por ventana $p_f = \frac{c \cdot f_1}{lp} = \frac{15 \cdot 500}{30} = 250$. Entonces los $t_1 = 3600000$ puntos por nodo se van a agrupar en ventanas de 250 puntos por nodo. De forma que el numero de ventanas es $v_f = t_i/p_f = 3600000/250 = 14400$.

Destacar que no consideramos en este estudio ningún solapamiento entre ventanas de tiempo, por el exceso de datos generados que eso provocaría. Al haberse obtenido en este estudio información relevante se considerará seguir trabajando con solapamiento de ventanas (como se explicará en la última sección del capítulo 5 de conclusiones),

especialmente para tratar de identificar la ventana o ventanas de tiempo con la que se obtiene una tasa de precisión ($accP$) muy alta.

Entonces $X_{\{i,j\}} = (X_{\{i,j,1\}}, ..., X_{\{i,j,t_i\}})$ se divide en $X_i = (X_{\{i,j\}_1}, ..., X_{\{i,j\}_{v_f}})$, donde cada $X_{\{i,j\}_s}$ con $s \in \{1, ..., v_f\}$ es igual a $X_{\{i,j\}_s} = (X_{\{i,j,s_1\}}, ..., X_{\{i,j,s_p\}})$, utilizando la notación s_p como equivalente a s_{p_f}, para simplificar, con p_f la variable paso, que representa las p_f observaciones dentro de la ventana de tiempo s. De nuevo, la imagen de cada vector dentro de la matriz $X_{\{i,j\}_s}$ se encuentra en los reales.

Para los pasos 2 y 3 se utilizará un método de preprocesado realizado por el co-director de esta trabajo para un estudio previo que no se puede citar, ni incluir en esta memoria, por que aún no está publicado. El resumen de lo que hace este método de preprocesamiento es el siguiente. Para el paso 2 se filtra el conjunto de los datos $X_{\{i,j\}_s}$ de forma que solo quede señal eléctrica de la banda de frecuencia que se requiera, explicadas dentro de la Tabla 2.3 que son 0.5-3; 3-7; 7-13; 13-30 y >30 Hz. En la Figura 2.3 se muestra el mapa de calor (los valores de la señal en una leyenda por colores) para la señal del ejemplo del Paciente 6, para mostrar como se diferencia la señal de los mismos datos después de haberse preprocesado en cada una de las bandas de frecuencia.

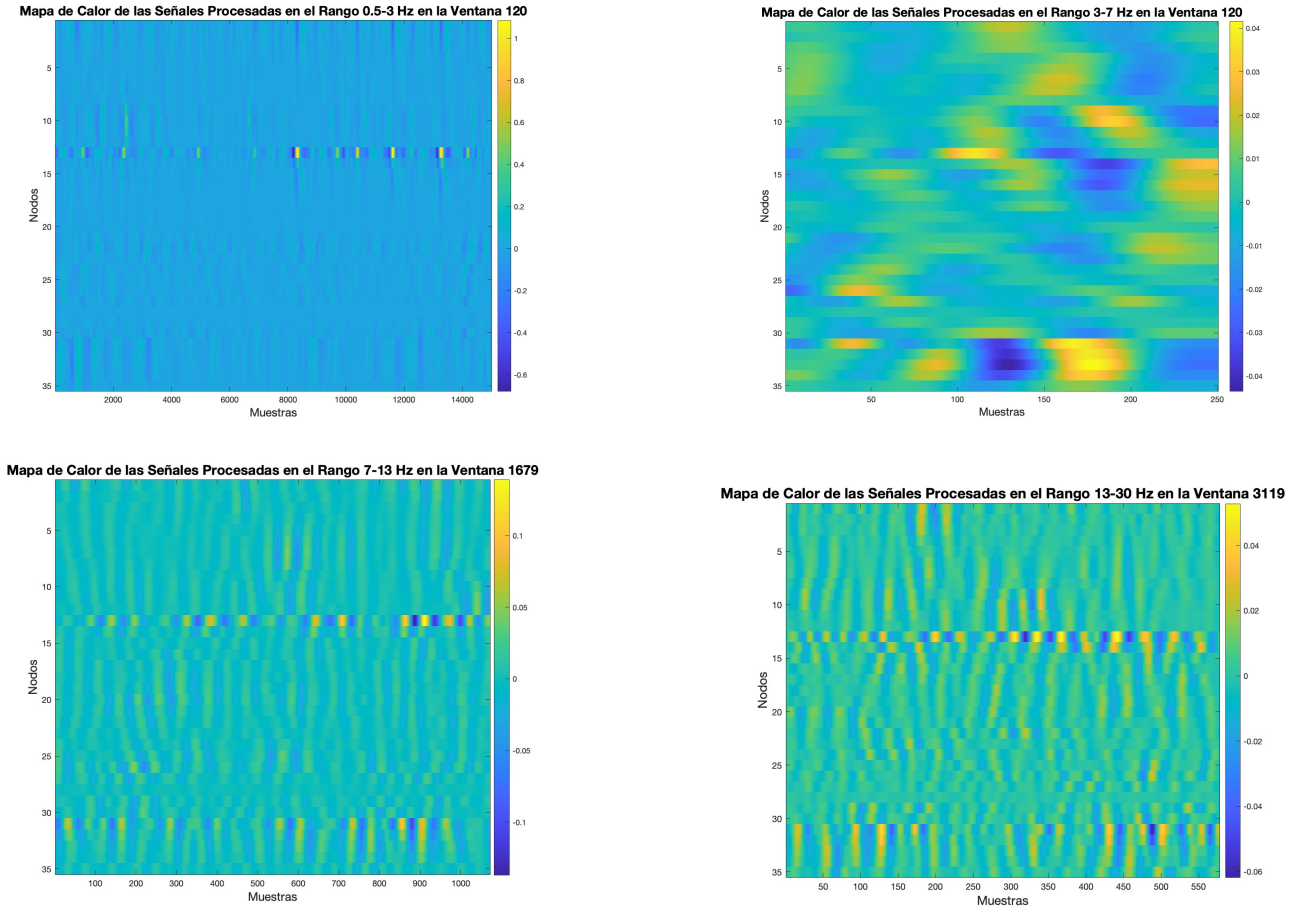

Figura 2.3: Mapa de Calor en una ventana cualquiera para cada banda de frecuencia, escogida aleatoriamente de entre toda la muestra para el Paciente 6.

Después, para el paso 3 lo que hace este método de preprocesamiento es eliminar el ruido o interferencia que hace la señal del aparato de medición (por ser un aparato electrónico) dentro de los datos, este ruido se considera que se encuentra en los 50 y 150 Hz.

Una vez procesados los datos, se tiene la posibilidad de incluir un último paso, que también pertenece a este método de preprocesamiento que comentamos, que es la eliminación de la señal común a todos los nodos utilizando regresión lineal múltiple, en el primer procedimiento que se desarrolla en este estudio este paso será un parámetro, pues se realizará para los 11 pacientes aplicando este paso (lo que denominaremos como preprocesado con regresión) o sin aplicar este último paso (preprocesado sin regresión). La eliminación de la señal común por un lado, es interesante

para poder detectar las características únicas de cada nodo, sin embargo, esto puede llevar a un exceso de pérdida de correlación entre los nodos por lo que se va a hacer el estudio para ambos casos. En el segundo procedimiento solo se estudiarán los datos con el preprocesado con regresión, debido a que las métricas utilizadas así lo exigen, como veremos en la sección 3.2.

Para realizar este último paso se utiliza una función de regresión lineal. En la siguiente Figura 2.4 se muestra la correlación de la señal de cada electrodo sin eliminar esta señal común (a la izquierda) y una vez eliminada (a la derecha), para el caso concreto del Paciente 4 con 64 nodos.

Para mostrar como afecta este paso del preprocesamiento a la señal original para el caso concreto del Paciente 4 X_4, se va a mostrar la señal tipificada Z_4, como se mostró en la Figura 2.2 en una ventana de tiempo cualquiera, s, es decir, voy a representar $Z_{\{4,j\}_s}$ en la Figura 2.5. En ella, encontraremos, a la izquierda, la señal tipificada antes de aplicar la función de regresión para todos los nodos, y además en negrita se muestra lo que se considera como señal común a todos los nodos. A la derecha encontramos la señal después de haber eliminado la común, es decir, después de haber aplicado la función con regresión.

Como resultado final del preprocesamiento se obtiene por cada paciente 30 conjuntos de datos $X_{\{i,j\}_s}$, 15 preprocesados con regresión y 15 sin ella y dentro de estos 15, 3 por cada banda de frecuencia escogida, uno para 5 ciclos por ventana de tiempo, otro para 10 y otro para 15.

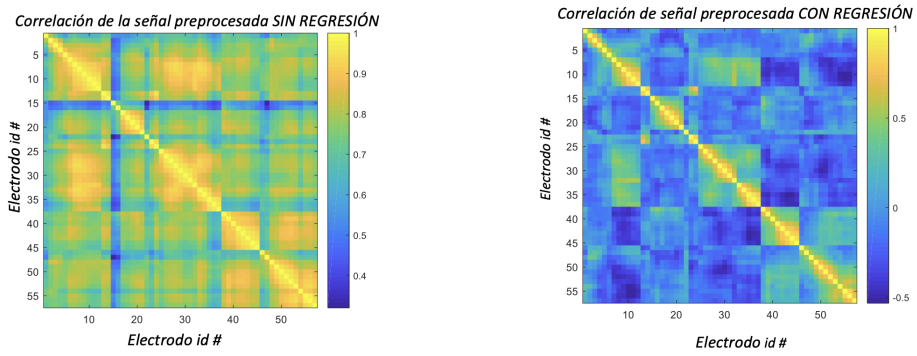

Figura 2.4: Correlación de la señal antes de aplicar la función de regresión (a la izquierda) y después de haberlo aplicado (a la derecha).

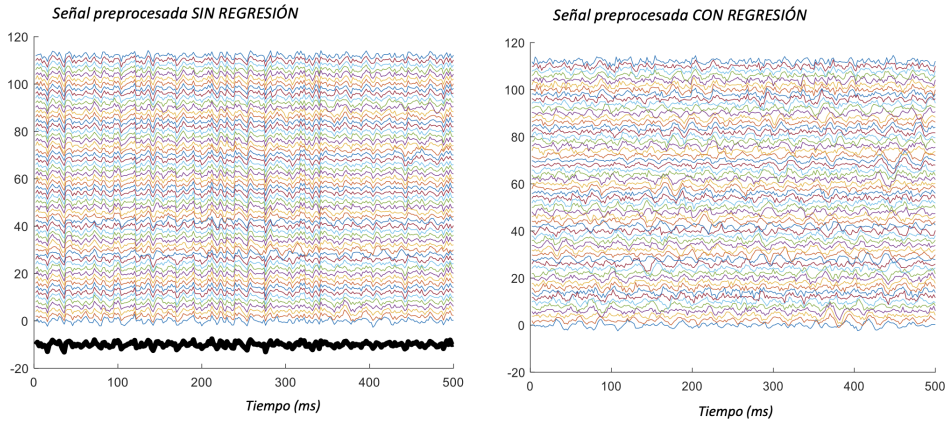

Figura 2.5: Señal eléctrica tipificada $Z_{\{4,j\}_s}$ antes de aplicar la función de regresión (a la izquierda) y después de haberla aplicado (a la derecha). Además en la figura de la izquierda se muestra en negrita lo que se considera como señal común a todos los nodos, que es precisamente la que se va a eliminar para obtener la figura de la derecha.

Capítulo 3: Metodología

Se van a llevar a cabo varios procedimientos de clasificación no supervisada o aprendizaje automático no supervisado, en inglés denominado *clustering*, que por eso en español también se suele denominar como análisis cluster, notación que usaremos a partir de ahora, que durante el grado se estudia en la asignatura de Advanced Statistics.

El análisis cluster consiste en la búsqueda de grupos naturales dentro de un conjunto de datos, tratando de que los datos dentro del mismo grupo sean similares entre sí. Existen varios métodos para realizar esta división en grupos, en particular, se va a trabajar con métodos que utilizan las distancias entre los puntos para realizar grupos, como es el método de k-medias [10], que basándose en k centros aleatorios, siendo k el número de grupos que se busca, se agrupan los puntos por cercanía a estos centros, una vez hechos los grupos se recalculan los centros y se repite el algoritmo hasta que este converge, es decir, hasta que los centros de los grupos no cambian significativamente de una iteración a otra.

Para el caso de este estudio en concreto, previamente al análisis cluster se van a calcular varias métricas, formando un vector de dimensión nodal para cada una, en cada ventana de tiempo y a partir de combinaciones de estos vectores se creará la matriz de métricas que se utilizará en el método de k-medias para realizar la clasificación.

El objetivo de esta clasificación es encontrar para cada muestra i, 3 grupos dentro del conjunto de nodos k_i basándonos en la hipótesis del estudio [6] de que las zonas del cerebro de cada paciente con epilepsia se pueden dividir en tres grupos correspondientes a zonas que no están involucradas en las crisis epileptogénicas, denominadas como *Not involved networks* (NIZ, Nivel 1 verde), aquellas que propagan la crisis epiléptica *Propagation Network* (PZN, Nivel 2 amarillo) y aquellas en las que se encuentra la lesión denominadas *Epileptogenic zone Network* (EZN, Nivel 3 rojo).

Se van a realizar dos procedimientos con este objetivo, el primero de ellos a partir del estudio de las ventanas de tiempo por separado, este procedimiento se va a repetir para el conjunto de todos los casos, es decir, para cada paciente i, procesado en cada uno de los 5 rangos de frecuencia escogidos, para cada uno de los 3 valores de ciclos escogidos, todo ello preprocesado con y sin regresión. Es decir en total, para el primer procedimiento se repetirá el proceso para cada paciente 30 veces con estas variaciones y se analizarán los grupos obtenidos del análisis cluster comparándolos con los grupos esperados y_i para determinar la eficiencia de cada uno de los parámetros utilizados y sacar conclusiones. El segundo procedimiento consistirá en el estudio de la red compleja a partir del conjunto total de ventanas, en este caso solo se repetirá para distintas bandas de frecuencia y número de ciclos pero en todo momento se usarán los datos preprocesados con regresión.

3.1. Primer Procedimiento: Agrupamiento por medidas de centralidad en cada ventana de tiempo.

El objetivo de este primer procedimiento es obtener una división en tres grupos de los nodos de cada paciente. Para ello, se calculará para cada ventana de tiempo varias medidas de centralidad, de forma que cada variable $X_{\{i,j\}_s}$ con i representando al paciente, j al nodo y s a la ventana de tiempo, tenga asociado un punto m-dimensional, siendo m el número de métricas calculadas.

Para cada ventana de tiempo, sobre estos k_i puntos m-dimensionales se realizará un análisis cluster por el método de k-medias. Una vez se obtienen los 3 grupos cluster por el método (llamados C1, C2 y C3) para todas las ventanas de tiempo se hará un análisis (explicado en la sección 3.1.3) de qué correspondencia encaja más con lo esperado entre los grupos C1, C2 y C3 y los niveles epileptogénicos 1,2 y 3 que contiene la variable y_i.

Entonces se calculará la tasa de precisión ($accP$) a partir de (2.2), así como una tasa de precisión aleatoria a través de una simulación de montecarlo (de 1000 casos) y se hará una selección de las ventanas de tiempo cuya tasa de precisión tiene una probabilidad mayor del 95 % de ser mayor que la aleatoria. Esto se explicará también en la sección 3.1.3.

Se repetirá este proceso de obtención de la clasificación nodal pero solo para un subconjunto concreto de las métricas que pasen el Test de Kruskal-Wallis [11] que se explicará en la sección 3.1.5, para así optimizar los resultados.

A continuación, se detalla cada parte del procedimiento 1.

3.1.1. Cálculo de Métricas

Una vez se han procesado los datos, se obtiene $X_i = (X_{\{i,j\}_1}, ..., X_{\{i,j\}_{v_f}})$. A partir de esto, para cada $X_{\{i,j\}_s}$ con $s \in \{1, ..., v_f\}$ se van a calcular varias medidas de centralidad [12] que aportarán información sobre la topología de la red. La métricas escogidas son las llamadas *Fuerza, Fuerza al cuadrado, Centralidad de Intermediación de los nodos, Centralidad del vector propio* y *Centralidad de subgrafo*. Antes de definirlas vamos a introducir algunos conceptos que se necesitará obtener para poder calcularlas. Estos conceptos previos son:

Matriz de Correlación. Para cada matriz $X_{\{i,j\}_s} = (X_{\{i,j,s_1\}}, ..., X_{\{i,j,s_p\}})$, con i y s fijo, pero j variable entre 1 y k_i, que son las columnas y con p el número de filas (observaciones por ventana de tiempo definido como p_f aunque ahora omitimos el subíndice para simplificar la notación), se calcula la matriz de correlación de Pearson muestral R, que tendrá dimensión nodal, es decir k_i x k_i, [13]. Entonces, sean j_1 y j_2 las filas y columnas de R, j_1 y $j_2 \in \{1, ..., k_i\}$, cada valor de la matriz será (en notación matricial):

$$R(j_1, j_2) = \frac{(X_{\{i,j\}_s} - \overline{X}_{j_2})^T (X_{\{i,j\}_s} - \overline{X}_{j_1})}{\sqrt{(X_{\{i,j\}_s} - \overline{X}_{j_1})^2}\sqrt{(X_{\{i,j\}_s} - \overline{X}_{j_2})^2}} \tag{3.1}$$

Donde $\overline{X}_j$ es la media muestral de la columna j, calculada de la siguiente forma:

$$\overline{X}_j = \sum_{x=1}^{p} \frac{1}{p} X_{\{i,j,s_x\}} \tag{3.2}$$

A continuación, en la Figura 3.1 se muestra un ejemplo de la matriz de correlación para el Paciente 1, con la señal preprocesada en el rango de frecuencias entre 30 y 249 Hz con una división en 14400 ventanas de tiempo. Se muestran 4 ventanas de tiempo distintas, las dos de arriba son ventanas de tiempo en las que se consigue una tasa de precisión alta (calculada a partir de (2.2)), de hecho máxima, entre $Y_{\{1,j\}_s}$ e y_1, con $s = 10612$ y $s = 673$, y las 2 de abajo muestran ventanas de tiempo en las que se consigue una tasa de precisión de 0, con $s = 14299$ y $s = 14371$. En esta figura se puede observar lo que parece un patrón, por el cual las ventanas con más correlación (arriba) dan mejores resultados, y las de menos correlación (abajo) dan peores resultados.

Matriz de Adyacencia. Una vez se tiene la matriz anterior se define una matriz de Adyacencia, que se utilizará para el cálculo de varias métricas. Se obtiene tomando la parte positiva de la matriz de correlación, de forma que se toma el máximo entre cada valor y 0. Como se observa en la Figura 3.2, para un análisis paralelo puede ser interesante tener en cuenta esta misma matriz pero escogiendo los valores negativos, es decir, tomando el valor absoluto del mínimo entre cada valor y 0. De esta forma para la ventana de tiempo de máxima precisión que se obtiene en el caso anterior se muestran ambas matrices de adyacencia en la Figura 3.2. Esta matriz se denotará como $A^+_{j_1,j_2}$ en el primer caso y $A^-_{j_1,j_2}$ en el segundo y se define como:

$$A^+_{j_1,j_2} = \begin{cases} R(j_1, j_2), & \text{si } R(j_1, j_2) > 0 \\ 0, & \text{si } R(j_1, j_2) \leq 0 \end{cases} \tag{3.3}$$

$$A^-_{j_1,j_2} = \begin{cases} 0, & \text{si } R(j_1, j_2) > 0 \\ |R(j_1, j_2)|, & \text{si } R(j_1, j_2) \leq 0 \end{cases} \tag{3.4}$$

Como veremos en la sección 3.1.4 utilizar $A^+_{j_1,j_2}$ o $A^-_{j_1,j_2}$ para calcular las métricas será un parámetro del modelo, es decir, se calcularán las dos versiones de las métricas por separado (utilizando estas dos definiciones de matriz de adyacencia) y se compararán los resultados. Por eso de aquí en adelante la notación en las definiciones será A_{j_1,j_2} y en el capítulo 4 de resultados se especificará en qué ocasiones se ha utilizado la matriz de adyacencia basada en la parte positva de la matriz de correlación $A^+_{j_1,j_2}$ y cuando la basada en la parte negativa de la matriz de correlación $A^-_{j_1,j_2}$.

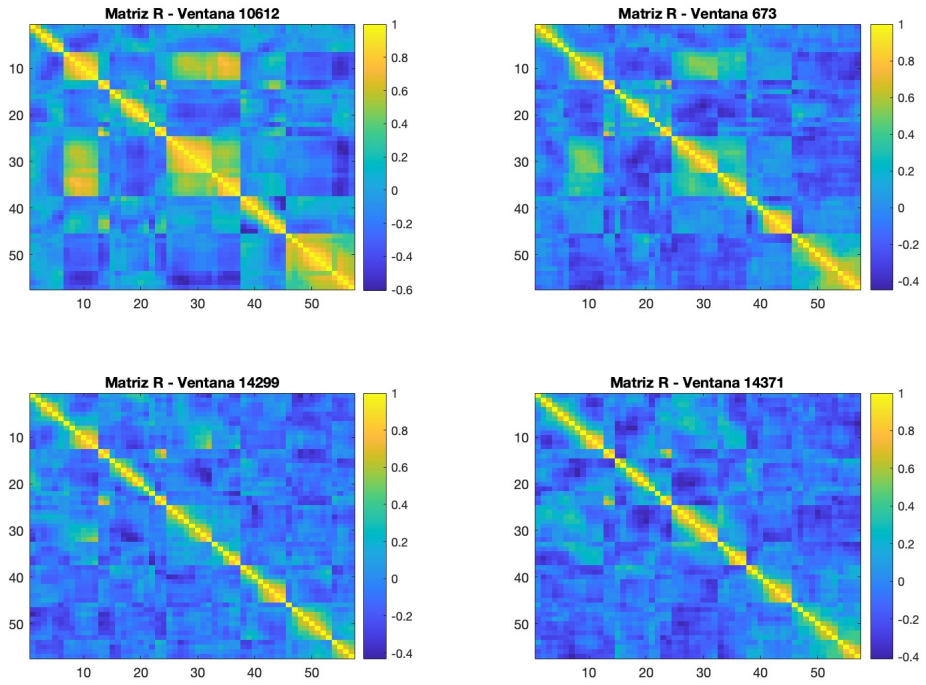

Figura 3.1: Muestra las diferencias entre la matiz de correlación para dos ventanas de tiempo con las que se obtiene una tasa de precisión máxima (arriba, ventanas 10612 y 673) y dos con la mínima, es decir, 0 (abajo, ventanas 14299 y 14371).

