Universidad de Valladolid
Facultad de Ciencias

Trabajo Fin de Grado

Grado en Estadística

Extracción de características y clasificación para la implementación de un sistema de reconocimiento biométrico mediante dispositivos ponibles (wearables)

Autor:

Dña. Irene Salvador Ortega

Tutor:

D. Miguel Alejandro Fernández Temprano

Agradecimientos

Este Trabajo de Fin de Grado ha sido fruto de la ayuda de muchas personas que han estado directa o indirectamente a mi lado durante estos cinco años, desde profesores que han conseguido que hoy sepa todo lo que sé hasta familiares y amigos que me han apoyado y hecho que no me sintiera sola en ningún momento. Me gustaría dedicaros unas palabras de agradecimiento.

A mi tutor, Miguel Alejandro Fernández, sin tus consejos y ayuda no hubiera podido llegar tan lejos. Gracias por haberme guiado y apoyado, en todo momento. A Carlos Vivaracho y Mª Aránzazu Simón por haberme enseñado el mundo de la biometría y de la investigación desde dentro. Siempre lo recordaré como algo difícil pero bonito y gratificante. Juntos hemos formado un gran equipo que, sin duda, volvería a elegir.

A los profesores que han prestado su colaboración, Félix Prieto en toda la parte de Android, Pablo de la Fuente con la planificación y Luis Ángel García con sus consejos y documentación, gracias.

A mi familia, especialmente a mis padres, por haberme apoyado en cada decisión y haberla hecho posible, habéis conseguido hacerme la vida mucho más fácil. A mi hermano por ayudarme en todos los momentos difíciles y haber disfrutado de mis logros, como si fuesen suyos.

A mis compañeros de Universidad y en especial al equipo que nunca olvidaré, Raúl, Adrián, gracias por haber compartido juntos tanto los momentos de sufrimiento como las alegrías. En cada trabajo hemos conseguido ver la parte positiva y negativa y hemos disfrutado con ellos. Espero que sigamos celebrando cada acontecimiento que ocurra en nuestras vidas.

A mis abuelos, aunque ahora mismo sean dos estrellas que brillan en el cielo, estoy segura de que me habéis guiado y ayudado a aprender una lección tras cada decisión he tomado y ha salido tanto bien como mal.

A todos los que habéis estado presentes en mi vida durante este tiempo, muchas gracias.

Resumen

Los sistemas biométricos para el reconocimiento de las personas están en continuo uso y crecimiento, lo que crea la necesidad de mejorar los sistemas actuales y proponer nuevos y originales enfoques, que puedan mejorarlos o complementarlos.

En el presente proyecto se va a trabajar con una biometría basada en las características del comportamiento del ser humano que permite verificación no intrusiva, continua, fácil de conseguir y difícil de robar o falsificar. El objetivo principal es determinar si el uso de los sensores presentes en los dispositivos ponibles comerciales puede permitir o no la verificación biométrica de personas mediante su forma de caminar.

Este trabajo se realiza en conjunto con otro del Grado en Ingeniería Informática, siendo ambos una continuación de trabajos previos ya realizados por el grupo de investigación, en los que se desarrolló el sistema móvil de captura de datos y se obtuvo un corpus con el que trabajar. Aquí, se va a realizar un análisis profundo de dichos datos: demostrando la existencia de periodicidad, extrayendo características, probando diferentes técnicas de preprocesamiento, ajustando los valores de diversos parámetros y el número de características a utilizar, proponiendo como resultado final un sistema de reconocimiento, que se probará en diferentes sensores y dispositivos en el trabajo del Grado en Ingeniería Informática. De manera que estos trabajos permitirán tener unas bases sólidas en las que asentar futuras investigaciones, sirviendo como aproximación inicial de lo que se puede llegar a conseguir en esta biometría.

Palabras claves

Análisis de Fourier, Biometría, dispositivos comerciales, dispositivos ponibles, dominio de la frecuencia, dominio del tiempo, forma de andar, reconocimiento biométrico, selección de características

Abstract

Biometric systems for the recognition of people are in continuous use and growth, which creates the need to improve existing systems and propose new and original approaches that can improve or complement them.

In this project we will work with a biometrics based on the characteristics of human behavior that allows an unobtrusive, continuous, easy to obtain and difficult to steal or falsify verification. The main objective is to determine if the use of the sensors present in the wearable devices can allow or not the biometric verification of people by their gait.

This work is done in conjunction with another of the Degree in Computer Engineering, both being a continuation of previous work already carried out by the research group, in which the mobile data capture system was developed and a corpus was obtained to work with. Here, an in-depth analysis of said data is going to be carried out: demonstrating the existence of periodicity, extracting characteristics, testing different preprocessing techniques, adjusting the values of various parameters and the number of characteristics to be used, proposing as a final result a recognition system, which will be tested in defferent sensors and devices in the work of the Degree in Computer Engineering. So these works will provide a solid basis on which to base future research, serving as an initial approximation of what can be achieved in this biometrics.

Keywords

Fourier Analysis, biometrics, commercial devices, wearable devices, frequency domain, time domain, gait, biometric recognition, feature selection

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Hoy en día, cada vez más personas aprovechan las ventajas que proporcionan los dispositivos electrónicos. Hace años, cuando querías comunicarte con otra persona que vivía lejos, era necesario comprar un sobre, un sello, escribir la carta y buscar un buzón de correos para enviarla, esperando que con suerte al día siguiente o a los dos días, el receptor recibiera la carta e hiciera el mismo procedimiento para poder conocer su respuesta. En cambio, ahora, basta con coger el teléfono móvil o el ordenador y realizar una llamada o enviar un correo electrónico que el receptor recibirá en segundos. Hay estudios que indican que en 2019 se envían 293.6 mil millones de correos electrónicos cada día. Y por si pareciera poco, se espera que en 2023 se envíen más de 347.3 mil millones de correos electrónicos diariamente [1]. Se está incrementando el número de actividades que se realizan a través de la red como comprar ropa e incluso artículos frescos (fruta, verdura...) en un supermercado, sector que está en continuo avance para hacer el servicio de las personas más cómodo y atractivo. Ya existen empresas que trabajan en lanzar una solución para la compra online de productos frescos que permitirá, a través de la tecnología de visión artificial y las cámaras de alta tecnología de seguimiento vinculadas a cada sección del supermercado, al cliente online coger su turno para ser atendido, informándole del tiempo de espera, y en el momento de iniciar su turno, le permitirá interactuar con el dependiente, ya sea para indicar el producto que desea y cómo quiere que se lo prepare como para solicitar recomendaciones o hacerle preguntas como si estuviera en la tienda física [2].

En el momento en que un usuario desea llevar a cabo cualquier tipo de comunicación o transacción a través de Internet, necesita algún tipo de autenticación. Los sistemas más utilizados actualmente requieren que el usuario se registre y recuerde contraseñas, provocando que este termine realizando prácticas poco seguras, como emplear contraseñas simples, repetir una misma contraseña en varios sitios web, guardarlas en algún archivo, no cambiarlas con el paso del tiempo, etc. También hay que añadir problemas administrativos derivados de posibles pérdidas de claves, o en el caso de que el usuario necesite llevar consigo alguna tarjeta o dispositivo, esta también se puede perder, ser robada o transferirse.

Una alternativa es el uso de sistemas basados en biometría, puesto que el usuario no necesita recordar ni llevar consigo nada, empleando únicamente sus características intrínsecas (físicas o

de la forma de actuar). Aunque la biometría se lleva aplicando desde finales del siglo XIX para la identificación de las personas con métricas como la huella dactilar, la continua existencia de limitaciones e inconvenientes ha provocado que a lo largo de los últimos años los estudios basados en sistemas de reconocimiento biométrico hayan cobrado una mayor relevancia.

En los últimos años, los métodos más empleados en reconocimiento biométrico estaban relacionados con las características físicas del individuo como su cara o su huella dactilar. Estas características dan buenos resultados, pero requieren que el usuario ponga su cara o su dedo en algún dispositivo, proceso que termina siendo incómodo para el individuo. Actualmente, son cada vez son más los objetos conectados (Internet of Things, IoT) que incorporan sistemas de reconocimiento biométrico. Estos sistemas tratan de medir las características del cuerpo humano, su estructura e incluso determinados comportamientos que permitan interacciones naturales entre humanos y máquinas. Estos sistemas inteligentes son una de las tendencias de Inteligencia Artificial (AI) que más se están desarrollando y utilizando en 2019. Actualmente ya se utilizan para la verificación de asistencia, acceso a lugares físicos, aplicaciones o información, comunicación con asistentes virtuales, etc [3]. Por eso, en la actualidad y en este proyecto se busca emplear el comportamiento de la forma de actuar del individuo para el reconocimiento biométrico, ya que es un método menos intrusivo.

Se va a aprovechar el desarrollo y gran éxito en ventas en los últimos años de los dispositivos ponibles comerciales (pulseras de actividad, relojes inteligentes, etc.) para ver si es posible verificar a una persona a partir de sus datos. Estos dispositivos ya tienen aplicaciones muy diversas, desde recibir y contestar a notificaciones del teléfono móvil hasta monitorizar el rendimiento de los deportistas para ayudarles durante su entrenamiento. Cuentan con una gran variedad de sensores, capaces de capturar distintos tipos de datos del usuario como su movimiento, ritmo cardíaco, sudor... lo que hace posible el reconocimiento biométrico.

Teniendo en cuenta el tipo de sensores que incorporan actualmente los dispositivos ponibles comerciales, y como continuación de dos trabajos previos del Grado en Ingeniería Informática, se va a centrar la atención en verificar a una persona a partir de su *forma de andar*. En el primer trabajo previo [4] se hizo un estudio más detallado de todos los sensores de distintos dispositivos ponibles y se llegó a la conclusión de que los únicos que proporcionaban una información susceptible de ser usada en biometría para la forma de andar son los aquí se van a usar: el acelerómetro y el giroscopio. También se construyó una aplicación Android que permitía la recogida de los datos de dos dispositivos seleccionados. El segundo trabajo [5] recogió datos de diversos usuarios y realizo un estudio preliminar con ellos que demostraba la existencia de periodicidad en la señal de los datos y resultados positivos que indicaban su posible uso en biometría.

La diferencia entre la forma de abordar el problema en este proyecto y el resto de los trabajos de la bibliografía, es que se van a utilizar dispositivos comerciales, cuando en la bibliografía se utilizan dispositivos creados ad hoc para el propio propósito del proyecto o smartphones.

La biometría puede ser aplicada para la identificación o la autenticación de las personas.

- **Identificación biométrica:** consiste en la capacidad de identificar quién es el usuario que

está empleando un sistema. Aquí el sistema debería ser capaz de comparar las muestras de un usuario a identificar con los patrones almacenados en el sistema, debido a ello, el sistema obtendrá una serie de resultados que rebasen cierto valor umbral. Es un problema de clasificación multiclase que busca respuesta a la pregunta *Who am I? (¿Quién soy yo?)*.

- **Verificación biométrica:** consiste en la capacidad de demostrar que cierto usuario de un sistema es quien dice ser. En un sistema como este, el usuario tendría que presentar su identificación y una muestra de datos. El sistema se encargaría de preprocesar los datos de las muestras y compararlos con el patrón que tenga almacenado para ese identificador. Si los datos coinciden, entonces se trata de un usuario *auténtico*, en caso contrario, de un *impostor*. Es un problema de clasificación de una clase que busca respuesta a la pregunta *Am I who I claim I am? (¿Soy yo quien digo que soy?)*

En este proyecto, nos vamos a centrar únicamente en la verificación, cuyas fases se podrían resumir como se muestra a continuación.

1. Adquisición de la señal de un usuario a través de un sensor presente en un dispositivo.
2. Fase de preprocesamiento y extracción de características de dicha señal.
3. Generación de los ficheros de salida necesarios para aplicar el clasificador.
4. Módulo de decisión: el clasificador generará una métrica que servirá para evaluar el rendimiento del sistema construido y se podrá utilizar para tomar una decisión: soy o no soy yo el usuario.

En este trabajo nos vamos a centrar en la "Fase de procesamiento" (punto 2), haciendo un análisis profundo de la etapa de preprocesamiento y sus distintos parámetros, estudiando distintos tipos de conjuntos de características. Ofreciendo como resultado final un sistema de reconocimiento, cuya aplicación se realizará en otro Trabajo Fin de Grado (TFG), correspondiente al Grado en Ingeniería Informática, que se ha realizado en paralelo con este [6].

La Base de Datos a utilizar va a ser la obtenida en el TFG del Grado en Ingeniería Informática previo a este [5]. El hecho de que el tamaño de la base de datos no sea muy grande, no es ningún problema, ya que el objetivo de este trabajo no es obtener un sistema de reconocimiento basado en ponibles, si no profundizar en el conocimiento y las especiales características de esta biometría, trabajo no realizado hasta ahora en la bibliografía. Además, el hecho de ser una base de datos no muy numerosa hace que sea factible un análisis particular del comportamiento de cada individuo, lo que, como se verá, nos va a permitir extraer conclusiones muy interesantes para trabajos futuros.

1.1. Biometría

Para comprender mejor este trabajo, en este apartado, se van a explicar una serie de conceptos comunes en biometría y los sistemas biométricos que aparecerán de manera recurrente a lo largo

del trabajo [7].

La **biometría** es el estudio estadístico de los fenómenos o procesos biológicos. Tiene muchas aplicaciones posibles, pero dentro de las tecnologías de la información, la más destacada es el estudio del reconocimiento de los seres humanos a partir de sus características, que se suelen clasificar en dos tipos:

- **Características fisiológicas:** son características físicas de los individuos. Dentro de este grupo cabe destacar la huella dactilar, el iris, etc. Se caracterizan por ser estáticas, es decir, no cambian con el tiempo.
- **Características del comportamiento:** son propiedades de la forma de actuar de los individuos. Dentro de este grupo se encuentra el modo con el que interactúan con los dispositivos, su voz, firma, forma de andar, etc. Se caracterizan por ser dinámicas, es decir, pueden cambiar con el paso del tiempo.

Un **sistema de reconocimiento biométrico** es una aplicación informática con la capacidad de identificar o verificar a una persona a partir de sus características, bien sean fisiológicas o de comportamiento.

Los sistemas de reconocimiento necesitan algún tipo de patrón para poder identificar o verificar a los individuos. Un **patrón** es un modelo creado mediante capturas o datos del usuario para representarle.

Evidentemente, no todas las características de un individuo pueden ser empleadas para el reconocimiento biométrico. Según [8–10], para que una característica biométrica pueda ser considerada como tal, ésta ha de cumplir las siguientes propiedades.

- **Universalidad:** todas las personas han de tener dicha característica biométrica.
- **Unicidad:** no ha de haber dos personas que sean idénticas atendiendo únicamente a esa característica.
- **Permanencia:** o biológicamente constante, es decir, la característica no tiene que variar con el tiempo.
- **Recolectable:** la característica ha de poder ser medible cuantitativamente.

Buscando conseguir un sistema de reconocimiento biométrico que tenga las siguientes características.

- **Rendimiento:** precisión que tiene el sistema biométrico empleado a la hora de identificar o verificar a un individuo.
- **Aceptabilidad:** el grado en que el público se muestra positivo a utilizar el sistema biométrico.

- **Invulnerabilidad:** el grado de facilidad del sistema a ser engañado mediante el uso de técnicas fraudulentas.

En la biometría basada en comportamiento como la forma de andar, en ocasiones, la propiedad de *permanencia* no se cumple, denominando a este tipo de biometrías como suaves o débiles *(Soft Biometrics)*.

Por otro lado, la posibilidad de verificar o identificar a un individuo a través de su forma de andar está sujeto a cuestiones éticas, teniendo una serie de ventajas e inconvenientes.

Entre sus ventajas se encuentran que el usuario sólo tiene que andar, sin necesidad de interacción durante el proceso de reconocimiento, el cuál es continúo (el propietario se mantiene automáticamente autorizado), discreto (sin molestar al usuario), difícil de robar o falsificar, pudiéndose capturar a distancia. Pero existen muchos factores que influyen en la forma de andar de las personas, tanto externos, como podrían ser las condiciones de la superficie, meteorológicas, la ropa o zapatos que lleve el usuario, etc., como internos relacionados con el estado físico, mental o enfermedades del usuario. Por otro lado, los conjuntos de datos contienen información muy sensible, que podría ser utilizada para identificar de forma única a las personas y dependiendo del tipo de sensor usado, se podría incluir información que pudiera revelar las condiciones médicas de los usuarios. Además, al no requerirse consentimiento del individuo se podría estar observando y extrayendo dicha información sin que el usuario lo sepa.

Esto lleva a enfrentarnos con un problema muy presente actualmente en nuestra sociedad, la privacidad de los datos. Hay que tener mucho cuidado porque aunque aparentemente sólo estemos observando su forma de andar, puede existir gente que de manera maliciosa aproveche esa información y consiga conocer la identidad física, fisiológica o psíquica de los usuarios. Como se ha podido ver en las ventajas e inconvenientes, está información se puede utilizar en el campo de la *medicina* de manera positiva, en la prevención de enfermedades y/o la posible actuación temprana de las mismas, o de manera negativa, revelando las condiciones médicas de los usuarios y utilizándolo para perjudicarlos. Para intentar evitar problemas, ya existe la primera guía de pautas éticas para hacer el uso de la Inteligencia Artificial (AI) más responsable, producidas por el Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial (AI HLEG) a nivel de la Unión Europea (EU) [11]. En ella se reconoce el enorme impacto positivo que la AI tiene a nivel mundial, tanto comercial como socialmente, siendo una tecnología tanto transformadora como disruptiva que ha ido evolucionando en los últimos años produciendo enormes cantidades de datos digitales, creando una importante innovación científica y de ingeniería. Aseguran que la AI continuará impactando a la sociedad y a los ciudadanos de una manera que aún no podemos imaginar. Por ello, consideran importante que se preste la debida atención a garantizar un entendimiento y un compromiso para construir una AI digna de confianza y han redactado las directrices para que esto sea así, asegurando el propósito ético. Y aunque, afirman que la AI puede provocar daños no intencionados, han desarrollado un marco para implementar la AI confiable, ofreciendo una orientación concreta para su logro, proponiendo métodos técnicos y no técnicos de ayuda para

su realización e implementación. De todas formas, en este trabajo se utilizará la información anonimizada.

1.2. Motivación

Desde el momento en el que supe que podía elegir el tema que yo quisiera, siempre que cumpliera unos contenidos mínimos, para llevar a cabo mi TFG, tuve claro lo que quería: un tema original, novedoso en el que hubiera datos complejos, no inmediatos, que analizar y con los que pudiera tener un amplio árbol de posibilidades para trabajar.

Entre las asignaturas del Grado en Estadística que más me habían gustado y con las que había conseguido mi afición por los datos, se encontraban:

- **Análisis de Datos:** consiguiendo ese primer contacto con las técnicas de análisis de datos, tanto para la selección de características como para la aplicación de modelos de clasificación.
- **Análisis Multivariante:** como continuación de la asignatura anterior, más centrada en profundizar los problemas de inferencia multivariante. Empezando un primer contacto con técnicas de Machine Learning como redes neuronales, máquinas de vectores soporte, árboles de clasificación o Random Forests.

La biometría también estuvo siempre entre mis motivaciones personales, incluso antes de elegir la carrera que ahora termino, por parecerme un campo innovador en el que se había trabajado mucho a lo largo de los años, pero sobre el que todavía quedaba mucho trabajo por hacer. Si alguien un día fue capaz de crear un sistema de reconocimiento por huella dactilar o por reconocimiento facial, que además ahora estaban empezando a utilizarse cada vez más para desbloquear los smartphones, ¿por qué no existían otros métodos más discretos de reconocimiento? Si realmente existen estudios médicos que indican que hay 24 componentes diferentes en la forma de andar humana que la hacen única [12], ¿por qué hay tan pocas investigaciones que tratan de considerar esas 24 componentes y empezamos a desbloquear nuestros teléfonos móviles de esta manera?

No obstante, a lo largo de la titulación nunca había trabajado con datos biométricos ni conocía el mundo de la biometría, pero cuando se me ofreció este tema que se encontraba incompleto, decidí aceptar el reto y empezar a trabajar en ello.

1.3. Objetivos

A continuación, se indica el objetivo general y se enumeran los objetivos específicos en que se divide el presente proyecto.

1.3.1. Objetivo general

El objetivo principal de este trabajo es estudiar el posible uso de dispositivos ponibles comerciales para el reconocimiento biométrico de personas mediante la forma de andar.

1.3.2. Objetivos específicos

Para poder cumplir el objetivo general, se han creado una serie de objetivos específicos que se desarrollarán de forma progresiva. Estos objetivos servirán para determinar si merece la pena continuar futuros estudios en este tema o, por el contrario, si es mejor abandonar esta línea de investigación. Los objetivos específicos que se han planteado llevar a cabo han sido los siguientes:

1. Realizar un análisis de los datos para ver qué características tienen y si pueden ser susceptibles o no de ser usados en biometría y de qué manera.
2. Realizar una búsqueda del preprocesamiento que mejores resultados produce a los datos.
3. Hacer un estudio prospectivo para valorar los parámetros que existen en el experimento y tratar de buscar sus mejores valores.
4. Realizar una primera aproximación de sistema de reconocimiento.
5. Realizar una selección de características para ver si produce mejoras sobre los resultados.

En el trabajo que se ha realizado en paralelo con este [6], se puede ver la evaluación del sistema de reconocimiento sobre distintos sensores y dispositivos, así como la puesta en marcha de la aplicación utilizada para la captura de los datos.

1.4. Estructura de la obra

Esta memoria se encuentra dividida en una serie de capítulos y secciones basándose en la estructura presentada en la guía docente para la asignatura TFG del Grado en Estadística de la *Universidad de Valladolid*.

De esta manera siguiendo al actual capítulo introductorio donde se habla de biometría, la motivación por este trabajo y los objetivos, se encuentran los siguientes capítulos.

Capítulo 2. Estado del Arte: En este capítulo se analizarán los trabajos previos existentes en el campo de la biometría considerando los diferentes sensores existentes de manera breve y centrándose en los que se van a utilizar en el presente proyecto.

Capítulo 3. Datos y análisis: se explicarán los datos con los que se va a trabajar y se realizará un análisis de autocorrelación para estudiar si existe periodicidad en la señal.

Capítulo 4. Configuración experimental: En este capítulo se explicarán los parámetros estáticos que se vayan a fijar en el sistema de reconocimiento final. Las decisiones serán sobre qué características se van a extraer de los datos y en qué dominios, cómo se van a medir los resultados, qué procedimiento experimental se va a seguir y qué algoritmo de clasificación se va a utilizar, justificando cada una de las decisiones.

Capítulo 5. Preprocesamiento: En este capítulo se van a probar diversas técnicas de preprocesamiento, con el objetivo de tomar decisiones, simplificando el problema a resolver.

Capítulo 6. Experimentos: Análisis de parámetros: aprovechando las decisiones del capítulo anterior, aquí se van a analizar, en profundidad, los diferentes parámetros que afectan al problema de biometría que se está resolviendo. Se explicarán los parámetros que se van a analizar y a continuación, se estudiarán y se tomarán las decisiones más apropiadas que darán lugar al sistema de reconocimiento final.

Capítulo 7. Experimentos: Selección de características: En este capítulo se realizará un análisis inicial para ver cómo funciona la selección de características sobre el sistema de reconocimiento construido.

Capítulo 8. Conclusiones y trabajo futuro: Este es el capítulo final donde se expondrán las conclusiones obtenidas y las posibles alternativas a probar en un futuro.

Para finalizar se encuentra una sección donde se explican los acrónimos y abreviaturas utilizados a lo largo de la memoria, un índice alfabético, los anexos del trabajo y la bibliografía.

Capítulo 2

Estado del Arte

A lo largo de este capítulo se va a exponer la situación actual de la biometría, centrándonos en el estudio de aquellas cosas que nos afectan, como la manera de actuar con los datos o las técnicas de aprendizaje automático que se emplean, haciendo un resumen de los resultados obtenidos en investigaciones similares utilizando dispositivos ponibles y el comportamiento de las personas como patrón.

El reconocimiento biométrico consiste en aplicar técnicas estadísticas y matemáticas sobre las características fisiológicas o del comportamiento de un individuo para su reconocimiento, ya sea identificación o verificación. Estos sistemas, como se ha dicho en el capítulo 1 de la Introducción, presentan una serie de ventajas, tales como que los usuarios no necesitan recordar claves complejas para su autenticación ni llevar consigo llaves, tarjetas u otros objetos físicos, que pueden perderse o transferirse.

Existen sistemas biométricos tradicionales y portátiles. Los sistemas portátiles, por su naturaleza, están siempre con el usuario pudiendo almacenar los datos dentro del dispositivo, siendo capaces de leer la señal del sujeto en cualquier momento y por tanto, permitiendo la autenticación continua, mientras que los sistemas biométricos tradicionales son generalmente colocados en un lugar fijo, menos susceptibles de deteriorarse así como más fácilmente reemplazables, haciendo uso de procesos más costosos computacionalmente, ya que pueden utilizar fuentes externas de energía [7]. Ejemplo de sistema biométrico tradicional es una característica de Windows 10 llamada *Windows Hello* [13] que permite al usuario autenticarse usando la cara, el iris o la huella digital.