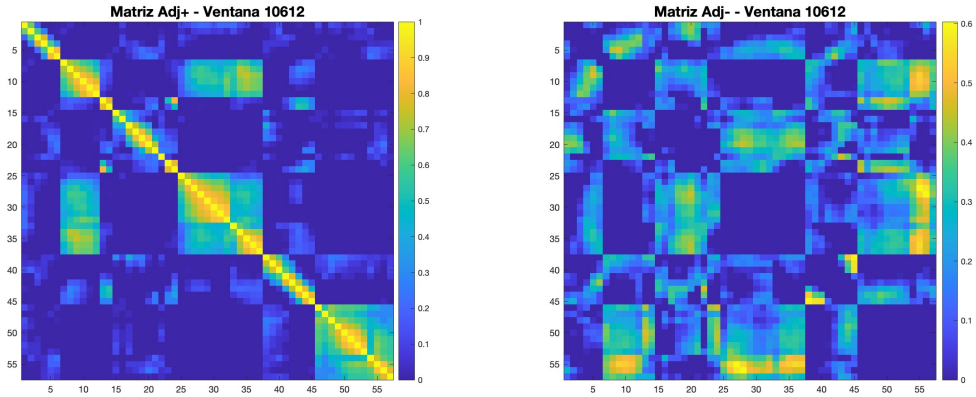

Figura 3.2: Muestra las diferencias entre la matiz de adyancencia tomando los valores positivos de la matriz de correlación (a la izquierda) y la matriz de adyacencia tomando los valores negativos de la matriz de correlación, y posteriormente tomando su valor absoluto (a la derecha). Para la ventana de máxima precisión del Paciente 1 en la banda de frecuencia de 30 a 249 Hz, con 15 ciclos y con regresión.

Grado. Describe el número de nodos conectados a uno en concreto. Se va a obtener a través de la suma por filas de la matriz de adyacencia binaria (de dimensión k_i x k_i). El grado de cada nodo vendrá determinado por la siguiente ecuación:

$$d(j_2) = \sum_{j_1=1}^{k_i} A^B_{j_1,j_2}. \tag{3.5}$$

Donde $A^B_{j_1,j_2}$ es la matriz de adyacencia binaria, se calcula en base a la matriz de adyacencia, de forma que cada valor de la matriz será 1 si la el valor de A_{j_1,j_2} es positivo y será 0 cuando el valor de A_{j_1,j_2} sea 0.

Como nota, cabe destacar que se consideró utilizar la métrica de *Grado*, como una más para el procedimiento, ya que es muy habitual en el contexto de problemas de red, sin embargo, se descartó ya que la densidad de la conectividad[1] es de alrededor del 50 %, por lo que esta métrica que se calcula a partir de la Ecuación (3.5), da siempre valores muy uniformes en todos los nodos. Se ha definido porque se necesita para calcular la métrica de *Centralidad de PageRank* (Ecuación (3.10)).

Ahora pasamos a definir las métricas que se van a emplear en el posterior algoritmo de k-medias:

Fuerza. Se trata de obtener el peso total de cada nodo dentro de la red y se calcula sumando las filas de la matriz de adyacencia. De forma que para cada nodo se tendrá el valor:

$$w(j_2) = \sum_{j_1=1}^{k_i} A_{j_1,j_2}. \tag{3.6}$$

Donde A_{ij} es la matriz de adyacencia anteriormente definida.

Fuerza al cuadrado. Con el objetivo de enfatizar las correlaciones más fuertes entre los dos nodos se va utilizar esta variación de la métrica anterior. Será de gran utilidad en esos pacientes en los que la correlación es mas baja de lo habitual en todos los nodos.

$$wd(j_2) = \sum_{j_1=1}^{k_i} A^2_{j_1,j_2}. \tag{3.7}$$

Donde A_{j_1,j_2} es la matriz de adyacencia anteriormente definida.

Centralidad de Intermediación de los nodos. Métrica que indica si un nodo j se encuentra en el camino mas corto entre otros nodos. Se puede calcular como:

$$b(j_2) = \sum_{s \neq j_2 \neq t} \frac{\sigma_{st}(j_2)}{\sigma_{st}}. \tag{3.8}$$

Donde $\sigma_{st}(j_2)$ es el número de caminos más cortos que empiezan en el nodo s y acaban en el t pasando por j_2, mientras que σ_{st} es el número total de caminos de s a t, estas variables se calculan a partir de una función ya existente obtenida del siguiente repositorio [14].

Centralidad del vector propio. Se obtiene a partir del calculo de los autovectores de la matriz de Adyacencia de dimensión k_i x k_i. De forma que para cada nodo j_2, $v(j_2)$ corresponda a la j-ésima componente del autovector e asociado al mayor valor propio del nodo j_2, λ_{max}.

$$v(j_2) = \frac{1}{\lambda_{max}} \sum_{j_1=1}^{n} A_{j_1,j_2} \cdot e_j. \tag{3.9}$$

[1] Medida que describe cuán interconectados están los nodos de una red, relacionando el número de conexiones existentes y el número de conexiones posibles máximas [12]

Centralidad de Pagerank. Esta métrica es una variante de la anterior. Siendo A la matriz de adyacencia, D matriz diagonal con los valores de los grados $d(j_2)$ (que se pueden calcular como se definió en (3.5)), a un factor de amortiguamiento y k_i el numero total de nodos. Se define en notación matricial el vector $p \in \mathbb{R}^{k_i}$ como:

$$p = (-f \cdot A \cdot D^{-1})^{-1}\Big((1-a)\cdot\frac{1}{k_i}\Big). \tag{3.10}$$

Se utiliza el valor más común de factor de amortiguamiento que es 0.85 [12], además el término final de $\frac{1}{k_i}$ corresponde con elegir que la distribución inicial de probabilidad en el grafo sea uniforme.

Centralidad del subgrafo de un nodo. Suma ponderada del numero caminos cerrados de diferentes longitudes en toda la red que comienzan y termina en ese nodo. Para su cálculo se descompone espectralmente la matriz de adyacencia A; como $A = \vee\Lambda\vee^T$, donde $\vee$ es la matriz de vectores propios (con v_{ij} sus elementos) en columnas de A y Λ es la matriz diagonal de valores propio λ_j.

$$Cs(j_2) = \sum_{j_1=1}^{n} v_{ij}^2 e^{\lambda_j}. \tag{3.11}$$

Una vez calculadas estas 6 métricas, se pasa de tener para cada ventana de tiempo la matriz $X_{\{i,j\}_s}$, a tener por cada ventana de tiempo una matriz de m x k_i dimensiones, siendo m el número de métricas calculadas y k_i el número de nodos, para esta matriz a partir de ahora se usará la notación de matriz de métricas M_m.

3.1.2. K-medias

De esta forma, en cada ventana de tiempo, tendremos k_i observaciones de dimensión m siendo k_i el número de nodos del paciente y m el número de métricas (denotaremos estas observaciones como puntos $x_1, ..., x_{k_i}$). A estas observaciones se les aplica el método de k-medias y se obtiene una asignación como grupo cluster 1, 2 o 3. El método de k-medias [10] trata de resolver el problema de la identificación de grupos para unos puntos pertenecientes a un espacio multidimensional, se estudia en Advanced Stadistics, pero como yo no la cursé lo voy a explicar.

Se desea encontrar K grupos o clusters en los que se puedan agrupar los k_i puntos m-dimensionales de la matriz de métricas M_m, siendo K un número dado (en nuestro caso $K = 3$). Se tomarán μ_k vectores pertenecientes al espacio multidimensional como los centroides de estos K clusters y se desea que la distancia de cada punto x_j a su correspondiente centroide μ_k (con $k \in \{1, ..., K\}$) sea mínima en comparación con la distancia de este mismo punto al resto de centroides para todos los puntos x_j (con $j \in \{1, ..., k_i\}$).

Se va a tomar como distancia, la distancia euclídea. Se denotará para cada punto $x_j \in \{x_1, ..., x_{k_i}\}$ un indicador binario $r_{jk} \in \{0, 1\}$ donde $k \in \{1, ..., K\}$ indicando a que cluster se asigna el punto x_j, este indicador será 1 cuando este punto está asignado al cluster k y 0 en otro caso. Entonces la función objetivo a minimizar será:

$$F = \sum_{j=1}^{k_i}\sum_{k=1}^{K} r_{jk}\|x_j - \mu_k\|^2 \tag{3.12}$$

Para la minimización se va a seguir un proceso iterativo, que comienza con la elección aleatoria de k centros y la asignación de los puntos a cada centro que se encuentra más cerca. Tras esta primera asignación se recalculará el centro de tal forma que la distancia a todos los puntos asignados sea mínima, entonces se asignarán nuevamente los puntos al centro que tenga más cercano y así sucesivamente hasta que se de la convergencia. De esta forma, se irán variando los valores r_{jk} (asignación del punto x_j al cluster k) y μ_k (centro del cluster k) para que F sea mínima.

A continuación, en la Figura 3.3 se muestra un ejemplo de las asignaciones originales de k-medias para la ventana estudiada en las Figuras 3.1 y 3.2, así como su matriz de confusión, en unidades de porcentaje de acuerdo. Se ha utilizado un espacio de dimensión 2 para facilitar su representación, con dos métricas: *Fuerza* y *Centralidad del vector propio* (más favorables según la selección de métricas realizada con el Test de Kruskal-Wallis y el criterio FDR que se explicará en la sección 3.1.6). Se observa que puede que la división en grupos sea correcta pues hay zonas con porcentajes de acuerdo altas, pero no encajan las etiquetas del método k-medias (C1,C2 y C3), con los niveles en el contexto de la epilepsia 1, 2 o 3 de referencia, pues estos porcentajes no se encuentran en la diagonal de la matriz. Es por esto que para calcular la matriz de confusión se buscar corregir primero las etiquetas como se explica en la siguiente sección.

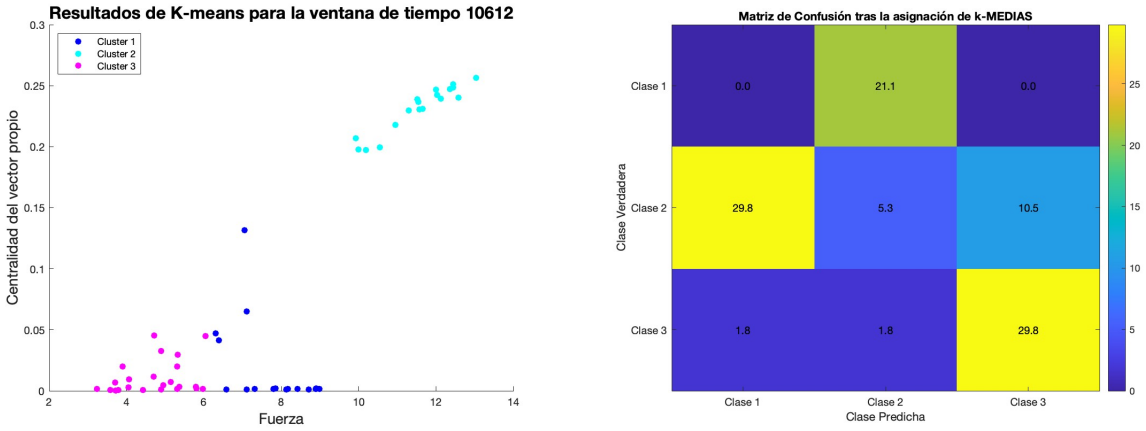

Figura 3.3: Representación de las asignaciones obtenidas del método k-medias, así como su matriz de confusión teniendo en cuenta los valores de referencia para el Paciente 1. Donde Clase verdadera hace referencia a los niveles de epilepsia N1,N2 y N3 y Clase predicha corresponde a los grupos formados por k-medias C1,C2 y C3.

3.1.3. Asignación de etiquetas

Si se compara la Figura 3.3 con la representación esperada, con los datos de y_1 (Figura 3.4), se observa claramente que la asignación realizada por k-medias como Cluster 1 encaja mejor con los nodos del Nivel 2 (PZN). Por lo que se necesitaría realizar un ajuste de etiquetas de forma que C1=N2, C2=N1; C3=N3.

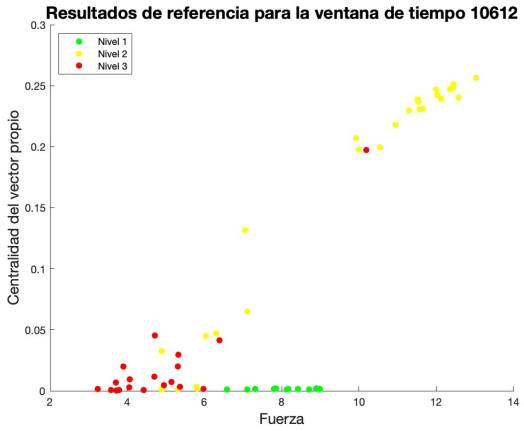

Figura 3.4: Representación de las asignaciones de referencia (y_1) para la misma ventana de tiempo que la Figura 3.3.

Para mejorar esta asignación, se analiza en cada ventana como encajan los resultados obtenidos con los esperados en cuanto a la asignación de grupos a los nodos. Partimos de la asignación Cluster 1, Cluster 2 o Cluster 3 ($Y_{\{i,j\}_s}$) en cada nodo $j \in \{1, ..., k_i\}$ para cada ventana de tiempo (obtenido por k-medias) así como los datos de referencia y_i con la asignación Nivel 1, Nivel 2 o Nivel 3 en cada nodo para cada paciente.

En primer lugar, se va a tratar de asignar una etiqueta Nivel 1, 2 o 3 a cada grupo cluster, para cada ventana de tiempo. Para esto, se calcula la matriz de confusión con las 6 posibilidades de etiquetas[2] (6, por ser la combinación de 3 valores en 3 posiciones), y se escoge la que obtenga mayor precisión ($accP$ número de nodos acertados frente al total definido por (2.2)).

Puede ocurrir que todas las posibles asignaciones tengan tasa de precisión 0, en ese caso las etiquetas se mantienen como se han obtenido en k-medias y la ventana se considerará como no significativa.

[2]Las posibilidades son: 1 2 3; 2 1 3; 3 2 1; 1 3 2; 2 3 1; 3 1 2.

Se muestra, a continuación en la Figura 3.5, las diferencias en las asignaciones para la misma ventana de tiempo que en la Figura 3.3 y como mejora la matriz de confusión.

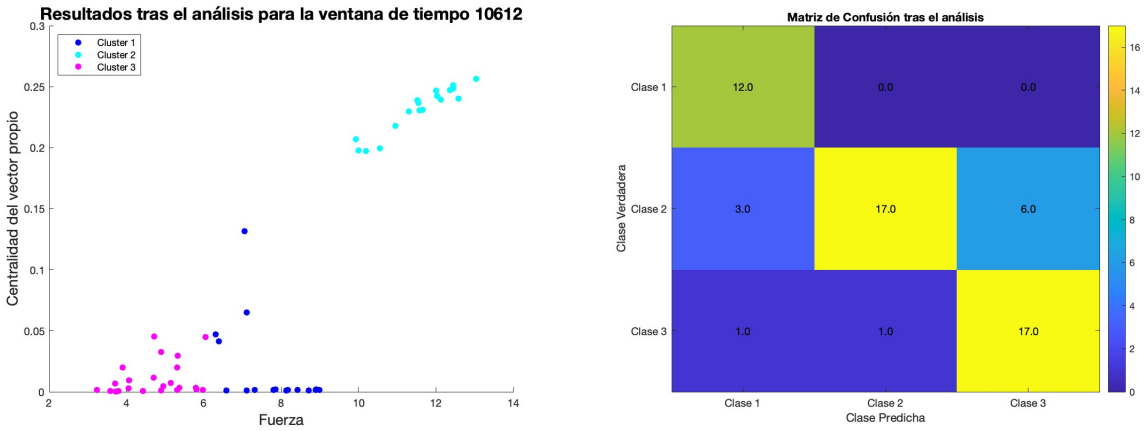

Figura 3.5: Representación de las asignaciones tras el ajuste de las asignaciones, así como su matriz de confusión teniendo en cuenta los valores de referencia para el Paciente 1.

De esta forma, se tiene un vector de asignaciones Y_i para cada ventana de tiempo con una tasa de precisión respecto a lo esperado ($accP$, porcentaje de nodos que coinciden en Y_i e y_i calculada a partir de (2.2)). Ahora para eliminar la posible aleatoreidad de este resultado se genera una sucesión de montecarlo de 1000 iteraciones. Para cada iteración r se genera un vector de asignaciones aleatorias, luego calcula se matriz de confusión y su tasa de precisión $accPR_r$.

De esta manera se tiene una distribución $accPR$ de 1000 precisiones con etiquetas asignadas aleatoriamente. Se calcula la probabilidad de que la tasa de precisión sea mayor que la aleatoria y se eliminan los índices no concluyentes, es decir, con esta probabilidad menor del 0.95. Para esto se obtiene el número de valores dentro de $accPR$ que son menores que la precisión $accP$ de esa ventana de tiempo, ese número dividido por el total de iteraciones (1000) tiene que ser mayor de 0.95 para que se considere que la ventana de tiempo tiene una tasa de precisión significativa. Si no es mayor de 0,95 se modifica $accP = 0$. Este umbral de significancia u se define como:

$$u = \frac{\sum_{r=1}^{1000} ac_r}{1000} > 0{,}95 \qquad \text{donde} \qquad ac_r = \begin{cases} 1, & \text{si } accP > accPR_r \\ 0, & \text{si } accP \leq accPR_r \end{cases} \tag{3.13}$$

Se obtiene entonces el número de indices (ventanas de tiempo) con $accP$ significativa y el porcentaje que representa dentro del número de indices totales.

Las variables que utilizaremos para comparar los resultados de los pacientes en el Capítulo 4 serán por un lado, el porcentaje de ventanas significativas, que como habrá un porcentaje por cada uno de los 30 casos de preprocesamiento se escogerá el máximo porcentaje dentro de los 30 casos, asociándolo a los parámetros de: con o sin regresión, número de ciclos por ventana y banda de frecuencia. Y por otro lado, la segunda variable será la tasa de precisión máxima dentro de ese procentaje de ventanas significativas, es decir, el porcentaje máximo de nodos acertados en una ventana de tiempo dentro del conjunto de todas las ventanas.

Habrá pacientes para los cuales el porcentaje de ventanas significativas sea bajo o que en varios casos de preprocesamiento tengan valores similares, es por eso que se va a tener en cuenta la corrección de Bonferroni, por la cual se considera que hay un 5 % de ventanas que pasarán el test estadístico de la simulación de montecarlo simplemente por casualidad. Por eso, no se considera diferencia significativa para escoger un caso de preprocesamiento concreto si la diferencia es menor de ese 5 %, para hacer esta diferenciación vamos a utilizar la segunda variable como desempate, por lo tanto, si dos casos son similares se mirará la máxima precisión dentro del conjunto de ventanas significativas y se escogerá la banda en la que el resultado sea mayor.

La correción de Bonferroni se explica de la siguiente manera, sea $H_1, ..., H_m$ una familia de hipótesis y $p_1, ..., p_m$ los p-valores que se desea considerar en correspondencia. Sea m el número total de hipótesis nulas y m_0 el número de hipótesis nulas verdaderas.

Si se considera un nivel de significancia $\alpha = 0{,}05$, existe un 5 % de probabilidad de rechazar la hipótesis nula siendo esta verdadera, lo que se considera error de tipo I (probabilidad de falso positivo), en un único test, esta probabilidad es baja, sin embargo al aumentar el número de test estos errores tienen un tamaño significativo, lo que supone un problema para las comparaciones múltiples [15]. De esta forma, si se realizan k comparaciones independientes, la probabilidad de que al menos ocurra un falso positivo (FWER) es:

$$FWER = 1 - (1 - \alpha)^k. \tag{3.14}$$

La corrección de Bonferroni rechaza la hipótesis nula para cada $p_i \leq \frac{\alpha}{m}$, controlando así el FWER en $\leq m$ por la desigualdad de Boole. [15]

3.1.4. Parámetros del modelo.

Durante este procedimiento se van a considerar varios parámetros que generaran diferencias en los resultados del modelo, con el objetivo de tratar de obtener un algoritmo de asignación concluyente. Estos parámetros afectarán a las fases de preprocesado (como ya se ha explicado, bandas de frecuencia, numero de ciclos y con o sin regresión), pero también al resto de fases que explicaremos en la siguiente sección de selección de métricas 3.1.6.

Utilizando los valores positivos o negativos de la matriz de correlación

Al estudiar varias ventanas de tasa de precisión ($accP$) máxima se observó que la matriz de correlación no solo tenía correlaciones positivas grandes entre los nodos, sino que para determinados caso existían también correlaciones negativas fuertes, como se mostrará en la sección de resultados 4.1 (Tabla 4.2). Es por esto, y para no perder información característica de los nodos, que se calculan las métricas de centralidad basándose en los valores positivos y negativos de la matriz de correlación. El uso de estas métricas, en un porcentaje alto de casos, mejoró los resultados.

Se muestra a continuación el ejemplo del Paciente 4, para la banda de frecuencia 0.5-3 Hz con 15 ciclos por ventana. En la Figura 3.6 se muestran los resultados obtenidos al escoger las dos ventanas de mayor tasa de precisión ($accP$ calculada a partir de (2.2)) y compararlas con su matriz de adyacencia positiva y su asignación esperada.

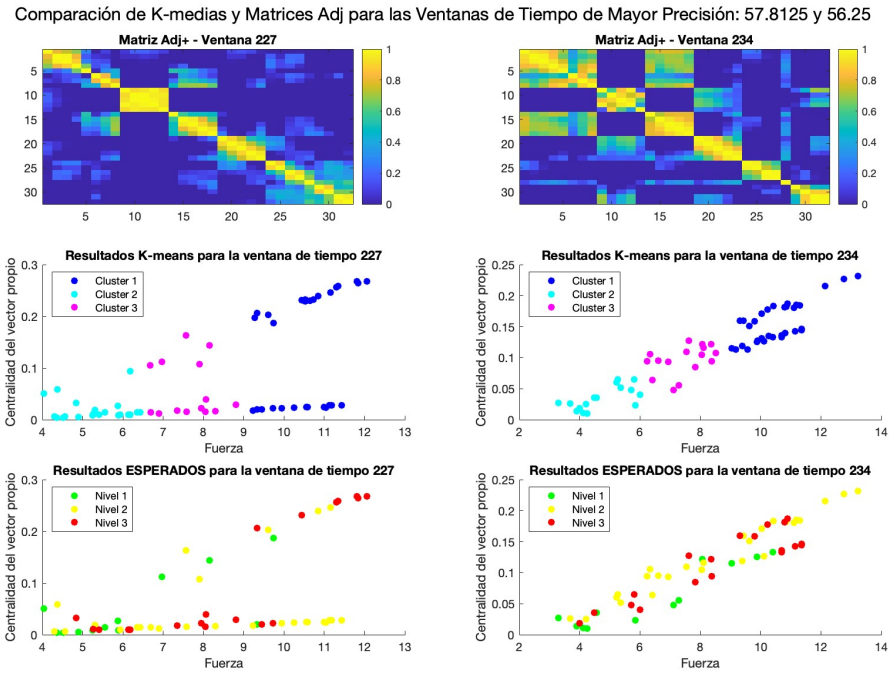

Figura 3.6: Matriz de Adyacencia positiva, asignación y asignación esperada para las dos ventanas de máxima tasa de precisión.

Ahora se repite el proceso a partir de la métricas calculadas con la matriz de adyacencia negativa. Los resultados se muestran en la Figura 3.7, se observa que la tasa de precisión ($accP$ calculada a partir de (2.2)) aumenta significativamente, por lo que se van a considerar ambos casos para todos los pacientes.

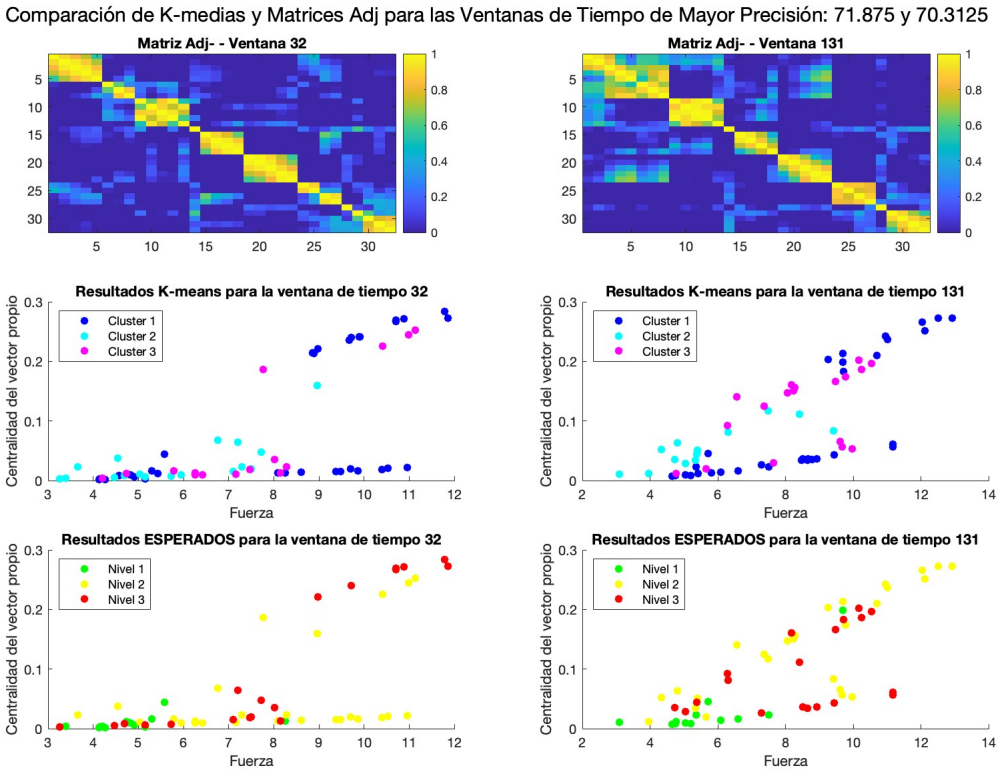

Figura 3.7: Matriz de Adyacencia negativa, asignación y asignación esperada para las dos ventanas de máxima tasa de precisión.

Test de hipótesis $\alpha = 0{,}05$ ó $\alpha = 0{,}01$

Al realizar el test de umbral a partir de la simulación de montecarlo (Ecuación (3.13)) se estipuló que para que una ventana pasara el test estadístico su precisión debe ser mayor que la aleatoria con una probabilidad del 95 %, en términos de la Ecuación (3.13), que $u > 0,95$. Tras múltiples análisis de resultados se observó que se podía conseguir en varios pacientes un porcentaje alto (entorno al 90 %) de ventanas que pasarán los test, sin embargo, la media de la precision a lo largo de estas ventanas ($\overline{accP}$),

$$\overline{accP} = \frac{\sum_{s=1}^{v_f} accP_s}{v_f}, \tag{3.15}$$

que pasaban rondaba en torno al (55 %). Es por esto que otro parámetro que se planteó es el de aumentar el umbral del test de hipótesis u, de forma se deba cumplir que $u > 0,99$, con el objetivo de que menos ventanas se consideren favorables pero que la precisión media aumente.

A continuación, en la Figura 3.8 se muestra la diferencia de resultados obtenidos para el Paciente 1, se observa que pese al cambio en el umbral la media de las tasas de precisión a lo largo de las ventanas de tiempo no aumenta de forma significativa, esto ocurre para los 11 pacientes por lo que se descarta el umbral $\alpha = 0{,}01$.

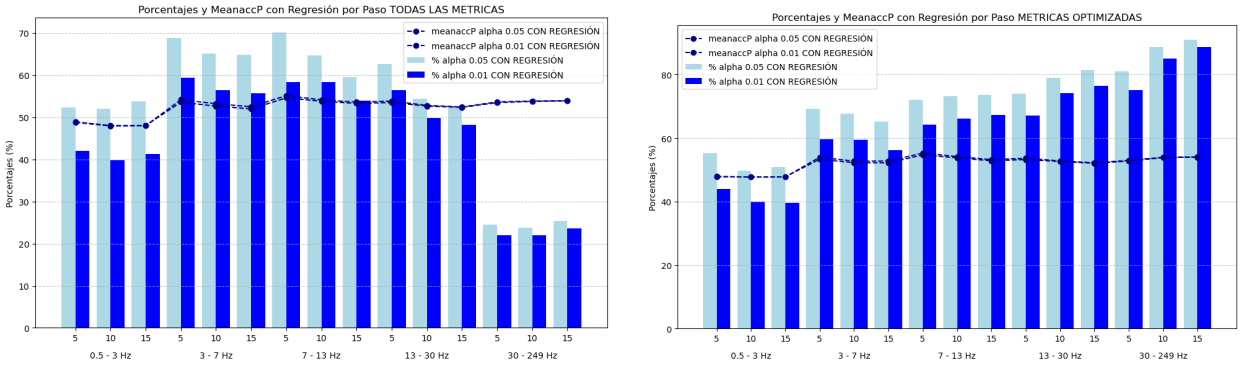

Figura 3.8: Representación del porcentaje de ventanas favorables con regresión asi como su precision media para el estudio con todas las metricas (izquierda) y para las metricas con p-valor menor de 0.05 (derecha), utilizando como limite de significancia $\alpha = 0{,}05$ y $\alpha = 0{,}01$.

3.1.5. Test Kruskal-Wallis

Una vez se realiza todo el proceso con el total de las métricas, se desea conocer si las utilizadas son favorables para este análisis cluster. Es decir, ver cuales de estas 12 métricas contienen información que pueda discriminar los nodos por sus diferencias en el campo de la epilepsia como zonas NIZ (Nivel 1), PZN (Nivel 2) o EZN (Nivel 3). El número es 12 porque tenemos 6 métricas, por un lado calculadas a partir de la parte positiva de la matriz de correlación y por otro con la parte negativa.

Para ver cual de estas 12 métricas contiene información significativa se utiliza como criterio de selección (*feature selection*) el Test de Kruskal-Wallis. Se trata de un test estadístico tipo ANOVA de factor clásico no paramétrico[3], que es especialmente útil para datos que tienen varios grupos de distribución. Su diferencia principal con un test de ANOVA de factor clásico es que el estadístico H es de tipo chi-cuadrado [16].

En una primera fase se calculan por cada ventana de tiempo las métricas, teniendo como resultado la matriz M_m con $m = 12$ donde cada métrica es un vector de k_i observaciones. El test de Kruskal-Wallis obtiene el p-valor para la hipótesis nula (H_0). Esta consiste en que las distribuciones de los grupos comparados a través del calculo de las métricas tienen funciones de distribución idénticas por lo que no sirven para distinguir entre las categorías nodales que indica el vector y_i. Es decir que siendo F_z la función de distribución de la métrica z con $z \in \{1, ..., m\}$:

$$H_0 : F_1 = ... = F_m. \tag{3.16}$$

Siendo la hipótesis alternativa (H_a) que al menos dos grupos tienen distribuciones diferentes por lo que las métricas si que sirven para distinguir entre categorías nodales.

$$H_a : \exists z_1, z_2 \in \{1, ..., m\} \qquad \text{tal que } F_{z_1} \neq F_{z_2}. \tag{3.17}$$

Para sacar el p-valor, en primer lugar, se calcula el estadístico H a partir de la Ecuación (3.18) [16], donde v_f es el número total de observaciones, K es el número total de grupos, $\mathcal{R}_i$ es la suma de los rangos de las observaciones para el grupo i y n_i es el número de observaciones en el grupo i.

$$H = \frac{12}{v_f(v_f+1)} \sum_{i=1}^{K} \frac{\mathcal{R}_j^2}{n_i} - 3(v_f + 1). \tag{3.18}$$

Proposición. Este estadístico H sigue una distribución chi-cuadrado con $K - 1$ grados de libertad bajo H_0 cuando el número de observaciones es elevado. [11]

Demostración. Para hacer esta demostración me baso en el libro [11] donde no se encuentra esta demostración completa sino que se desarrolla la Ecuación (3.18) a partir de una variable de distribución chi-cuadrado con $K - 1$ grados de libertad.

[3]Prueba estadística en la que no se asumen los supuestos paramétricos como que los datos provienen de distribuciones normales con varianzas homogéneas, por lo que se trabaja con medias, comparacion de distribuciones y medianas [11].

En primer lugar, destacamos que bajo la hipótesis nula los K grupos tienen la misma distribución, por eso la esperanza de la suma de los rangos del grupo i ($E(\mathcal{R}_i)$) va a ser n_i veces el rango esperado de una observación j. Sea $\mathcal{R}_{i,j}$ el rango de una observación j dentro del grupo i, su esperanza va a ser el promedio de todos los posibles rangos, desarrollando esto, junto con la fórmula de la suma de los v_f primeros naturales se tiene que:

$$E(\mathcal{R}_{i,j}) = \frac{1+2+...+v_f}{v_f} = \frac{v_f(v_f+1)}{2}\frac{1}{v_f} = \frac{v_f+1}{2}. \tag{3.19}$$

De esta forma, la esperanza de la suma de los rangos j dentro del grupo i es:

$$E(\mathcal{R}_i) = \sum_{j=1}^{n_i} E(\mathcal{R}_{i,j}) = n_j\frac{v_f+1}{2}. \tag{3.20}$$

Ahora se tiene en cuenta, que bajo la hipótesis nula, los rangos se asignan aleatoriamente a las observaciones, por lo que se puede utilizar que la $Var(\mathcal{R}_{i,j})$ es la varianza de una secuencia de enteros consecutivos de 1 a v_f.

$$Var(\mathcal{R}_{i,j}) = \frac{v_f^2-1}{12}. \tag{3.21}$$

Utilizando esto y que v_f es muy grande se tiene que: [11]

$$Var(\mathcal{R}_i) = E(\mathcal{R}_i^2) - E(\mathcal{R}_i)^2 \approx \frac{n_i(v_f+1)(v_f-n_i)}{12}. \tag{3.22}$$

Ahora consideramos la variable normalizada de los rangos $\mathcal{R}_i$, Z_i:

$$Z_i = \frac{\mathcal{R}_i - E(\mathcal{R}_i)}{\sqrt{Var(\mathcal{R}_i)}}. \tag{3.23}$$

Por el Teorema central del Límite, Z_i sigue una distribución normal de media 0 y desviación típica 1.

Si sustituimos $\mathcal{R}_i = Z_i\sqrt{Var(\mathcal{R}_i)} + E(\mathcal{R}_i)$ en la Ecuación (3.18) se tiene lo siguiente:

$$H = \frac{12}{v_f(v_f+1)}\sum_{i=1}^{K}\frac{\left(E(\mathcal{R}_i) + Z_i\sqrt{Var(\mathcal{R}_i)}\right)^2}{n_i} - 3(v_f+1). \tag{3.24}$$

Desarrollamos ahora término a término la expresión del sumatorio.

Primer término. $\sum_{i=1}^{K}\frac{E(\mathcal{R}_i)^2}{n_i} = \frac{(v_f+1)^2}{4}\sum_{i=1}^{K} n_i = \frac{(v_f+1)^2}{4}\cdot v_f$.

Segundo término. $\sum_{i=1}^{K}\frac{Z_i^2 Var(\mathcal{R}_i)}{n_i} = \sum_{i=1}^{K} Z_i^2\frac{(v_f-n_i)(v_f+1))}{12(v_f-1)}$.

Tercer Término. $\sum_{i=1}^{K}\frac{2E(\mathcal{R}_i)Z_i\sqrt{Var(\mathcal{R}_i)}}{n_i}$, por seguir Z_i una distribución normal de media 0 y desviación típica 1, la esperanza de este término es 0 y por el desarrollo obtenido en [11] el término se anula.

Sustituyendo esto en la Ecuación (3.24):

$$\begin{aligned} H &= \frac{12}{v_f(v_f+1)}\left(\frac{(v_f+1)^2}{4}\cdot v_f + \sum_{i=1}^{K} Z_i^2\frac{(v_f-n_i)(v_f+1)}{12(v_f-1)} + 0\right) - 3(v_f+1) = \\ &= 3(v_f+1) + \frac{1}{v_f-1}\sum_{i=1}^{K} Z_i^2(v_f-n_i) - 3(v_f+1) = \\ &= \frac{1}{v_f-1}\sum_{i=1}^{K} Z_i^2(v_f-n_i). \end{aligned} \tag{3.25}$$

Se debe tener en cuenta que el cuadrado de una variable normal estándar sigue una distribución chi-cuadrado con un grado de libertad [11]. En el caso de la Ecuación (3.25) se tiene una combinación lineal de K variables aleatorias independientes χ_1^2 con la restricción de que la suma total de los rangos es fija ($\frac{v_f(v_f+1)}{2}$), lo cual impone

una restricción sobre Z_i (reduciendo el número de grados de libertad a K-1). De forma que cuando se normaliza esta suma dividiéndola por $v_f - 1$, lo que se obtiene es una distribución χ^2_{K-1}.

Fin de la demostración.

Una vez calculado el estadístico H y sabiendo valor crítico de la distribución χ^2_{K-1} para $\alpha = 0{,}05$ y $K - 1 = 2$ grados de libertad, el p-valor será la probabilidad de que el valor crítico χ^2_{K-1} sea mayor que el estadístico H.

A partir de esto para cada paciente, se calcula un p-valor para cada métrica en cada ventana de tiempo. se considera que la métrica es significativamente favorable para este tipo de asignación cluster, cuando el p-valor es menor de 0.05. Como se obtiene un p-valor para cada ventana de tiempo, escogemos un criterio para seleccionar la métrica para el conjunto completo de las ventanas de tiempo de un determinado paciente con un preprocesado concreto (que consta de rango de frecuencias y número de ciclos), este criterio se explica a continuación.

3.1.6. Selección de métricas

En este punto del proceso, para cada paciente nos encontramos con un conjunto de v_f p-valores, siendo v_f en número de ventanas de cada caso (entendemos por caso, el preprocesamiento con la banda de frecuencia y número de ciclo concreto elegido), para cada una de las 12 métricas. Como el número de ventanas puede llegar hasta 14000 este test pasa a la categoría de test múltiple de gran escala, en este caso no se considera buena opción usar los criterios estándar basados en la corrección de errores de tipo I (ratio de falsos positivos) ya que la dimensión alta de las comparaciones los vuelve demasiado conservadores, impidiendo detectar diferencias reales. Es por esto que se escoge el criterio de control del FDR (*false discovery rate*) o proporción esperada de falsos positivos de entre todos los test considerados significativos) [17].

El objetivo de controlar el FDR es establecer un límite para el conjunto de test, de forma que la proporción de hipótesis nulas verdaderas no supere un determinado valor. Escogemos el método descrito por Benjamini & Hochberg, de forma que para considerar que una métrica aporta información sobre las asignaciones nodales seguimos el siguiente algoritmo. Sea v_f el número de ventanas y $\alpha \cdot 100$ el porcentaje que queremos que no supere el FDR se hace lo siguiente:

- Ordenar los v p-valores de menor a mayor. $(p_1, ..., p_{v_f})$.
- Calcular los umbrales $q_i = \frac{i}{v} \cdot \alpha$.
- Encontrar el valor k tal que $p_k \leq q_k$.
- Se consideran significativos para rechazar la hipótesis nula todos los p-valores hasta esa posición k $(p_1, ..., p_k)$.

En este caso se dice que el criterio de control del FDR al $(\alpha \cdot 100)\,\%$ considera significativos estos k p-valores. Con esta información debemos decidir si el número de p-valores significativos es lo suficientemente grande respecto del total para considerar la métrica como significativa para la optimización del método, escogemos un umbral del $\theta\,\%$ de p-valores que pasan el test para escoger la métrica donde θ se establece como 50, salvo para 4 excepciones dentro del conjunto de los 22 casos (11 pacientes con pre procesamiento incluyendo regresión y sin incluirla).

Estas 4 excepciones se dividen en dos. El primer caso resalta por la falta de p-valores seleccionados, este es el caso del Paciente 3 con regresión para el cual se obtiene un 0 % de p-valores seleccionados en todas las métricas salvo en la métrica *fuerza* en la que obtenemos un 31 %. Es por esto que seleccionamos esta métrica. También destaca el caso del Paciente 4 con regresión, para este caso todas las métricas obtienen alrededor de un 5-6 % de p-valores. Se considera que puede haber un 5 % de p-valores que han podido ser seleccionados por error, aleatoriamente, por lo que con este procesamiento seleccionamos las métricas que superan este rango.

La otra excepción es el caso contrario, en el que casi todas las métricas obtienen valores por encima del 90 % de p-valores seleccionados, por lo que con el umbral de $\theta = 50$, estaríamos seleccionando 10 métricas de 12, y podríamos caer en redundancia, pues muchas métricas están correlacionadas. Para el caso del Paciente 1 con regresión se seleccionan por encima de un 98 % de p-valores para 5 métricas calculadas con la parte positiva de la matriz de adyacencia; sin embargo, esas 5 métricas calculadas con la parte negativa de la matriz obtienen un 75 % de p-valores seleccionados, es por esto que se escogen las 5 con la parte positiva de la matriz. La otra excepción es el Paciente 9 con regresión para el cual 6 métricas rondan el 98 % y 5 el 90 %, escogemos las 6 primeras.

Una vez realizado este análisis de métricas se repite el proceso anterior en el total de las casuísticas anteriores pero solo con las métricas seleccionadas, lo que provoca una mejora en los resultados como comentaremos en el capítulo 4.

3.2. Segundo Procedimiento: Agrupamiento por medidas de centralidad sobre la conectividad funcional de los nodos.

Para el segundo procedimiento se van a calcular métricas que aportan información sobre la variabilidad de la conectividad funcional de los nodos (*links variability*) y de las propriedades que tienen estos como red (como la centralidad). La conectividad funcional de los nodos se define como la dependencia temporal de la actividad neuronal entre regiones cerebrales anatómicamente separadas [18], es por esto que el cálculo de la métrica va a tener en cuenta la relación entre todas las ventanas de tiempo procesadas en la fase inicial, a través de una matriz media de la correlación funcional de los nodos en todas las ventanas de tiempo y una desviación típica de la media para cada ventana en particular. Por tanto como resultado se obtienen k_i puntos m-dimensionales, siendo m el número de métricas que se utilizan en el procedimiento, para cada paciente i en conjunto.

Con estos puntos se repite el método seguido en el primer procedimiento aplicando el algoritmo de k-medias y posteriormente escogiendo las mejores etiquetas correspondientes a los niveles epileptogénicos, basándonos en las asignaciones esperadas y_i. En este caso se calculará únicamente la tasa de precisión ($accP$) (calculada a partir de (2.2)) para el único vector de asignaciones que se obtiene por paciente Y_i y se dará por concluyente si tiene una probabilidad mayor del 95 % de ser mayor que la precisión aleatoria fruto de 1000 simulaciones de montecarlo. Si $accP$ supera esta test se mantiene el valor obtenido, sino la tasa de precisión ($accP$) para este paciente i pasa a ser 0. Al igual que se hace para el primer procedimiento como se explica en las secciones 3.1.2 y 3.1.3.

En este caso, el estudio preeliminar de todas las métricas en conjunto no aporta ninguna información relevante debido a que el número total de métricas ($m = 55$) es demasiado grande, generando demasiado ruido entre los datos, de forma que el método de k-medias no es capaz de discernir entre la información debida a la hipótesis de los niveles epileptogénicos y la que no tiene que ver con esta hipótesis. Es por esto que se aplica, un criterio de selección de métricas.