La creciente popularidad de los dispositivos portátiles está llevando a nuevas formas de interactuar con otros dispositivos inteligentes y con otras personas. Los *wearables* equipados con una serie de sensores son capaces de capturar los rasgos fisiológicos y de comportamiento del propietario, resultando apropiados para biometría, siendo éstos los que se van a utilizar en el presente proyecto. Los sensores predominantes en los dispositivos portátiles actuales son [7]:

- **Sensores de luz:** dependiendo de la resolución del sensor, pueden ser utilizados para medir la intensidad de la luz, como por ejemplo los sensores fotopletismográficos (PPG) [14]

que miden el volumen de cambio sanguíneo dentro del tejido microvascular, o proporcionar imágenes completas, como es el caso de los lectores de huellas dactilares [15] o cámaras digitales [16] que pueden capturar las características fisiológicas como la cara u otras características corporales como la forma de andar de los individuos a través de lo que se llaman *técnicas de visión.*

- **Sensores de fuerza:** mide la fuerza que afecta al dispositivo de medición, ya sea originada por el movimiento, ejemplo de ello es el acelerómetro tridimensional [17, 18] o por la fuerza de Coriolis como hace el giroscopio o el campo magnético de la Tierra con el magnetómetro o la presión del aire con el barómetro.

- **Sensores eléctricos:** mide la actividad eléctrica de algunas partes del cuerpo, como por ejemplo, un electrocardiograma para el corazón [19] o cómo cambia una corriente cuando se aplica al cuerpo, como por ejemplo, la conductividad de la piel con un sensor de respuesta galvánica de la piel [20].

- **Sensores de temperatura:** funcionan como una cámara infrarroja. Se captura la energía infrarroja y se transforma en una señal digital que representa la temperatura. Los sensores de temperatura de la piel generalmente se colocan a una distancia muy corta o en contacto directo con la piel. La miniaturización de la tecnología ha permitido el desarrollo de pequeños sensores de temperatura de la piel que pueden incorporarse en casi cualquier dispositivo electrónico, como los dispositivos ponibles [21].

- **Sensores de sonido:** un micrófono traduce las ondas de sonido que viajan por el aire en una señal eléctrica. Hay micrófonos comerciales que están preparados para capturar la voz humana a una distancia razonable (60dB a 1 metro), ya que la voz de una persona se define por las características fisiológicas del sistema respiratorio de la persona [22].

- **Sensores de localización:** El Sistema de Posicionamiento Global (GPS) consta de 32 satélites y cualquier número de receptores GPS ubicados en la superficie de la Tierra. Un receptor GPS utiliza la señal de cuatro satélites de línea de visión diferentes para triangular la ubicación del dispositivo, ofreciendo sus coordenadas (longitud y latitud), proporcionando información de comportamiento solo con respecto a la ubicación del sujeto.

En este trabajo, como se explicará en el apartado 3.1, se va a trabajar con sensores portátiles de fuerza. Pero la biometría es un problema difícil en continuo estudio, con cada vez más tipos de sensores diferentes, que algún día podrán ser usados de forma complementaria para conseguir mejores resultados.

Una vez adquirida una muestra de datos del usuario mediante el sensor existen dos formas de abordar el trabajo: considerando toda la muestra adquirida o dividiendo esa muestra en marcos temporales. Los artículos encontrados trabajan de la segunda forma ya que justifican que de esta manera se captura la variabilidad del individuo con el tiempo, pero dependiendo de cuál lo hace de diferente forma. Todos ellos consideran ciclos de marcha y que la forma de caminar humana

es un movimiento periódico, compuesto por un paso de la pierna derecha y un paso de la pierna izquierda. Es decir, un ciclo de marcha empieza cuando un pie toca el suelo y termina cuando el mismo pie toca el suelo nuevamente como se muestra en la figura 2.1. Dentro de los trabajos leídos cabe destacar [23] por utilizar solamente la dimensión Z del acelerómetro para hacer la partición del ciclo de la forma de andar, ya que afirma existir una asociación entre la fuerza de reacción del suelo y la fuerza de la señal de este eje, que forma picos de gran magnitud y busca esos cambios del eje Z para dividir la señal en ventanas. Otros como [12] utilizan el periodo de la señal para detectar los ciclos y hacer la división. Por último, en [24] se hace una revisión extensa del enfoque de ventanas, mostrando el tamaño utilizado, en segundos, de distintas publicaciones en las que se realizan diversas actividades, no solo la de caminar y se sitúan distinto número de acelerómetros en distintas posiciones, que también se indica. Considera la creación de ventanas, para cada actividad, en función del flujo de datos del sensor y los cambios que se producen, pudiendo identificar dichos cambios a través de un análisis de variaciones en las características de la frecuencia de la señal; o bien detectando el contacto inicial y final del pie con el suelo a través de la aceleración lineal del pie. Introduce la superposición entre ventanas adyacentes, lo que llamaremos "solapamiento" y demuestra que su efecto es beneficioso para el reconocimiento de actividades periódicas como caminar o correr, y estáticas como estar de pie o sentado, pero de utilidad cuestionable para la detección de actividades esporádicas, en las que su naturaleza es más compleja e intercalada. La publicación [25] considera ventanas con 20 % de solapamiento y [26, 27] consideran un 50 %.

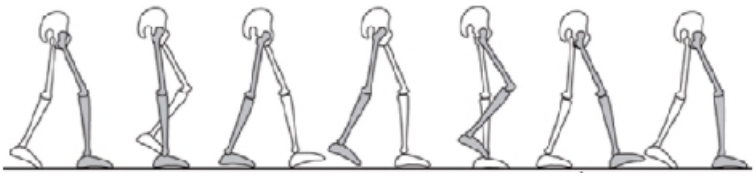


Figura 2.1: Esquema de un *ciclo de marcha*.

Por otro lado, casi nunca se utiliza la señal cruda de los datos, ya que un buen preprocesamiento puede ayudar a mejorar los resultados. Lo que todos los artículos hacen es eliminar el ruido, destacando [12, 28–31] por hacerlo asignando pesos a los datos a través del filtro Weighted Moving Average (WMA), en [29, 30] también se eliminan los falsos mínimos a través del ciclo medio, calculando aquellos puntos fuera del rango ($media \pm desviacion\_estandar$) o bien con filtros de la mediana como en [32, 33] o filtros *lowpass* o *highpass* para eliminar las interferencias fuera de la banda como hacen [23, 28, 34]. La segunda técnica más aplicada es la de la interpolación por tener los datos disponibles en intervalos de tiempo irregulares, [29, 30, 35] aplican una interpolación de spline, mientras que [12, 23, 28] justifican que utilizar una interpolación lineal es suficiente y más sencillo. Por último, la mayoría de los estudios analizados normalizan los datos, tanto si trabajan en el Dominio del Tiempo como si lo hacen con las amplitudes de Fourier en el Dominio de la Frecuencia, destacando [23, 29, 30, 34, 35], pero ninguno de ellos compara el efecto de lo que ocurre si no se normalizan los datos.

Una vez se ha decidido si trabajar con toda la muestra o con una división de ella en ventanas y el

preprocesado que aplicar a los datos, hay que decidir si trabajar con la señal preprocesada cruda o si realizar una extracción de características que represente al usuario. Cuando se trabaja con la señal preprocesada cruda se suele aplicar el método de Dynamic Time Warping (DTW) para obtener la distancia entre las ventanas o las señales, como ocurre en [7]. Pero la mayoría de las investigaciones se centran en la extracción de características, utilizando tanto el Dominio del Tiempo como el Dominio de la Frecuencia. En el Dominio de la Frecuencia, lo más utilizado es la transformada de Fourier, donde destacan las publicaciones [12, 28, 32, 34, 36]. No obstante también se aplica la Transformada del Coseno Discreta (DCT) en [7, 12, 18] donde se utilizan sus coeficientes para intentar representar al usuario. En el Dominio del Tiempo lo más utilizado son medidas estadísticas como la media y la mediana en [29,30,37,38], la desviación estándar en [23,38,39], el mínimo y el máximo en [18,28,38,40,41]. También se extraen otras características como la curtosis en [18,37], el coeficiente de asimetría en [36] o los ratios medios de las componentes XY, XZ o YZ como representantes de la gravedad, en [28]. Otro tipo de características incluidas son las correlaciones en [12,34,35] o la información sobre los ángulos de los ejes en [40]. Respecto a las características, cabe destacar [40], donde se demuestra como la precisión mejora cuando se agregan valores estadísticos adicionales al vector de características y se indica que el vector óptimo es aquel que contiene tantos pocos datos como sea posible sin perder ningún criterio de información discriminativo. Entre las técnicas para reducir la dimensión del vector de características destacan Principal Component Analysis (PCA), Independent Component Analysis (ICA), Linear Discriminant Analysis (LDA) y Support Vector Machines (SVM), por aplicarse en [32], mientras que PCA también se aplica en [23].

Se ha hecho un resumen de varios artículos en los que se han utilizado dispositivos wearables, para los que se va a indicar la siguiente información. En todos los casos se está utilizando el sensor de un teléfono móvil (smartphone).

- **Técnica ML:** Técnica de aprendizaje automático utilizada.
- Referencia al **artículo**, junto con el nombre del autor y el año.
- Tipo de **sensor** utilizado entre los explicados anteriormente y que nos interesan, que son los de fuerza.
- **Modo** de reconocimiento biométrico: identificación (I) o verificación (V).
- Los resultados, indicando la tasa de equierror **(EER)**, la precisión **(H)** y/o la tasa de falsos positivos **(FPR)**, según qué información esté disponible.
- Número de **sujetos** en la Base de Datos.
- El número de **características** utilizadas.

Los algoritmos de Machine Learning usados en biometría intentan resolver dos problemas diferentes.

- **Identificación biométrica:** modo I, es un problema de clasificación multiclase.
- **Verificación biométrica:** modo V, es un problema de clasificación de una clase. Más frecuente en los sistemas portátiles.

La salida de los algoritmos ejecutados es un valor numérico que mide el grado de similitud entre la señal consultada y un sujeto registrado. Después de obtener este resultado, generalmente se aplica un umbral t para determinar la decisión final. La variación de t ajusta las tasas de falsos positivos y falsos negativos (FPR y FNR, respectivamente), generando lo que se llama las curvas Receiver Operating Characteristic (ROC). Las métricas más usadas en biometría y que se van a utilizar en el estudio de los diversos artículos son además de la tasa FPR, la tasa de equierror (EER) que es el punto de la curva ROC en el que FPR es igual a FNR, y una medida más general de la precisión (H) que corresponde con el número de veces que el sistema produce la decisión correcta.

Hay muchas técnicas pero las más frecuentes para este tipo de problemas, y cuyos resultados aparecen en la tabla 2.1 son:

- ***K*-Nearest Neighbour (KNN):** ó k-vecinos más próximos es un método de aprendizaje perezoso por almacenar vectores de características en el conjunto de entrenamiento y retrasar todo el procesamiento hasta la clasificación. Muy utilizado y popular por su simplicidad y efectividad.
- **Support Vector Machines (SVM):** ó máquinas de vectores soporte es un método estadístico que construye un hiperplano que separa de manera óptima las diferentes clases de las muestras de entrenamiento. La efectividad de SVM depende del kernel seleccionado y de un parámetro de margen que describe la influencia de una sola muestra en el hiperplano. En [39, 42] los investigadores utilizaron el acelerómetro para verificar la identidad de los sujetos mientras realizaban gestos, pudiendo verificar la identidad del sujeto solo mientras caminaban en un ambiente muy restringido. En cambio, en [43] propusieron un sistema multimodal que consistía en acelerómetro, giroscopio y posicionamiento GPS para verificar la identidad de un sujeto, consiguiendo resultados prometedores, pero utilizando únicamente 3 personas. Por otro lado, destaca [44] por obtener el 100 % de precisión utilizando los datos del acelerómetro, pero su estrategia de clasificación no fue exactamente un sistema biométrico, sino una mezcla de identificación y verificación, con una población también muy limitada de 5 individuos.
- **Gaussian Mixture Model (GMM):** o modelo de Mixtura Gaussiana, es un modelo probabilístico que asume que todas las muestras del mismo sujeto pueden generarse por una suma ponderada de un número finito de distribuciones gaussianas. Los pesos de cada distribución y sus parámetros se obtienen a través de diferentes métodos de ajuste, por ejemplo, el más común en la literatura es el de maximización de las expectativas (EM). En un sistema de verificación, se debe establecer un umbral de probabilidad para seleccionar muestras como válidas para ese GMM. En un sistema de identificación biométrica, la muestra de consulta

se pasa a través de todos los GMM de los sujetos, y se selecciona el que tiene más probabilidad. Pero es un método más utilizado para los sonidos emitidos por el cuerpo, como por ejemplo, del corazón, donde los experimentos [45, 46] han logrado precisiones entre 0.86 y 1 con poblaciones de sujetos de tamaño medio (entre 10 y 80 individuos). También se ha empleado en verificación utilizando el acelerómetro y la respuesta galvánica de la piel [17, 18] con peores resultados de EER y FPR por encima de 0.14 en todos los casos.

- **Hidden Markov Model (HMM):** o modelo oculto de Markov es un tipo particular de red Bayesiana, donde el sistema realiza la transición de un estado a otro según las observaciones y un conjunto de probabilidades de transición que se desconocen previamente. Los HMM se han utilizado ampliamente en varios problemas de aprendizaje automático pero son especialmente conocidos por sus aplicaciones en reconocimiento de voz [47], donde se ha logrado EERs promedio de 0.10 con 48 individuos diferentes.

- **Decision Trees (DTree):** o árboles de decisión, donde cada nodo evalúa una característica y las hojas del árbol especifican la decisión a tomar. El algoritmo más utilizado y conocido se llama C4.5 [48], en su funcionamiento va (i) calculando la característica que proporciona la mayor ganancia de información en las muestras, (ii) crea un nodo de decisión utilizando el atributo que mejor divide el conjunto de datos de entrenamiento, (iii) crea listas secundarias de muestras utilizando los criterios de decisión creados y (iv) crea un Decision Trees (DTree) para todas las listas secundarias a partir del nodo de decisión. El algoritmo se detiene cuando todas las muestras en una lista secundaria pertenecen a una clase específica, que es cuando el algoritmo crea un nodo de decisión para esa clase. En [49] se encontró que los árboles de decisión podrían identificar a los sujetos con alta precisión usando datos del acelerómetro, pero utilizó una población muy pequeña de sólo 5 individuos.

Técnica ML	**Referencia**	**Sensor**	**Modo**	**EER**	**H**	**FPR**	**Sujetos**	**Caract.**
SVM	[Casale et al. 2012] [42]	ACC	V	-	-	0.01	20	18
	[Hestbek et al. 2012] [39]	ACC	I	0.1	-	-	36	12
	[Ho et al. 2012] [44]	ACC	V	-	1	-	36	-
	[Sugimori et al. 2011] [49]	ACC	I	-	0.98	-	5	2
GMM	[Lu et al. 2014] [18]	ACC	V	0.14	-	-	12	87
	[Meharia and Agrawal 2015] [17]	ACC	V	-	0.8	0.14	10	-
KNN	[Nickel et al. 2012] [50]	ACC	I	-	0.82	-	36	52
HMM	[Nickel et al. 2011] [51]	ACC	I	0.1	-	-	48	26
DTree	[Sugimori et al. 2011] [49]	ACC	I	-	0.98	-	5	2

Tabla 2.1: Resultados de artículos utilizando un smartphone.

En la tabla 2.1, las máquinas de vectores soporte consiguen buenos resultados, tanto en el

caso de identificación como de verificación, pero utilizando Bases de Datos muy pequeñas, con un máximo de 36 individuos. Destaca [49], ya que consigue una precisión alta de 0.98; y cuando se vuelve a utilizar con árboles de decisión consigue los mismos resultados, pero sigue ocurriendo lo mismo, con 5 individuos y 2 características no se pueden extraer muchas conclusiones. El modelo de Mixtura Gaussiana (GMM) parece obtener peores resultados que SVM, además llama la atención [18], ya que con 12 individuos se están usando 87 características, pudiendo existir un problema de sobreajuste que este perjudicando a los resultados. K-vecinos más próximos consigue una precisión de 0.82, peores resultados que SVM, pero de nuevo se están utilizando muchas más características, por lo que podría existir también aquí un problema de sobreajuste. Por último, en el modelo oculto de Markov (HMM) se obtiene una tasa de equierror razonable, de 0.1, y es la Base de Datos que contiene más sujetos, de un total de 48, con un número de características más pequeño que antes, de 26. Sin duda y como se puede ver en la tabla, el sensor más utilizado en la literatura es el acelerómetro, aunque hablan sobre el giroscopio, no lo utilizan tanto para obtener resultados. Pero existen artículos como [52], que comparan ambos sensores utilizando señales PPG durante la realización de ejercicio físico. Estas señales recogen el estado del corazón y otros órganos. Considerando que los acelerómetros por sí solos no pueden diferenciar entre aceleración debida al movimiento o a la gravedad y que las correlaciones presentes en las diferentes señales de movimiento (3-ejes del acelerómetro y los 3-ejes del giroscopio) recogen diferente información.

Un problema del uso de *smartphones* se encuentra en la ubicación del dispositivo, ya que no todos los usuarios lo llevan siempre en el mimo sitio y posición. Destaca el trabajo [31], donde se emplearon dispositivos ponibles diseñados específicamente para capturar los datos del movimiento mediante un acelerómetro, probando diferentes posiciones del dispositivo: en el pie, la cadera, el bolsillo del pantalón y la muñeca de los distintos usuarios, y se consiguieron los resultados de la tabla 2.2. Como en cada localización se utiliza distinto número de usuarios, se va a realizar una especie de EER por usuario, utilizando el cociente del EER y el número de individuos para poder compararlos. Parece que los mejores resultados se consiguen con el dispositivo en la cadera y en el bolsillo del pantalón, zonas muy similares, donde el acelerómetro podría capturar mejor el movimiento, mientras que en el tobillo los resultados son peores, probablemente porque sea una zona con más ruido, al ir demasiado en contacto con el pie y la muñeca parece capturar peor la información consiguiendo peores resultados. Pero hace falta recoger el EER medio utilizando las 4 localizaciones y el mismo número de individuos para que las conclusiones sean más realistas, ya que los 30 individuos usados para obtener el EER con el dispositivo en la muñeca pueden ser los que peores resultados estén dando por tener movimientos muy similares entre ellos.

No obstante, los sistemas biométricos rara vez alcanzan una precisión perfecta en la práctica debido a muchos factores, tales como el ruido, entrenamiento incompleto o un algoritmo de aprendizaje automática no ideal, lo cual afecta a la tasa obtenida de falsos positivos y falsos negativos. Todos estos resultados y la investigación realizada ha sido utilizando, la mayoría de las veces, smartphones o dispositivos creados para el propio propósito del estudio cuando en el presente proyecto se utilizan *wearables* comerciales, donde un algoritmo muy preciso podría drenar la batería rápidamente o tomar demasiado tiempo para tomar una decisión, lo cual no es factible.

Localización dispositivo	EER	Nº individuos	EER por individuo
Tobillo	5 %	21	0.24
Cadera	13 %	100	0.13
Bolsillo del pantalón	7.3 %	50	0.15
Muñeca	10 %	30	0.33

Tabla 2.2: Resultados cambiando la posición del dispositivo.

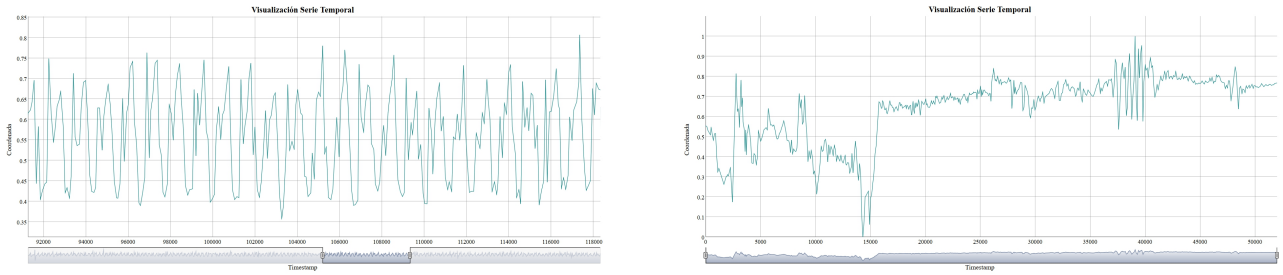
Capítulo 3

Datos y análisis

El tiempo invertido en conocer los datos, su naturaleza y el problema que se quiere resolver, junto con el dedicado a la limpieza de los datos y a garantizar una cierta calidad de estos, es realmente la parte inicial más importante y que puede llevar a finalizar con éxito o fracaso un proyecto.

Se va a realizar un análisis de autocorrelación, buscando ver si existe periodicidad y/o patrones concretos en los datos y si estos, se repiten a lo largo de las diferentes muestras, correspondientes al mismo o distinto usuario.

Lo ideal sería que todas las muestras de datos tomadas para todos los usuarios fuesen periódicas, con el mismo patrón repetido entre todas las muestras correspondientes al mismo usuario, mientras que aquellas que corresponden a otro usuario distinto, tuvieran otro patrón periódico pero diferente. Un ejemplo de serie periódica puede verse en la figura 3.1 (a). El caso opuesto sería una serie no periódica, cuyo ejemplo se puede ver en la figura 3.1 (b).



(a) Periódica (b) No Periódica

Figura 3.1: Tipos posibles de series de datos.

3.1. Base de datos

Se ha utilizado una Base de Datos que había sido recogida durante la realización de un TFG de Ingeniería Informática anterior [5], con el objetivo de no perder tiempo en recoger nuevos datos y

centrarse en el análisis de los ya disponibles. En ella se disponía de 21 usuarios. Cada uno realizó un recorrido andando de minuto y medio, aproximadamente, durante 2 sesiones en diferentes días. Y dependiendo del usuario, cada día realizó el recorrido una única vez o dos.

En la tabla 3.1 se muestra la información disponible de cada uno de los usuarios: 13 hombres, 7 mujeres y 1 usuario sin identificar en edades comprendidas entre 16 y 57 años, utilizando el reloj y la pulsera en la mano dominante o la opuesta en función del usuario. En la misma tabla se muestra el número de datos recogidos de cada usuario. El número total de datos es 66.[1]

*Usuarios que en lugar de realizar el recorrido una vez en cada sesión, lo realizaron las dos veces en la misma sesión: la primera.

Usuario	Sexo	Edad	Mano dominante	Mano portadora	Nº de datos
usuario0	-	-	-	-	2
usuario1	Hombre	21	Derecha	Izquierda	4
usuario2	Hombre	57	Derecha	Izquierda	4
usuario3	Hombre	50	Derecha	Izquierda	4
usuario4	Hombre	50	Derecha	Izquierda	2
usuario5	Mujer	53	Derecha	Izquierda	2*
usuario6	Hombre	21	Derecha	Derecha	2
usuario7	Mujer	16	Derecha	Izquierda	2
usuario8	Mujer	56	Derecha	Derecha	4
usuario9	Mujer	46	Derecha	Izquierda	4
usuario10	Mujer	19	Derecha	Izquierda	4
usuario11	Mujer	46	Derecha	Derecha	4
usuario12	Hombre	16	Derecha	Derecha	4
usuario13	Hombre	49	Derecha	Derecha	4
usuario14	Hombre	20	Derecha	Izquierda	4
usuario15	Hombre	22	Derecha	Derecha	4
usuario16	Mujer	48	Derecha	Izquierda	2
usuario17	Hombre	53	Derecha	Derecha	2*
usuario18	Hombre	22	Derecha	Izquierda	4
usuario19	Hombre	23	Derecha	Derecha	2
usuario20	Hombre	21	Derecha	Izquierda	2
TOTAL MUESTRAS DE DATOS DISPONIBLES					66

Tabla 3.1: Metadatos de los usuarios en la Base de Datos inicial.

Se dispone de la misma cantidad de datos en los 2 dispositivos comerciales y con 2 sensores en cada uno de ellos. Los dispositivos son un reloj *Motorola Moto 360* (Moto) y una pulsera

[1]El número de datos disponibles son 2 si realizó el recorrido una única vez en cada una de las dos sesiones o 4 si lo realizó dos veces, salvo las excepciones marcadas con asterisco.

Microsoft Band 2 (Micro) que habían sido ya adquiridos para un TFG anterior [4] cuyo objetivo fue desarrollar la aplicación móvil de recogida de datos, que fue empleada posteriormente en otro TFG [5], cuyos datos van a ser utilizados más ampliamente en este trabajo. Los dispositivos se pueden ver en la figura 3.2.

(a) Microsoft Band 2 (b) Motorola Moto 360

Figura 3.2: Dispositivos disponibles.