Para que este proceso de selección sea más eficiente se va a realizar la selección para el conjunto de las métricas calculadas en todos los casos de preprocesamiento, es decir, se calculan estas 55 métricas para cada una de las 5 bandas de frecuencia, para cada uno de los 3 valores de ciclos escogidos, todo para el preprocesamiento con regresión, necesario para obtener resultados concluyentes en este tipo de métricas.

En primer lugar, se repetirá la primera parte del primer procedimiento (secciones 3.1.2, 3.1.3 y 3.1.5), adaptado al caso particular de 1 ventana de tiempo y 1 métrica, de forma que para cada paciente i y para cada métrica m dentro del conjunto total de 825, se tiene el vector de asignaciones, su precisión y el p-valor de la métrica utilizada.

La idea a continuación es repetir otra vez el método del primer procemiento pero contando con las métricas de forma acumulada, es decir, primero calculamos la tasa de precisión de la mejor (en el siguiente párrafo explicaré como se determinará que una métrica es mejor que otra), luego de la primera y la segunda y así sucesivamente, hasta encontrar la máxima tasa de precisión. Al ir añadiendo estas métricas se realizará un estudio de correlación para evitar añadir métricas que no aporten información nueva, de forma que se encontrará un subconjunto de m_f métricas favorables.

Ahora queda por definir a que criterio se considera que una métrica es la mejor. La primera parte será contando con las métricas ordenadas por su tasa de precisión individual, es decir, primero se repite el procedimiento con la que ha obtenido mejor precisión individualmente, luego con esa y con la de segunda mejor tasa de precisión, y así sucesivamente. Esto se denotará como Criterio 1.

El segundo criterio planteado, es coger como mejor métrica la que tiene menor p-valor, de forma que se vayan acumulando las métricas que tienen los p-valores más pequeños (Criterio 2).

De esta forma, se escogerá un conjunto de m_f métricas de un conjunto de 825 para cada paciente, y una tasa de precisión obtenida con este conjunto, que será máximo. El valor de m_f rondará las 4 métricas tras el estudio de la correlación, cosa esperada ya que muchas métricas son muy similares, como vemos en la siguiente sección. Como notación llamaremos al vector resultante del cálculo de la métrica i sobre el conjunto de los datos como M_i.

3.2.1. Métricas

Se calculan por tanto 55 métricas para cada conjunto de datos procesados de la forma descrita en el Capítulo 2 $X_i = (X_{\{i,j\}_1}, ..., X_{\{i,j\}_{v_f}})$. Pese a que estas métricas tienen como resultado un solo vector de asignaciones Y_i por caso, los datos que son los procesados en ventanas de tiempo al igual que para el primer procedimiento. Esto se

debe a que el resultado final tiene en cuenta la relación entre los nodos relacionando todas las ventanas de tiempo como un dato en conjunto, no como datos independientes como se hace en el primer procedimiento.

En primer lugar se van a calcular parámetros acerca de las ventanas de tiempo para luego con combinaciones de los mismos obtener las 55 métricas.

Matriz de correlación Funcional (iFC). Para cada ventana de tiempo se calcula la correlación funcional instantánea utilizando la correlación de Pearson (3.1) de las variables X_i tras el preprocesamiento de los datos explicado en la sección 2.2, de forma que se obtiene un conjunto de v_f matrices, siendo v_f el número total de ventanas para cada caso, $iFCS = \{iFC_1, ..., iFC_{v_f}\}$, cada una de estas iFC_s con $s \in \{1, ..., v_f\}$, tiene una dimensión de k_i x k_i [18]. A partir de esto se obtiene por un lado el promedio de esta correlación funcional a lo largo de todas las ventanas de tiempo ($avFC$) y la desviación de cada una de estas matrices iFC con respecto a esta media ($divFC_s$, manteniendo el subíndice s para hacer referencia a cada ventana de tiempo) a partir de las siguientes ecuaciones (3.26) y (3.27). avFC es de dimension k_i x k_i, de forma que cada coordenada de la matriz resultado del numerador se divide por el número de ventanas de tiempo y $divFC_s$ es de dimension k_i x k_i x v_f.

$$avFC = \frac{\sum_{s=1}^{v_f} iFC_s}{v_f}. \tag{3.26}$$

$$divFC_s = iFC_s - avFC. \tag{3.27}$$

Métricas de conectividad funcional (iBC, iEC, iCC). Se calculan 3 métricas en cada ventana de tiempo s para tratar de caracterizar la red funcional que forma el conjunto de los k_i nodos de cada paciente i. La primera *Centralidad de Intermediación* (en inglés: *Betweenness Centrality*, de ahí la notación BC), que identifica el número de veces que un nodo se comporta como enlace más corto entre dos nodos, se calcula en cada ventana de tiempo $s \in \{1, ..., v_f\}$ a partir de la Ecuación (3.28), (tal y como se calcula en el primer procedimiento (3.8)), utilizando una función del repositorio [14]. Entonces para cada nodo v:

$$iBC(v) = \sum_{i \neq v \neq j} \frac{\sigma_{ij}(v)}{\sigma_{ij}}. \tag{3.28}$$

Donde $\sigma_{ij}(v)$ es el número de caminos más cortos que empiezan en el nodo i y acaban en el nodo j y σ_{ij} es el número total de caminos de i a j.

A continuación, se calculará la *Centralidad del Vector Propio* (en inglés, *Eigenvector Centrality*, de ahí la notación EC) que representa la importancia de cada nodo dentro de la red, que se calcula utilizando la Ecuación (3.29), donde se toma el autovector asociado al máximo autovalor de la matriz de correlación Funcional iFC [14].

$$iFC \cdot iEC = \lambda_{max} \cdot iEC. \tag{3.29}$$

Y por último, el *Coeficiente de Clusterización* que mide la probabilidad de un nodo de formar grupos con los nodos adyacentes conectados, a partir de la Ecuación (3.30), donde V_j son los vecinos de cada nodo j y T_j es en numero de conexiones en las que participa el nodo j [14].

$$iCC = \frac{2T_j}{V_j(V_j - 1)}, \quad \text{donde } V_j = \sum_{i=1}^{n} iFC_{ij}, \quad \text{y } T_j = \frac{1}{2} \sum_{i=1}^{V_j} \sum_{j=1}^{V_j} iFC_{V_i, V_j}. \tag{3.30}$$

De la misma forma que para la matriz de correlación iFC se obtendrá una métrica por cada ventana de tiempo obteniendo al final un conjunto de v_f matrices que denotaremos como iBCS, iECS y iCCS. Utilizando las ecuaciones (3.26) y (3.27) con estos tres conjuntos de matrices se calcula la media de todas y la desviación de cada una con respecto a la media que denotamenos como avBC, avEC, avCC y divBC, divEC, divCC.

A continuación se describen las 55 métricas calculadas a partir de estos parámetros. Debido al gran volumen de métricas, cada una se denotará como *Métrica* m (M_m) con $m \in \{1, ..., 55\}$. Estas métricas, en su mayoría no son resultado de la consulta de ninguna referencia bibliográfica concreta sino que son variaciones de diferentes métricas de conectividad funcional que el co-director de este trabajo Paolo Bonifazi, basándose en su experiencia, consideró oportuno utilizar en un estudio preliminar a este trabajo, que no se puede citar al no estar publicado aún. Es por esto que el número de métricas es elevado, porque en principio, no se tiene una referencia de qué métricas se podrían

utilizar y el objetivo es encontrar aquellas variaciones que aportan información acerca de la hipótesis a comprobar de los niveles epileptogénicos.

Nota: antes de empezar el desarrollo de las ecuaciones, aclarar la notación utilizada. Como se va a partir de estructuras matriciales, para no complicar la notación se utilizará $divFC_{ijs}$ para referirnos a la matriz con i filas y j columnas correspondiente a la ventana s dentro del conjunto de las v_f matrices totales que forma $divFC$. Cuando se elimine el subíndice s, eso representará que se esta tomando la media de todas las v_f matrices normalmente por columnas.

Métricas 1, 2, 3 y 4. Las cuatro primeras métricas están relacionadas con el cálculo de la media de los valores de cada fila de la matriz avFC con distintas variantes, como que estos valores sean en valor absoluto (M_2), sean solo los positivos (M_3) o solo los negativos (M_4), se calcularán por las siguientes ecuaciones, donde k_i corresponde al número de nodos, k_i^+ es el conjunto dentro del total cuyo valor en la matriz $avFC$ en la fila i es positivo y k_i^- cuyo valor es negativo.

$$M_1 = \left(\{\frac{1}{k_i}\sum_{i=1}^{k_i} avFC_{i1}\}, ..., \{\frac{1}{k_i}\sum_{i=1}^{k_i} avFC_{ik_i}\} \right), \tag{3.31}$$

$$M_2 = \left(\{\frac{1}{k_i}\sum_{i=1}^{k_i} |avFC_{i1}|\}, ..., \{\frac{1}{k_i}\sum_{i=1}^{k_i} |avFC_{ik_i}|\} \right), \tag{3.32}$$

$$M_3 = \left(\{\frac{1}{k_i^+}\sum_{i=1}^{k_i^+} avFC_{i1}\}, ..., \{\frac{1}{k_i^+}\sum_{i=1}^{k_i^+} avFC_{ik_i}\} \right), \tag{3.33}$$

$$M_4 = \left(\{\frac{1}{k_i^-}\sum_{i=1}^{k_i^-} avFC_{i1}\}, ..., \{\frac{1}{k_i^-}\sum_{i=1}^{k_i^-} avFC_{ik_i}\} \right). \tag{3.34}$$

Métricas 5, 6 y 7. Estas métricas representan la media de la desviación estándar relativa, la desviación estándar relativa y la varianza relativa de la conectividad funcional. Esta información proviene de un análisis a lo largo de todas las ventanas de tiempo. Para la media se calcula el promedio de la desviación absoluta de las variables (divFC) por filas y se divide entre el promedio absoluto es decir la media por filas de la matriz avFC. Para la desviación estándar y la varianza lo que se hace es sustituir el facto promedio de la desviación absoluta de las variables (divFC) y se calcula la desviación estandar y la varianza de las matrices a los largo de todas las ventanas de tiempo. De esta forma, se busca obtener información acerca de variabilidad a lo largo de las ventanas de tiempo y la estabilidad de la iFC.

$$M_5 = \left(\left\{ \frac{\frac{1}{k_i}\sum_{i=1}^{k_i}\left(\frac{1}{v_f}\sum_{s=1}^{v_f}|divFC_{ijs}|\right)}{\frac{1}{k_i}\sum_{i=1}^{k_i}|avFC_{i1}|} \right\}, ..., \left\{ \frac{\frac{1}{k_i}\sum_{i=1}^{k_i}\left(\frac{1}{v_f}\sum_{s=1}^{v_f}|divFC_{ijs}|\right)}{\frac{1}{k_i}\sum_{i=1}^{k_i}|avFC_{ik_i}|} \right\} \right), \tag{3.35}$$

$$M_6 = \left(\left\{ \frac{\frac{1}{k_i}\sum_{i=1}^{k_i}\sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f}(divFC_{ijs} - divFC_{i1})^2}}{\frac{1}{k_i}\sum_{i=1}^{k_i}|avFC_{i1}|} \right\}, ..., \left\{ \frac{\frac{1}{k_i}\sum_{i=1}^{k_i}\sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f}(divFC_{ijs} - divFC_{ik_i})^2}}{\frac{1}{k_i}\sum_{i=1}^{k_i}|avFC_{ik_i}|} \right\} \right), \tag{3.36}$$

$$M_7 = \left(\left\{ \frac{\frac{1}{k_i}\sum_{i=1}^{k_i}\frac{1}{v_f}\sum_{s=1}^{v_f}(divFC_{ijs} - divFC_{i1})^2}{\frac{1}{k_i}\sum_{i=1}^{k_i}|avFC_{i1}|} \right\}, ..., \left\{ \frac{\frac{1}{k_i}\sum_{i=1}^{k_i}\frac{1}{v_f}\sum_{s=1}^{v_f}(divFC_{ijs} - divFC_{ik_i})^2}{\frac{1}{k_i}\sum_{i=1}^{k_i}|avFC_{ik_i}|} \right\} \right). \tag{3.37}$$

Métricas 8, 9 y 10. En este caso, repetimos lo anterior calculando en vez de las métricas en relativo, normalizando las mismas, es decir, se calcula la media de la desviación absoluto normaliza por nodo, la media de la desviación estándar normalizada por nodo y la media de la varianza normalizada por nodo. En este caso se busca conocer también la variabilidad de iFC, pero centrándose en cada nodo en particular.

$$M_8 = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \frac{\left(\frac{1}{v_f} \sum_{s=1}^{v_f} |divFC_{ijs}| \right)}{|avFC_{i1}|} \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \frac{\left(\frac{1}{v_f} \sum_{s=1}^{v_f} |divFC_{ijs}| \right)}{|avFC_{ik_i}|} \right\} \right), \tag{3.38}$$

$$M_9 = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \frac{\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ijs} - divFC_{i1} \right)^2}}{|avFC_{i1}|} \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \frac{\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ijs} - divFC_{ik_i} \right)^2}}{|avFC_{ik_i}|} \right\} \right), \tag{3.39}$$

$$M_{10} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \frac{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ijs} - divFC_{i1} \right)^2}{|avFC_{i1}|} \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \frac{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ijs} - divFC_{ik_i} \right)^2}{|avFC_{ik_i}|} \right\} \right). \tag{3.40}$$

Métricas 11, 12 y 13. Ahora, se busca tener en cuanta solo las ventanas de tiempo en las que medias calculadas en las tres métricas anteriores sea mayores que uno, este valor corresponde a un nodo cuya variabilidad temporal es significativa frente a la media, por lo que puede ayudar a identificar nodos que generen conexiones dinámicas.

$$M_{11} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{\frac{1}{v_f} |divFC_{i1s}|}{|avFC_{i1}|} > 1 \right) \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{\frac{1}{v_f} |divFC_{ik_is}|}{|avFC_{ik_i}|} > 1 \right) \right\} \right), \tag{3.41}$$

$$M_{12} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}} \right)^2}}{|avFC_{i1}|} > 1 \right) \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}} \right)^2}}{|avFC_{ik_i}|} > 1 \right) \right\} \right), \tag{3.42}$$

$$M_{13} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}} \right)^2}{|avFC_{i1}|} > 1 \right) \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}} \right)^2}{|avFC_{ik_i}|} > 1 \right) \right\} \right). \tag{3.43}$$

Métricas 14, 15 y 16. En estas tres métricas solo se va a tener en cuenta la desviación con respecto a la media de cada iFC_s midiento la media de estas desviaciones, la desviación estándar y la varianza.

$$M_{14} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{1}{v_f} \sum_{s=1}^{v_f} |divFC_{i1s}| \right) \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{1}{v_f} \sum_{s=1}^{v_f} |divFC_{ik_is}| \right) \right\} \right), \tag{3.44}$$

$$M_{15} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}} \right)^2} \right) \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}} \right)^2} \right) \right\} \right), \tag{3.45}$$

$$M_{16} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}} \right)^2 \right) \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left(\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}} \right)^2 \right) \right\} \right). \tag{3.46}$$

Métricas 17, 18, 19, 20, 21 y 22. La siguientes 6 métricas van calcularse a partir de la misma Ecuación (3.47), en que se obtiene la media de la matriz de adyacencia umbralizada para el coeficiente de variación ruidoso $U_{\{ij\}}$.

$$M_m = \frac{1}{k_i(k_i-1)} \sum_{i \neq j} U_{\{ij\}_m}, \quad \text{donde} \quad U_{\{ij\}_m} = \begin{cases} \text{CV}_{ij_m} & \text{si } \text{CV}_{ij_m} \geq \text{CV}_{[0{,}05 \times k_i^2]} \\ 0 & \text{si } \text{CV}_{ij_m} < \text{CV}_{[0{,}05 \times k_i^2]}. \end{cases} \tag{3.47}$$

La matriz de adyacencia umbralizada se obtendrá seleccionando el máximo 5 % de los valores del coeficiente de variación ruidoso CV_m. Este coeficiente de variación ruidoso se define como una medida de la variablidad relativa de los valores de centralidad en una red neuronal [18]. Esta matriz CV_m se puede calcular de diferentes maneras por lo que para cada una de las m métricas ($m \in \{17, ..., 22\}$) es diferente. Estas métricas sirven para identificar los enlaces entre nodos más variables (altos valores de CV), en el caso de las 3 primeras teniendo en cuenta los valores promedios; y para las otras 3 en términos absolutos sin tener en cuenta los promedios.

$$CV_{ij_{17}} = \frac{\frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{ijs}|}{|avFC_{ij}|}, \qquad CV_{ij_{18}} = \frac{\sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ijs} - \overline{divFC_{ij}}\right)^2}}{|avFC_{ij}|^2}, \tag{3.48}$$

$$CV_{ij_{19}} = \frac{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ijs} - \overline{divFC_{ij}}\right)^2}{|avFC_{ij}|^2}, \qquad CV_{ij_{20}} = \frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{ijs}|, \tag{3.49}$$

$$CV_{ij_{21}} = \sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ijs} - \overline{divFC_{ij}}\right)^2}, \qquad CV_{ij_{22}} = \frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ijs} - \overline{divFC_{ij}}\right)^2. \tag{3.50}$$

Métricas 23, 24 y 25. Se va a utilizar la Ecuación (3.8) para obtener la Centralidad de Intermediación de 3 variables. Es decir, en lugar de utilizar la matriz iFC para la función del repositorio [14], se van a utilizar otras 3 variantes. Para la métrica 23 se obtiene directamente para avFC, para la métrica 24 se utiliza la media de divFC sobre las ventanas de tiempo dividida entre el valor absoluto de avFC: ($\overline{divFC_{ij}}/avFC$), y para la métrica 25 utiliza solamente la media de divFC sobre las ventanas de tiempo: $\overline{divFC_{ij}}$.

Métricas 26, 27 y 28. Se va a utilizar la Ecuación (3.9) para obtener la métrica de la Centralidad del Vector Propio de 3 variables. Para la métrica 26, de nuevo, se obtiene directamente para avFC, para la métrica 26 se utiliza la media de divFC sobre las ventanas de tiempo dividida entre el valor absoluto de avFC, sin tener en cuenta los valores infinitos y para la métrica 27 esto mismo pero sin discernir entre valores infinitos o finitos.

Métricas 29, 30 y 31. Se va a calcular la métrica de *PageRank*, nombre proveniente del inglés, esta métrica mide la influencia que tienen los nodos en la red neuronal, basándose en la cantidad de enlaces entre ellos y la fuerza de estos enlaces. Para un nodo tener un valor alto en esta métrica significa que este nodo tiene más probabilidad de aparecer en un enlace entre nodos cualesquiera que se controla con un factor de amortiguamiento d. Para calcular estas métricas se va utilizar la Ecuación (3.51) donde I representa la matriz identidad, d va a ser igual a 0.85 (al ser el parámetro más habitual en este contexto [14]), D_{k_i} corresponde a la matriz diagonal con el valor $1/deg(i)$ en la posición (i, i), donde $deg(i)$ es el grado del nodo i, y $b = (1 - d)$, A representa la matriz de adyacencia que va a ser diferente para cada caso. En la métrica 29 será avFC, en la 30 la media de divFC sobre las ventanas de tiempo dividida entre el valor absoluto de avFC teniendo en cuenta solo los valores finitos (para así descartar cualquier posible valor obtenido de una división por números muy cercanos al 0) y para la 31 será solo la media de divFC sobre las ventanas de tiempo. [14]

$$PC = (I - d \cdot \mathbf{A} \cdot D_{k_i})^{-1} \cdot b. \tag{3.51}$$

Métricas 32, 33 y 34. Se va a calcular también la métrica de *Centralidad de subgrafos*, en inglés conocida como *Subgraph centrality*, que mide el número de caminos entre nodos que empiezan y acaban en el mismo nodo. Para calcularlo se parte de una matriz de adyacencia que denominaremos como A, se calcula su matriz diagonal de autovalores y su matriz de autovectores, a continuación se obtiene un vector con todos los autovalores, λ, y se elevan al cuadrado, denominándolo como V^2, y por último para obtener la centralidad de subgrafos que obtiene la

parte real de multiplicar V^2 por e^{λ}; es decir $SC = Re(V^2 e^{\lambda})$ [14]. Para la métrica 32, la matriz de adyacencia A será directamente $avFC$, para la métrica 33 A seguirá la Ecuación (3.52), y para la métrica 34, la Ecuación (3.53).

$$A_{33} = \frac{\frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{ijs}|}{|avFC_{ij}|}, \tag{3.52}$$

$$A_{34} = \frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{ijs}|. \tag{3.53}$$

Métricas 35, 36 y 37. Ahora se va a medir la media, la desviación estándar y la varianza de la desviación de la conectividad funcional $divFC$ en valor absoluto ponderada en todos los casos con la media $avFC$.

$$M_{35} = \left(\left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{i1s}| \right) \cdot avFC_{i1} \right| \right\}, ..., \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{ik_is}| \right) \cdot avFCik_i \right| \right\} \right), \tag{3.54}$$

$$\begin{aligned} M_{36} = \Bigg(& \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}}\right)^2} \cdot avFC_{i1} \right| \right\}, ..., \\ & \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}}\right)^2} \cdot avFC_{ik_i} \right| \right\} \Bigg), \end{aligned} \tag{3.55}$$

$$\begin{aligned} M_{37} = \Bigg(& \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}}\right)^2 \right) \cdot avFC_{i1} \right| \right\}, ..., \\ & \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}}\right)^2 \right) \cdot avFC_{ik_i} \right| \right\} \Bigg). \end{aligned} \tag{3.56}$$

Métricas 38, 39 y 40. Se repiten las tres métricas anteriores esta vez tomando el valor absoluto solo de la media, la desviación estándar y la varianza y luego ponderándolo multiplicándolo por $avFC$.

$$M_{38} = \left(\left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{i1s}| \right) \right| \cdot avFC_{i1} \right\}, ..., \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f}\sum_{s=1}^{v_f} |divFC_{ik_is}| \right) \right| \cdot avFC_{ik_i} \right\} \right), \tag{3.57}$$

$$\begin{aligned} M_{39} = \Bigg(& \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}}\right)^2} \right| \cdot avFC_{i1} \right\}, ..., \\ & \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \sqrt{\frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}}\right)^2} \right| \cdot avFC_{ik_i} \right\} \Bigg), \end{aligned} \tag{3.58}$$

$$\begin{aligned} M_{40} = \Bigg(& \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}}\right)^2 \right| \cdot avFC_{i1} \right\}, ..., \\ & \left\{ \frac{1}{k_i}\sum_{i=1}^{k_i} \left| \frac{1}{v_f}\sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}}\right)^2 \right| \cdot avFC_{ik_i} \right\} \Bigg). \end{aligned} \tag{3.59}$$

Métricas 41, 42 y 43. Para estas tres métricas se continua con lo anterior pero ponderando en este caso con la media sobre las filas de $avFC$.

$$M_{41} = \left(\left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f} \sum_{s=1}^{v_f} |divFC_{i1s}| \right) \right| \cdot \sum_{i=1}^{k_i} |avFC_{i1}| \right\}, ..., \left\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f} \sum_{s=1}^{v_f} |divFC_{ik_is}| \right) \right| \cdot \sum_{i=1}^{k_i} |avFC_{ik_i}| \right\} \right), \tag{3.60}$$

$$\begin{aligned} M_{42} = \Bigg(\Bigg\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left| \left(\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}} \right)^2} \right) \right| \cdot \sum_{i=1}^{k_i} |avFC_{i1}| \Bigg\}, ..., \\ \Bigg\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left| \left(\sqrt{\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}} \right)^2} \right) \right| \cdot \sum_{i=1}^{k_i} |avFC_{ik_i}| \Bigg\} \Bigg), \end{aligned} \tag{3.61}$$

$$\begin{aligned} M_{43} = \Bigg(\Bigg\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{i1s} - \overline{divFC_{ij}} \right)^2 \right) \right| \cdot \sum_{i=1}^{k_i} |avFC_{i1}| \Bigg\}, ..., \\ \Bigg\{ \frac{1}{k_i} \sum_{i=1}^{k_i} \left| \left(\frac{1}{v_f} \sum_{s=1}^{v_f} \left(divFC_{ik_is} - \overline{divFC_{ij}} \right)^2 \right) \right| \cdot \sum_{i=1}^{k_i} |avFC_{ik_i}| \Bigg\} \Bigg). \end{aligned} \tag{3.62}$$

Métricas 44, 45 y 46. A continuación, se calculan la media de los valores de cada fila de la matrices avBC, avCC y avEC utilizando las siguientes ecuaciones donde k_i corresponde al número de nodos de Paciente i.

$$M_{44} = \left(\{ \frac{1}{k_i} \sum_{i=1}^{k_i} avBC_{i1} \}, ..., \{ \frac{1}{k_i} \sum_{i=1}^{k_i} avBC_{ik_i} \} \right), \tag{3.63}$$

$$M_{45} = \left(\{ \frac{1}{k_i} \sum_{i=1}^{k_i} avCC_{i1} \}, ..., \{ \frac{1}{k_i} \sum_{i=1}^{k_i} avCC_{ik_i} \} \right), \tag{3.64}$$

$$M_{46} = \left(\{ \frac{1}{k_i} \sum_{i=1}^{k_i} avEC_{i1} \}, ..., \{ \frac{1}{k_i} \sum_{i=1}^{k_i} avEC_{ik_i} \} \right). \tag{3.65}$$

Métricas 47, 48, 49, 50, 51, 52, 53, 54 y 55. Para finalizar para las 3 variables $divBC$, $divEC$ y $divCC$ se va a obtener la media, la desviación estándar y la varianza.

$$M_{47} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} divBC_{i1s} \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} divBC_{ik_is} \right) \right\} \right), \tag{3.66}$$

$$M_{48} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} divEC_{i1s} \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} divEC_{ik_is} \right) \right\} \right), \tag{3.67}$$

$$M_{49} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} divCC_{i1s} \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} divCC_{ik_is} \right) \right\} \right), \tag{3.68}$$

$$M_{50} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sqrt{\sum_{s=1}^{v_f} (divBC_{i1s} - \overline{divBC_{ij}})^2} \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sqrt{\sum_{s=1}^{v_f} (divBC_{ik_is} - \overline{divBC_{ij}})^2} \right) \right\} \right), \tag{3.69}$$

$$M_{51} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sqrt{\sum_{s=1}^{v_f} (divEC_{i1s} - \overline{divBC_{ij}})^2} \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sqrt{\sum_{s=1}^{v_f} (divEC_{ik_is} - \overline{divBC_{ij}})^2} \right) \right\} \right) \tag{3.70}$$

$$M_{52} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sqrt{\sum_{s=1}^{v_f} (divCC_{i1s} - \overline{divBC_{ij}})^2} \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sqrt{\sum_{s=1}^{v_f} (divBC_{ik_is} - \overline{divCC_{ij}})^2} \right) \right\} \right), \tag{3.71}$$

$$M_{53} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} (divBC_{i1s} - \overline{divBC_{ij}})^2 \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} (divBC_{ik_is} - \overline{divBC_{ij}})^2 \right) \right\} \right), \tag{3.72}$$

$$M_{54} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} (divEC_{i1s} - \overline{divBC_{ij}})^2 \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} (divEC_{ik_is} - \overline{divBC_{ij}})^2 \right) \right\} \right), \tag{3.73}$$

$$M_{55} = \left(\left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} (divCC_{i1s} - \overline{divBC_{ij}})^2 \right) \right\}, ..., \left\{ \sum_{i=1}^{k_i} \left(\sum_{s=1}^{v_f} (divBC_{ik_is} - \overline{divCC_{ij}})^2 \right) \right\} \right). \tag{3.74}$$

Siguiendo las ecuaciones anteriores, como ya se ha dicho, se calcularán estas 55 métricas para cada paciente i, en cada uno de los casos de preprocesamiento de datos. La manera de ordenar estos vectores ha sido por banda de frecuencia, es decir, las primeras 55 columnas de la matriz corresponden a la banda de frecuencia 0.5-3 Hz para 5 ciclos, las siguientes 55 para la misma frecuencia, pero con 10 ciclos y así sucesivamente, de forma que se obtiene un conjunto de 825 vectores de dimensión k_i.

3.2.2. Kruskal Wallis, K-medias y análisis

Una vez se tiene este conjunto de métricas, se va a seguir la misma metodología que para el primer procedimiento explicado en las secciones 3.1.2, 3.1.3 y 3.1.5; pero con pequeñas variaciones que se irán detallando a continuación.

En primer lugar, se realiza el Test de Kruskal-Wallis para este conjunto de 825 casos, de nuevo se trata de comprobar la hipótesis de que las métricas contienen información acerca de los grupos categóricos (y_i) que se encuentran en cada conjunto de nodos. Se obtiene un p-valor para cada vector de métricas (dentro del total de 825), a continuación se aplica el método de k-medias para cada una de las métricas por separado y se calcula la tasa de precisión del vector de asignaciones obtenido, tal y como se hace en el primer procedimiento pero esta vez usando las métricas de una en una.

A continuación, se ordenan las métricas a raíz de la tasa de precisión obtenida (criterio 1) y se seleccionan las 80 primeras. Se intuye que algunas de estas métricas son muy similares entre si, para confirmar esta hipótesis se calcula la matriz de correlación de Pearson para estas 80 métricas con mejor tasa de precisión como se muestra a continuación en la Figura 3.9.

En principio, se buscaría ir aplicando el algoritmo de k-medias utilizando la primera métrica, luego la primera y la segunda y así sucesivamente hasta utilizar las 80 (buscando en algún punto obtener una tasa de precisión máxima). Sin embargo, como se ve en la figura anterior, dentro de las 80 métricas con tasas de precisión más altas, hay varias con una correlación muy elevada, por lo que el siguiente paso es establecer un umbral de correlación α de forma que si dos métricas superan esta correlación se considera que la segunda no aporta información nueva para la asignación de nodos y no se añade.

El procedimiento para hacer esto es el siguiente, se realiza un bucle sobre las 80 métricas comentadas anteriormente, en la primera iteración se añade al nuevo vector de métricas Z, aquella que ha obtenido la mayor tasa de precisión, se normaliza de forma que:

$$Z_{norm_{\{i,j\}}} = \frac{Z_{\{i,j\}}}{\text{máx}_i \, X_{\{i,j\}}},$$

siendo i el número de nodos y j el número de ventanas, y se repite el algoritmo para obtener las asignaciones y su tasa de precisión con la métrica normalizada.

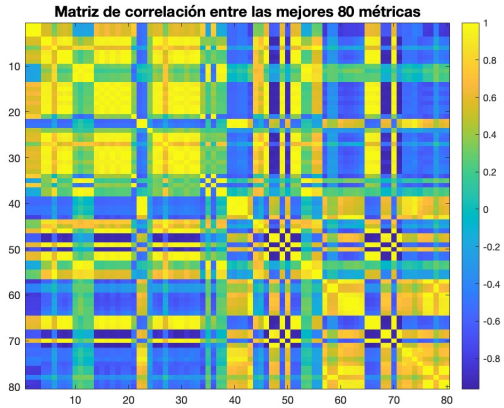

Figura 3.9: Matriz de correlación de Pearson entre las 80 métricas con precisión más alta para el Paciente 1.

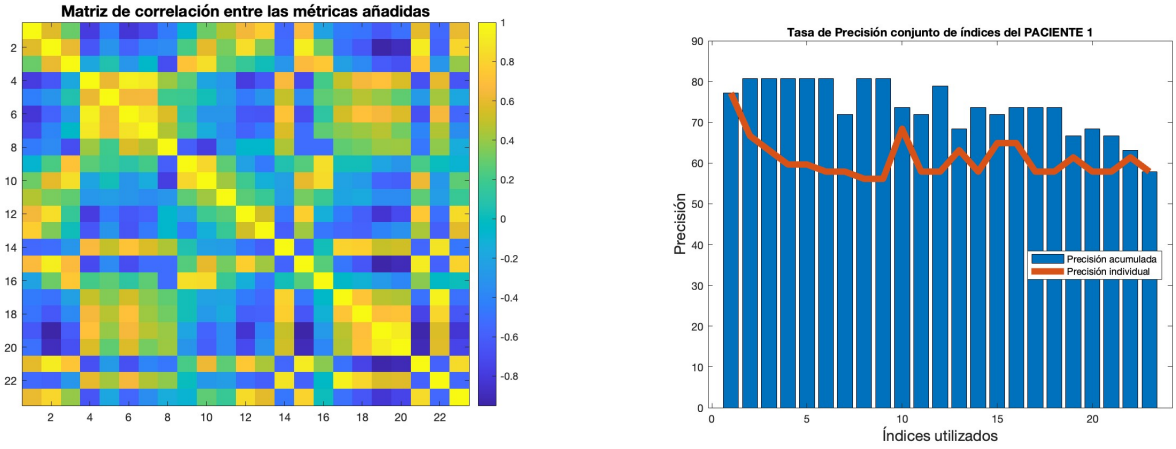

Figura 3.10: Matriz de correlación de las 23 métricas consideradas significativas, así como la representación de la tasa de precisión obtenida con la acumulación de estas métricas. Además en rojo se muestra la tasa de precisión que se obtiene con cada métrica por separado.

Luego para el resto de iteraciones (hasta 80) se seleccionan las métricas que no se han añadido a Z del conjunto total, se normalizan como se ha hecho en la primera iteración y se realiza un bucle de test para cada una de ellas. En la primera iteración, se crea una matriz Z_{test} con esta métrica, se establece el valor de máxima precisión ($accP_{max}$) como 0, se calcula la precisión y los coeficientes de correlación. Si la precisión es mayor que $accP_{max}$ y el valor máximo de correlación entre la nueva métrica y las anteriores es menor que el umbral $\alpha = 0{,}95$ la métrica se añade a Z y $accP_{max}$ pasa a ser igual a la tasa de precisión calculada. A continuación, la siguiente métrica se añade a Z_{test} y se repite el proceso, ahora con el valor de máxima precisión actualizado, si la nueva métrica de la iteración cumple el criterio de que la precisión es mayor que $accP_{max}$ y el valor máximo de correlación entre la nueva métrica y las anteriores es menor que el umbral $\alpha = 0{,}95$, se añade y se actualiza la $accP_{max}$ con la precisión nueva. Con este criterio nos aseguramos que solo se añaden las métricas si la precisión va en aumento.

Si al final de este bucle test ninguna métrica se ha añadido se terminan las iteraciones del bucle global, por lo que el número de iteraciones será el número de métricas que maximizan la precisión sin caer en redundancia. Si al final de este bucle test si que se han añadido métricas, se obtiene una matriz con todas las métricas seleccionadas, se normalizan y se va calculando las tasas de precisión acumuladas, y se pasa a la siguiente iteración.

Con este proceso lo que se consigue es obtener la mayor tasa de precisión posible minimizando la redundancia entre las métricas. A continuación se muestra lo obtenido para el Paciente 1 en la Figura 3.10 donde se muestra la matriz de correlación de las 26 métricas consideradas significativas, así como su tasa de precisión acumulada y en rojo la tasa de precisión de cada métrica por separado.

Criterio 2

Como se ha comentado en la introducción de la Sección 3.2, y con la idea de seguir la estructura del primer procedimiento, se podría utilizar el Test de Kruskal Wallis [16], como criterio de selección de métricas. La hipótesis inicial que se planteó fue usar como primera métrica la de menor p-valor, e ir añadiendo las demás en orden ascendente, teniendo en cuenta el umbral de correlación del que hablamos en la sección anterior y teniendo como tope la última métrica con p-valor menor de 0.05.

Una vez se tienen los resultados de tasa de precisión con ambos criterios se obtuvo que para este criterio 2, en dos pacientes mejoraban los resultados, en dos empeoraban y en el resto se mantenían (tenían una variación de $\pm 5\,\%$) todo esto comparado con los resultados obtenidos por el criterio 1. Solo con esta información podría establecerse como un criterio que da resultados similares, sin embargo, se descarta, por los siguientes factores.

Para la mayoría de los casos en los que los resultados de precisión se mantienen en porcentaje, se observa que este se alcanza con la combinación de más de 13-14 métricas, cuando para el criterio 1 para obtener ese mismo resultado se usan 2-3 de promedio. Este es el caso del Paciente 6 por ejemplo, cuya comparación entre los resultados se muestra en la Figura 3.11, utilizando el criterio 1 (izquierda) y utilizando el criterio 2 (derecha) .

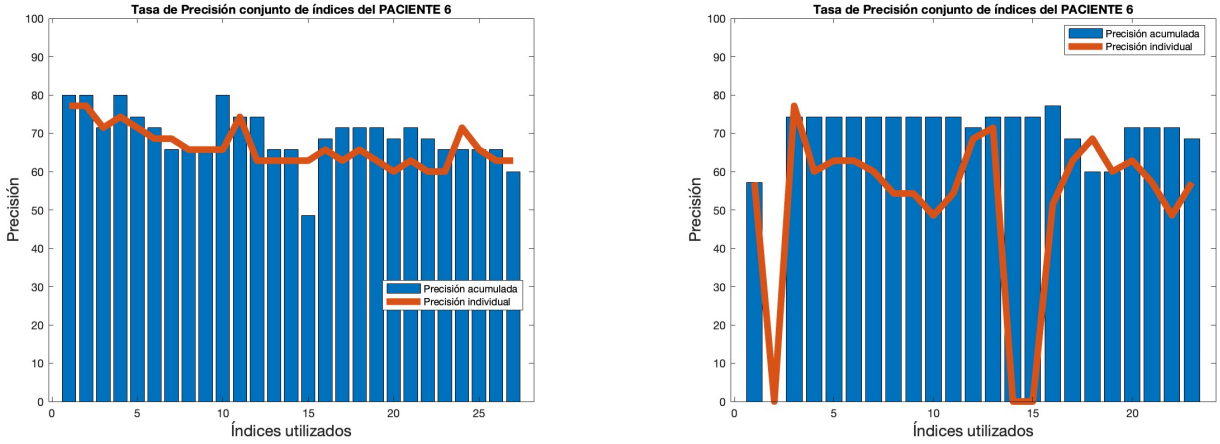

Figura 3.11: Resultados para el Paciente 6 a la izquierda utilizando el criterio 1 y a la derecha con el criterio 2. Los resultados son un $80\,\%$ de precisión para el criterio 1 en la primera métrica y un $77,5\,\%$ para el criterio 2, con la combinación de 16 métricas.

Además, hay otro factor que es la posibilidad de que la mejor opción sea una sola métrica y esta no tenga el p-valor más bajo. En este caso lo que ocurre es que la precisión individual máxima (línea roja) supera a la acumulada máxima (barra azul) en algún punto como se ve en la figura anterior de la derecha. Esto claramente demuestra que no es un criterio óptimo, pues en este caso lo más apropiado es utilizar esta métrica de precisión máxima y luego ver si añadiendo las demás esto mejora, es decir, justo lo que se hace cuando se utiliza el criterio 1.

De esta forma el criterio 2 queda descartado y para este segundo procedimiento no se utiliza el Test de Kruskal-Wallis para la selección de métricas.

Nota: se observa que para la Figura 3.11 de la izquierda en el primer valor la tasa de precisión individual no es igual a la acumulada, que en el caso de la primera métrica debería ser igual pues en ambos casos solo se usa una métrica. Esto es porque como se detalla en la sección anterior, primero se calcula la precisión para todas las métricas por separado y se guardan (esto es la línea roja de las figuras) y luego se realiza el bucle de ir calculando la precisión para cada métrica con todas las anteriores, en este caso para la primera iteración se repite el calculo. Entonces, debido a la aletoreidad de la inicialización de k-medias el resultado varía ligeramente.

Capítulo 4: Resultados obtenidos

4.1. Primer Procedimiento

En primer lugar, se ha realizado un análisis cluster o aprendizaje automático no supervisado mediante la técnica de k-medias, obteniéndose un vector de dimensión nodal con la asignación correspondiente a los tres niveles epileptogénicos (verde, amarillo o rojo) en cada ventana de tiempo, para los 11 pacientes (sección 3.1.2). Tras un análisis (sección 3.1.3) se obtiene el porcentaje de índices de ventanas de tiempo cuyos resultados tienen una tasa de precisión significante con respecto al vector esperado y_i y además no se deben a una asignación aleatoria favorable.

Se observa que debido a la inicialización aleatoria de k-medias los resultados no son exactamente iguales cuando se repite el proceso. Se tiene que el porcentaje tiene una desviación de $\pm 2\,\%$. Esto se debe tener en cuenta a la hora de valorar que bandas de frecuencia son más favorables en los casos en los que los valores no se diferencien en menos de un $2\,\%$. Además se tiene en cuenta la corrección de Bonferroni por la cual se va a considerar que el $5\,\%$ de las ventanas favorables son significantes por una asignación favorable aleatoria.

Una vez obtenidos los primeros resultados se escoge una banda de frecuencia a trabajar en cada paciente, la que presente un porcentaje de ventanas favorables mayor. Se realiza el test de Kruskal-Wallis con un control del FDR del $5\,\%$ para seleccionar las métricas que son más significativas en este análisis.

El procedimiento es realizar el test con las 6 métricas (*Fuerza, Fuerza al cuadrado, Centralidad de Intermediación de nodos, Centralidad del vector propio, Centralidad de Pagerank y Centralidad de subgrafo*) basadas en la parte positiva de la matriz de correlación y con las mismas 6 pero basadas en la parte negativa de la matriz de correlación. Se repite el análisis cluster, en todas las bandas de frecuencia pero solo para las métricas seleccionadas y se observan los resultados. Este procedimiento se realiza dos veces utilizando los datos preprocesados con y sin regresión.

A continuación, se presentan los resultados obtenidos para cada paciente. Los resultados se muestran en dos Figuras de la misma estructura para todos los pacientes, a la izquierda se representa el porcentaje de índices favorables para cada numero de ciclos en cada rango de frecuencia, realizando el preprocesamiento sin y con la función regresión. Y a la derecha lo mismo pero utilizando solo las métricas que han sido consideradas como favorables.

Se va a describir para cada paciente, que banda de frecuencia y número de ciclos se ha utilizado para realizar la selección de métricas tanto para el preprocesado con como sin regresión, que métricas se han seleccionado, y una vez se ha repetido el método con la selección de métricas, que banda de frecuencia y número de ciclos obtiene mejor resultado, como ya se ha comentado hay una porcentaje de margen del $5\,\%$ tanto por la aletoreidad de la inicialización de k-medias como por la correción de Bonferroni, de forma que en caso de diferencias menores de este margen entre dos bandas de frecuencia el criterio prioritario será ver en cual de las dos se obtiene la mayor tasa de precisión máxima dentro del conjunto de ventanas significativas.

Al final, se mostrará una tabla resumen con esta información para cada paciente así como con cual es el valor máximo de porcentaje de ventanas en el caso escogido y dentro de este porcentaje de ventanas cual es la tasa de precisión máxima que se obtiene. De forma que se puedan comparar todos los resultados, facilitando el sacar conclusiones.

Paciente 1

Para realizar el test de Kruskall Wallis y la posterior selección de métricas se utilizan 5 ciclos en la banda de frecuencia de 7-13 Hz para el caso con regresión y 15 ciclos para la banda de frecuencia de 13 a 30 Hz para el caso sin regresión. La selección de métricas a utilizar coincide en ambos casos y es *Fuerza*, *Fuerza al cuadrado* *Centralidad del vector propio*, *Centralidad de Pagerank* y *Centralidad de subgrafo* con la parte positiva de la matriz de correlación.

Al repetir el método utilizando solo las métricas más significativas vemos que los mejores resultados se obtienen para el caso con regresión para 15 ciclos en la banda de frecuencia de 30 a 249 Hz. Los resultados se muestran en la Figura 4.1.