Los sensores utilizados son tanto el acelerómetro (ACC) como el giroscopio (GYR) tridimensional que poseen los dispositivos usados en la captura.

- **Acelerómetro:** mide la orientación de una plataforma fija respecto a la superficie terrestre. En esta situación podría verse como la rapidez con que algo se acelera.
- **Giroscopio:** mide la velocidad de rotación sobre un eje determinado.

Las 3 componentes son X, movimiento hacia la izquierda o derecha; Y, movimiento hacia adelante o hacia atrás; Z, movimiento hacia arriba o hacia abajo.

De manera resumida, al realizar cada recorrido, se va guardando en la Base de Datos la siguiente información.

- Identificador del usuario.
- International Mobile Equipment Identity (IMEI) del teléfono móvil o la herramienta utilizado para la adquisición de los datos. El IMEI es un código que identifica al aparato de forma exclusiva a nivel mundial.
- Dispositivo que se está utilizando (Micro o Moto).
- Tipo de sensor al que pertenece el dato (ACC o GYR).
- Timestamp: contiene tanto la fecha, como la hora con una precisión en milisegundos.
- Las coordenadas X, Y y Z del sensor indicado.
- Nombre del usuario.

- Número de la tarea, la sesión y la muestra para distinguir entre las diferentes tomas de datos del mismo usuario.

Con ello, se construye un fichero en formato CSV para cada toma de datos de cada usuario. El fichero contiene únicamente la información necesaria, que se va a utilizar a lo largo de este trabajo.

- Una primera columna con el tiempo relativo, que es la diferencia de tiempo entre una captura de las coordenadas X, Y, Z y la anterior. Los datos se almacenan con este valor temporal porque es más compacto que almacenar el timestamp.
- Tres columnas para las coordenadas X, Y, Z correspondientes a la captura de datos que marque el tiempo relativo. En la Base de datos tienen el nombre de dato1, dato2 y dato3 para hacer referencia a las coordenadas X, Y, Z respectivamente.

El recorrido dura, aproximadamente, minuto y medio, por lo que se tienen bastantes capturas para cada toma de datos de cada usuario. En la figura 3.3 se muestra un ejemplo de toma de datos, con el formato final con el que se va a trabajar.

	A	B	C	D	E
1	tiempoRelativo	dato1	dato2	dato3	
2	0	0,800781	-0,551514	0,602539	
3	68	1,062012	-0,692627	0,31665	
4	111	1,089844	-0,79248	0,296143	
5	102	0,954346	-0,553955	0,223877	
6	105	0,973145	-0,55835	0,19873	
7	94	1,097656	-0,663818	0,200439	
8	83	0,681885	-0,542725	0,268799	
9	148	1,078857	-0,369629	0,190674	
10	63	0,808838	-0,330078	0,204102	
11	71	0,764648	-0,465332	0,207764	
12	118	0,772461	-0,49585	0,30127	
13	99	0,766357	-0,543213	0,398438	
14	79	0,880615	-0,638672	0,471191	
15	136	0,775879	-0,593506	0,387939	
16	51	0,700195	-0,594727	0,356201	
17	101	0,719727	-0,477051	0,350586	
18	121	0,575928	-0,40625	0,341309	
19	79	0,584229	-0,411133	0,350098	

Figura 3.3: Formato de los datos que se van a utilizar.

Se ha realizado un análisis visual de los datos, corrigiendo anomalías. En ambos dispositivos y sensores ocurre lo mismo, presencia de valores de tiempo relativo negativo que se manifiestan en determinadas ocasiones, desconocidas. Se ha diseñado un algoritmo para eliminarlos, tras verse que al construir el tiempo relativo acumulado y eliminar esos valores negativos, la señal se reconstruía perfectamente. También, se ha detectado que el usuario0 era ficticio para probar la aplicación móvil. Si se quiere conocer más sobre este estudio, se encuentra en el proyecto del Grado en Ingeniería Informática realizado en paralelo y como complemento a este [6].

3.2. Análisis de autocorrelación de los datos

El análisis de autocorrelación de los datos se va a utilizar para comprobar si existe o no periodicidad en la señal. La autocorrelación es un coeficiente que mide la correlación cruzada de una señal consigo misma [5, 53]. Se puede relacionar una señal periódica con una señal con coeficiente de autocorrelación en valor absoluto alto y, en consecuencia, una señal que tiene un patrón y que es susceptible de poder ser usada en biometría.

Más concretamente, aquí pretendemos reconocer al usuario mediante su forma de andar, por lo que buscamos en la señal adquirida patrones que se repitan a cada paso. Eso significa que debemos encontrar valores de autocorrelación altos para ventanas de tamaño temporal aproximado de un segundo, que es el tiempo que dura, aproximadamente, un paso al andar.

La fórmula empleada para calcular la autocorrelación de un proceso discreto X con n observaciones $X_1, X_2, ..., X_n$ es (3.1).

$$R(k) = \frac{1}{n \cdot \sigma^2} \cdot \sum_{t=1}^{n-k} (X_t - \mu) \cdot (X_{t+k} - \mu) \tag{3.1}$$

Donde n es el número de muestras de la señal, X_t es el valor t-ésimo de X, μ y σ^2 representan respectivamente la media y la varianza de los valores de X, y el entero positivo $k<n$ es el desfase o desplazamiento en número de muestras para el cual queremos calcular la autocorrelación.

Los valores de *R(k)* están acotados entre -1 y 1. Un valor de autocorrelación de 1 indica que existe una correlación perfecta, mientras que un valor de -1 indica que hay una anticorrelación perfecta. Por otro lado, si el valor de autocorrelación es 0, indicará ausencia de correlación. Idealmente, se tienen que obtener valores altos en valor absoluto que indicarán que la señal sigue un patrón que se repite a lo largo de la misma.

El cálculo de la autocorrelación se puede ver como si dividiéramos la señal en ventanas de tamaño k muestras y obtuviéramos cuánto se parecen esas ventanas entre sí. La relación entre el valor de k y la duración de la ventana en segundos, τ, la podemos ver en la fórmula (3.2), donde T es el periodo de muestreo de la señal. Como la señal se muestreó a una frecuencia de $Fm = 12s.$, el valor de T ($\frac{1}{Fm}$) es de, aproximadamente, 83 ms.

La duración de un paso andando es de, como se ha comentado, aproximadamente un segundo, valor alrededor del cual se deben obtener los mayores valores de autocorrelación. Para comprobar esto, ese coeficiente se ha calculado para valores de k entre 11 y 16, que despejando en (3.2) nos da tamaños de ventana, τ, en milisegundos de entre 830 y 1245.

$$k = \frac{\tau}{T} + 1 \tag{3.2}$$

Daniel González Alonso, autor del TFG anterior [5] tomado como referencia realizó también un estudio de la autocorrelación. En su caso, analizó cada usuario en los dos dispositivos: Micro y Moto y los dos sensores: ACC, GYR, probando valores de desfase, τ, entre 500 y 1496 ms, con

saltos de 83 ms, verificándose que el periodo con el máximo valor de autocorrelación era 998 ms. La tabla 3.2 muestra un ejemplo del trabajo que realizó. En ellas marco en color rojo y con fondo verde los valores de autocorrelación superiores a 0.6.

τ	Autocorrelaciones											
	Micro						Moto					
	ACC			GYR			ACC			GYR		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
500	0,19	−0,83	−0,50	−0,61	−0,82	−0,83	0,33	−0,73	−0,28	−0,64	−0,68	−0,74
583	−0,13	−0,69	−0,45	−0,56	−0,68	−0,69	0,09	−0,62	−0,22	−0,55	−0,51	−0,62
666	−0,38	−0,40	−0,29	−0,38	−0,40	−0,42	−0,30	−0,39	−0,10	−0,38	−0,23	−0,39
749	−0,39	−0,03	−0,09	−0,11	−0,07	−0,07	−0,50	−0,11	−0,04	−0,18	−0,03	−0,14
832	−0,15	0,35	0,18	0,21	0,29	0,31	−0,35	0,18	0,00	0,06	0,12	0,13
915	0,29	0,67	0,49	0,53	0,65	0,67	0,12	0,49	0,20	0,38	0,37	0,48
998	0,71	0,84	0,71	0,77	0,88	0,90	0,64	0,72	0,42	0,70	0,67	0,81
1081	0,72	0,83	0,71	0,76	0,86	0,88	0,76	0,77	0,44	0,81	0,73	0,89
1164	0,26	0,62	0,47	0,48	0,61	0,62	0,41	0,63	0,26	0,61	0,48	0,65
1247	−0,17	0,29	0,15	0,18	0,25	0,24	−0,12	0,35	0,07	0,29	0,14	0,27
1330	−0,38	−0,09	−0,12	−0,11	−0,11	−0,13	−0,49	0,05	−0,07	0,01	−0,08	−0,05
1413	−0,36	−0,45	−0,33	−0,38	−0,43	−0,46	−0,47	−0,23	−0,10	−0,22	−0,23	−0,30
1496	−0,11	−0,70	−0,47	−0,55	−0,67	−0,70	−0,17	−0,50	−0,22	−0,43	−0,42	−0,51

Tabla 3.2: Autocorrelaciones obtenidas para la muestra 1 y sesion 1 del usuario19.

3.2.1. Resultados del análisis de autocorrelación

En este trabajo se ha repetido el análisis de autocorrelación, pero de manera más detallada y centrada en el procesamiento que aquí vamos a realizar y que veremos de manera pormenorizada más adelante.

Como se verá, una de las cosas analizadas en el presente trabajo es el parámetro "tamaño de la señal usada para reconocer al usuario". Vamos a explicarlo un poco más aquí, aunque, como se ha comentado, volveremos más adelante sobre ello.

De cada individuo hemos adquirido distintas muestras de duración aproximada un minuto y medio. Una de las etapas más importantes de cualquier sistema biométrico es la extracción de características, que serán las que representen al usuario. Pues bien, a la hora de abordar esta extracción de características se puede hacer de dos maneras: usando toda la muestra adquirida o dividiendo esa muestra en marcos temporales; en el primer caso tendremos un único vector de características por muestra, en el segundo, una secuencia de ellos. Dada la naturaleza variable de cualquier rasgo biométrico, es importante poder capturar esa variabilidad, por lo que aquí hemos optado por la segunda opción: dividir la muestra en marcos temporales o ventanas. Como suele ser habitual en biometría, para obtener una secuencia de vectores de características más "suavizada", esas ventanas se suelen solapar.

Esta forma de proceder hace que aparezca, como parámetro que puede influir en el rendimiento del sistema, el tamaño de esas ventanas, por lo que tendrá que ser analizado. Aquí vamos a mostrar parte de ese estudio, el relacionado con la autocorrelación.

De las distintas opciones que barajamos para realizar la división de la señal en ventanas, optamos por tomar como criterio el número de ciclos; entendiendo aquí como "ciclo" el patrón de la señal que se repite en el tiempo. La duración de ese ciclo se obtuvo para cada usuario mediante el análisis de autocorrelación indicado en la sección anterior; como se vio, es de aproximadamente un segundo, con ligeras variaciones dependiendo del usuario, dispositivo y coordenada del sensor.

Este análisis se ha realizado para la señal completa, sin eliminar el ruido ni realizar ningún tipo de preprocesamiento. Se ha calculado la autocorrelación de cada marco temporal en que se divide la señal, para distintos valores de tamaño de ventana de autocorrelación, como se hizo en el anterior apartado, es decir, modificando el valor de k en la ecuación (2). Los tamaños de ventana de autocorrelación probados fueron de $830ms. \leq \tau \leq 1245ms$, con saltos de 83 ms. En las tablas se mostrará el valor máximo de autocorrelación con respecto a k.

Para simplificar los resultados y conseguir una mejor visualización, se van a utilizar tablas que van a mostrar por cada fila, un usuario junto con el valor de la ventana con máxima y mínima autocorrelación en valor absoluto de cada una de sus componentes, teniendo en cuenta todas las sesiones y muestras disponibles de ese usuario.[2] Estos valores son representativos de lo que se quiere analizar aquí, presencia de señales periódicas con poco ruido. Por lo que se va a hablar de *mejores resultados* cuando la autocorrelación sea más alta, ya que indicará que la señal es más periódica. Los números van a estar redondeados a un máximo de 4 cifras decimales y marcados en rojo cuando el valor de la autocorrelación en valor absoluto sea inferior a 0.6.

3.2.1.1. Microsoft Acelerómetro

Utilizando como dispositivo la pulsera *Microsoft* y como sensor el acelerómetro, se tienen las tablas 3.4, 3.5 y 3.6 correspondientes a tamaño de ventana 3, 8 y 10 ciclos respectivamente. Como se puede ver, a mayor tamaño de la ventana, los valores de la autocorrelación son más altos. Veremos que esto se repite cuando analicemos (sección 6) el rendimiento del sistema con respecto al mismo parámetro, pudiendo ya ver una relación entre autocorrelaciones altas y buenos resultados en reconocimiento. Con 3 periodos existen usuarios como el 2, 3, 4, 8, 9, 15 y 17 donde su autocorrelación mínima está próxima al 0, que es el peor resultado posible, indicativo de ausencia de correlación y, por tanto, de periodicidad.

La tabla 3.3 muestra el número de usuarios distintos, sin distinguir en cuántas sesiones y muestras, con autocorrelación máxima inferior al valor umbral fijado, viéndose de manera resumida la conclusión obtenida: a mayor tamaño de ventana, en general, o se mantienen o se obtienen mejores resultados, con autocorrelaciones más altas. Pero como se puede observar ya aquí, y se corroborará más adelante cuando se mida el rendimiento en reconocimiento (sección 6), hay un límite en el tamaño de la ventana a partir del cual ya no se observan mejoras en los valores de autocorrelación, incluso pueden empeorar en algunas de las componentes del sensor.

[2]No se han hecho tablas donde cada fila hiciera referencia a un usuario/sesión/muestra dado que el tamaño resultante era muy grande y difícil de visualizar.

MICRO ACC	X	Y	Z	Modulo
3 periodos	2	2	1	2
8 periodos	0	1	0	1
10 periodos	0	2	0	0

Tabla 3.3: Resumen del número de usuarios distintos con autocorrelaciones máximas pequeñas (Micro ACC)

Datos Enteros – 3 periodos – 20 solapamiento								
usuario	Min X	Max X	Min Y	Max Y	Min Z	Max Z	Min modulo	Max modulo
1	0.1126	0.6928	0.0316	0.6647	0.0577	0.6571	0.136	0.6998
2	0.0665	0.6394	0.0718	0.6567	0.078	0.6282	0.0917	0.657
3	0.0526	0.6868	0.0636	0.6343	0.1822	0.6815	0.1388	0.6951
4	0.0793	0.6875	0.0971	0.6598	0.0782	0.6886	0.0842	0.6853
5	0.1003	0.6495	0.1085	0.6537	0.0912	0.6285	0.1536	0.6626
6	0.1191	0.6642	0.0895	0.6531	0.0609	0.6886	0.2273	0.6671
7	0.1137	0.6714	0.0955	0.6251	0.0629	0.6202	0.0706	0.6761
8	0.0674	0.6579	0.0481	0.6867	0.0843	0.6635	0.0782	0.6555
9	0.035	0.4856	0.0623	0.4514	0.1021	0.4993	0.1097	0.4881
10	0.1368	0.683	0.0794	0.6356	0.0444	0.6363	0.1135	0.6827
11	0.0927	0.6357	0.1013	0.5919	0.0815	0.6133	0.0978	0.6338
12	0.1023	0.6571	0.1033	0.6195	0.1104	0.6851	0.0757	0.6588
13	0.1107	0.6652	0.1393	0.6768	0.0903	0.6933	0.1724	0.6675
14	0.0919	0.6889	0.0396	0.6419	0.1091	0.6367	0.0728	0.6897
15	0.0642	0.718	0.1094	0.6955	0.0921	0.6845	0.0567	0.7386
16	0.1594	0.6896	0.117	0.6493	0.1321	0.696	0.1965	0.6944
17	0.0774	0.6601	0.0598	0.6414	0.1511	0.6833	0.1091	0.6495
18	0.133	0.674	0.1163	0.6172	0.1081	0.6752	0.2185	0.6809
19	0.0968	0.666	0.0854	0.6347	0.1689	0.6451	0.0779	0.6708
20	0.0881	0.5029	0.1361	0.6093	0.2763	0.6114	0.2265	0.569

Tabla 3.4: Máxima y mínima autocorrelación con 3 periodos (Micro ACC).

Datos Enteros – 8 periodos – 20 solapamiento								
usuario	Min X	Max X	Min Y	Max Y	Min Z	Max Z	Min modulo	Max modulo
1	0.2182	0.8558	0.1628	0.8045	0.2037	0.7671	0.3427	0.8541
2	0.1319	0.8056	0.0645	0.8264	0.1638	0.8165	0.3253	0.8228
3	0.1364	0.8287	0.1207	0.737	0.3051	0.8426	0.1731	0.8311
4	0.2882	0.8497	0.1545	0.8067	0.1492	0.8028	0.459	0.8531
5	0.3127	0.765	0.3868	0.8048	0.4206	0.7898	0.3783	0.8092
6	0.1529	0.7901	0.0905	0.7292	0.1204	0.7908	0.3021	0.7995
7	0.4486	0.819	0.1356	0.7769	0.1714	0.7434	0.4495	0.8166
8	0.1044	0.8193	0.0874	0.8463	0.1044	0.8074	0.0695	0.8236
9	0.0981	0.6148	0.0797	0.4939	0.1319	0.6542	0.105	0.5969
10	0.4332	0.8438	0.1736	0.7559	0.136	0.7454	0.4767	0.8404
11	0.0782	0.6887	0.1418	0.6649	0.025	0.6503	0.1611	0.7013
12	0.4039	0.8238	0.0867	0.7499	0.3261	0.8153	0.4202	0.8238
13	0.1718	0.8162	0.1386	0.8743	0.2258	0.8859	0.4312	0.8255
14	0.2664	0.7865	0.197	0.747	0.0927	0.7773	0.2761	0.7843
15	0.2816	0.81	0.1545	0.8512	0.1706	0.8123	0.1634	0.8127
16	0.3952	0.8412	0.2986	0.8394	0.273	0.8119	0.2689	0.8468
17	0.1438	0.8113	0.0453	0.7342	0.1737	0.8427	0.2119	0.8149
18	0.2107	0.8269	0.1467	0.7215	0.2063	0.8476	0.2929	0.8294
19	0.3513	0.7904	0.215	0.7741	0.2251	0.7678	0.4067	0.7913
20	0.1721	0.6547	0.2366	0.7739	0.3872	0.7881	0.148	0.6724

Tabla 3.5: Máxima y mínima autocorrelación con 8 periodos (Micro ACC).

Datos Enteros – 10 periodos – 20 solapamiento								
usuario	Min X	Max X	Min Y	Max Y	Min Z	Max Z	Min modulo	Max modulo
1	0.3795	0.8692	0.2256	0.8244	0.2544	0.7911	0.3987	0.8779
2	0.1041	0.8367	0.1175	0.8191	0.1475	0.7463	0.392	0.8383
3	0.167	0.8518	0.1475	0.7758	0.4366	0.8571	0.3	0.8572
4	0.3422	0.8796	0.2149	0.7962	0.1764	0.814	0.58	0.8815
5	0.3332	0.7277	0.4201	0.7706	0.315	0.796	0.3757	0.7727
6	0.1781	0.8125	0.1681	0.7309	0.095	0.7694	0.3175	0.8294
7	0.5222	0.8573	0.1129	0.805	0.1385	0.7609	0.528	0.8594
8	0.056	0.8229	0.1176	0.8734	0.1342	0.8368	0.0776	0.8247
9	0.1082	0.6392	0.0732	0.551	0.1342	0.6336	0.0962	0.6237
10	0.4931	0.8538	0.2012	0.7713	0.124	0.7715	0.4901	0.8617
11	0.128	0.638	0.1169	0.5457	0.1384	0.657	0.0907	0.6679
12	0.4217	0.7895	0.1691	0.7524	0.2935	0.7716	0.4157	0.7883
13	0.2008	0.8221	0.1687	0.8541	0.295	0.8657	0.4662	0.8225
14	0.3047	0.8196	0.2035	0.7874	0.2475	0.7831	0.3533	0.8188
15	0.2765	0.8557	0.2411	0.8705	0.2423	0.8502	0.3231	0.8533
16	0.5102	0.8578	0.3311	0.841	0.3196	0.8331	0.3932	0.8602
17	0.1771	0.8273	0.1026	0.757	0.1722	0.8607	0.2906	0.8296
18	0.3202	0.8395	0.0899	0.7058	0.2844	0.8675	0.3468	0.846
19	0.4372	0.8216	0.2254	0.793	0.1023	0.7618	0.5071	0.8182
20	0.1781	0.6735	0.3242	0.7653	0.4558	0.7661	0.178	0.6743

Tabla 3.6: Máxima y mínima autocorrelación con 10 periodos (Micro ACC).

3.2.1.2. Microsoft Giroscopio

Al cambiar de sensor dentro del mismo dispositivo, los resultados son similares, consiguiendo autocorrelaciones más altas y por tanto, mejores resultados, con tamaños de ventana más grandes. En la tabla 3.7 se muestra lo mismo que antes, el número de usuarios distintos con autocorrelaciones máximas inferiores a 0.6. Dado que se está haciendo sobre los datos originales, se puede ver que la componente Y y la Z son más periódicas en el giroscopio que en el acelerómetro; mientras que con la componente X y el módulo ocurre justo lo contrario. Esto podría ser indicativo de que una combinación de ambos sensores podría ser apropiada en esta biometría.

MICRO GYR	X	Y	Z	Modulo
3 periodos	2	1	1	4
8 periodos	2	0	0	2
10 periodos	2	0	0	2

Tabla 3.7: Resumen del número de usuarios distintos con autocorrelaciones máximas pequeñas (Micro GYR)

3.2.1.3. Motorola Acelerómetro

El reloj *Motorola* muestra peores resultados, con valores de autocorrelación en ambos sensores, más pequeños. La tabla 3.8 está calculada de la misma manera que en el otro dispositivo y muestra que a mayor tamaño de ventana mejores resultados, pero no tan buenos como conseguía la pulsera de *Microsoft*, y por tanto, los datos de este dispositivo parecen tener más ruido.

MOTO ACC	X	Y	Z	Modulo
3 periodos	5	7	5	5
8 periodos	2	2	1	3
10 periodos	3	2	1	3

Tabla 3.8: Resumen del número de usuarios distintos con autocorrelaciones máximas pequeñas (Moto ACC)

3.2.1.4. Motorola Giroscopio

Al cambiar de sensor en el dispositivo de *Motorola*, la cosa mejora en dos componentes, la Y y la Z. Mientras que la X y el módulo siguen siendo más ruidosas y peores en más usuarios, como indica la tabla 3.9. Sin embargo, las conclusiones son las mismas en ambos dispositivos y sensores:

- A mayor tamaño de ventana, en general, o se mantienen o se obtienen mejores resultados, consiguiéndose autocorrelaciones más altas y por tanto ventanas de usuarios más periódicas.
- Existe un límite en el tamaño de la ventana a partir del cual ya no se observan mejoras en los valores de autocorrelación, pudiendo empeorar en algunas componentes.
- Una combinación de ambos sensores podría ser apropiada, ya que en ambos dispositivos ocurre lo mismo. Se consiguen mejores resultados en la componente Y y Z del giroscopio. Mientras que la componente X y el módulo funcionan mejor en el acelerómetro.

MOTO GYR	X	Y	Z	Modulo
3 periodos	6	4	4	7
8 periodos	4	0	0	5
10 periodos	4	0	0	4

Tabla 3.9: Resumen del número de usuarios distintos con autocorrelaciones máximas pequeñas (Moto GYR)

Capítulo 4

Configuración experimental

La realización de un experimento requiere tomar muchas decisiones, las cuales pueden terminar siendo un éxito o un fracaso. A lo largo de este capítulo se explican los parámetros estáticos que se han fijado y utilizarán en las siguientes secciones junto con las razones por las cuáles se han elegido. La configuración experimental está muy relacionada con los objetivos del trabajo. Aquí no pretendemos llegar a un sistema de reconocimiento biométrico basado en ponibles, sino que, como es lógico en una primera aproximación al problema, lo que se pretende es analizar los distintos elementos que entran en juego en el sistema, para entenderlos mejor y ver su relación en el rendimiento del reconocimiento del usuario. En definitiva, se busca plantar unas bases sólidas sobre las que seguir trabajando, ya sí, en un sistema eficiente de reconocimiento.

4.1. Extracción de características

Tras un análisis detallado de los artículos y la documentación actual más relevante en el campo de la biometría, y en concreto de los dispositivos ponibles, resumen de ello se puede encontrar en el capítulo "Estado del Arte" de [6], se ha decidido dividir la señal en marcos temporales o ventanas, y sobre cada una de ellas, extraer características en el dominio del tiempo y en el dominio de la frecuencia, para estudiar las diferencias. Hasta donde conocemos, nunca se ha realizado este estudio comparativo.

Posteriormente, cuando se estudie el rendimiento final, se estudiará la influencia de hacer una selección de características para eliminar aquellas que sean irrelevantes, redundantes o altamente correlacionadas entre sí con el objetivo de mejorar los resultados y eliminar posible sobreajuste.

4.1.1. Dominio del tiempo

Los datos capturados son muestras de la evolución de una señal en el tiempo, donde se ha obtenido para cada toma de datos de cada usuario el instante de tiempo *(timestamp)* y los datos

en los sensores. Tras un análisis bibliográfico de las características extraídas directamente sobre estas señales, es decir, en el dominio del tiempo se decidió probar las siguientes:

- Periodo de cada una de las componentes en el que se consigue la máxima autocorrelación entre las pruebas hechas mencionadas en la subsección 3.2.
- Valor de la autocorrelación en cada una de las componentes donde se consigue el máximo.
- Medidas estadísticas en cada una de las componentes (X/Y/Z):
  - Media: medida de tendencia central que representa el centro de gravedad de la distribución de la variable.
  - Mediana: medida de tendencia central que representa el valor de la variable en la posición central entre un conjunto de datos ordenados.
  - Máximo: el valor más grande entre el conjunto de valores.
  - Mínimo: el valor más pequeño entre el conjunto de valores.
  - Desviación estándar: medida de dispersión que se utiliza para cuantificar la variación de un conjunto de datos.
  - Rango: intervalo entre el valor máximo y el valor mínimo, proporcionando una idea de la dispersión de los datos.
  - Kurtosis: característica de forma de la distribución de probabilidad/frecuencias de los datos que indica que tan apuntada o achatada se encuentra una distribución respecto a un comportamiento normal (distribución normal). Valores grandes indican mayor concentración de valores de la variable tanto muy cerca de la media de la distribución (pico) como muy lejos de ella (colas), al tiempo que existe una relativamente menor frecuencia de valores intermedios, no implicando con ello una mayor varianza [54].
  - Quantil 25 % y 75 %: puntos tomados a intervalos regulares de la función de distribución de la variable aleatoria. Lo que se ha utilizado es dividir la distribución en cuatro partes correspondientes a los cuantiles 25 %, 50 % (media) y 75 %.
  - Coeficiente de asimetría: representa el grado de simetría (o asimetría) de la distribución de probabilidad de la variable aleatoria. Considerando como eje de simetría la recta paralela al eje de ordenadas que pasa por la media de la distribución, una distribución es simétrica si existe el mismo número de valores a la derecha que a la izquierda de la media y por tanto el mismo número de desviaciones con signo positivo que con signo negativo. Mientras que hay asimetría positiva si hay valores más separados de la media por la derecha y asimetría negativa si hay valores más separados de la media por la izquierda [55].
- Energía conjunta de las 3 componentes: proporciona una idea de la disposición de los datos tridimensional.

$$Energia = \frac{1}{N} \cdot \sum_{n=1}^{N} (\sqrt{X[n]^2 + Y[n]^2 + Z[n]^2})^2 \tag{4.1}$$

- Ratio medio de las componentes XY: media de todos los cocientes de X e Y. Proporciona una medida de la gravedad de la distribución conjunta de ambas componentes de la variable.
- Ratio medio de las componentes XZ: media de todos los cocientes de X y Z.
- Ratio medio de las componentes YZ: media de todos los cocientes de Y y Z.

4.1.2. Dominio de la frecuencia

Para pasar al dominio de la frecuencia se ha usado, como es habitual, la transformada de Fourier.

La transformada de Fourier (FT) descompone la función del tiempo original en las frecuencias que la constituyen. Es un conjunto de números complejos, cuya magnitud (módulo) representa la amplitud de cada frecuencia presente en la función original y cuyo argumento es el desfase de la onda.

Para que los resultados de la transformada de Fourier sean interpretables, es decir, se pueda calcular la frecuencia de cada componente extraído, la señal original debe estar muestreada a frecuencia constante. De estudios realizados en trabajos anteriores [5], se vio que muestrear a más de 12 Hz. era innecesario; se fijó ese valor para la frecuencia de muestreo. El problema que tenemos con la señal capturada es que, al ser dispositivos reales, el periodo entre dos muestras consecutivas no es constante. Por lo tanto, lo primero que se tuvo que hacer para aplicar la FT es un remuestreo de la señal a 12 Hz. La bibliografía muestra distintas alternativas, de las cuales la más sencilla es usar la interpolación lineal, que se explicará en el apartado 5.2. Alternativa que suele dar un rendimiento similar o superior a algoritmos más complejos, por lo que fue la técnica de remuestreo usada aquí.

La ecuación (4.2) muestra la expresión de la Transformada de Fourier Discreta (TDF) [56].

$$X_k = \sum_{n=0}^{N-1} X_n exp\left(\frac{-i2\pi kn}{N}\right) \tag{4.2}$$

Donde:

- X_k: cantidad de frecuencia k en la señal; cada valor k^{th} es un número complejo que incluye amplitud (fuerza) y cambio de fase.
- N: número de muestras.
- n: muestra, $n\epsilon\{0...N-1\}$.
- k: frecuencia entre *0 Hz.* y *N-1 Hz.*
- $1/N$: tamaño real de los picos de tiempo.
- n/N: porcentaje de tiempo que ha pasado.

- $2\pi k$: velocidad en *radianes/segundo*.
- $\exp^{-ix}$: camino circular hacia atrás que indica cuánto nos hemos movido, para esta velocidad y tiempo.

El resultado de la TDF, X_k, es un número complejo del que nos interesa su amplitud (la amplitud carece de información biométrica) y la frecuencia asociada. Ambos valores se calculan de la siguiente manera:

- $Amplitud = \sqrt{Re(X_k)^2 + Im(X_k)^2}$ donde Re e Im son la parte real e imaginaria del número complejo X_k.
- $Frecuencia = 0 : (length(X_k) - 1) * f/length(X_k)$

La teoría dice que la división de una señal en ventanas tiene un efecto sobre sus componentes frecuenciales. Este efecto es mayor cuando la TDF se aplica sobre la ventana extraída directamente de la señal. Para suavizar o paliar este efecto, se suelen usar las denominadas *funciones ventana*. Estas son funciones matemáticas que tratan de evitar discontinuidades al principio y al final de la señal. Al multiplicar la señal $s(t)$ por una función ventana $h(t)$, se genera $S_h(t) = s(t) \cdot h(t)$ que será la señal sobre la que se aplique la TDF.

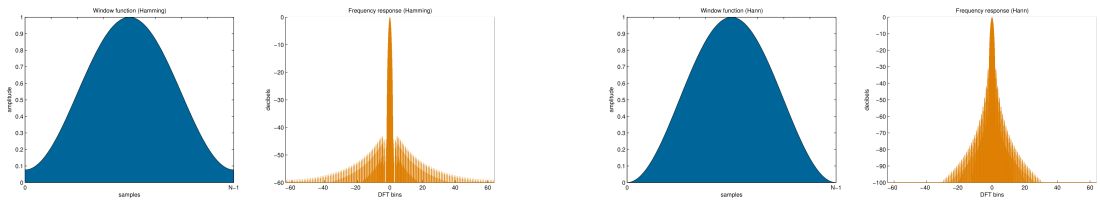
Hay muchos tipos de ventanas, las más usadas son:

- Hamming: se define con la función matemática de (4.3) y tiene la forma que muestra la figura 4.1 (a).

$$v(n) = 0.53836 - 0.46164 \cdot cos\left(\frac{2\pi n}{N-1}\right) \tag{4.3}$$

- Hanning: se define con la función matemática de (4.4) y tiene la forma que muestra la figura 4.1 (b).

$$v(n) = 0.5 - 0.5 \cdot cos\left(\frac{2\pi n}{N-1}\right) \tag{4.4}$$



(a) Hamming Window (b) Hanning Window

Figura 4.1: Ventanas de suavizado aplicadas a la transformada de Fourier.

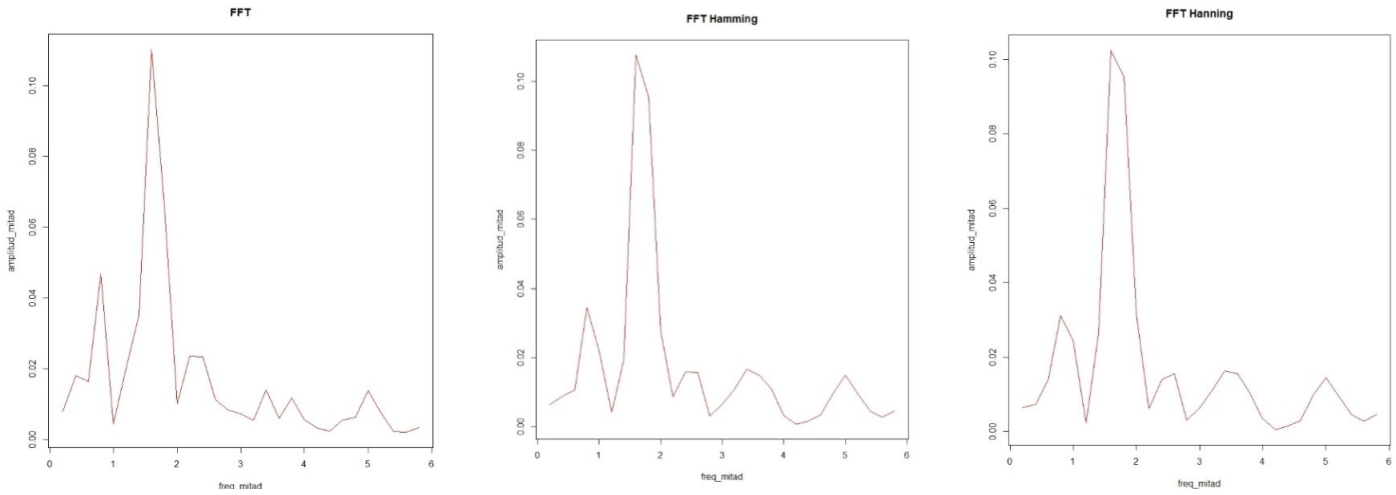


Figura 4.2: Aplicación de Fourier con y sin ventanas al usuario1, S1, M1, 1ª ventana

Tras analizar las gráficas de Fourier sin aplicar las ventanas de suavizado y después aplicándolas, se ha podido ver como no existían diferencias significativas. Un ejemplo de ello puede verse en la figura 4.2. Por esta razón, la decisión ha sido no usar estas ventanas.

Para eliminar la componente de continua (componente 0 de la TDF), que no aporta información, se ha aplicado el filtro de restar la media en cada ventana, tal y como hace el artículo [27]. A partir de lo indicado en la bibliografía, se han extraído las siguientes características en cada una de las componentes X/Y/Z:

- Primera y segunda amplitud dominante: representa los dos valores más altos obtenidos entre las amplitudes resultantes del Análisis de la transformada de Fourier en cada una de las componentes de los datos.
- Primera y segunda frecuencia dominante: representa los dos valores de la frecuencia correspondientes a los dos puntos donde se consiguen las amplitudes anteriores.
- Área bajo la curva de Fourier (AUC) basado en splines: utiliza una interpolación de splines para calcular la cantidad de área bajo la curva formada por las amplitudes del Análisis de Fourier.
- Las mismas medidas estadísticas que en el dominio del tiempo, quitando el máximo y el mínimo.

4.1.3. Señal combinada

Se ha trabajado combinando la señal a través del módulo (4.5), tal y como se hace en la bibliografía [30, 31, 33, 39]. Una vez aplicado el módulo, se extraerán las características mostradas

tanto en el dominio del tiempo como de la frecuencia.

$$Modulo = \sqrt{X^2 + Y^2 + Z^2} \tag{4.5}$$

Otra alternativa, menos utilizada, es el uso del arcoseno (4.6) [35, 38]. La bibliografía muestra resultados similares al módulo, por lo que fue la alternativa probada en el presente trabajo.

$$Arcoseno = \arcsin \frac{Z}{\sqrt{X^2 + Y^2 + Z^2}} = \arcsin \frac{Z}{Modulo} \tag{4.6}$$

Otra alternativa que se barajó y estudio en este proyecto fue fusionar las coordenadas a nivel de características, es decir, creando un vector de características resultante de juntar las de las coordenadas X, Y y Z. Esto nos daba un vector de 79 características. Las pruebas prospectivas realizadas no mostraron un buen rendimiento de esta alternativa, que, junto con el más alto coste computacional debido al mayor tamaño del vector de características, nos hizo desechar esta vía de trabajo.

4.2. Medición del error

Otra decisión importante es cómo evaluar los modelos implementados con el objetivo de poder compararlos y buscar la mejor solución final.

Entre las medidas más utilizadas en los sistemas biométricos se encuentran las curvas ROC *(Receiver Operating Characteristic)*, éstas son una representación gráfica de la sensibilidad frente a la especificidad para un sistema de clasificación binario según se varía el umbral de decisión.

Nuestro problema se corresponde con el de clasificación binaria, dado que, para cada usuario, se considera a dicho usuario como auténtico y al resto como usuarios impostores.

Las medidas de error básicas usadas en este tipo de problemas son:

- Falsos positivos (False Positives o FP) o falsa aceptación: ocurre cuando se identifica a una persona no autorizada como autorizada. De manera que, si el sistema trata de verificar la identidad de una persona, un usuario impostor podría acceder de forma no autorizada.
- Falsos negativos (False Negatives o FN) o falso rechazo: ocurre cuando se impide el acceso a una persona autorizada.
- Verdaderos positivos (True Positives o TP): ocurre cuando el sistema trata de verificar la identidad de una persona y un usuario auténtico (verdadero) accede de forma correcta y es autorizada.
- Negativos verdaderos (True Negatives o TN): ocurre cuando el sistema trata de verificar la identidad de una persona y un usuario impostor es rechazado.

- Sensibilidad (True Positive Rate o TPR): proporción de usuarios auténticos que se consideran correctamente como autorizados, con respecto a todos los usuarios auténticos. En función de los términos anteriores, se puede calcular con la fórmula (4.7).

$$Sensibilidad = \frac{TruePositive}{FalseNegative + TruePositive} = \frac{TP}{FN + TP} \tag{4.7}$$

- Especificidad (False Positive Rate o FPR): proporción de usuarios impostores que se consideran erróneamente como autorizados con respecto a todos los usuarios impostores, cuyo resultado se puede obtener con la fórmula (4.8).

$$Especificidad = \frac{FalsePositive}{FalsePositive + TrueNegative} = \frac{FP}{FP + TN} \tag{4.8}$$

Tanto la sensibilidad como la especificidad tienen valores en el rango [0,1], generando una curva ROC en estos rangos donde su área se denomina AUC. Los valores de AUC se interpretan de manera que cuanto mayor sea el valor del AUC, mejor es el rendimiento del modelo.

Otra medida del rendimiento muy utilizada en biometría es la tasa de equierror, que es el punto de intersección entre ambas tasas: sensibilidad y especificidad, conocido como *Equal Error Rate (EER)*. Cuanto menor sea su valor, mejor será el sistema.

La figura 4.3 muestra la especificidad en el eje de abscisas y la sensibilidad en el eje de ordenadas, generando la curva sobre su área (AUC) marcado en gris. El valor de la tasa de equierror se produce con FPR=0.2 y TPR=0.8.

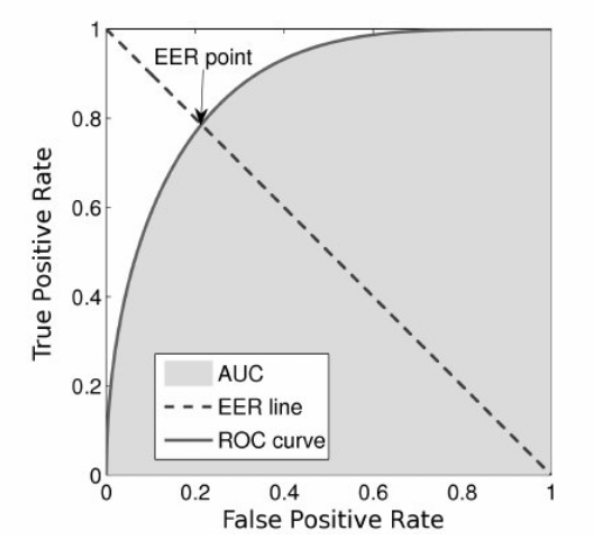


Figura 4.3: Ejemplo de EER a partir de la curva ROC y el AUC.

Como resultado final, tenemos dos maneras de mostrar el error:

- De manera gráfica: como se muestra en la figura 4.3, los valores de la sensibilidad y la especificidad para distintos valores umbrales.

- Mediante valor numérico: utilizando el área bajo la curva ROC o la tasa de equierror explicada. Pero estos valores se pueden calcular de manera individual para cada usuario o de manera global como la media de todos los usuarios disponibles.

Dado que nuestro objetivo aquí es comparar resultados, la opción gráfica es poco práctica en este caso, siendo la más habitualmente utiliza, y la que decidimos adoptar, la de utilizar valores numéricos, considerando ambas métricas, aunque mayoritariamente se va a utilizar el EER. Para la toma de decisiones se utilizará su valor medio con respecto a todos los usuarios, pero posteriormente se aprovechará la ventaja de tener pocos usuarios para hacer un estudio detallado de cada uno de ellos.

4.3. Experimentos

Teniendo en cuenta el contenido de la base de datos o corpus que estamos usando, se tienen:

- Diversos **usuarios**.
- Dos **sesiones** posibles en que se recogieron datos. S1 y S2 hacen referencia a la sesión 1 y 2 respectivamente.
- Un máximo de dos **muestras** de datos tomadas por sesión y pulsera a cada usuario, representándose como M1 y M2 para referenciar a la muestra 1 y 2 respectivamente.

Para cada usuario i vamos a tener los siguientes conjuntos de datos:

- Conjunto de entrenamiento *(train)*: contiene los datos del usuario auténtico que se usarán para crear su patrón.
- Conjunto de prueba *(test)*: distinguiendo entre:
  - **Muestras auténticas:** Serán muestras del usuario distintas a las usadas para el entrenamiento. Se usarán para calcular la tasa de falsos negativos.
  - **Muestras impostores:** Serán muestras de otros usuarios distintos al usuario i. Simularán ataques al sistema, por lo tanto, se usarán para calcular la tasa de falsos positivos.

Con respecto a la sesión y muestra, tenemos las siguientes pruebas:

1. Monosesión-Monomuestra (MonoMono): compara los datos dentro de la misma sesión y muestra, es decir, las muestras usadas para entrenamiento y para *prueba auténtico* del usuario son tomadas en la misma sesión. Es el caso más favorable y el que primero abordaremos para analizar los parámetros del sistema.

- **Train:** S1, M1, usuario i
- **Test:**
  - Test Auténticos: S1, M2, usuario i
  - Test Impostores: S1, M2, usuario j $\neq$ i

2. Multisesión-Monomuestra (MultiMono): Las muestras usadas para entrenamiento y *prueba auténtica* son tomadas en distintas sesiones. Aquí se quiere probar la variabilidad del rasgo biométrico con el tiempo.

   - **Train:** S1, M1, usuario i
   - **Test:**
     - Test Auténticos: S2, M1, usuario i
     - Test Impostores: S1, M2, usuario j $\neq$ i

3. Multisesión-Multimuestra (MultiMulti): En biometría se ha demostrado que la variabilidad del rasgo con el tiempo es un problema que afecta al rendimiento del sistema. Una forma de paliarlo es intentar incluir en el modelo del usuario esta variabilidad. Una manera de hacerlo es usar para entrenamiento muestras de distintas sesiones. Esto es lo que se prueba aquí.

   - **Train:** S1 y S2, M1, usuario i
   - **Test:**
     - Test Auténticos: S1 y S2, M2, usuario i
     - Test Impostores: S1 y S2, M2, usuario j $\neq$ i

El protocolo experimental seguido para cada caso, *Monosesión-Monomuestra, Multisesión-Monomuestra* y *Multisesión-Multimuestra*, es el indicado. Ahora bien, otra forma que puede parecer más razonable de actuar es considerar las distintas posibilidades dentro de *Monosesión-Monomuestra*, que serían la indicada junto con las siguientes tres:

1. **Train:** S2, M1, usuario i

   - Test Auténticos: S2, M2, usuario i
   - Test Impostores: S2, M2, usuario j $\neq$ i

2. **Train:** S1, M2, usuario i

   - Test Auténticos: S1, M1, usuario i
   - Test Impostores: S1, M1, usuario j $\neq$ i

3. **Train:** S2, M2, usuario i

   - Test Auténticos: S2, M1, usuario i

- Test Impostores: S2, M1, usuario j $\neq$ i

De la misma forma, en *Multisesión-Monomuestra* existirían además de la mencionada las siguientes tres:

1. **Train:** S1, M2, usuario i
   - Test Auténticos: S2, M2, usuario i
   - Test Impostores: S1, M1, usuario j $\neq$ i
2. **Train:** S2, M1, usuario i
   - Test Auténticos: S1, M1, usuario i
   - Test Impostores: S2, M2, usuario j $\neq$ i
3. **Train:** S2, M2, usuario i
   - Test Auténticos: S1, M2, usuario i
   - Test Impostores: S2, M1, usuario j $\neq$ i

Y por último, en *Multisesión-Multimuestra* existiría, además de la mencionada otra más que es:

1. **Train:** S1 y S2, M2, usuario i
   - Test Auténticos: S1 y S2, M1, usuario i
   - Test Impostores: S1 y S2, M1, usuario j $\neq$ i

Utilizando todas las posibilidades se haría una especie de validación cruzada. Esto, que parece una buena idea, no lo pudimos hacer debido a las deficiencias de los datos de que disponemos, ya que no todos los usuarios tienen todas las sesiones y dos muestras en cada sesión. Actuar de esa manera nos obligaría a quedarnos solo con los que tienen todo, lo que supone un subconjunto demasiado pequeño. En nuestros datos se tienen: 11 usuarios completos, 2 usuarios con solo una sesión con dos muestras, 7 usuarios con 1 sola muestra en cada sesión y 1 usuario con 3 datos, dos de la primera sesión y sólo uno de la segunda.

Por otro lado, entre todas las posibilidades, se ha selecciona la primera opción del procedimiento experimental mencionado, que utiliza como entrenamiento la primera sesión y muestra, siguiendo las pautas y bases fijadas en biometría que intentan simular el comportamiento de la vida real. Interpretando que los datos se utilizan en orden y que la primera sesión S1 y muestra M1 que se obtiene es la que forma parte del conjunto de *train*, la que se usa para lo que en biometría se denomina *inscribir* al usuario.

Los sensores utilizados son tanto el acelerómetro como el giroscopio tridimensional que poseen los dispositivos usados en la captura. Esto permite 4 posibilidades *(Microsoft acelerómetro, Microsoft giroscopio, Motorola acelerómetro, Motorola giroscopio)*, de las cuales se pueden comparar si los resultados de los sensores son similares y apropiados para trabajar de manera complementaria o si sería mejor centrarse en uno; igual que en los dispositivos, para poder ver si las conclusiones se pueden generalizar y existen posibilidades de encontrar un sistema de reconocimiento bueno para cualquier dispositivo comercial.

4.4. Clasificación

Cuando se tiene un problema, se tiende a utilizar muchos algoritmos de clasificación diferentes o con pequeñas variaciones para conseguir resolverlo obteniendo el mejor resultado final. En este trabajo, no se pretende resolver el problema completo, sino realizar un buen estudio prospectivo que siente las bases de esta novedosa biometría, y dado que existen muchas incógnitas: *¿con qué tamaño de ventana extraer características?, ¿qué características extraer?, ¿qué tipo de preprocesado beneficiará más a los datos?, ¿qué valor umbral fijar para permitir ventanas con autocorrelación alta?*, se ha elegido un clasificador sencillo que no introduce muchos parámetros al problema.

La elección es el algoritmo basado en distancias de **k-vecinos más próximos**. Un método simple, fácil de programar y entender si se necesita explicar a un público amplio. Además, solo necesita muestras del usuario para crear su patrón; la mayoría de los clasificadores discriminantes necesitan para su entrenamiento muestras de la clase auténtica y de la clase impostor, lo que introduce la variabilidad asociado a qué muestras de la clase impostor usar. Para nuestro problema, se puede resumir su funcionamiento a través de un bucle de usuarios auténticos y otro de usuarios impostores.

Partiendo de los datos divididos en *conjunto de entrenamiento* y *conjunto de prueba*, tal como se explica en la subsección 4.3 y con el objetivo de encontrar las distancias correspondientes a los usuarios auténticos, se tendrán que seguir los siguientes pasos *para cada usuario* i a estudiar:

1. Para cada ventana en el conjunto de datos de prueba del usuario i, calcular la distancia entre esta ventana y cada una de las ventanas en el conjunto de entrenamiento del usuario i que se esté estudiando.

2. Con las distancias obtenidas en el paso anterior, seleccionar las k distancias más pequeñas. Como distancia se ha usado la euclídea. Se pueden utilizar otras distancias, pero está es la más general y típicamente usada.

3. La distancia final o "score" de la muestra de prueba se obtiene mediante un estadístico (media, mediana, etc.) sobre las k distancias del paso 2.