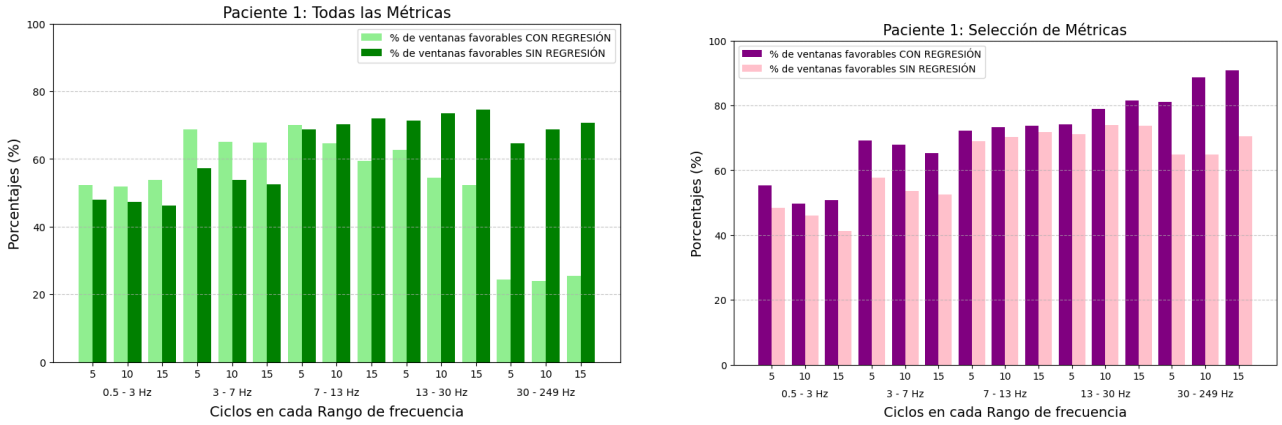

Figura 4.1: Resultados del primer procedimiento (C3VT) para el Paciente 1.

Paciente 2

Se han tenido en cuenta los datos procesados con 5 ciclos en la banda de frecuencia de 13 a 30 Hz con regresión y 15 ciclos en la misma banda de frecuencia sin regresión. La selección de métricas a utilizar ha sido *Fuerza*, *Fuerza al cuadrado* y *Centralidad de Pagerank* con la parte positiva de la matriz de correlación con regresión y solo *Fuerza* para el caso sin regresión.

Si se repite el método vemos que los resultados son mas favorables para el caso con regresión y destaca claramente que el porcentaje siempre mejora para el caso de los 15 ciclos, llegando a obtener resultados muy similares (ambos alrededor del 88 %) en las dos primeras bandas de frecuencia. Se escoge la banda de 3-7 Hz. Los resultados se muestran en la Figura 4.2.

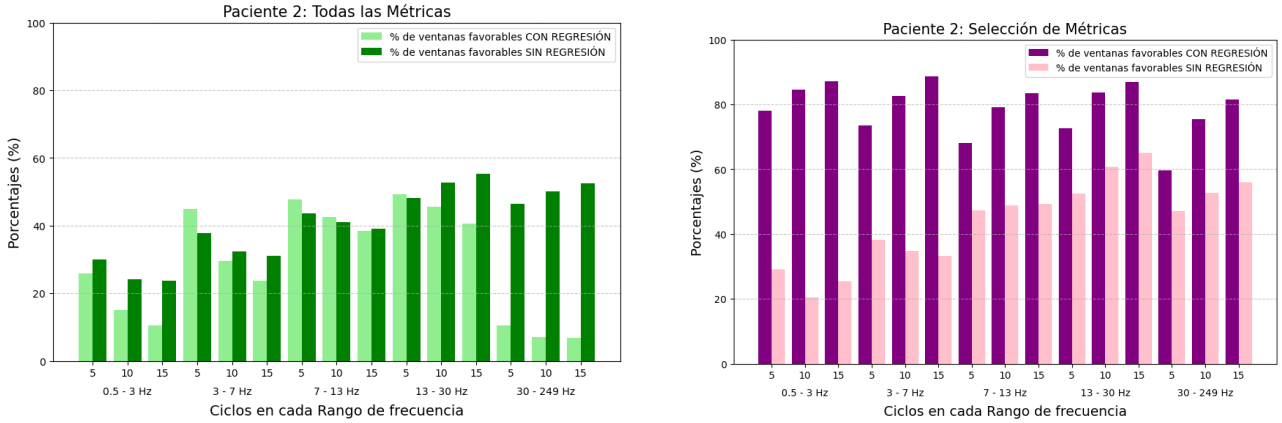

Figura 4.2: Resultados del primer procedimiento (C3VT) para el Paciente 2.

Paciente 3

Observamos que claramente para el caso sin regresión la mejor banda es la última con 15 ciclos, por lo que se utiliza esta para realizar la selección de métricas, donde se escogen *Fuerza*, *Fuerza al cuadrado* y *Centralidad de Pagerank* con la parte positiva de la matriz de correlación. En cambio, para el caso con regresión todas las bandas obtienen resultados muy bajos e igualados, se escoge la banda de 13-30 Hz con 15 ciclos que es ligeramente mayor, en este caso solo se selecciona la métrica de *Fuerza al cuadrado* con la parte positiva de la matriz de correlación.

Se repite el método solo para estas métricas y se observa que los mejores resultados se encuentran para el preprocesado con regresión en la banda de frecuencia de 3 a 7 Hz con 15 ciclos. Los resultados se muestran en la Figura 4.3.

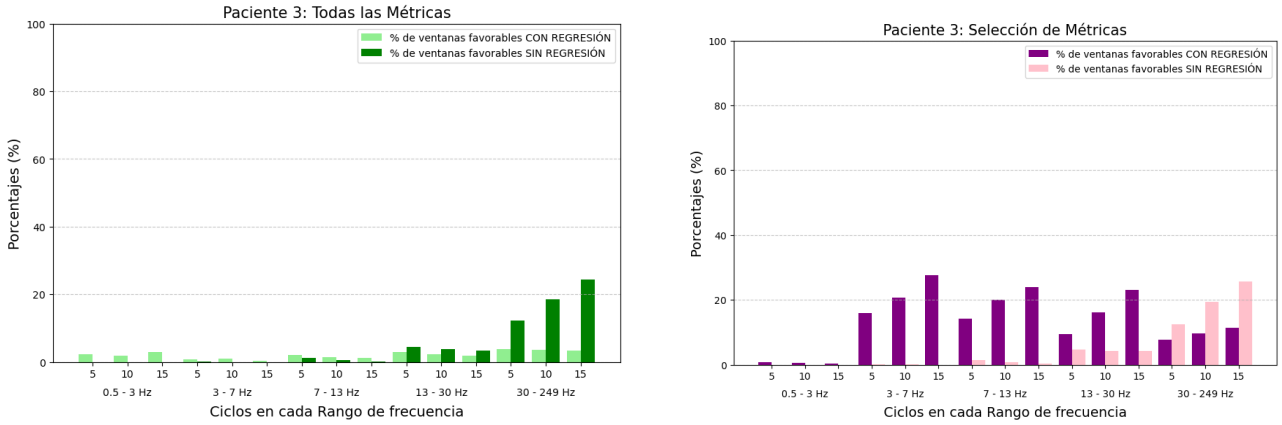

Figura 4.3: Resultados del primer procedimiento (C3VT) para el Paciente 3.

Paciente 4

Para este paciente el método original da resultados similares para todas las bandas de frecuencia, aunque destaca ligeramente el caso de 5 ciclos de 7-13 Hz para el caso con regresión y de 7 a 13 Hz con 15 ciclos sin regresión, así que son estos los que se utilizan para la selección de métricas. Para el caso con regresión se obtienen las métricas de *Fuerza*, *Fuerza al cuadrado*, *Centralidad de vector propio*, *Centralidad de Pagerank* y *Centralidad de subgrafo*, esta vez para la parte negativa de la matriz de correlación. Para el caso sin regresión como ya hemos comentado los porcentajes de p-valores significativos de cada métrica son muy bajos y similares rondando el $0-2\,\%$, pero destaca que todas las métricas con la parte positiva de la matriz obtienen 0 p-valores, en cambio las mencionadas para el caso de regresión rozan el $2\,\%$ por lo que se escogen las mismas.

Al repetir el método con esta selección de métricas, destaca claramente la primera banda de frecuencia con regresión que tiene resultados muy por encima del resto para todos los ciclos. Es por esto que se escoge esta primera banda con 15 ciclos con regresión. Los resultados se muestran en la Figura 4.4.

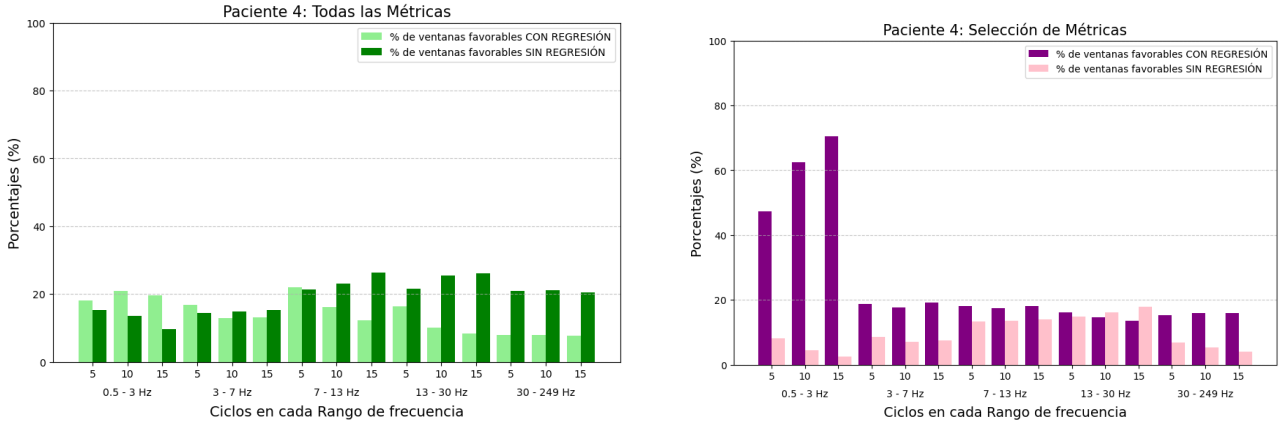

Figura 4.4: Resultados del primer procedimiento (C3VT) para el Paciente 4.

Paciente 5

Se han tenido en cuenta los datos procesados con 15 ciclos en la banda de frecuencia de 0.5 a 3 Hz para los datos con regresión, para los cuales se ha obtenido una sola métrica como significativa que es *Centralidad del vector propio* para la parte positiva de la matriz de correlación. Destaca que al utilizar la mejor banda (7-13 Hz con 10 ciclos) para el caso sin regresión lo que se obtiene de la selección de métricas es que destacan todas las métricas de la parte positiva de la matriz de correlación salvo *Centralidad de intermediación*.

Se repite el método para estas métricas y se observa que los mejores resultados se dan con regresión en la última banda de frecuencia con 15 ciclos. Los resultados se observan en la Figura 4.5.

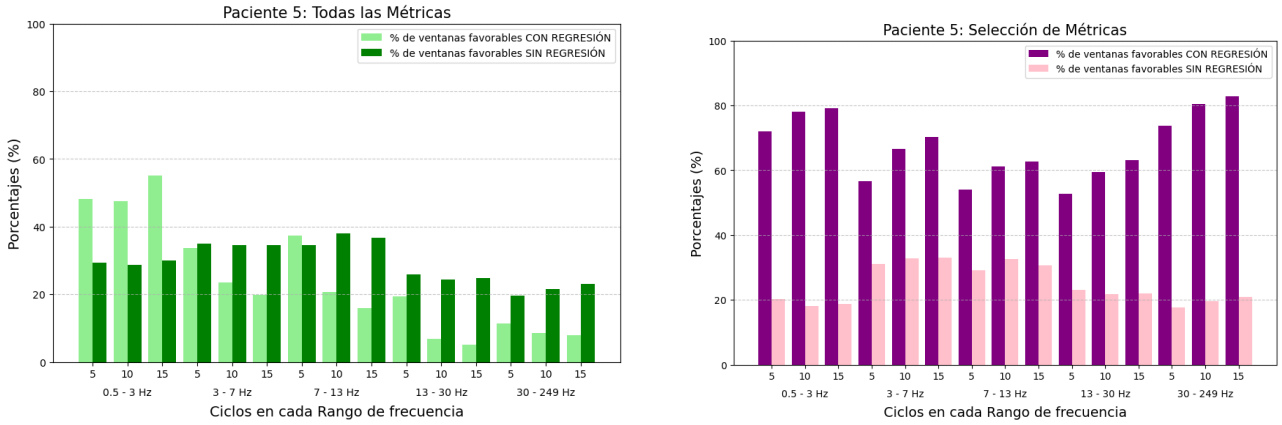

Figura 4.5: Resultados del primer procedimiento (C3VT) para el Paciente 5.

Paciente 6

Se han tenido en cuenta los datos procesados con 15 ciclos en la banda de frecuencia de 3 a 7 Hz para los datos sin regresión y la banda de 7 a 13 Hz con 5 cilos con regresión. Para ambos casos la selección de métricas es de *Fuerza*, *Fuerza al cuadrado*, *Centralidad del vector Propio*, *Centralidad de Pagerank* y *Centralidad de subgrafo* con la parte positiva de la matriz de correlación.

Al repetir el método con esta selección, vemos que los resultados son muy altos tanto para con como sin regresión, pero el caso con regresión es ligeramente mejor, por lo que es el escogido, para la banda de 7 a 13 Hz con un 92 % de ventanas favorables. Destaca al igual que para el paciente 2 y 5, que claramente los resultados mejoran en cualquier caso para 15 ciclos. Los resultados se muestran en la Figura 4.6.

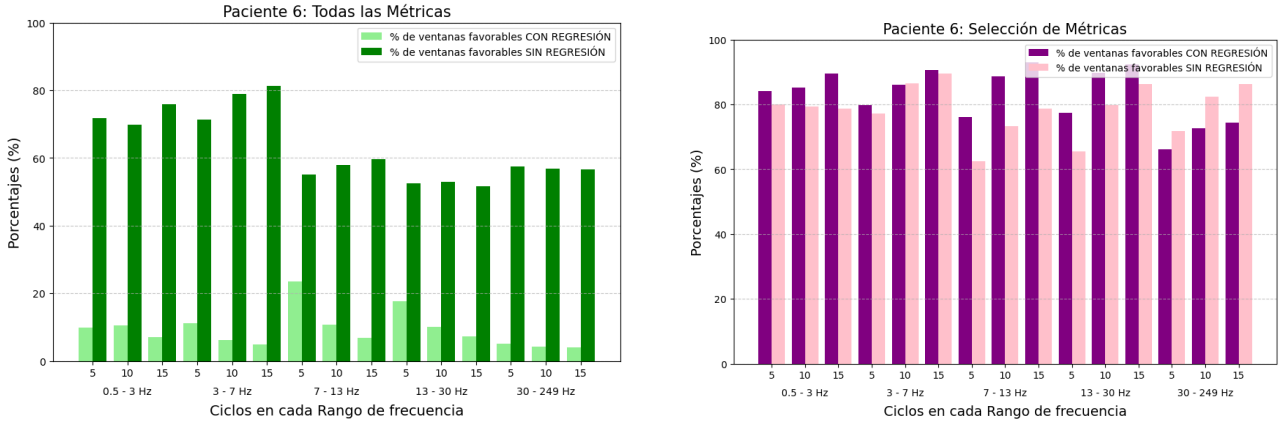

Figura 4.6: Resultados del primer procedimiento (C3VT) para el Paciente 6.

Paciente 7

Se han tenido en cuenta los datos procesados con 15 ciclos en la banda de frecuencia de 7 a 13 Hz y 15 ciclos con regresión y la banda de 13 a 30 Hz con 5 para los datos sin regresión. El test de Kruskal-Wallis indica que las métricas a utilizar son *Centralidad del vector propio* para la parte positiva de la matriz de correlación y para el caso sin regresión todas las métricas de la parte negativa de la matriz de correlación menos *Centralidad de intermediación*.

A la vista de lo obtenido se escoge la banda de frecuencia de 13 a 30 Hz con 15 ciclos para el preprocesado con regresión. Los resultados se muestran en la Figura 4.7.

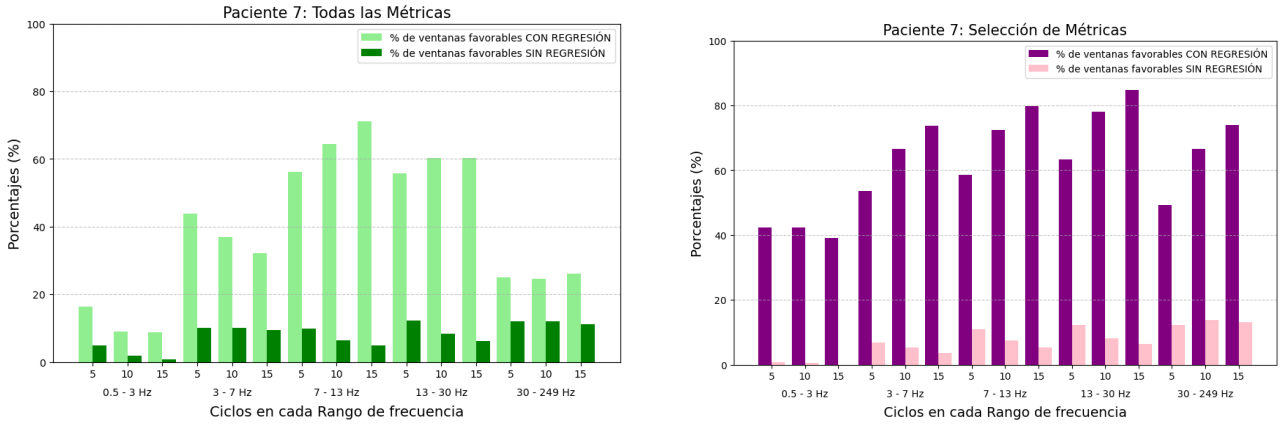

Figura 4.7: Resultados del primer procedimiento (C3VT) para el Paciente 7.

Paciente 8

Se han tenido en cuenta los datos procesados con 5 ciclos en la banda de frecuencia de 13 a 30 Hz para el preprocesado con regresión, donde se obtienen como favorables la métrica de *Centralidad del vector propio* con la parte positiva de la matriz de correlación. Las métricas que se obtienen en el caso sin regresión son *Centralidad del vector propio* y *Centralidad de subgrafo* pero con la parte negativa de la matriz de correlación, en la misma banda de frecuencia pero con 15 ciclos.

Se observa que el porcentaje de ventanas aumenta en todos los casos, destacando la última banda de frecuencia con regresión con 15 ciclos.Los resultados se muestran en la Figura 4.8.

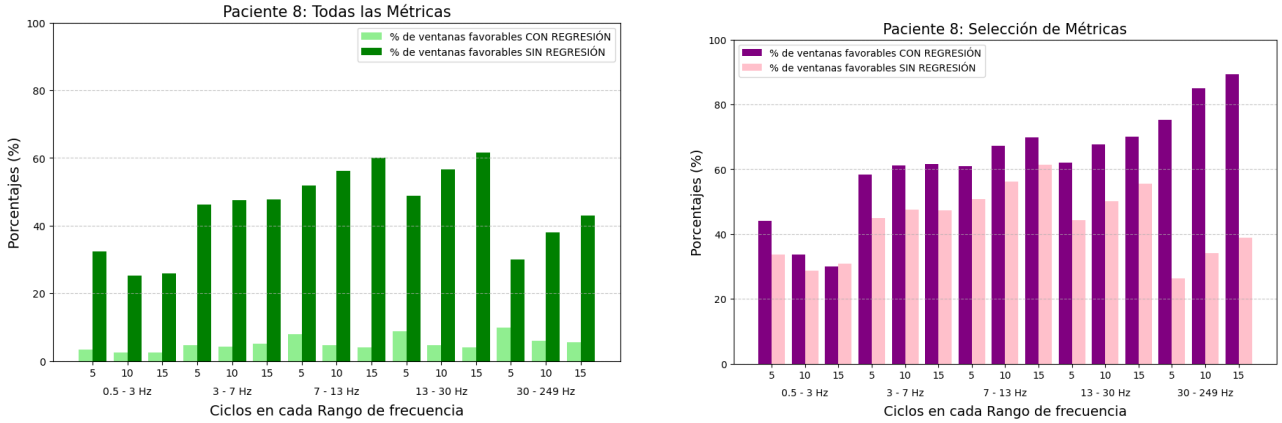

Figura 4.8: Resultados del primer procedimiento (C3VT) para el Paciente 8.

Paciente 9

Se han tenido en cuenta los datos procesados con regresión con 5 ciclos en la banda de frecuencia de 0.5 a 3 Hz y con los mismos ciclos pero en la ventana de 7-13 Hz en el caso de los datos preprocesados sin regresión. En ambos casos se obtiene un porcentaje de p-valores muy alto para todas las métricas, por lo que vamos a seleccionar solo las 5 que rondan el 98 % de p-valores que son *Fuerza*, *Fuerza al cuadrado*, *Centralidad de vector propio*, *Centralidad de subgrafo* y *Centralidad de Pagerank* para la parte positiva de la matriz.

Se observa que se llega hasta a un 97 % de ventanas con 15 ciclos en la banda de frecuencia de 0.5 a 3 Hz con regresión. Los resultados se muestran en la Figura 4.9.

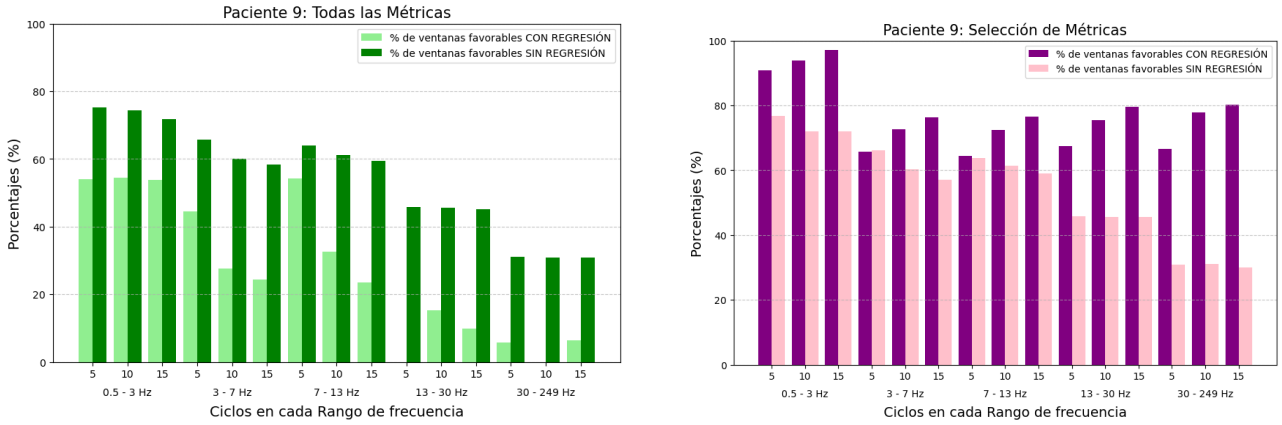

Figura 4.9: Resultados del primer procedimiento (C3VT) para el Paciente 9.