Con el bucle anterior, se obtienen los scores para el conjunto de muestras auténticos, pero hay que realizar otro bucle que obtenga las distancias de los usuarios impostores j. En este bucle, se

repiten las mismas operaciones que en el anterior, pero ahora, para las muestras de prueba del resto de usuarios, es decir, $\forall j \neq i$.

El pseudocódigo de ambos bucles puede verse en los algoritmos 1 y 2, donde las tablas de entrada del conjunto de datos *train* y *test* contienen los atributos de interés, eliminando aquellas columnas que indican el usuario, la sesión y la muestra, ya que se va calculando una distancia entre variables, todas ellas, numéricas y se está trabajando con datos de la misma clase, sin indicar la variable respuesta.

La función *distanciaEuclidea(x,y)* es la que calcula la distancia euclídea entre dos vectores con n características numéricas siguiendo (4.9).

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{4.9}$$

Algorithm 1: Algoritmo para obtener los scores de los usuarios auténticos

Input Tabla de datos de entrenamiento de todos los usuarios $train_i$
Tabla de datos de prueba de todos los usuarios $test_i$
Usuarios disponibles en la forma experimental $usuariosDisponibles$
Output Scores auténticos de todos los usuarios $scoreAutentico_i$

```
for user in usuariosDisponibles do
    n_ventanas_train ← nrow(train_user)
    n_ventanas_test ← nrow(test_user)

    for i ← 1 to n_ventanas_test do
        scores_auxiliares ← vector()

        for j ← 1 to n_ventanas_train do
            salida ← distanciaEuclidea(test_user[i,], train_user[j,])
            score_auxiliares ← c(score_auxiliares, salida)

        end
        seleccion ← min(score_auxiliares)
        scoreAutentico_user ← c(scoreAutentico_user, seleccion)

    end
end
```

El único parámetro por fijar en este algoritmo es el valor de k, el cual depende fundamentalmente de los datos. De manera general, valores grandes de k reducen el efecto de ruido en la clasificación, pero crean límites entre clases parecidas.

Algorithm 2: Algoritmo para obtener los scores de los usuarios impostores

Input Tabla de datos de entrenamiento de todos los usuarios $train_i$
Tabla de datos de prueba de todos los usuarios $test_i$
Usuarios disponibles en la forma experimental $usuariosDisponibles$
Output Scores impostores de todos los usuarios $scoreImpostores_i$

```
for user_train in usuariosDisponibles do
    for user_test in usuariosDisponibles do
        if user_test ≠ user_train then
            n_ventanas_train ← nrow(train_{user_train})
            n_ventanas_test ← nrow(test_{user_test})

            for i ← 1 to n_ventanas_test do
                scores_auxiliares ← vector()

                for j ← 1 to n_ventanas_train do
                    salida ← distanciaEuclidea(test_{user_test}[i,], train_{user_train}[j,])
                    score_auxiliares ← c(score_auxiliares, salida)

                end
                seleccion ← min(score_auxiliares)
                scoreImpostores_{user_train} ← c(scoreImpostores_{user_train}, seleccion)

            end
        end
    end
end
```

Para los datos originales, sin ningún tipo de preprocesamiento, únicamente eliminando el ruido de manera manual, se han probado valores impares de k entre 1 y 25, tanto en el dominio del tiempo como de la frecuencia utilizando o el módulo o las 3 componentes XYZ juntas. Los resultados se pueden ver en los gráficos de las figuras 4.4 y 4.5 para el dominio del tiempo y 4.6 y 4.7 para el dominio de la frecuencia.

El eje X de los gráficos de la figura 4.4, 4.5, 4.6 y 4.7 indica el valor de k y el eje Y el valor de la tasa de equierror. Idealmente es mejor cuánto EER más pequeño. Aunque hay excepciones, suele ser mejor utilizar *k=1*, resultando el mejor de manera global. Otra cosa que se observó analizando cada usuario es que los usuarios que tienen comportamientos extraños y diferentes frente al resto, son aquellos que llevaban el reloj/pulsera en la mano dominante. Ejemplo de ello es el usuario 8 que obtiene resultados malos, llegando incluso en el dominio de la frecuencia con el módulo a obtener EER de siempre 0.5 (igual que aleatorio) (gráfico 4.7); lo mismo le ocurre al usuario 15 obteniendo valores de 0.5 tanto en el módulo como en XYZ del dominio de la frecuencia (gráficos 4.6 y 4.7). Y resultados que aunque no son de 0.5, si son excesivamente malos con *k=1* ocurre en el usuario 11 para el dominio del tiempo y las 3 componentes XYZ (gráfico 4.4).

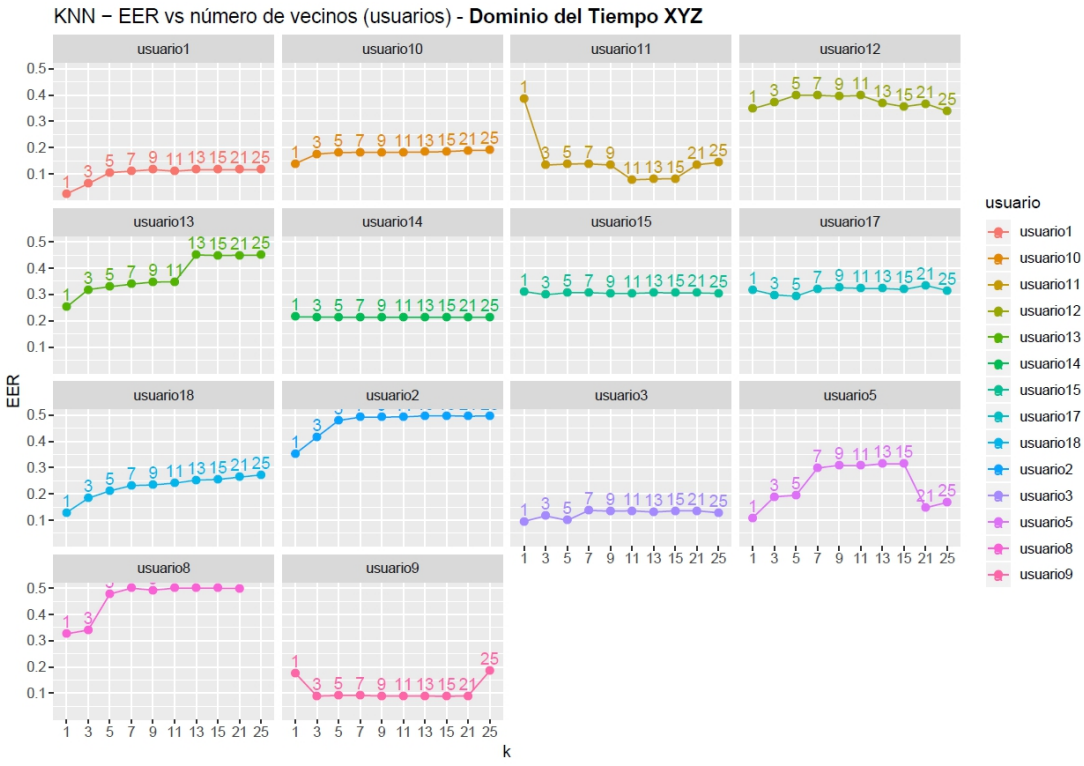


Figura 4.4: EER variando k en KNN, para cada usuario, Dominio del Tiempo XYZ.

Finalmente, por su sencillez y rapidez, ya que necesita menor tiempo de cómputo, y por rendimiento se ha seleccionado el valor impar de *k=1*, llamando en este caso al algoritmo como *Nearest Neighbor Algorithm* o algoritmo del vecino más cercano.

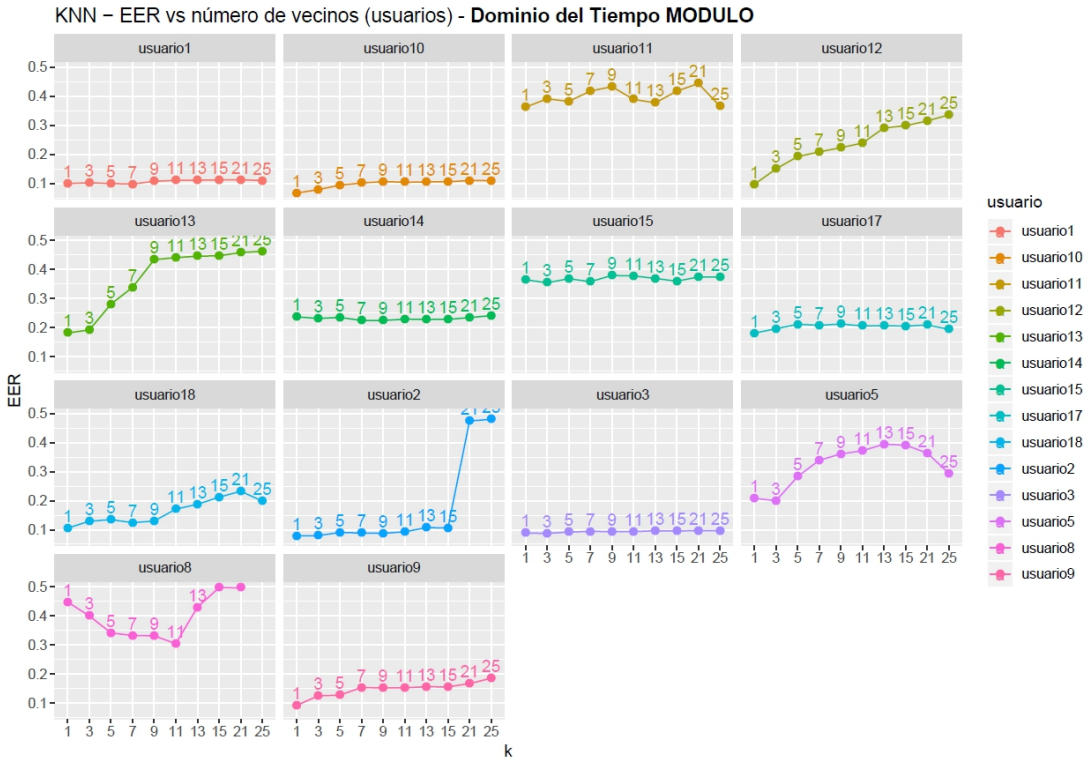


Figura 4.5: EER variando k en KNN, para cada usuario, Dominio del Tiempo Módulo.

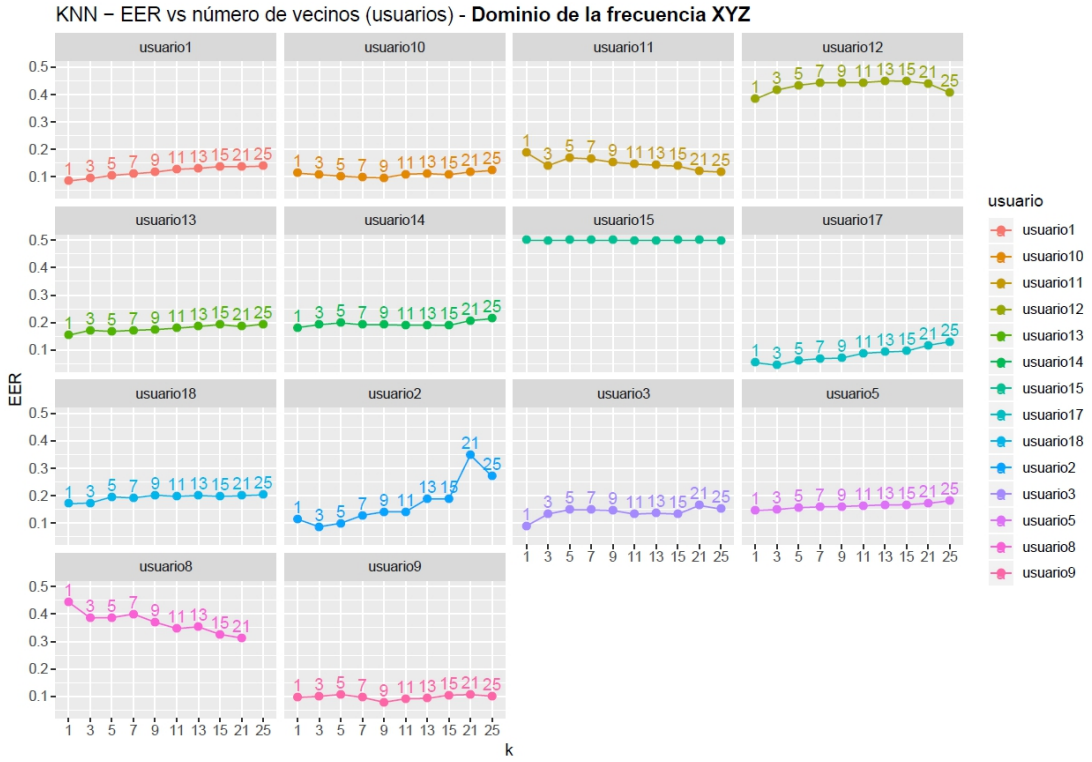


Figura 4.6: EER variando k en KNN, para cada usuario, Dominio de la Frecuencia XYZ.

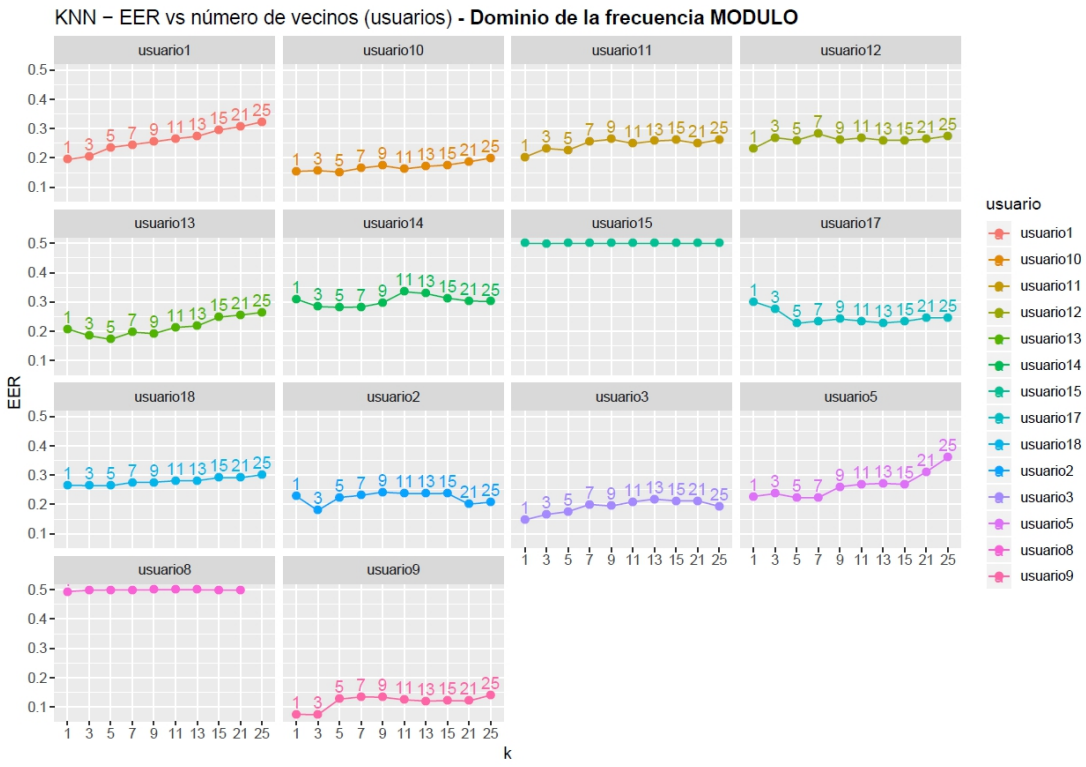


Figura 4.7: EER variando k en KNN, para cada usuario, Dominio de la Frecuencia Módulo.

Capítulo 5

Preprocesamiento

Los problemas de aprendizaje no se resuelven solamente aplicando distintos algoritmos y estimando su tasa de error. La adaptación y modificación de los datos puede facilitar e incluso en algunos casos, hacer posible el aprendizaje. Para ello, se van a probar las técnicas de preprocesamiento que se explican a lo largo del presente capítulo, con el objetivo de ayudar a los datos a explicar la mayor cantidad de información posible.

Se van a mostrar los resultados de un primer estudio prospectivo de los datos, con tamaño de ventana 3 ciclos y procedimiento experimental *Monosesión-Monomuestra*, en el que se ha buscado ver su comportamiento y reducir la dimensión del problema para centrarse en las técnicas más importantes a la hora de buscar los mejores parámetros en el siguiente capítulo.

5.1. Detección de las zonas de interés

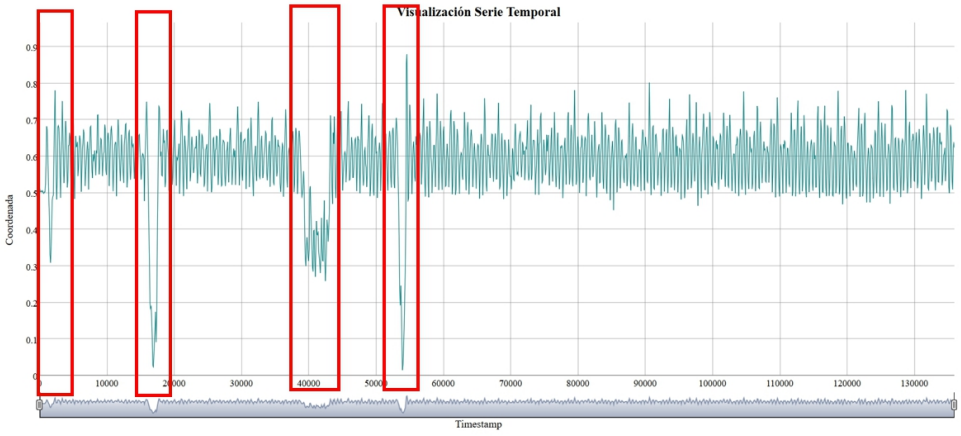
Junto con la limpieza de los datos, la eliminación de ruido y extracción de las zonas de interés es un paso importante con el fin de quedarnos con las zonas donde haya periodicidad, para que puedan ser empleadas como patrón para el reconocimiento biométrico.

Para llevarlo a cabo se utilizó la librería *dygraphs* de R [57], la cual permite visualizar los datos en un gráfico de líneas con el tiempo acumulado en el eje de abscisas y la componente de los datos que se esté utilizando en el eje de ordenadas. Esta librería permite ampliar y reducir la visualización, así como cortar zonas para un análisis más detallado y con el cursor en la línea, indica los valores exactos del punto a que corresponde. La salida del gráfico es en formato *html*.

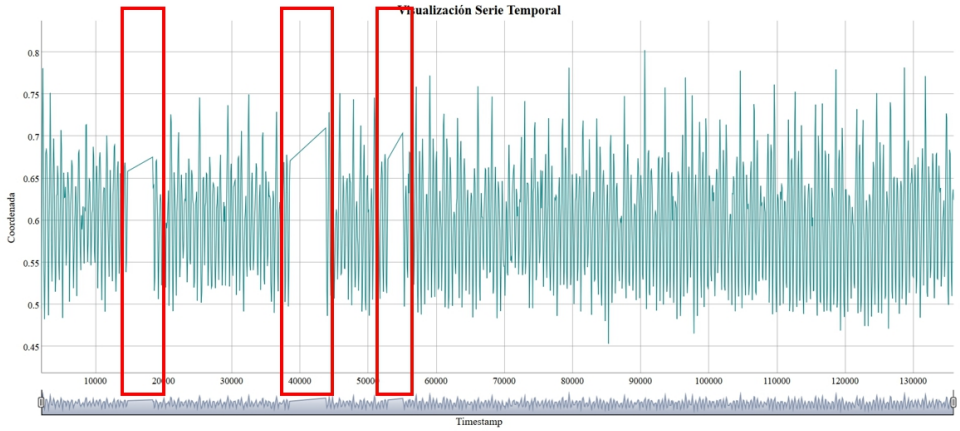
Por este motivo y con ayuda de la librería mencionada, se ha realizado una selección manual de las zonas de interés, es decir, de las zonas donde el comportamiento de la señal es periódico y característico del usuario que lo hace susceptible de ser usado en biometría. Se ha hecho para cada fichero de datos y eje de coordenadas. El resultado es un fichero con etiquetas que marca los instantes en milisegundos de inicio y de fin de las correspondientes zonas de interés. La razón para hacerlo inicialmente así era la ausencia del suficiente conocimiento de los datos como para

automatizarlo directamente y confiar en que esa decisión tomada fuera la más adecuada. Vamos a hacer un estudio básico del problema, para entenderlo mejor, para ello es importante estar seguros de que los datos a usar sean de la mayor calidad posible, por eso se optó, inicialmente, por la segmentación manual.

Debido a la subjetividad y el enorme trabajo que implica, después y con un mayor conocimiento del problema y de los datos, la selección manual ha sido sustituida por una selección automática basada en la autocorrelación, como se explicará más adelante.



(a) Señal con ruido



(b) Señal sin ruido

Figura 5.1: Tipos posibles de series de datos.

En la figura 5.1 (a) se puede ver la muestra de datos de un usuario concreto antes de eliminar el ruido y en la figura 5.1 (b) la misma muestra de datos eliminando el ruido del comienzo de la señal, y de tres movimientos extraños marcados en rojo, que pueden haber sido causados por pérdida de la señal entre el dispositivo y el teléfono móvil o un movimiento diferente de los brazos del usuario.

5.2. Interpolación

Una de las características que tiene el uso de dispositivos comerciales para el reconocimiento del usuario mediante ponibles es que no se tiene control sobre la frecuencia de muestreo. Dependiendo del dispositivo, se puede fijar un valor que permita la adquisición con mayor o menor frecuencia, pero siempre como referencia, ya que nunca se va a lograr que el ponible nos entregue los datos a una frecuencia fija. Esta es una de las diferencias entre nuestra forma de abordar el problema y el resto de trabajos en la bibliografía, donde usan dispositivos creados ad hoc. La consecuencia son datos obtenidos con periodos de muestreo (tiempo entre dos muestras consecutivas) variable.

Como se iba a trabajar tanto en el dominio del tiempo como en el dominio de la frecuencia, para conseguir un mejor análisis posterior de los datos, se decidió pasar a muestras con frecuencia de muestreo fijo, aplicando el algoritmo de interpolación lineal que se muestra en el pseudocódigo del algoritmo 3.

La función *calculaInterpolacion(t, x_1, y_1, x_2, y_2)* es la que calcula las coordenadas de un nuevo punto a partir de la recta formada por los puntos *(x_1, y_1)*, *(x_2, y_2)* y un valor conocido del eje de abscisas t, que en este caso es el tiempo. Para calcular este valor se utiliza interpolación lineal, como indica la fórmula (5.1).

$$y = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (t - x_1) \tag{5.1}$$

La frecuencia es una magnitud que mide el número de repeticiones por unidad de tiempo de cualquier fenómeno o suceso periódico. Para calcular la frecuencia de un suceso, se contabilizan un número de ocurrencias de éste, teniendo en cuenta un intervalo temporal, y luego estas repeticiones se dividen por el tiempo transcurrido. En unidades del Sistema Internacional (SI), la frecuencia se mide en hercios (Hz). Un hercio es la frecuencia de un suceso o fenómeno repetido por segundo. De la frecuencia de la señal se puede obtener el periodo, utilizando la letra T para referirnos a él y Fm para hacer referencia a la frecuencia [58].

Cuanto más alta sea la frecuencia de muestreo, los ficheros de datos generados serán más grandes e interpolar a una frecuencia superior a la máxima presente en los datos originales, no aportará nada. El análisis visual de los datos muestra que hay pocos con un valor de tiempo entre muestras consecutivas inferior a 83ms, no superando en media el valor de 100. Siguiendo el teorema de muestreo de *Nyquist-Shannon* [59], el cual dice que se puede reconstruir una señal original a partir de una señal interpolada si la frecuencia de muestreo de esta última es de al menos el doble que la máxima frecuencia presente en la señal original, y siguiendo el estudio realizado por Daniel González Alonso en su TFG [5], se ha utilizado 12Hz como frecuencia de muestreo.