Paciente 10

Se han tenido en cuenta los datos procesados con 15 ciclos en la banda de frecuencia de 7 a 13 Hz tanto para el caso con como sin regresión. La selección de métricas es igual en ambos casos y son *Fuerza*, *Fuerza al cuadrado*, *Centralidad de vector propio*, *Centralidad de subgrafo* y *Centralidad de Pagerank* para la parte negativa de la matriz.

Destaca claramente que los resultados son mejores para la banda de frecuencia de 0.5 a 3 Hz, los porcentajes son muy similares pero destaca ligeramente el caso sin regresión para 10 ciclos. Los resultados se muestran en la Figura 4.10.

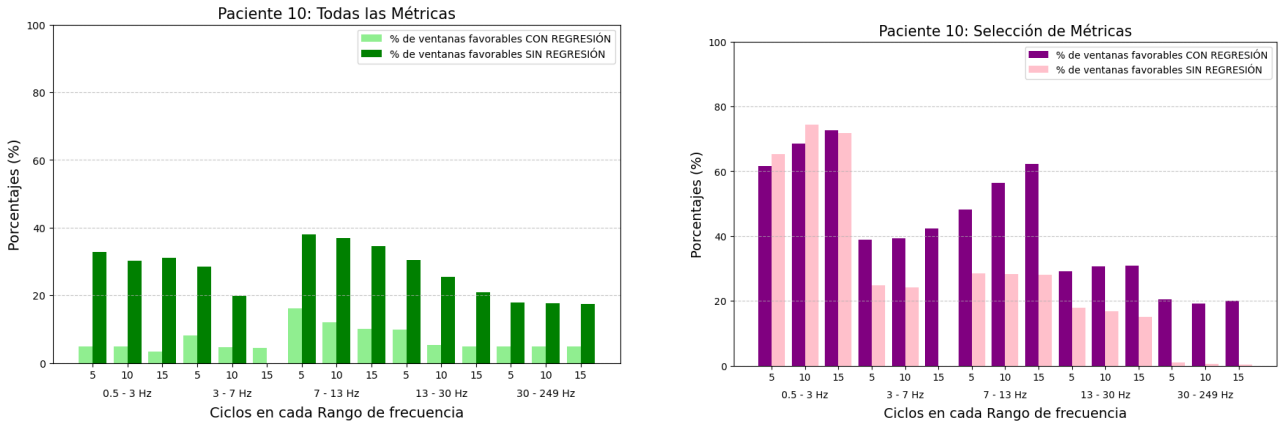

Figura 4.10: Resultados del primer procedimiento (C3VT) para el Paciente 10.

Paciente 11

Se han tenido en cuenta los datos procesados con regresión y 5 ciclos en la banda de frecuencia de 13 a 30 Hz y para 5 ciclos en la banda de 0.5 a 3 Hz para los datos sin regresión. Para los datos sin regresión la selección es de las métricas *Fuerza*, *Fuerza al cuadrado*, *Centralidad de vector propio*, *Centralidad de subgrafo* y *Centralidad de Pagerank* calculadas con la parte positiva de la matriz de correlación. Sin embargo, para los datos con regresión la selección se reduce a *Fuerza*, calculadas para la parte positiva de la matriz de correlación.

Se observa que para el caso con regresión los resultados son similares, sin embargo con regresión se obtiene hasta un 80 % de ventanas de tiempo siginificativas, para la banda de frecuencia de 30-249 Hz con 15 ciclos, aunque se escoge la banda con 10 ciclos ya se presenta una tasa de precisión máxima mayor. Los resultados se muestran en la Figura 4.11.

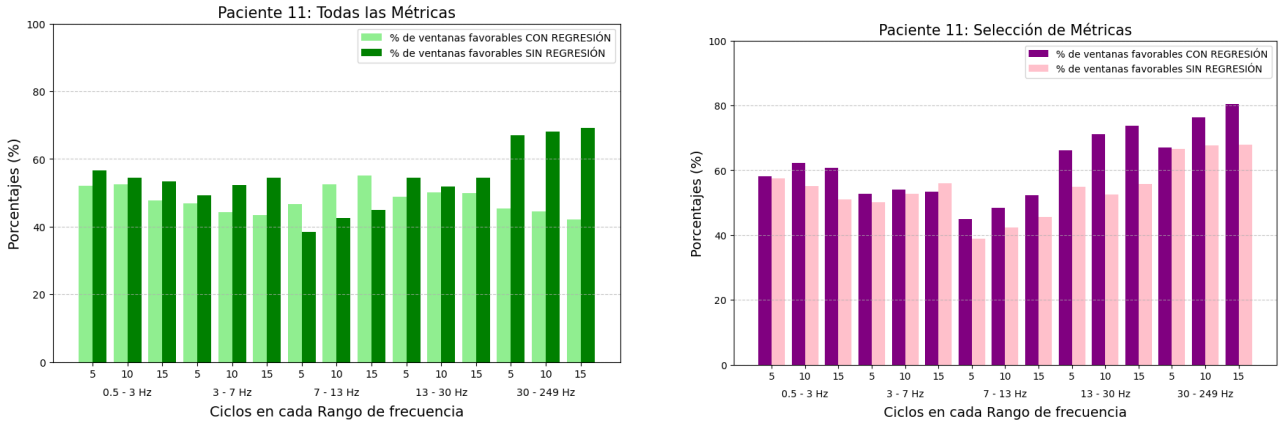

Figura 4.11: Resultados del primer procedimiento (C3VT) para el Paciente 11.

Resumen de Resultados

Ahora que se han descrito los resultados paciente por paciente se muestra la Tabla 4.2 como resumen de lo más relevante para sacar conclusiones. Uno de los parámetros que se incluye el parámetro de **Engel**, parámetro utilizado en el campo de la epilepsia para clasificar los cambios en los pacientes después de entre 12 y 24 meses del tratamiento quirúrgico basado en las asignaciones realizadas por los médicos y_i. Se representa en una escala de I al IV, de la siguiente forma. [19]

- Clase 1: Libre de crisis epilépticas incapacitantes.
- Clase 2: Crisis epilépticas poco frecuentes. (lo que podríamos identificar con casi libre de crisis.)
- Case 3: Mejora significativa.
- Clase 4: No hay mejora significativa.

Este parámetro ayuda a la interpretación de los resultados, aportando información a cerca de la fiabilidad de las asignaciones nodales que se tienen como referencia para cada Paciente i (y_i). De esta manera, se puede afirmar que los resultados para los pacientes 6 y 8 son muy favores, ya que rondan el 90 % de máxima precisión y tienen un porcentaje de ventanas significativas muy alto, así como un parámetro de Engel de I. El paciente 1 muestra resultados similares, sin embargo, su parámetro de Engel es de IV, lo que podría indicar que hay regiones epileptogénicas fuera de la zona monitorizada por los electrodos en estos pacientes en concreto.

Con este parámetro podremos dividir a los pacientes en dos grupos, por un lado el grupo de Engel I libre de crisis epilépticas y el grupo de Engel II, III y IV, para poder observar si existen diferencias en las distribuciones de los resultados, de forma que los mejores aparezca en los pacientes con Engel I o no.

Para realizar esta comparación se utilizará un test estadístico no paramétrico conocido como de Kolmogorov Smirnov para la comparación de dos distribuciones [11]. Para este test se tienen dos distribuciones basadas en observaciones independientes para cada población. Sean $X_1, X_2, ..., X_m$ y $Y_1, Y_2, ..., Y_n$ dos muestras independientes con funciones de distribución desconocidas F_X y G_Y. Sean $F_m(x)$ Y $G_n(x)$ las correspondientes funciones de distribución empíricas, queremos comprobar las siguientes hipótesis:

- $H_0 : F_X(x) = G_Y(x)\ \forall x$.
- $H_1 : F_X(x) \neq G_Y(x)$ para algún x.

Para comprobarlo se utiliza el siguiente estadístico $D_{m,n} = sup_x|F_m(x) - G_n(x)|$, el p-valor se calcula usando la distribución asintótica de $Kolmogorov - Smirnov$, y para este primer procedimiento, utilizando la variable de máxima tasa de precisión se obtiene lo siguiente, mostrado en la Tabla 4.1.

Proced. 1	**Grupo Engel I**	**Grupo Engel II, III, IV**
Media	84.41	71.93
Desviación típica	6,22	6.04
Estadístico KS	0.833	
p-valor	0.026	

Tabla 4.1: Resultados del test de Kolmogorov-Smirnov entre los grupos Engel I y Engel II, III y IV.

Estos resultados confirman que las distribuciones de las tasas de precisión son diferentes en ambos grupos, indicando mejores resultados para los casos del pacientes de Engel tipo I. Esto se considera un resultado muy favorable dentro de nuestro estudio, ya que Engel I indica que los médicos clasificaron y posicionaron el electrodo en la posición donde la epilepsia se genera; en el caso de Engel más bajo, la perdida de calidad de nuestros resultados es razonable porque los datos presentan mejor información acerca de las redes epilépticas y por eso los médicos tampoco han sido capaces de clasificar correctamente los electrodos.

Además de este parámetro, la Tabla 4.2 muestra la banda de frecuencia y número de ciclos con la que se obtiene el mayor número de ventanas significativas, vemos que las 5 bandas de frecuencia aparecen en algún momento, sin embargo, en casi todos los pacientes es mejor el uso de 15 ciclos por ventana, como ya se iba comentando. Respecto al preprocesamiento se observa que en la mayoría de los casos la optimización del método solo con las métricas favorables da mejores resultados en el caso de pre procesamiento con regresión.

Se muestran también las métricas seleccionadas para esta optimización y si han sido calculadas con la parte positiva o la parte negativa de la matriz de correlación, y cual ha sido el porcentaje de ventanas significativas para cada paciente. Además, se añade la información de cual ha sido la máxima precisión dentro de este conjunto de ventanas significativas, lo resultados son dispares pero llegan hasta una precisión del 92 % lo que se considera muy favorable.

Paciente	**nº ciclos**	**Banda f. (Hz)**	**Métricas**	**Con/Sin reg.**	**% vt fav.**	**max accP**	**Engel**
1	15	>30	Fuerza+, $Fuerza^2$+ C. vector propio+, C. Pagerank+, C. subgrafo+	CON	90,9	82,45	*IV*
2	15	3-7	Fuerza+, $Fuerza^2$+ C. Pagerank+	CON	87, 1	73,01	*III*
3	15	3-7	$Fuerza^2$+	CON	27,7	75	*IV*
4	15	0.5-3	Fuerza-,$Fuerza^2-$, C. vector propio-, C. Pagerank-, C. subgrafo-	CON	70,4	67,18	*II*
5	15	>30	C.Vector propio+	CON	82,8	78,57	*I*
6	15	7-13	Fuerza+,$Fuerza^2$+, C. vector propio+, C. Pagerank+, C. subgrafo+	CON	92,8	91,42	*I*
7	15	13-30	C. vector propio+	CON	84,7	63,38	*II*
8	15	>30	C. vector propio+	CON	89,3	90,62	*I*
9	15	0.5-3	Fuerza+, $Fuerza^2$+, C. vector propio+, C. Pagerank+, C. subgrafo+	CON	97,1	85,45	*I*
10	10	0.5-3	Fuerza-, $Fuerza^2-$, C. vector propio-, C. Pagerank-, C. subgrafo-	SIN	74,4	70,58	*IV*
11	10	>30	Fuerza+	CON	76,27	76	*I*

Tabla 4.2: Resumen de los resultados obtenidos para los 11 pacientes en el primer procedimiento, incluyendo la banda de frecuencias con el numero de ciclos por ventana en el que se obtiene el máximo porcentaje de ventanas significativas, así como las métricas utilizadas, si se han preprocesado los datos con o sin regresión y cual es la máxima tasa de precisión que se obtiene dentro del conjunto de ventanas de tiempo significativas. También se incluye en parámetro de Engel. Como notación se utiliza el signo '+' al lado de la métrica para indicar que esa métrica se ha calculado con la parte positiva de la matriz de correlación, y un '-' cuando se ha calculado con la parte negativa.

4.2. Segundo Procedimiento

Para este procedimiento se obtienen entre 20 y 50 vectores de asignaciones Y_i para cada Paciente i, estos vectores provienen de realizar el análisis cluster de k-medias y su posterior análisis estadístico de asignaciones a partir de las combinaciones de las mejores métricas que han sido seleccionadas como se describe en la metodología.

Se recuerda que las métricas que se escogen son sobre un total de 825 que corresponden con las 55 métricas definidas para este procedimiento, calculadas en cada uno de los 15 casos de bandas de frecuencia y ciclos de preprocesamiento de datos.

En la Figura 4.12 se muestran los resultados del Paciente 1, lo que se representa es la tasa de precisión obtenida con las métricas seleccionadas (línea roja) por separado (es decir, teniéndolas en cuenta de forma individual) y la tasa de precisión acumulada que se va obteniendo al utilizar cada métrica junto con todas las anteriores (barras de histograma en azul). Se observa que la máxima precisión del vector de asignaciones Y_1 frente a y_1 se obtiene alrededor de la segunda/tercera métrica y es del 80 % de los nodos correctamente clasificados. Dentro del conjunto de las 825 métricas estas dos con las que se obtiene la máxima precisión son la 1 con 5 ciclos y en la banda de 7 a 13 Hz, y la 41 con 10 ciclos y 30-249 Hz, como podemos ver en la figura de la derecha.

En ella se muestra un mapa con las 24 métricas que se han considerado en el procedimiento, para poder ver cuales son en términos de banda de frecuencia, número de ciclos e identificador dentro de las 55. De forma las primeras aparecen representadas en tonos azules como muestra la leyenda y las últimas en rojo.

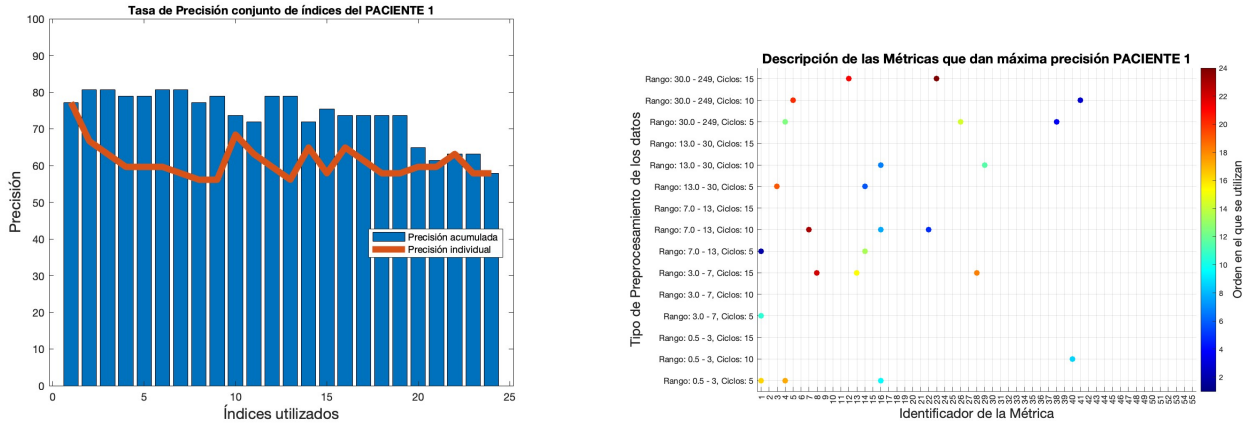

Figura 4.12: Se muestran las tasas de precisión acumuladas para el Paciente1, asi como la descripción de las métricas que se han utilizado para obtener esos resultados.

A continuación, se muestran en las Figuras 4.13, 4.14, 4.15, 4.16 y 4.17 los resultados en cuanto a tasa de precisión del resto de pacientes. Destaca que la máxima precisión se obtiene siempre con combinaciones de entre 1 y 4 métricas y no más, además la precisión en la mayoría de los casos (concretamente 9 de 11) de la asignación obtenida supera el 75 % lo que se considera un éxito dentro del procedimiento.

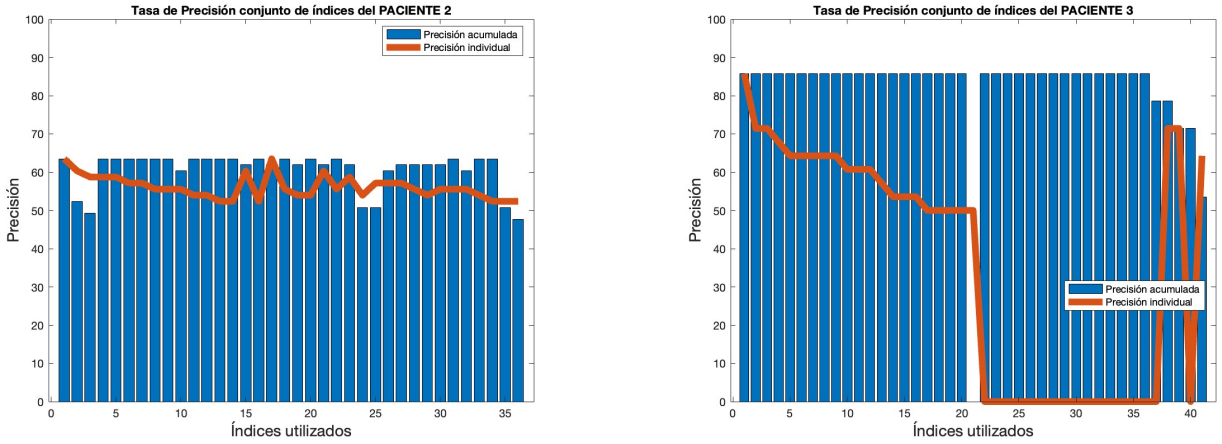

Figura 4.13: Resultados Paciente 2 y 3 para el segundo procedimiento

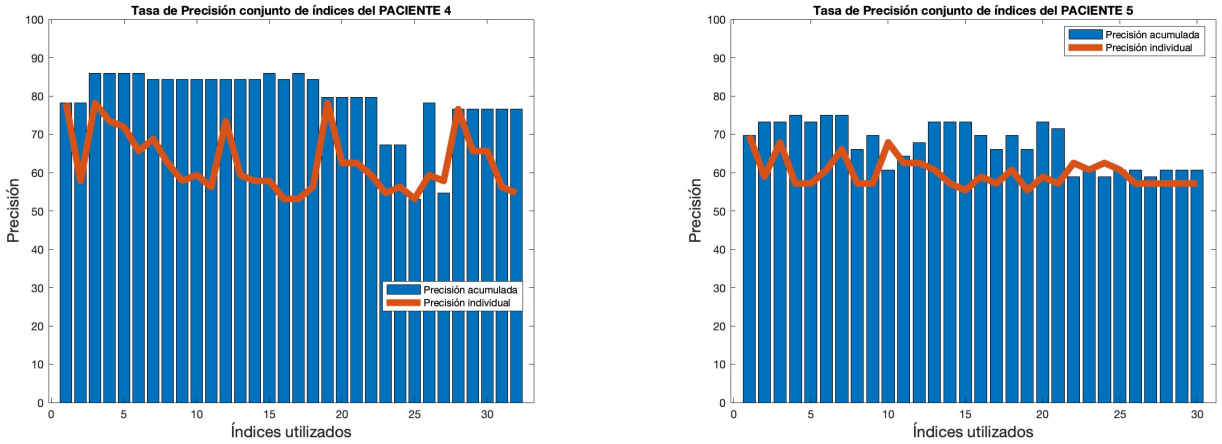

Figura 4.14: Resultados Paciente 4 y 5 para el segundo procedimiento

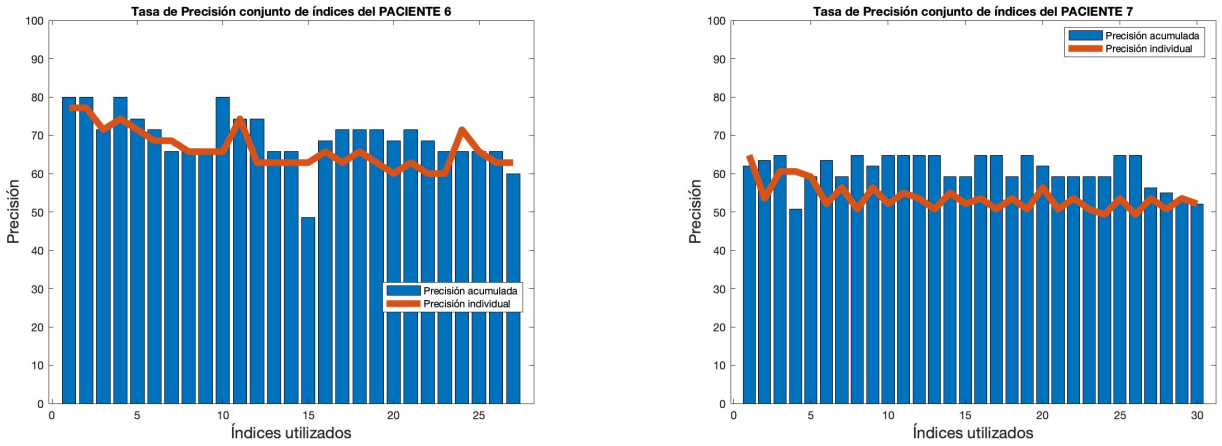

Figura 4.15: Resultados Paciente 6 y 7 para el segundo procedimiento

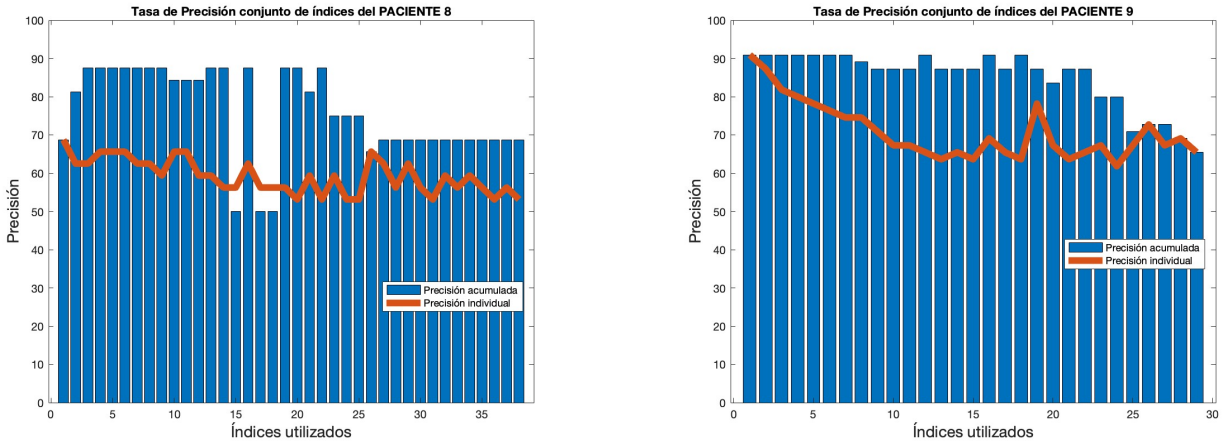

Figura 4.16: Resultados Paciente 8 y 9 para el segundo procedimiento

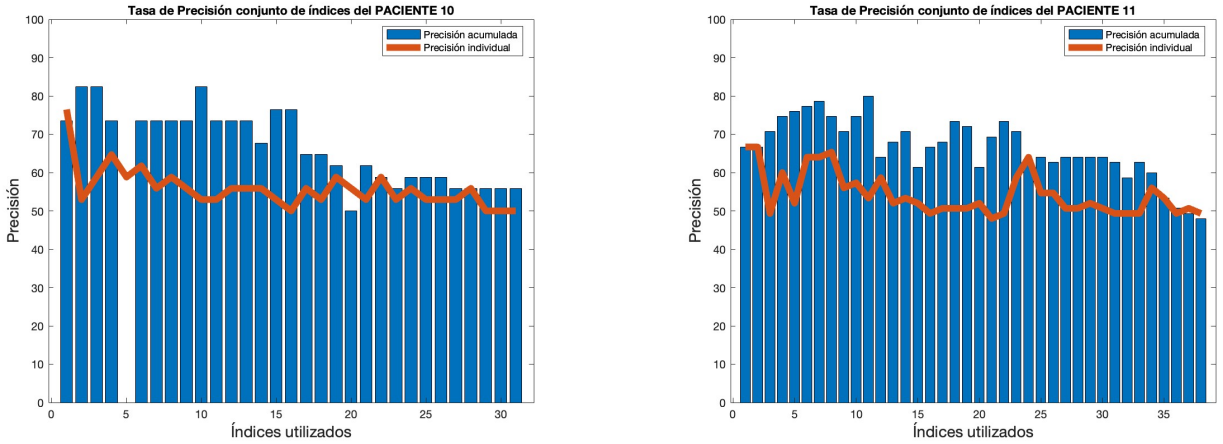

Figura 4.17: Resultados Paciente 10 y 11 para el segundo procedimiento

De nuevo, se muestra la Tabla 4.3 con el resumen del segundo procedimiento para todos los pacientes. Se incluye la información acerca de cual es la máxima precisión obtenida, cuantas y cuáles son las métricas con las que se obtiene esta precisión (identificador dentro del conjunto de las 55, banda de frecuencia y número de ciclos), así como, de nuevo, el parámetro de Engel.

Paciente	**max accP**	**Nº metricas**	**Descripción**	**Engel**
1	80,7	2	• 1 con 5 ciclos y 7-13 Hz • 41 con 10 ciclos y 30-249 Hz	*IV*
2	63,5	1	• 1 con 5 ciclos 0.5-3 Hz	*III*
3	85,7	1	• 8 con 5 ciclos, 30-249 Hz	*IV*
4	85,9	3	• 28 con 10 ciclos y 0.5-3 Hz • 20 con 10 ciclos 3-7 Hz • 21 con 5 ciclos y 3-7 Hz	*II*
5	75	4	• 14 con 15 ciclos y 30-249 Hz • 14 con 5 ciclos y 3-7 Hz • 16 con 10 ciclos y 3-7 Hz • 20 con 10 ciclos y 3-7 Hz	*I*
6	80	1	• 3 con 5 ciclos y 13-30 Hz	*I*
7	64,7	3	• 37 con 5 ciclos 0.5-3 Hz • 43 con 15 ciclos 0.5-3 Hz • 37 con 15 ciclos 0.5-3 Hz	*II*
8	87,5	3	• 13 con 15 ciclos y 30-249 Hz • 26 con 10 ciclos y 13-30 Hz • 1 con 5 ciclos y 30-249 Hz	*I*
9	90,9	1	• 12 con 5 ciclos y 0.5-3 Hz	*I*
10	82,35	2	• 1 con 10 ciclos 7-13 Hz • 21 con 15 ciclos 7-13 Hz	*IV*
11	77,3	4	• 13 con 15 ciclos 30-249 Hz • 7 con 10 cilos 13-30 Hz • 4 con 10 cilos 0.5-3 Hz • 40 con 10 cilos 30-249 Hz	*I*

Tabla 4.3: Resumen de los resultados para el segundo procedimiento.

Al igual que en el procedimiento anterior se desea comprobar si hay diferencias entre los pacientes de Engel *I* (más favorable) y el resto que se observen en este procedimiento utilizando el test no paramétrico de $Kolmogorov-Smirnov$. Los resultados para el segundo procedimiento se muestran a continuación en la Tabla 4.4.

Proced. 2	**Grupo Engel I**	**Grupo Engel II, III, IV**
Media	82.14	77.14
Desviación típica	6.07	9.4
Estadístico KS	0.4	
p-valor	0.6883	

Tabla 4.4: Resultados del test de Kolmogorov-Smirnov entre los grupos Engel I y Engel II-IV.

Estos resultados descartan que las distribuciones de las tasas de precisión sean diferentes, por lo que se concluye que con este procedimiento no hay diferencias significativas entre los resultados de los pacientes de Engel tipo *I* y el resto.

Capítulo 5: Discusión y Conclusiones

Una vez obtenidos los resultados de ambos procedimientos, pasamos a discutir los resultados obtenidos. En primer lugar, la hipótesis inicial planteada era si se podía obtener información sobre la clasificación epileptogénica de los nodos cerebrales a partir de datos recogidos en periodos sin crisis epilepticas. Se observa claramente que todos los pacientes tienen un significativo porcentaje de ventanas con información acerca de los grupos epileptogénicos. Además, encontramos pacientes en los que este porcentaje significativo ronda el 90 %, si combinamos los resultados obtenidos junto con el parámetro de Engel, para el primer procedimiento, los mejores resultados se obtienen en los siguientes pacientes:

Paciente	**% de vt fav.**	**max precisión**
6	92,8	91,42
8	89,3	90,62
9	97,1	85,45

Tabla 5.1: Recapitulación de los 3 mejores resultados del procedimiento 1 ordenados.

Para los tres casos coincide que el parámetro de Engel es el mejor I, lo que indica que entre 12 y 24 meses después de la intervención quirúrgica basada en la asignación realizada por los médicos, (y_6, y_8 e y_9), se ha logrado que las crisis epilépticas desaparezcan. Esto respalda los resultados obtenidos, ya que asegura que la asignación de referencia es adecuada, por lo que un 90 % de coincidencia se considera una buena asignación por parte del método, lo que respalda la hipótesis inicial, indicando que sí que es posible obtener esta clase de información a partir de datos en períodos de actividad basal.

Para el segundo procedimiento, los tres mejores resultados se obtienen en el Paciente 9 ($accP = 90,9$), Paciente 8 ($accP = 87,5$), ambos de Engel I y, en tercer lugar, en el Paciente 4 ($accP = 85,7$) de Engel II. De nuevo, el parámetro de Engel respalda que los resultados obtenidos altos sean suficientes para, confirmar la hipótesis inicial.

Sin embargo, cabe destacar que con el segundo procedimiento, basado en el análisis de las 2 horas de grabación en su totalidad, se obtienen resultados algo menos favorables en términos de coincidencia entre los grupos identificados en Y_i y los grupos identificados clínicamente y_i, lo que nos indica que no todas las ventanas de tiempo contienen información de la organización epileptogenica de la red, por eso cuando el estudio se focaliza en un número concreto de ventanas de tiempo (primer procedimiento), los resultados son más favorables.

Primer Procedimiento

A la vista de los resultados obtenidos resumidos en la Tabla 4.2, se puede llegar a las siguientes conclusiones. En primer lugar, acerca del preprocesamiento con o sin regresión, pese a que en una primera aplicación del método (con todas las métricas) varios pacientes presentan mejoras muy significativas con el preprocesamiento sin regresión (como el Paciente 8), una vez se optimiza el método utilizando solo las métricas favorables, el número de ventanas significativas es mayor en los casos con regresión en la mayoría de los casos, y en los que no son mayores, son muy similares (como es el caso del Paciente 10). Por lo que se concluye que el preprocesamiento con regresión da mejores resultados.

Otro factor que llama la atención es que en 9 de 11 de los casos se obtienen mejores resultados con ventanas de tiempo divididas con 15 ciclos, lo que lleva a elegir este tipo de preprocesamiento para futuros avances, así como incluso la posibilidad de trabajar con ventanas de tiempo divididas con más ciclos por ventana, como por ejemplo 30.

Esta casi unanimidad no ocurre, sin embargo, para el caso de los rangos de frecuencia, observamos que todos ellos aparecen en la Tabla 4.2, lo que se contradice con la idea intuitiva de que las características nodales que pueden aportar información sobre la clasificación epileptogénica se dan en un tipo de estado cerebral como se describe en la Tabla 2.2.

Concretamente, se ve que hay dos bandas de frecuencias más representadas, la primera de 0.5-3 Hz (Paciente 4, 9 y 10) que corresponde con estado cerebral dormido, y la última de >30Hz (Paciente 1, 5, 8 y 11) que corresponde con estado cerebral concentrado. Esto indica que estas características nodales que aportan información acerca de

la clasificación epileptogénica pueden aparecer en un tipo de estado cerebral característico para cada paciente, pudiendo así dividir a los pacientes a partir de este parámetro para futuras líneas de investigación. Más adelante cuando comparemos ambos procedimientos veremos que esta banda de frecuencia característica coincide, en algunos casos, en los dos procedimientos, confirmando aún más esta hipótesis.

En cuanto a las métricas, se puede concluir que la *Centralidad de Intermediación de los nodos* no es una métrica adecuada para este procedimiento, ya que en ningún caso es escogida por el método descrito para la selección de métricas. Luego, en cuanto al parámetro sobre utilizar la parte positiva de la matriz de correlación o la parte negativa para el cálculo de métricas, la mayoría de los pacientes estudiados en este trabajo presentan como métricas favorables las calculadas en base a la parte positiva de la matriz de correlación. Sin embargo, destacan casos como el Paciente 4, en el que en ningún caso las métricas con la parte positiva de la matriz de correlación generan buenos resultados, y sin embargo utilizando las métricas con la parte negativa el número de ventanas significativas aumenta claramente, por lo que se va a considerar como un parámetro característico de cada paciente.

Por último, podría destacar haber ordenado los pacientes en la Tabla 5.1 considerando el Paciente 8 como mejor que el Paciente 9, esto se debe a que se da más importancia al porcentaje de precisión que obtiene la asignación del método frente a la asignación médica, que al número de ventanas significativas dentro del conjunto total. La realidad es que la precisión media del porcentaje de ventanas significativas ronda el $55-60\,\%$, que no es muy alto; pero la idea consiste en que cuanto mayor sea el porcentaje de ventanas significativas más probabilidad habrá de encontrar ventanas con precisiones máximas.

Segundo Procedimiento

Basándonos en los resultados del segundo procedimiento resumidos en la Tabla 4.3, se puede llegar a las siguientes conclusiones. En primer lugar, acerca del número de métricas óptimo para conseguir la máxima tasa de precisión. Los resultados muestran que la máxima precisión se encuentra entre 1 y 4 métricas, aunque debido a la aleatoreidad de la inicialización de k-medias este número podría aumentar hasta 6, en cualquier caso y teniendo en cuenta que el total de métricas es de 825, consideramos que necesitamos pocas métricas para llegar a esta máxima precisión.

En cuanto a la descripción de las métricas que han sido seleccionadas vemos que, de nuevo, hay pacientes en los que destaca una banda de frecuencia en concreto, además de métricas concretas. En primer lugar, destaca la métrica número 1 (repetida en 4 pacientes), que aporta información sobre la media de las observaciones de avFC, es decir del promedio de la correlación funcional. Además, de otras como la 13, la 20 o la 21 (repetidas en 2 pacientes).

Destaca también el caso de pacientes como el 5 o el 7, en el que vemos como hay métricas que se repiten, pero preprocesadas en otra banda de frecuencia o con otro número de ciclos, lo que indica una correlación directa entre estas métricas y la información característica de los nodos sobre la epilepsia.

Comparación de ambos procedimientos

Los resultados obtenidos son similares para ambos procedimientos, aunque ligeramente mejores para el primero, sin embargo, cabe destacar la facilidad de aplicación de los resultados del segundo procedimiento, lo que hace que suba su valor pese a estas pequeñas diferencias. Con facilidad de aplicación nos referimos a que con el segundo procedimiento una vez escogidas las métricas óptimas tenemos un único vector de asignaciones Y_i por paciente. Sin embargo, para el primer procedimiento una vez seleccionadas las métricas óptimas se tiene un subconjunto de entorno al $70-90\,\%$ del total donde se debe seleccionar la ventana con la mejor asignación, es por esto que se necesitará continuar desarrollando el método como se comentará en la última sección del capítulo.

El comparar ambos métodos, nos ayuda a confirmar otra hipótesis que surgía al observar los resultados del primer procedimiento, de que cada paciente tiene una o dos bandas de frecuencia en las que se consiguen los mejores resultados. Es este el caso del Paciente 1, en el que las métricas elegidas en el procedimiento 2 están calculadas con un preprocesamiento de 7-13 Hz y >30Hz, mismas bandas que destacan en el primer procedimiento.

Este es el caso también del Paciente 2, 4 y 9 en el que en ambos procedimientos destacan las bandas de frecuencias bajas, y el caso del Paciente 8 destacan las bandas de frecuencia altas.

Siguientes líneas de investigación

Como se ha comentado, se da por aceptada la hipótesis planteada en la introducción de que a partir de actividad cerebral basal (sin crisis epilépticas) se puede obtener información acerca de los tipos de nodos epileptogénicos, lo que abre la puerta a obtener más datos de este tipo (basales). Por lo que se repetirán los procedimientos con registros obtenidos mediante resonancias magnéticas (RM) en vez de mediante sEEG, lo que podrá mejorar la precisión de los resultados.

Otra continuación consistirá en estudiar las ventanas de tiempo del primer procedimiento que tengan una mayor precisión. Para ello, utilizaremos técnicas de preprocesado con las que se crearán ventanas de tiempo alrededor de las que son mejores con el preprocesado actual, con diferentes longitudes (ciclos por ventana) y configuraciones (solapamiento) hasta encontrar una tasa de precisión superior, se espera poder mejorar la precisión hasta un $90 - 95\,\%$, mejorando los resultados de este estudio.

Además, como ya se ha comentado, para el primer procedimiento, la precisión media de las ventanas de tiempo significativas esta entre el 50 y el $60\,\%$, la cuestión a estudiar será si este porcentaje de precisión, es decir, si el porcentaje de nodos acertados por el modelo ocurre para la mayoría de las ventanas de tiempo en los mismos nodos. En resumen, si hay nodos (zonas del cerebro) cuyo nivel epileptogénico se puede predecir acertadamente en un número muy grande de ventanas de tiempo, mientras que en otros dentro del mismo paciente es más difícil.

Tras realizar el primer análisis se observa que este es favorable para un determinado grupo de pacientes por lo que es necesario obtener formas de dividir a los mismos ciegamente (sin tener en cuenta el dato de grupo Esperado). Además de buscar mejorar los resultados del resto de pacientes.

Con estos objetivos la estrategia va a consistir en añadir al primer procedimiento una división previa. Se van a considerar diferentes vectores X en cada ventana de forma que se pueda realizar un análisis cluster de ventanas con información epiléptica y sin ella (2 grupos). Una vez hecha esta división en el número de ventanas, se va a repetir el cluster de grupos esperados por nodos, esperando que en un grupo de ventanas los resultados sean buenos y en el otro no. Esto nos aportaría una forma de dividir las ventanas de tiempo, en significativas y no significativas sin tener en cuenta la información de asignaciones esperadas (y_i).

Para este nuevo análisis se toma la parte triangular superior de la matriz de correlación R y se crea un vector de dimensiones $\frac{k_i \cdot (k_i - 1)}{2}$ siendo k_i x k_i la dimensión de la matriz de correlación. La matriz X que se usará para el primer análisis cluster tendrá las columnas con estos vectores y tantas filas como ventanas de tiempo.

Como se comentaba en la sección anterior, queda bastante claro a la vista de los resultados que no todas las ventanas de tiempo contienen información acerca de la clasificación epileptogénica de la red, por lo que otra aproximación será repetir el procedimiento 2, solo en las ventanas de tiempo que se consideran significativas (es decir, las obtenidas en el procedimiento 1).

También, con el objetivo de mejorar los resultados del segundo procedimiento, lo que se hará es calculas la asignación y su tasa de precisión para un conjunto grade de métricas (alrededor de las 200 con mejor tasa de precisión individual) e ir eliminándolas de una en una, y observar las variaciones en los resultados.

Se pretenden publicar, tanto los resultados obtenidos en este trabajo, como las siguientes líneas de investigación desarrolladas en mi trabajo fin de máster como artículo científico.

Bibliografía

[1] Duncan, J., Sander, W., (2006). *Adult epilepsy.* The Lancet, Volume 367, Issue 9516, 1087 - 1100.

[2] BrightFocus Foundation (2021) *Anatomía cerebral y Sistema Límbico.* URL: https://www.brightfocus.org/espanol/alzheimer/anatomia-cerebral-y-sistema-limbico. Septiembre 2024.

[3] Bertram, E., (1998) *Funtional anatomy of limbic epilepsy: a proposal for central sinchronization of a diffusely.* Epilepsy Research, Volume 32, Issues 1-2.

[4] Bartolomei, F., Wendling, F. (2004). *Pre-ictal synchronicity in limbic networks of mesial temporal lobe epilepsy.* Epilepsy Research, Volume 61, Issues 1-3.

[5] Ball, T., Kern, M., Mutschler, I., Aertsen, A., Schulze-Bonhage, A., (2009). *Signal quality of simultaneously recorded invasive and non-invasive EEG.* NeuroImage, Volume 46, Issue 3.

[6] Bartolomei, F., Lagarde, S., Wendling, F., McGonigal, A., Jirsa, V., Guye, M., Bénar, C. (2017). *Defining epileptogenic networks: Contribution of SEEG and signal analysis.* Epilepsia, 58(7), 1131-1147. https://doi.org/10.1111/epi.13791

[7] Campani, G. (2023) *A network analysis of EEG signal of a patient with epilepsy.* Department of Physics and Astronomy. Alma Mater Studiorum. Universitá di Bologna.

[8] Başar E. (2013). *Brain oscillations in neuropsychiatric disease. Dialogues in clinical neuroscience.* https://doi.org/10.31887/DCNS.2013.15.3/ebasar

[9] Priyanka A. Abhang, Bharti W. Gawali, Suresh C. Mehrotra. *Chapter 2 - Technological Basics of EEG Recording and Operation of Apparatus.* Introduction to EEG- and Speech-Based Emotion Recognition, Academic Press, 2016, Pages 19-50, ISBN 9780128044902.

[10] Bishop, C. M. , (2006) *Pattern Recognition and Machine Learning.* Berkeley. ISBN-10 0-387-31073-8.

[11] H. Kvam, P., Vidakovic, B., (2007). *Nonparametric Statistics with Applications to Science and Engineering* Georgia Institute of Thechnology. ISBN 978-0-470-08147-1.

[12] Newman, M. E. J. (2018) *Networks, An Introduction.* Oxford University press, New York. ISBN 978-0-19-920665-0.

[13] Fisher, R.A. (1958) *Statistical Methods for Research Workers*, 13th Ed., Hafner.

[14] M. Rubinov y O. Sporns, *Brain Conectivity toolbox.* 2010, URL: https://sites.google.com/site/bctnet/ Septiembre 2023.

[15] Friedman, J., Hastie, T., Tibshirani, R., (2008) *The elements of Statistical Learning. Data mining, Inference and Prediction.* 2nd Edition. ISBN-10 0387848576.

[16] Choi, W., Lee, J. W., Huh, M.-H., Kang, S.-H. (2003). *An Algorithm for Computing the Exact Distribution of the Kruskal–Wallis Test. Communications in Statistics - Simulation and Computation*, 32(4), 1029-1040. https://doi.org/10.1081/SAC-120023876

[17] Benjamini, Y., Yekutiei, D., (2001) *The Control of the False Discovery Rate in Multiple Testing under Dependency.* The Annals of Statistics, Vol 29 nº4 pp 1165-1188.

[18] Proal, E., M.S., Alvarez-Segura, M., M. D., de la Iglesia-Vayá M., Ph.D., Martí-Bonmatí, Castellanos, F. X. M.D., Spanish Resting State Network (SRSN). *Actividad funcional cerebral en estado de reposo: REDES EN CONEXIÓN.* Rev Neurol. 2011 Mar 1; 52(0 1): S3-10.

[19] Massachusetts General Hospital, Neurology. (2016) *Epilepsy Surgery Outcome Scales* URL: https://seizure.mgh.harvard.edu/engel-surgical-outcome-scale/ Enero 2025

[20] Devroge, L., Gyorfi, L., Lugosi, G. (1996) *A probabilistic theory of pattern recognition.* Springer, New York. ISBN NO-387-94618-7.