De manera que:

$$Fm = \frac{1}{T} \tag{5.2}$$

Algorithm 3: Algoritmo para transformar datos tomados con frecuencia de muestreo variable a frecuencia de muestreo fija

Input Matriz *datos* con los datos capturados en cada eje
Vector *tiempos* con los instantes en que se tomó cada fila de *datos*
Periodo T al que se va a interpolar

Output Matriz *datos2* con los datos capturados en cada eje con frecuencia de muestreo fija
Vector *tiempos2* con los instantes en que se tomó cada fila de *datos2*

nColumnas ← número de columnas de *datos* (ejes de coordenadas)
nFilas ← número de filas de *datos* (filas en el fichero)
$t \leftarrow tiempos[1]$
$f \leftarrow 1$
while $t <$*tiempos[nFilas]* **do**
  $p1 \leftarrow 1$
  $p2 \leftarrow 2$
  for $i \leftarrow 2$ ***to*** *nFilas-1* **do**
    if *tiempos[i]*$<t \wedge$ *tiempos[i]*$>$*tiempos[p1]* **then**
      $p1 \leftarrow i$
      $p2 \leftarrow i + 1$
    end
  end
  $tiempos2[f] \leftarrow t$

  for $j \leftarrow 1$ ***to*** *nColumnas* **do**
    $datos2[f, j] \leftarrow calculaInterpolacion(t, tiempos[p1], datos[p1, j], tiempos[p2], datos[p2, j])$

  end
  $t \leftarrow t + T$
  $f \leftarrow f + 1$
end

Se está trabajando en milisegundos, por lo que conocida la frecuencia de muestreo de 12Hz, el periodo resulta ser:

$$T = \frac{1}{Fm} = \frac{1}{12}s \times \frac{1000ms}{1s} = 83,\widehat{3}ms \tag{5.3}$$

Redondeando hacía abajo se utilizan 83ms. Por otro lado, y según el artículo de la referencia [7] movemos el brazo a un máximo de 8.6Hz, haciendo el movimiento más rápido posible. Como nuestros datos son recogidos andando, una frecuencia de muestreo de 12 Hz. (frecuencia máxima en la señal 6 Hz) parece razonable.

5.2.1. Resultados Microsoft Acelerómetro

En un estudio inicial realizado se han utilizado los datos eliminando únicamente el ruido *(originales)* y aplicando, a mayores, la interpolación lineal *(interpolados)*. Se han estudiado ambos dominios y en cada uno de ellos, utilizar las componentes X, Y, Z por separado, juntas o con el módulo, comparando las métricas AUC y EER. Se ha aprovechado para en el dominio del tiempo, observar los atributos: periodo y autocorrelación, eliminándolos por separado y juntos. *Todo* significa que se están utilizando todas las características. En el dominio de la frecuencia se han usado las amplitudes de Fourier originales y escaladas. Los resultados se muestran en las tablas 5.1 y 5.2. Se han escalado las amplitudes de Fourier, llevándolas a un rango común y comparable, como se explicará en el siguiente apartado 5.3. Se ha querido incluir primero la interpolación y luego la normalización porque los resultados son más claros y visuales.

Para mejorar la visualización de las tablas, se muestran los números redondeados a un máximo de 4 cifras decimales y marcados en **negrita** los mejores resultados: valores altos de AUC y bajos de EER, que pueden no coincidir en las mismas componentes si las diferencias son pequeñas, ya que no miden la misma información.

En el dominio del tiempo se puede ver como el periodo aporta mucha información consiguiendo mejoras grandes en algunos resultados, mientras que la autocorrelación es indiferente y produce los mismos resultados. Por lo que de aquí en adelante se utilizarán todas las características mencionadas en la subsección 4.1, sin eliminar ninguna. Por otro lado, la interpolación no aporta mejoras significativas, empeorando en algunos casos. El mejor resultado se obtiene con el módulo.

El dominio de la frecuencia funciona mejor sin normalizar las amplitudes, no habiendo muchas diferencias entre aplicar la interpolación o no. El mejor resultado se obtiene utilizando las componentes X, Y, Z juntas.

5.3. Normalización

La normalización de los datos es importante en muchos métodos numéricos, cuando se busca dar igual importancia a cada atributo, pero hay otros casos en los que no merece la pena e incluso puede perjudicar a los resultados. Un ejemplo de ello es si se busca identificar jugadores de

		originales		interpolados	
		AUC	EER	AUC	EER
todo	DT-XYZ	0.8	0.2272	0.7826	0.2394
	DT-X	0.7575	0.2421	0.7494	0.2238
	DT-Y	0.7396	0.2534	0.7917	0.2144
	DT-Z	0.8351	0.1872	**0.8337**	**0.1969**
	modulo DT	**0.8439**	**0.186**	0.83	0.2011
No Periodo	DT-XYZ	0.648	0.296	0.6592	0.2946
	DT-X	0.6258	0.339	0.6103	0.3233
	DT-Y	0.7125	0.2628	0.6942	0.2772
	DT-Z	0.805	**0.1989**	**0.7871**	**0.2226**
	modulo DT	**0.8223**	0.2061	0.7629	0.2504
No Autocor	DT-XYZ	0.8	0.2272	0.7826	0.2394
	DT-X	0.754	0.2441	0.7482	0.2254
	DT-Y	0.7303	0.2559	0.791	0.2142
	DT-Z	**0.8301**	**0.1907**	**0.8305**	**0.1989**
	modulo DT	0.8288	0.195	0.8286	0.2002
No Periodo Autocor	DT-XYZ	0.6452	0.2965	0.6589	0.2946
	DT-X	0.6155	0.3474	0.613	0.3191
	DT-Y	0.6986	0.2714	0.7	0.2752
	DT-Z	**0.8009**	**0.2064**	**0.7874**	**0.2227**
	modulo DT	0.7957	0.2205	0.7677	0.2464

Tabla 5.1: Estudio inicial - Interpolación - Dominio del Tiempo.

		originales		interpolados	
		AUC	EER	AUC	EER
Todo – Amplitudes Fourier Escaladas	DF-XYZ	0.7104	0.3845	0.6795	0.362
	DF-X	**0.7243**	0.286	**0.6931**	0.3161
	DF-Y	0.6568	0.3669	0.6224	0.4298
	DF-Z	0.6923	0.3423	0.684	0.3295
	modulo DF	0.7186	**0.279**	0.6922	**0.3146**
Todo – Amplitudes Fourier No Escaladas	DF-XYZ	**0.8337**	**0.2299**	**0.8032**	0.2563
	DF-X	0.7898	0.2435	0.7483	0.2813
	DF-Y	0.7354	0.313	0.6917	0.3431
	DF-Z	0.8023	0.2338	0.7925	**0.2445**
	modulo DF	0.7545	0.2885	0.7276	0.289

Tabla 5.2: Estudio inicial - Interpolación - Dominio de la Frecuencia.

baloncesto entre una población grande de personas, donde está claro que la estatura es un atributo discriminante y con mayor importancia frente a otros como el peso.

Normalizar significa comprimir o extender los valores de la variable para que estén en un rango definido o para que sigan una distribución con valores concretos [60].

5.3.1. Escalado de variables

Los valores de la variable se comprimen a una escala con límites definidos. Se puede conseguir con (5.4).

$$X_{escalado} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{5.4}$$

Está técnica implica una distorsión de los datos y una amplificación del ruido, de manera que si se tiene una señal estable con muy poco ruido, este método podría no ser apropiado, pudiendo perjudicar a los resultados.

5.3.2. Estandarización de variables

Una alternativa al escalado es la estandarización, la cual considera la media y la desviación estándar de los valores de la variable y los estandariza a media 0 y desviación estándar 1, a través de (5.5). Al tratarse de un cambio de localización y escala, si los datos no se distribuyen normalmente, una vez transformados tampoco lo harán.

$$X_{estandarizado} = \frac{X - X_{media}}{X_{desviacin\_estndar}} \tag{5.5}$$

Está técnica es simple y no produce cambios en la correlación. No obstante, al tener todos los datos la misma media y desviación estándar, no se podrán utilizar como características para discriminar.

5.3.3. Resultados Microsoft Acelerómetro

Como estudio inicial de esta técnica, se ha probado la estandarización y el escalado sobre los datos originales en los que únicamente se ha eliminado el ruido de manera manual. Observándolo por separado para el dominio del tiempo y de la frecuencia, como se puede ver en las tablas 5.3 y 5.4.

Depende de la componente que se esté estudiando, pero de manera global, en ninguno de los dos dominios aporta mejoras aplicar las normalizaciones. Por otro lado, en el dominio del tiempo funciona mejor el módulo o la componente Z; mientras que en el dominio de la frecuencia funciona mejor utilizar las 3 componentes, o si no se quieren añadir tantos atributos, la componente Z o la X por separado funcionan bastante bien, ligeramente peor que las 3 componentes juntas, pero con pocas diferencias.

		originales		escalados		estandarizados	
		AUC	EER	AUC	EER	AUC	EER
Todo	DT-XYZ	0.8	0.2272	0.7885	0.2483	0.7284	0.2923
	DT-X	0.7575	0.2421	0.7654	0.2333	0.7418	**0.2657**
	DT-Y	0.7396	0.2534	0.7224	0.2856	0.6547	0.3443
	DT-Z	0.8351	0.1872	0.7692	0.2381	0.7214	0.3048
	modulo DT	**0.8439**	**0.186**	**0.793**	**0.2157**	**0.7635**	0.2707

Tabla 5.3: Estudio inicial - Normalización - Dominio del Tiempo.

		originales		escalados		estandarizados	
		AUC	EER	AUC	EER	AUC	EER
Todo	DF-XYZ	**0.8337**	**0.2299**	**0.753**	**0.2689**	0.6927	**0.2931**
	DF-X	0.7898	0.2435	0.6867	0.3038	**0.7093**	0.3159
	DF-Y	0.7354	0.313	0.6742	0.3407	0.5813	0.423
	DF-Z	0.8023	0.2338	0.6727	0.3475	0.6081	0.361
	modulo DF	0.7545	0.2885	0.721	0.2871	0.6525	0.3544

Tabla 5.4: Estudio inicial - Normalización - Dominio de la Frecuencia.

5.4. Filtrado

El filtrado es una técnica que permite la combinación de varios puntos para obtener un valor con mayor significación que los puntos individuales. Existen muchas técnicas de filtrado, pero en este proyecto se va a trabajar con el filtro de media móvil simple por ser uno de los más usados en la bibliografía, que, además, es intuitivo, fácil de implementar y rápido de calcular.

El resultado y lo que se busca conseguir es una señal suavizada que elimine parte del ruido de alta frecuencia. El tamaño de la ventana, lo que se va a llamar orden, va a tener influencia en el comportamiento del filtro, siendo más grande el suavizado de la señal, cuanto más grande sea su tamaño. A lo largo del proyecto se van a probar dos.

- Filtro de media móvil de orden 3:

$$X_t = \frac{X_{t-1} + X_t + X_{t+1}}{3} \tag{5.6}$$

- Filtro de media móvil de orden 5:

$$X_t = \frac{X_{t-2} + X_{t-1} + X_t + X_{t+1} + X_{t+2}}{5} \tag{5.7}$$

Al igual que la normalización, el filtro de media móvil tiene desventajas relacionadas con la debilidad del uso de la media como estimador de tendencia, siendo poco estable ante la aparición de puntos espurios (puntos anómalos muy alejados del valor real). En estos casos, resultaría más

conveniente utilizar un filtro de la mediana, o una combinación de ambos, pero no se va a entrar en ello [61].

En el estudio inicial no se ha probado está técnica porque se veía como algo necesario de probar una vez obtenidas las decisiones finales para comprobar si conseguía mejorarlas, es decir, tratando el filtro de la media móvil como una técnica de mejora. Se utilizará más adelante, en el siguiente capítulo.

5.5. Preprocesamiento final

Se han probado las técnicas anteriores sobre los datos, tanto de manera aislada, como ya se han mostrado, como de manera conjunta *(interpolación+normalización)*. Los resultados de todo ello se pueden ver en las tablas 5.5 y 5.7 correspondientes al dominio del tiempo y al dominio de la frecuencia, respectivamente. En el caso de no aplicar ninguna técnica *(datos originales)*, se ha probado a realizar diferentes tipos de normalización.

- Realizando la **normalización por columnas:** es lo más típicamente usado, dado que las columnas contienen características y las filas ventanas (muestras) de los distintos usuarios. Esta manera permite mantener las relaciones entre las distintas características.
  - Escalado usuario/sesión/muestra: consiste en realizar el escalado explicado en el apartado 5.3.1 para cada grupo posible de ventanas usuario/sesión/muestra de cada columna.
  - Estandarización usuario/sesión/muestra: lo mismo que el ítem anterior, pero aplicando la estandarización explicada en el apartado 5.3.2.
  - Cociente del máximo usuario/sesión/muestra: como solo se ha aplicado en este paso, no se ha explicado en el apartado 5.3. Consiste en dividir cada grupo de ventanas usuario/sesión/muestra distinto en cada columna por el valor máximo de dicha columna.
- Realizando la **normalización por filas:** se centra en cada vector de características, permitiendo normalizar ventana a ventana y llevándolas todas a una "zona común". No es muy común su uso, pero en biometría es bastante frecuente por considerar ventanas de usuarios por separado y haber demostrado buenos resultados en algunas situaciones.
  - Escalado: realiza el escalado explicado en el apartado 5.3.1 para cada fila de datos, correspondiente a una ventana de un usuario. La "zona común" es el área de un hipercírculo unidad.
  - Estandarización: lo mismo que el ítem anterior, pero aplicando la estandarización explicada en el apartado 5.3.2. La "zona común" es el borde de una hipercircunferencia de radio 1.
  - Cociente del máximo: consiste en dividir cada fila por el valor de la característica más alta. En este caso, correspondería a dividir cada fila por el valor del periodo, siendo

siempre, para todas las filas, valores muy parecidos. La "zona común" sería de nuevo, el borde de la hipercircunferencia unidad.

Los resultados de estas pruebas aplicadas a los datos originales se pueden ver en las tablas 5.6 y 5.8, correspondientes al dominio del tiempo y al dominio de la frecuencia.

En el dominio del tiempo, los peores resultados se obtienen con los datos solamente normalizados (escalados o estandarizados) e interpolados+estandarizados, siendo para el resto de configuraciones similares. El mejor resultado de la tabla 5.5 se consigue con los datos originales y el módulo, siendo un AUC de 0.8439 y EER de 0.186. Al aplicar las normalizaciones por filas y columnas a los datos originales, tabla 5.6, funciona bastante mejor la normalización por filas y la técnica menos conocida y usada del cociente máximo. En concreto, el mejor resultado es la normalización por filas aplicando el cociente máximo sobre el módulo que consigue AUC de 0.8352 y EER 0.1915, peor que si no se aplica ninguna normalización. Según lo expuesto, en el dominio del tiempo, se tomó la decisión de no aplicar nada y utilizar los datos originales.

En el dominio de la frecuencia, tabla 5.7, funciona bastante mejor utilizar los datos originales, con pocas diferencias con respecto a utilizar los datos interpolados, pero perjudicando mucho realizar las normalizaciones. El mejor resultado se consigue utilizando las 3 componentes *(XYZ)* y los datos originales, con AUC 0.8337 y EER 0.2299. En este caso, utilizar los datos interpolados perjudica un 3 % a los resultados. Por otro lado, al aplicar las normalizaciones por filas y columnas a los datos originales, tabla 5.8, se ve lo mismo que en dominio del tiempo, lo mejor es la normalización por filas, obteniendo el mejor resultado estandarizando las 3 componentes con AUC 0.8131 y EER 0.211, resultados similares, con peor AUC pero mejor EER. Parecería razonable tomar la decisión de utilizar los datos originales, pero al estar realizando un Análisis de Fourier a los datos para extraer las características, para poder tener un significado físico de las componentes obtenidas, se van a utilizar datos con frecuencia de muestreo fija, tomando la decisión de utilizar los datos interpolados de aquí en adelante para este dominio.

	originales		interpolados		escalados		estandarizados		interpolados+ escalados		interpolados+ estandarizados	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
DT-XYZ	0.8	0.2272	0.7826	0.2394	0.7885	0.2483	0.7284	0.2923	**0.8035**	**0.2278**	0.7191	0.2946
DT-X	0.7575	0.2421	0.7494	0.2238	0.7654	0.2333	0.7418	**0.2657**	0.7854	0.237	0.7326	0.2782
DT-Y	0.7396	0.2534	0.7917	0.2144	0.7224	0.2856	0.6547	0.3443	0.7833	0.248	0.6922	0.3154
DT-Z	0.8351	0.1872	**0.8337**	**0.1969**	0.7692	0.2381	0.7214	0.3048	0.7647	0.2515	0.7173	0.3029
modulo DT	**0.8439**	**0.186**	0.83	0.2011	**0.793**	**0.2157**	**0.7635**	0.2707	0.7858	0.2367	**0.7833**	**0.2601**

Tabla 5.5: Estudio inicial. Resumen del Dominio del Tiempo.

	Normalización por columnas						Normalización por filas					
	USM (escalado)		USM (estandarización)		USM (cociente máximo)		escalado		estandarización		cociente máximo	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
DT-XYZ	**0.6737**	0.3545	0.5442	0.427	0.5486	0.3754	0.6674	0.3184	0.69	0.2976	0.6959	0.2933
DT-X	0.6303	**0.3465**	0.5572	0.4197	0.5725	0.362	0.6268	0.3153	0.6386	0.3044	0.6777	0.2565
DT-Y	0.5453	0.405	0.5507	0.4194	0.5922	0.3532	0.6451	0.3353	0.6924	0.3039	0.7661	0.2239
DT-Z	0.6415	0.347	**0.5835**	0.4111	0.6	0.4024	0.6553	0.3354	0.6768	0.3234	0.8048	0.2101
modulo DT	0.6348	0.3557	0.5793	**0.4066**	**0.7899**	**0.2557**	**0.8239**	**0.2025**	**0.7879**	**0.2364**	**0.8352**	**0.1915**

Tabla 5.6: Estudio inicial. Resumen del Dominio del tiempo sobre los datos originales.

	originales		interpolados		escalados		estandarizados		interpolados+ escalados		interpolados+ estandarizados	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
DF-XYZ	**0.8337**	**0.2299**	**0.8032**	0.2563	**0.753**	**0.2689**	0.6927	**0.2931**	**0.7614**	**0.2452**	**0.6798**	0.3555
DF-X	0.7898	0.2435	0.7483	0.2813	0.6867	0.3038	**0.7093**	0.3159	0.6422	0.3711	0.6493	0.358
DF-Y	0.7354	0.313	0.6917	0.3431	0.6742	0.3407	0.5813	0.423	0.6693	0.3468	0.583	0.4625
DF-Z	0.8023	0.2338	0.7925	**0.2445**	0.6727	0.3475	0.6081	0.361	0.705	0.2998	0.6501	**0.3553**
modulo DF	0.7545	0.2885	0.7276	0.289	0.721	0.2871	0.6525	0.3544	0.7119	0.3166	0.5959	0.3947

Tabla 5.7: Estudio inicial. Resumen del Dominio de la Frecuencia.

	Normalización por columnas						Normalización por filas					
	USM (escalado)		USM (estandarización)		USM (cociente máximo)		escalado		estandarización		cociente máximo	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
DF-XYZ	**0.5984**	**0.3959**	0.5316	0.4459	0.6565	0.3337	0.7046	**0.2823**	**0.8131**	**0.2111**	**0.7976**	0.2556
DF-X	0.5582	0.4104	**0.5629**	**0.4208**	0.6126	0.3638	0.6944	0.3004	0.7929	0.2493	0.7537	0.2681
DF-Y	0.5847	0.4025	0.5113	0.4517	0.6124	0.3695	0.6609	0.3203	0.7239	0.2806	0.7088	0.3246
DF-Z	0.5624	0.4088	0.5414	0.4362	0.5729	0.3704	0.6726	0.3139	0.7774	0.2506	0.7909	**0.2338**
modulo DF	0.5526	0.4124	0.5456	0.4267	**0.6755**	**0.3175**	**0.7641**	0.286	0.7397	0.2672	0.7329	0.3079

Tabla 5.8: Estudio inicial. Resumen del Dominio de la Frecuencia sobre los datos originales.

Capítulo 6

Experimentos: Análisis de parámetros

En este capítulo se han aprovechado las decisiones tomadas a lo largo del documento para hacer un análisis más profundo de los parámetros del problema. Ahora únicamente se trabaja con los datos originales en el dominio del tiempo y con los datos interpolados sin aplicar ninguna normalización a las amplitudes del Análisis de Fourier en el dominio de la frecuencia, continuando con el procedimiento experimental *Monosesión-Monomuestra*. Se van a explicar los parámetros del experimento, los resultados obtenidos utilizando la visualización de gráficos de líneas y las decisiones tomadas respecto a éstos, utilizando la ayuda del *test de Wilcoxon* para apoyar las conclusiones desde el punto de vista estadístico. La necesidad de ir por esta vía ha resultado de un primer análisis aumentando el tamaño de las ventanas del módulo y el arcoseno en el estudio inicial. Los resultados se pueden ver en las tablas referenciadas como 6.1.

En ambos dominios mejoran los resultados al aumentar el tamaño de la ventana. El comportamiento del módulo es muy parecido al del arcoseno. No obstante, la fórmula del arcoseno es más compleja y difícil de entender, por esta razón, de ahora en adelante se va a trabajar únicamente con el módulo.

		originales	
		AUC	EER
modulo DT	3 ciclos	0.8439	0.186
	4 ciclos	**0.8579**	**0.1681**
	5 ciclos	0.8542	0.1692
Arcoseno DT	3 ciclos	0.8326	0.1944
	5 ciclos	**0.8584**	**0.1733**

(a) Dominio del tiempo

		interpolados	
		AUC	EER
modulo DF	3 ciclos	0.7276	0.289
	4 ciclos	0.7698	0.2477
	5 ciclos	**0.806**	**0.2318**
arcoseno DF	3 ciclos	0.7702	0.2576
	5 ciclos	**0.7915**	**0.2348**

(b) Dominio de la frecuencia

Tabla 6.1: Estudio Inicial. Variar tamaño de las ventanas con el módulo y arcoseno.

6.1. Parámetros del experimento

A la hora de mejorar los resultados y estudiar en profundidad un problema, hay que tener en cuenta las distintas variables que afectan al mismo. En este caso, se han identificado las siguientes:

1. El **tamaño de las ventanas/número de ciclos**: como se ha comentado, la señal se divide en ventanas, cuyo tamaño se mide en número de ciclos. Se ha estudiado un tamaño apropiado frente a utilizar el tamaño más pequeño posible, de 2 y el más grande, de 15.

2. Aplicar el **filtro de la media móvil** de orden 3: comprobar si esta medida de suavizado mejora significativamente los resultados.

3. **Calidad de la señal**: parece razonable pensar que cuanto mayor sea la autocorrelación de una ventana, mejor será el patrón incluido en ella. Se quiere probar esta premisa, usando en el reconocimiento solo ventanas de autocorrelación alta y ver su influencia en el rendimiento.

4. **Eliminación del ruido automática:** se parte de la misma premisa del punto anterior, y se probará si podemos sustituir la eliminación del ruido de manera manual, haciendo una segmentación automática basada en algún criterio de la autocorrelación sobre los datos originales.

5. **Fusión de varias ventanas**: no es exactamente un parámetro pero sí una técnica que se va a utilizar para intentar mejorar los resultados. Hasta ahora, para cada ventana de prueba obteníamos su distancia (score) a la muestra de entrenamiento y esa distancia era la salida del clasificador. La propuesta es usar como "salida del clasificador", ahora, la fusión de los scores de varias ventanas consecutivas. Esta propuesta tiene dos parámetros, cuyo valor será analizado: número de ventanas consecutivas a fusionar y solapamiento entre grupos (la fusión se puede realizar cogiendo n ventanas y luego las siguientes n y así sucesivamente, o solapando los grupos de ventanas usados en la fusión).

Buscando estudiar la dependencia entre cada uno de estos "parámetros" y los resultados. La significación estadística de los mismos será calculada con el *test de Mann-Whitney-Wilcoxon* que se explica en el siguiente apartado.

6.2. Test Estadístico de Mann-Whitney-Wilcoxon

Más conocido como "test de Wilcoxon" [62]. Es un test no paramétrico que contrasta si dos muestras proceden de poblaciones equidistribuidas. A través de la tasa de equierror de cada usuario y con el procedimiento experimental *Monosesión-Monomuestra*, se ha buscado ver si existen diferencias significativas entre cada par de poblaciones probadas.

Se fundamenta bajo la idea de que, si las dos muestras comparadas proceden de la misma población, al juntar todas las observaciones y ordenarlas de menor a mayor, cabría esperar que las observaciones de una y otra estuvieran intercaladas aleatoriamente. Por lo contrario, si una de las muestras pertenece a una población con valores mayores o menores que la otra población, al ordenar las observaciones, estas tenderán a agruparse de modo que las de una muestra queden por encima de las de la otra.

Se podría decir que la hipótesis nula H_0 y alternativa H_a de este test para dos poblaciones X e Y son las de las fórmulas (6.1) y (6.2).

$$H_0 : P(X > Y) = P(Y > X) \tag{6.1}$$

$$H_0 : P(X > Y) \neq P(Y > X) \tag{6.2}$$

Las condiciones que exige se muestran a continuación.

- Los datos tienen que ser independientes.
- Los datos tienen que ser ordinales o bien se tienen que poder ordenar de menor a mayor.
- Igualdad de varianza entre grupos (homocedasticidad). Se comprobará para cada par de poblaciones.

No es necesario asumir que las muestras se distribuyan de manera normal o que procedan de poblaciones normales, lo cual hace que sea menos potente que otros test como el *t-test*, implicando tener menos probabilidad de rechazar la hipótesis nula cuando realmente es falsa. Aun así, se va a utilizar como técnica de ayuda para la toma de decisiones.

Al realizar el test, para estudiar la influencia de cada parámetro, se va a utilizar una de dos maneras posibles:

- Sobre el pico: considerando el mejor resultado, que corresponderá a un tamaño de ventana concreto.
- Sobre un rango de ventanas: utilizando un rango de tamaños de ventanas, sobre el que se conocerá el inicio y el final, debido a que el tamaño óptimo será desconocido.

Respecto a los resultados y a qué se va a considerar p-valor pequeño o grande, se va a hacer de manera global entre todos los resultados de cada prueba. Considerando, generalmente, un valor de 0.1 como límite.

6.3. Resultados obtenidos

Para facilitar la toma de decisiones, en lugar de utilizar los resultados de la métrica en cada usuario, se va a seguir utilizando la media de todos los usuarios posibles disponibles, que son un total de 14. A través del análisis visual de los gráficos de líneas y el test de Wilcoxon se van a tomar las decisiones finales, con el objetivo de generar un pequeño sistema de reconocimiento inicial que ayude a comparar entre distintos procedimientos experimentales.

Los gráficos de líneas van a mostrar en el eje de abscisas el tamaño de la ventana y en el eje de ordenadas el resultado de la métrica. Para no incluir demasiados gráficos se va a utilizar únicamente la tasa de equierror (EER) como medida de comparación. Se van a mantener siempre los mismos colores, los cuales pueden verse en la figura 6.1.

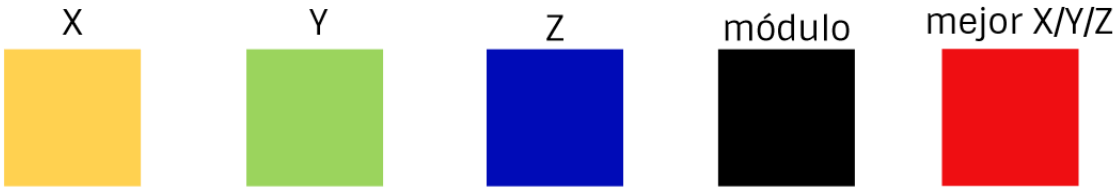


Figura 6.1: Colores componentes usados en los gráficos.

Mejor X/Y/Z resulta de utilizar para cada usuario, su mejor coordenada entre las 3 posibles: X, Y o Z. Serían los resultados que se obtendrían si fuésemos capaces de predecir a priori la mejor componente de cada usuario.

Se va a utilizar la expresión S_i^j para hacer referencia a la suposición i con significado j.

6.3.1. Tamaño de la ventana

Con el propósito de ver si un tamaño de ventanas intermedio produce mejores resultados que utilizar el más pequeño posible, que será 2, se han calculado los resultados en tamaños de ventana desde 2 hasta 15 ciclos. Probando también si se consiguen mejoras significativas entre aplicar el tamaño intermedio y el máximo de 15. Como tamaño máximo se ha utilizado 15, por parecer visualmente razonable y porque para utilizar tamaños más grandes de ventana se necesitarían muchos más datos por muestra, es decir, en cada toma de datos, se necesitaría que los usuarios estuvieran andando durante más tiempo.

En ambas situaciones, con el mínimo y el máximo, se va a aplicar el test de Wilcoxon sobre el pico, seleccionando el mejor tamaño de ventana y comparándolo con usar el más pequeño o el más grande en las mismas condiciones. Los mejores tamaños de ventana se han buscado entre el preprocesamiento de los datos seleccionado en cada dominio y el mismo aplicando el filtro de la media móvil de orden 3 (MM3).

6.3.1.1. Mejor tamaño de ventana vs mínimo

Primero, para la comparación del mejor tamaño de ventana y el más pequeño se obtienen los siguientes resultados.

En el dominio del tiempo, gráficos de la figura 6.2, el filtro de la media móvil de orden 3 (línea discontinua) parece producir mejores resultados. No obstante, se ve que mejora los resultados si la componente es buena. Pero si la componente es mala, como la Y, el suavizado no ayuda y produce peores resultados.

En el dominio de la frecuencia, gráficos de la figura 6.3, todas las componentes producen resultados bastante peores que los del dominio del tiempo. El filtro de la media móvil de orden 3 (línea discontinua) perjudica aún más a los resultados. Viéndose de nuevo que las técnicas de suavizado funcionan cuando los resultados ya son buenos, pero no cuando son malos.

Los mejores tamaños de ventana se encuentran entre 6 y 10 ciclos, hay ciertos tamaños en que se producen mejoras, pero no parece necesario probar tamaños más grandes que 15.

En ambos casos, en la figura (b) se muestra la evolución en los diferentes tamaños de ventana de su mejor caso. En el dominio del tiempo aplicando el filtro de suavizado y en el dominio de la frecuencia sin filtro.

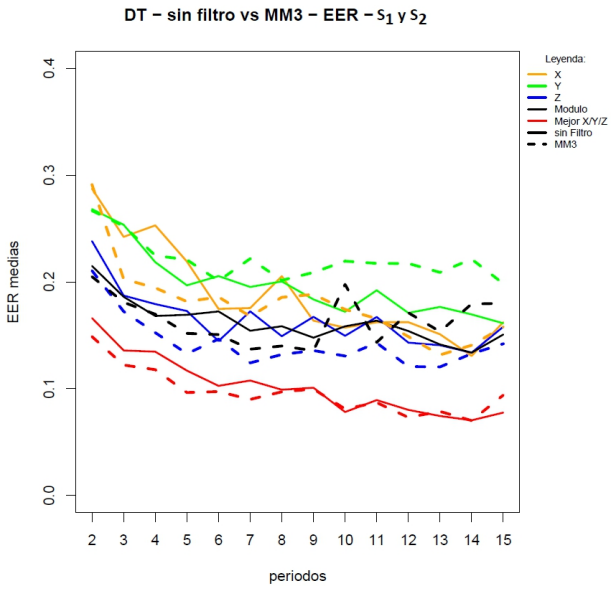
Visualmente sí parece que tamaños intermedios funcionan mejor que utilizar tamaño 2, especialmente en el dominio del tiempo, que es donde se obtienen los mejores resultados. El test de Wilcoxon, cuyos resultados se encuentran en la tabla 6.2 (a), obtiene p-valores más bien pequeños en el dominio del tiempo indicativos de que sí hay diferencias estadísticas significativas y por tanto, de que merece la pena utilizar tamaños de ventana más grandes. En el dominio de la frecuencia, la componente X y el módulo indican lo mismo, pero en las componentes Y y Z los p-valores son muy altos, indicando justo lo contrario, la no existencia de diferencias significativas. En todos los casos menos uno, el test de homocedasticidad, que se encuentra en la tabla 6.2 (b), obtiene p-valores altos, verificando que no hay evidencias en contra de la igualdad de varianzas.

Respecto a los colores de las tablas del test de Wilcoxon, se utiliza fondo verde y *cursiva* para mostrar los resultados del Dominio del Tiempo y fondo blanco sin cursiva para mostrar los del Dominio de la Frecuencia. El fondo rojo con letras rojas en **negrita** se reserva para p-valores inferiores al 5 % y únicamente **negrita** para indicar p-valores pequeños frente al resto, generalmente considerando valores inferiores a prácticamente el 10 %.

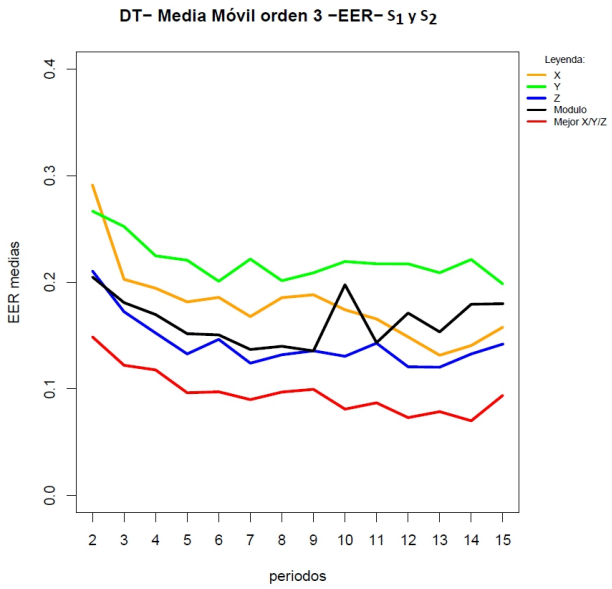
6.3.1.2. Mejor tamaño de ventana vs máximo

De manera conjunta con el mínimo, se ha buscado ver si seguir aumentando el tamaño de las ventanas sigue produciendo mejoras significativas en los resultados.

Los gráficos de la figura 6.2 para el dominio del tiempo y 6.3 para el dominio de la frecuencia, muestran que aunque sí hay algunas diferencias entre el mejor caso y utilizar tamaño 15, éstas no son muy grandes. El test de Wilcoxon, cuyos resultados pueden verse en la tabla 6.3 (a),

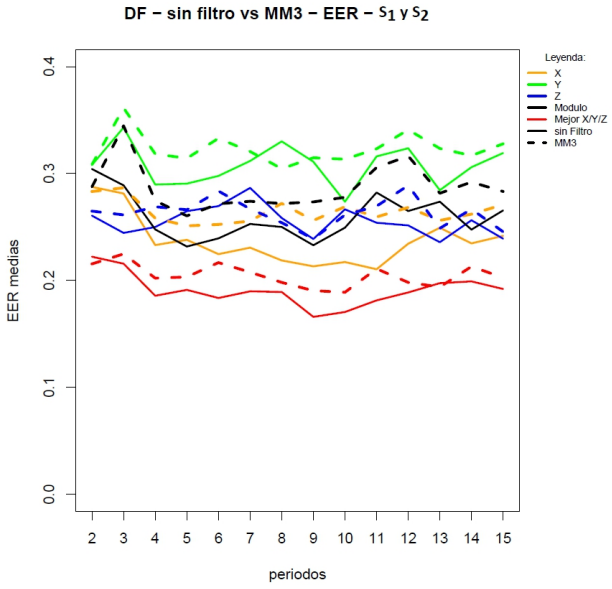


(a) datos originales vs con filtro MM3

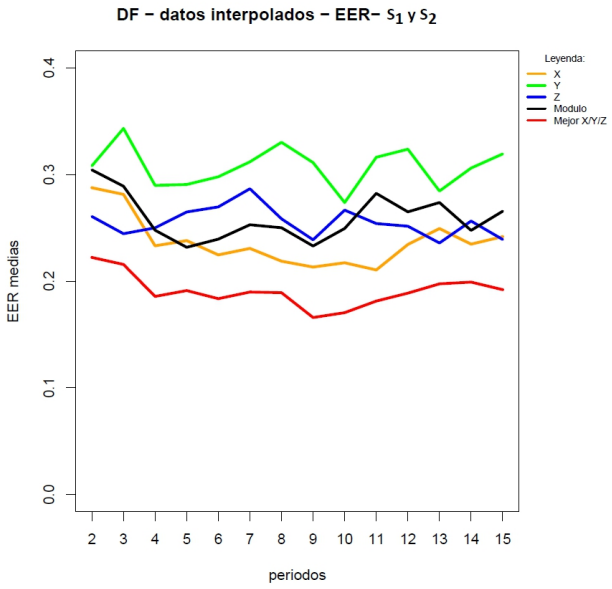


(b) datos con filtro MM3

Figura 6.2: Resultados para distintos tamaños de ventana en el Dominio del Tiempo.



(a) datos interpolados vs con filtro MM3



(b) datos interpolados

Figura 6.3: Resultados para distintos tamaños de ventana en el Dominio de la Frecuencia.

	P-valor $S_1^{mínimo}$
X + DT	**0.0694**
Y + DT	**0.0848**
Z + DT	**0.0108**
MÓDULO + DT	**0.1129**
X + DF	0.0241
Y + DF	0.5714
Z + DF	0.8036
MÓDULO + DF	**0.0556**

(a) p-valor test Wilcoxon S_1^{minimo}

	Homocedasticidad $S_1^{mínimo}$
X + DT	**0.0334**
Y + DT	0.893
Z + DT	0.2191
MÓDULO + DT	0.632
X + DF	0.6994
Y + DF	0.3049
Z + DF	0.7952
MÓDULO + DF	0.9698

(b) Homocedasticidad S_1^{minimo}

Tabla 6.2: Resultados test Wilcoxon - Mejor tamaño de ventana vs mínimo.

obtienen p-valores grandes en ambos dominios, indicando que como se veía, no existen diferencias estadísticamente significativas entre utilizar el mejor tamaño de ventana, generalmente en un rango de valores entre 6 y 10, que utilizar el tamaño más grande, de 15. Se supera el test de homocedasticidad, no habiendo evidencias en contra de la igualdad de varianzas en ninguno de los casos, se puede ver en la tabla 6.3 (b).

	P-valor $S_1^{máximo}$
X + DT	0.4886
Y + DT	0.8357
Z + DT	0.4761
MÓDULO + DT	0.5639
X + DF	0.2852
Y + DF	0.8036
Z + DF	1
MÓDULO + DF	0.1936

(a) p-valor test Wilcoxon S_1^{maximo}

	Homocedasticidad $S_1^{máximo}$
X + DT	0.3004
Y + DT	0.4833
Z + DT	0.7499
MÓDULO + DT	**0.0802**
X + DF	0.9491
Y + DF	0.8525
Z + DF	0.8459
MÓDULO + DF	0.5965

(b) Homocedasticidad S_1^{maximo}

Tabla 6.3: Resultados test Wilcoxon - Mejor tamaño de ventana vs máximo.

6.3.2. Filtrado: filtro de la media móvil de orden 3 vs no aplicar el filtro

Como complemento a las dos suposiciones anteriores, y como ya se estaba utilizando en ellas, se ha querido probar si hay diferencias estadísticamente significativas entre aplicar el suavizado de la media móvil de orden 3 a los datos o no. En este caso, se ha utilizado la prueba del test de Wilcoxon aplicada a un rango de ventanas entre 4 y 10. La razón se encuentra en que los resultados no tienen un crecimiento o decrecimiento constante, pudiéndose decir que se producen picos continuos, y de cara a tomar decisiones para automatizar todo el proceso, no se sabrá cuál es el tamaño de ventanas óptimo.

Visualmente se veía que el filtro de la media móvil de orden 3 beneficiaba a los resultados cuando ya eran buenos. En el dominio del tiempo, de manera general, sí parecía tener mejoras en algunas componentes, mientras que en el dominio de la frecuencia parecía perjudicar siempre. El test de Wilcoxon, como se ve en la tabla 6.4 (a), muestra p-valores grandes en todas las componentes, indicativos de que el filtro de la media móvil es indiferente, no produciendo mejoras estadísticamente significativas. En cambio, en el dominio de la frecuencia, para la componente X y el módulo obtiene p-valores pequeños; indicativos de que sí hay diferencias y de verdad el filtro está perjudicando a los resultados. Para las componentes Y y Z los p-valores son grandes e indicarían que las diferencias no son significativas. El test de homocedasticidad se mostrará, en todos los casos, en la tabla de la derecha y si no existen problemas, no se dirá nada.

Como había que tomar una decisión, tratando de buscar un sistema final razonablemente bueno, se ha decidido aplicar el filtro de la media móvil de orden 3 en el Dominio del Tiempo, dado que, aunque no produzca mejoras estadísticamente significativas, sí consigue EER mejores en la mayoría de las componentes; y en el Dominio de la Frecuencia continuar con los datos interpolados sin aplicar el filtro de la media móvil, ya que además de perjudicar, en algunos casos esas pérdidas sí eran significativas.

6.3.3. Calidad de la señal

Con el objetivo de comprobar la eliminación del ruido de manera manual, se ha probado a sobre ello, eliminar ventanas con baja autocorrelación para estudiar la influencia de coger solo ventanas de autocorrelación alta. Se han considerado las siguientes dos opciones.

- Eliminar ventanas con **valor de autocorrelación fijado**: utilizando los valores de 0.3 y 0.4. Es decir, para cada usuario y toma de datos se han eliminado todas las ventanas con autocorrelación inferior a 0.3 o 0.4, respectivamente, independientemente de que dicho usuario no tuviera ventanas con valores superiores a dicho valor.
- Eliminar ventanas con **valor de autocorrelación fijado en porcentaje sobre su máximo**: se han utilizado $\frac{2}{4}$, $\frac{2}{3}$ y $\frac{3}{4}$, de manera que para cada usuario, sesión, muestra y compo-

	P-valor $S_2^{filtroMM3}$
X + DT	0.9459
Y + DT	0.6347
Z + DT	0.4013
MÓDULO + DT	0.9459
X + DF	**0.069**
Y + DF	0.1936
Z + DF	0.8743
MÓDULO + DF	**0.0939**

(a) p-valor test Wilcoxon $S_2^{filtroMM3}$

	Homocedasticidad $S_2^{filtroMM3}$
X + DT	0.9066
Y + DT	0.7824
Z + DT	0.5325
MÓDULO + DT	0.6013
X + DF	0.9662
Y + DF	0.3434
Z + DF	0.6793
MÓDULO + DF	0.6267

(b) Homocedasticidad $S_2^{filtroMM3}$

Tabla 6.4: Resultados test Wilcoxon - Filtro de la media móvil de orden 3 vs no filtro.

nente se ha buscado el valor máximo de la autocorrelación en todas sus ventanas y se han eliminado aquellas con valores inferiores al porcentaje de ese valor.

De esta manera, si los resultados mejoran podrían ser indicativos de que no se ha eliminado el ruido demasiado bien y todavía quedan cosas por eliminar.

Los gráficos de las figuras 6.4 y 6.5 muestran para el dominio del tiempo y de la frecuencia, respectivamente, la evolución de tamaños de ventana de 3 a 11 y el efecto de eliminar únicamente el ruido de manera manual (línea continua) y a mayores eliminar ventanas con baja autocorrelación según los criterios mencionados.

En ambos dominios las conclusiones son las mismas. Dejando fijo el valor de la autocorrelación se terminan perdiendo usuarios a medida que aumenta el tamaño de la ventana. En caso de perder usuarios, el resultado es *NA* (missing) y en el gráfico desaparece la línea. Esto es algo que no se puede permitir ya que, al tener distinto número de usuarios, los resultados no serían comparables. Además, la base de datos utilizada tiene muy pocos usuarios como para permitir perder más. Por otro lado, utilizando porcentajes de los valores máximos en cada usuario/sesión/muestra y componente, se puede ver que, a mayor porcentaje, mejores resultados y cómo a medida que aumenta el tamaño de las ventanas, funciona aún mejor.

Visualmente, los resultados obtenidos eliminando las ventanas con autocorrelación inferior al producto de $\frac{3}{4}$ y la máxima autocorrelación de dicho usuario/sesión/muestra y componente son los mejores, produciendo bastantes mejoras en las componentes malas y dejando tal cual o empeorando ligeramente las componentes buenas. Se ha realizado el *test de Wilcoxon* por rangos de ventanas entre 6 y 10, para los casos en que no se pierden usuarios, que son cuando el valor de la autocorrelación es variable. Los resultados que pueden verse en la tabla 6.5 (a), muestran p-valores altos en todos los casos, indicando que en ninguno de los dos dominios hay diferencias

estadísticamente significativas y por tanto, da igual eliminar las ventanas de baja autocorrelación o no, pero aun así es una manera de poder llegar a conseguir mejores tasas de equierror.

6.3.4. Eliminación del ruido automática

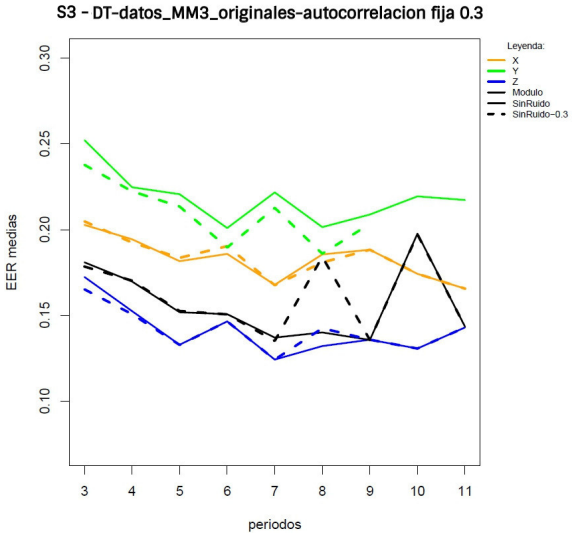
Buscando paliar la subjetividad de eliminar el ruido de manera manual y tras ver los resultados beneficiosos de eliminar ventanas con baja autocorrelación habiendo ya eliminado el ruido, se ha decidido probar si existían diferencias significativas entre eliminar el ruido manualmente y utilizar los datos tal cual, eliminando solamente aquellas ventanas con baja autocorrelación. Esto disminuiría enormemente el trabajo de preprocesamiento y estar más cerca de poder encontrar un sistema de reconocimiento automático. Los criterios para fijar el valor de autocorrelación son los mismos que en el apartado anterior.

Los gráficos de las figuras 6.6 y 6.7 muestran para el dominio del tiempo y de la frecuencia, respectivamente, la evolución de tamaños de ventana de 3 a 11. En ambos dominios las conclusiones son las mismas. En este caso, dejando fijo el valor de la autocorrelación no se pierden usuarios, pero para nuevos datos podrían perderse ya que es un criterio muy fuerte. Se ve que cuando la componente es mala, los resultados eliminando el ruido con el criterio de la autocorrelación son similares o incluso mejores; pero cuando la componente es buena se ve perjudicada, obteniendo peores resultados. No obstante, los mejores resultados se vuelven a obtener eliminando las ventanas con autocorrelación inferior al producto de $\frac{3}{4}$ y la máxima autocorrelación de dicho usuario/sesión/muestra y componente, existiendo menos diferencias en las componentes buenas y resultando visualmente una buena opción.

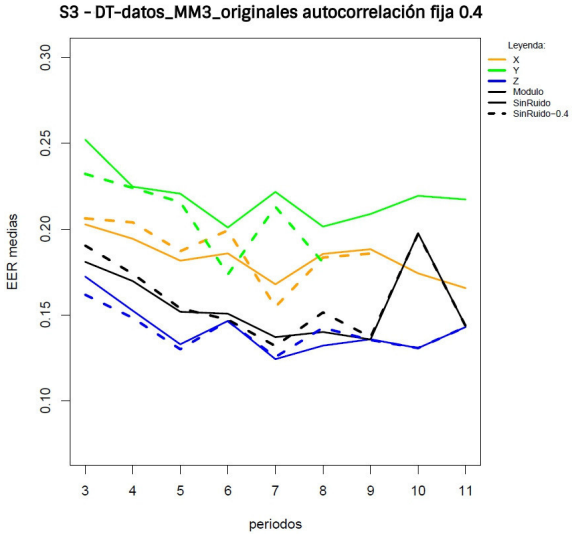
El test de Wilcoxon se ha vuelto a realizar por rangos de ventanas entre 6 y 10. Los resultados pueden verse en la tabla 6.6 (a), pero para todos los criterios el p-valor es grande, indicando que en ninguno de los dos dominios hay diferencias estadísticamente significativas y por tanto, daría igual eliminar el ruido eliminando las ventanas con autocorrelación inferior a $\frac{3}{4}$ del valor máximo, en cada caso. No obstante, y aunque los p-valores son muy altos, se tienen pocos datos como para fiarse y utilizar como única razón para tomar decisiones el test no paramétrico de Wilcoxon.

6.3.5. Fusión de ventanas

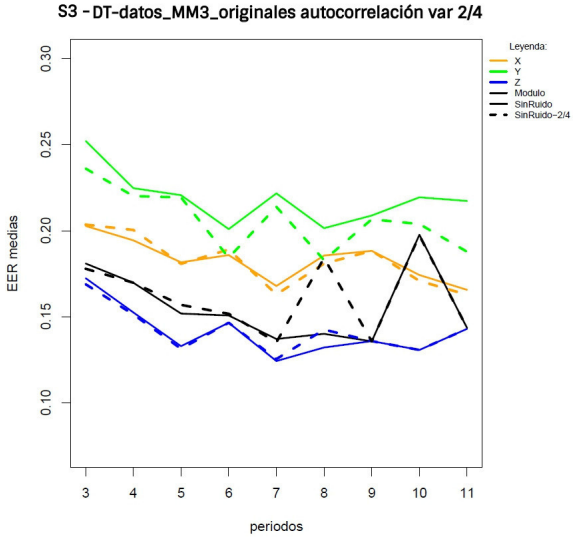
Se trata de fusionar los scores obtenidos con el clasificador de k-vecinos más próximos para un tamaño de ventana concreto. La idea es, fijado un nuevo tamaño de ventana para la fusión y el número de ellas solapadas, calcular nuevos scores aplicando algún estadístico. Su funcionamiento para tamaño de ventana 3 y solapamiento 1 puede verse en la figura 6.8, donde por separado, para los usuarios auténticos e impostores, se obtienen los nuevos scores finales usados para calcular la métrica. Los scores finales, en este ejemplo, se calcularían como se indica en (6.3).



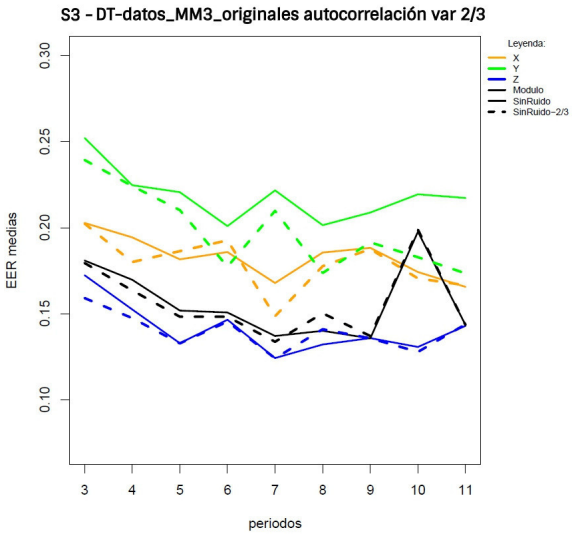
(a) Valor fijo 0.3 $S_3^{calidadSeñal}$



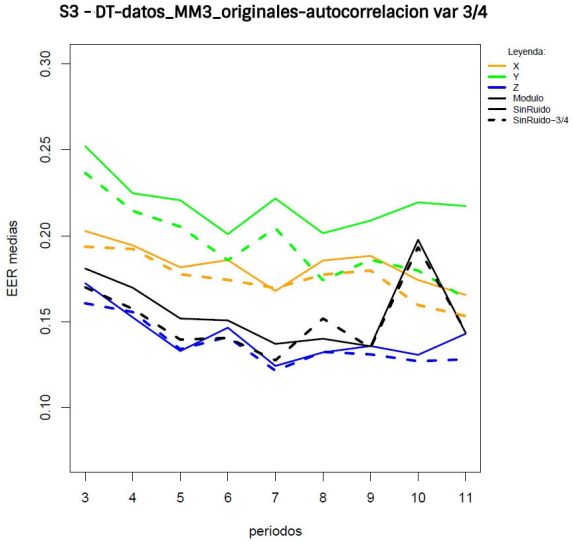
(b) Valor fijo 0.4 $S_3^{calidadSeñal}$



(c) Valor variable 2/4 $S_3^{calidadSeñal}$

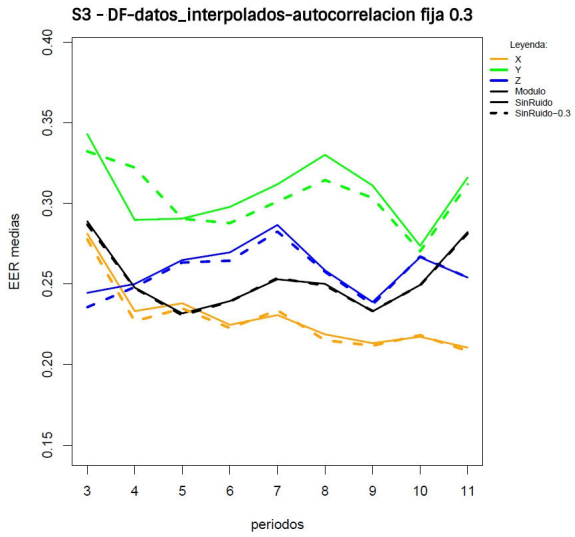


(d) Valor variable 2/3 $S_3^{calidadSeñal}$

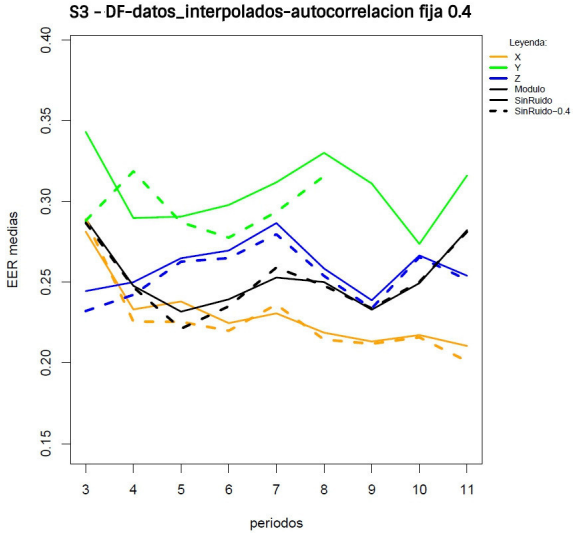


(e) Valor variable 3/4 $S_3^{calidadSeñal}$

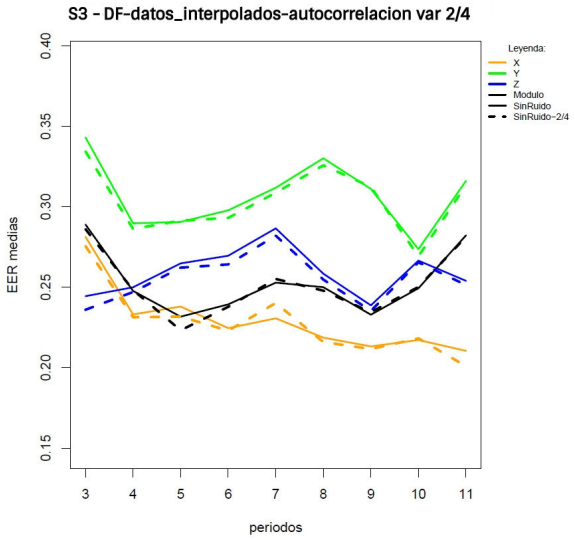
Figura 6.4: Dominio del Tiempo - Eliminar autocorrelación manual y ventanas de baja autocorrelación.



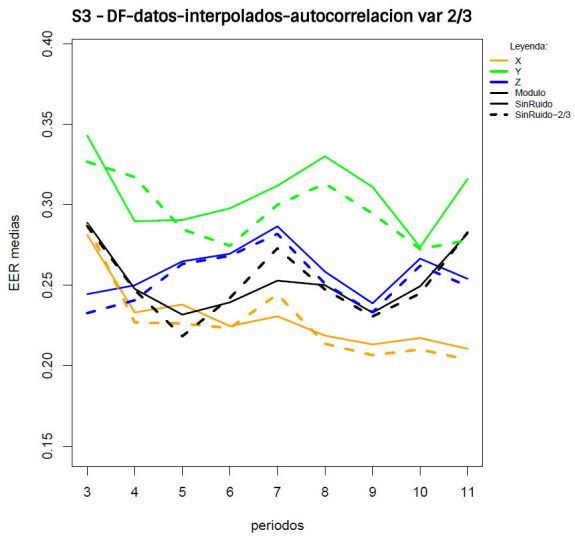
(a) Valor fijo 0.3 $S_3^{calidadSeñal}$



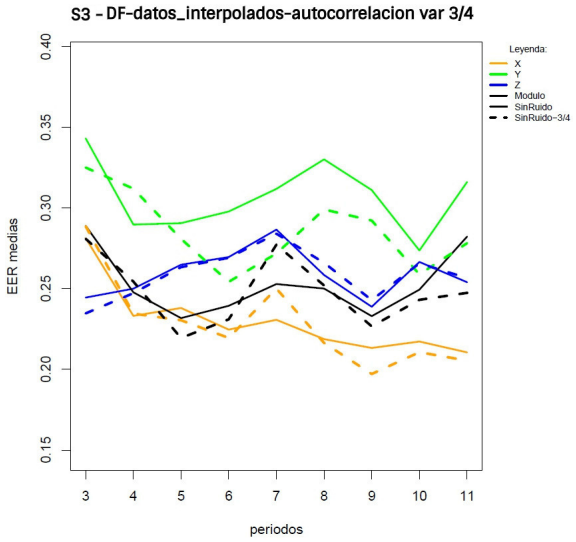
(b) Valor fijo 0.4 $S_3^{calidadSeñal}$



(c) Valor variable 2/4 $S_3^{calidadSeñal}$



(d) Valor variable 2/3 $S_3^{calidadSeñal}$



(e) Valor variable 3/4 $S_3^{calidadSeñal}$

Figura 6.5: Dominio de la Frecuencia - Eliminar autocorrelación manual y ventanas de baja autocorrelación.

	P-valor $S_3^{calidadSeñal}$ (RV: 6 a 10)		
	S_3 2/3	S_3 2/4	S_3 3/4
X + DT	1	0.8388	0.91
Y + DT	0.7131	0.8903	0.7131
Z + DT	0.982	0.9816	0.982
MÓDULO + DT	0.8036	0.7688	0.91
X + DF	0.982	0.8743	0.7688
Y + DF	0.7345	0.8388	0.6027
Z + DF	0.8388	0.8743	1
MÓDULO + DF	0.9459	0.8743	1

(a) p-valor test Wilcoxon $S_3^{calidadSeñal}$

	Homocedasticidad $S_3^{calidadSeñal}$ (RV: 6 a 10)		
	S_3 2/3	S_3 2/4	S_3 3/4
X + DT	0.9145	0.9772	0.9871
Y + DT	0.9197	0.9647	0.8457
Z + DT	0.8048	0.9855	0.7713
MÓDULO + DT	0.9214	0.8572	0.8921
X + DF	0.5874	0.8777	0.6382
Y + DF	0.868	0.9527	0.6771
Z + DF	0.9557	0.8839	0.6204
MÓDULO + DF	0.7704	0.9091	0.6496

(b) Homocedasticidad $S_3^{calidadSeñal}$

Tabla 6.5: Resultados test Wilcoxon - Eliminar autocorrelación manual y ventanas de baja autocorrelación.

$$\begin{aligned} S_1^{a*} &= estadistico(S_1^a, S_2^a, S_3^a) \\ S_2^{a*} &= estadistico(S_3^a, S_4^a, S_5^a) \\ &\dots \end{aligned} \tag{6.3}$$

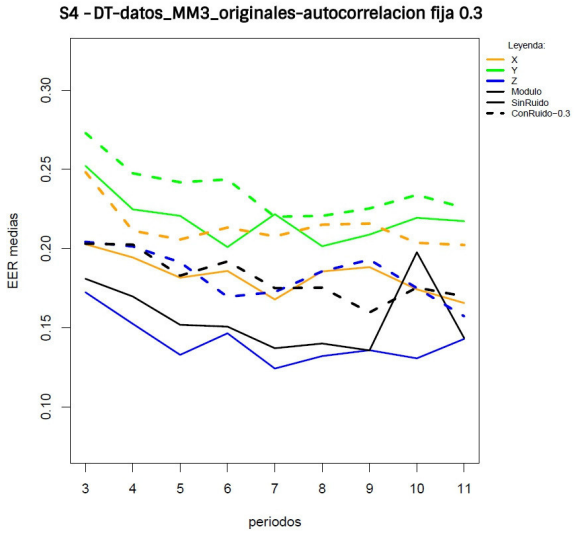
Si esto funcionará bien significaría que en esta biometría beneficia tener muchos datos para poder extraer ventanas, no necesariamente grandes, con las que obtener características e irlas fusionando con este criterio. Pasamos a mostrar los resultados con los distintos estadísticos probados.

6.3.5.1. Aplicar varias ventanas a los scores con el estadístico de la media vs no modificar los scores

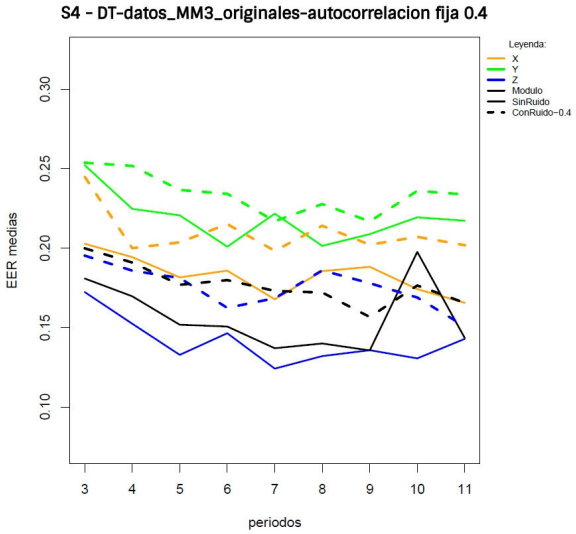
En este primer caso se ha utilizado como estadístico la **media** y dada la poca cantidad de muestras que se tienen en cada dato de los usuarios, se han empleado el siguiente número de ventanas para la fusión y solapamiento.

- Tamaño de ventana para la fusión 3, solapamiento 1.
- Tamaño de ventana para la fusión 4, solapamiento 1.
- Tamaño de ventana para la fusión 4, solapamiento 2.

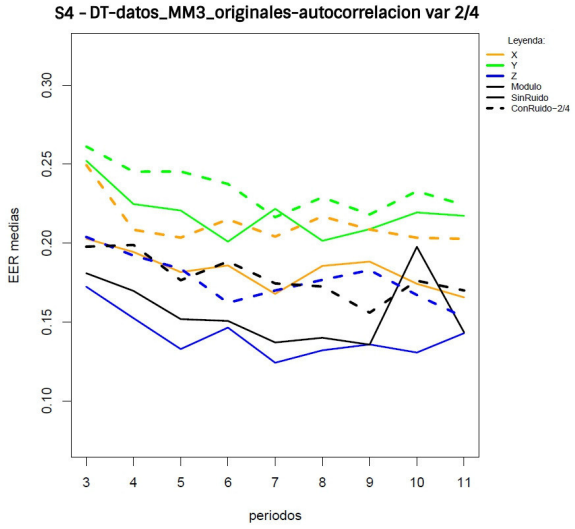
Visualmente, en los gráficos de las figuras 6.9 y 6.10, correspondientes al dominio del tiempo y de la frecuencia respectivamente, se muestra la evolución de tamaños de ventana de 3 a 11 utilizando está fusión (línea discontinua) y con los datos eliminando únicamente el ruido de manera manual (línea continua). En ambos dominios, todas las componentes consiguen mejores resultados tras la fusión. Las mejoras aumentan al incrementar el tamaño de la ventana y el solapamiento, resultando mejor tamaño 4 y solapamiento 2.



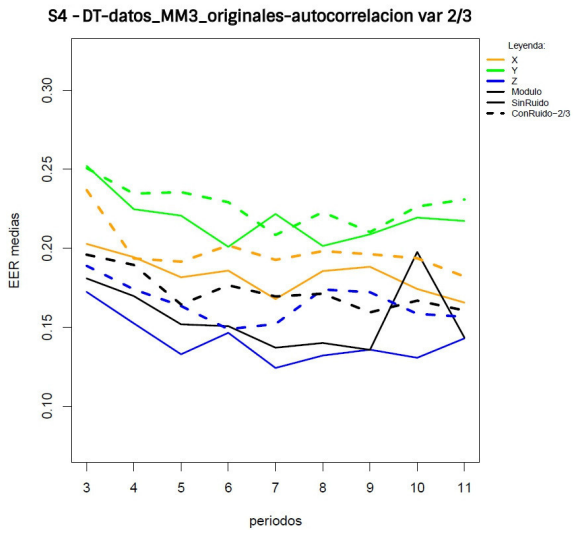
(a) Valor fijo 0.3 $S_4^{ruidoAutomático}$



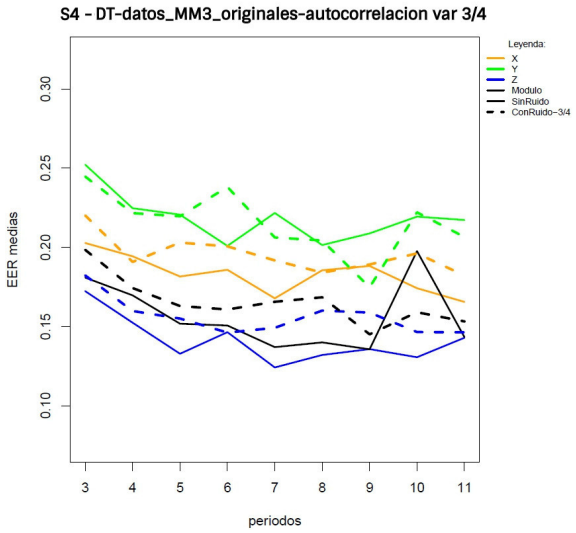
(b) Valor fijo 0.4 $S_4^{ruidoAutomático}$



(c) Valor variable 2/4 $S_4^{ruidoAutomático}$

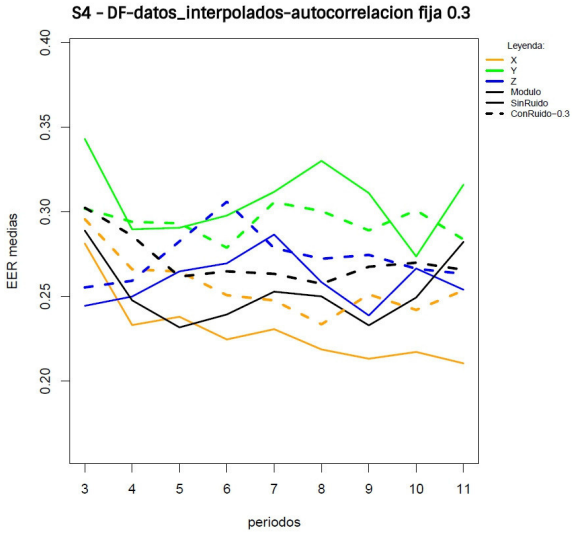


(d) Valor variable 2/3 $S_4^{ruidoAutomático}$

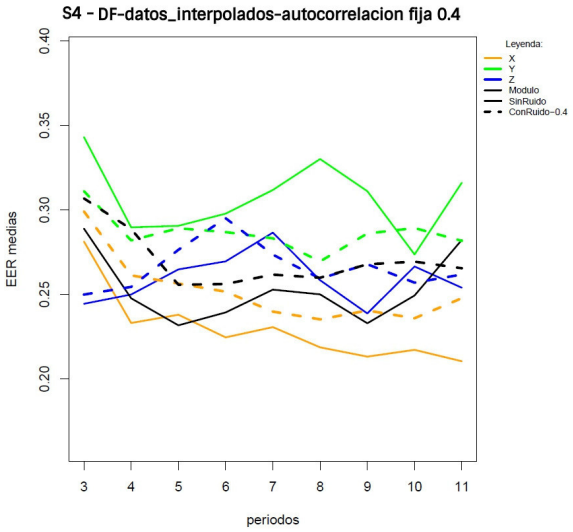


(e) Valor variable 3/4 $S_4^{ruidoAutomático}$

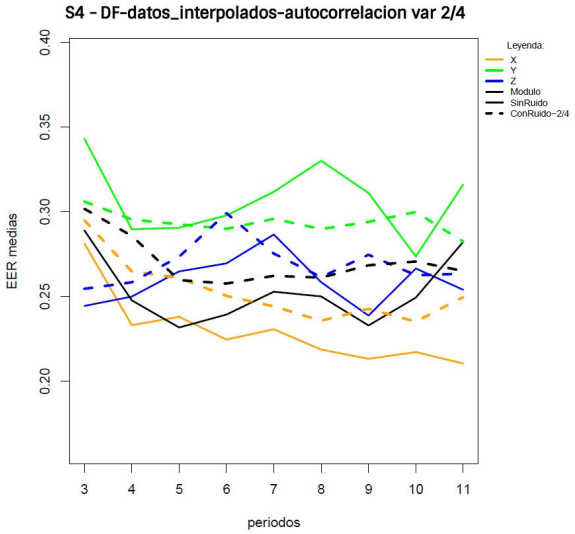
Figura 6.6: Dominio del Tiempo - Eliminar ruido manualmente y datos originales eliminando baja autocorrelación.



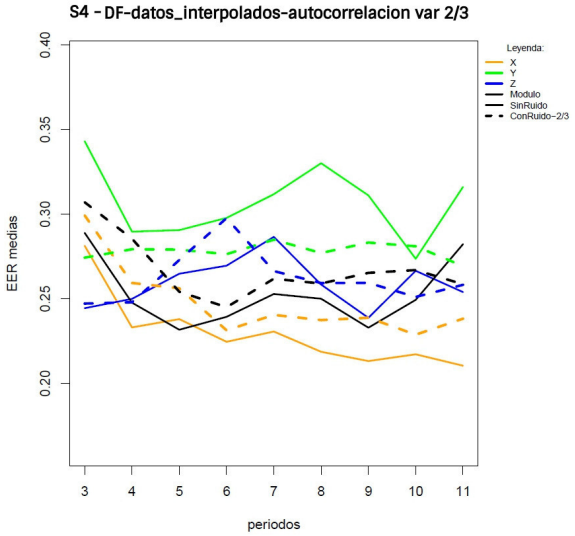
(a) Valor fijo 0.3 $S_4^{ruidoAutomático}$



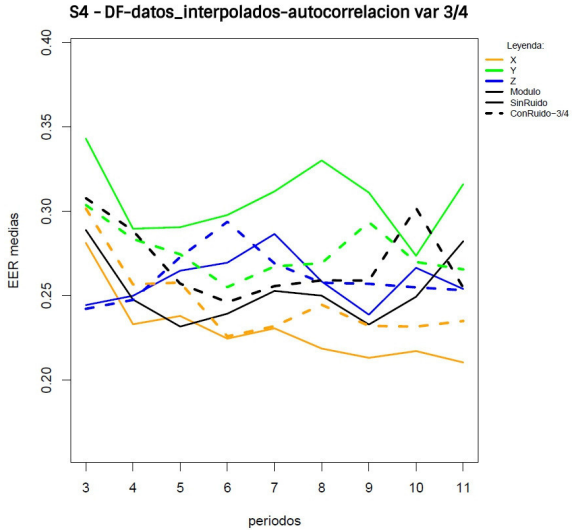
(b) Valor fijo 0.4 $S_4^{ruidoAutomático}$



(c) Valor variable 2/4 $S_4^{ruidoAutomático}$



(d) Valor variable 2/3 $S_4^{ruidoAutomático}$



(e) Valor variable 3/4 $S_4^{ruidoAutomático}$

Figura 6.7: Dominio de la Frecuencia - Eliminar ruido manualmente y datos originales eliminando baja autocorrelación.

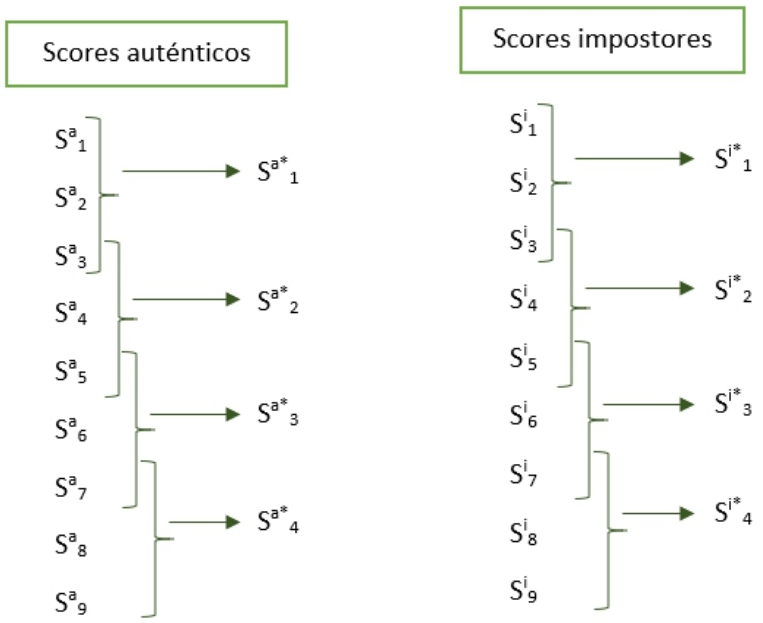
	P-valor $S_4^{ruidoAutomático}$ (RV: 6 a 10)				
	S_4 FIJO 0.3	*S_4 FIJO 0.4*	*S_4 var 2/3*	*S_4 var 2/4*	*S_4 var 3/4*
X + DT	*0.5112*	*0.5409*	*0.6347*	*0.5112*	*0.6673*
Y + DT	*0.7688*	*0.8388*	*0.8036*	*0.7345*	*0.9459*
Z + DT	***0.1636***	*0.21*	*0.3064*	*0.1936*	*0.4824*
MÓDULO + DT	*0.4013*	*0.4824*	*0.5112*	*0.4544*	*0.7345*
X + DF	0.2852	0.5409	0.5714	0.4544	0.4824
Y + DF	0.7006	0.8743	0.9459	0.6347	0.91
Z + DF	0.7345	0.8036	0.91	0.8036	0.8743
MÓDULO + DF	0.3519	0.3287	0.4274	0.3761	0.7345

(a) p-valor test Wilcoxon $S_4^{ruidoAutomático}$

	Homocedasticidad $S_4^{ruidoAutomático}$ (RV: 6 a 10)				
	S_4 FIJO 0.3	*S_4 FIJO 0.4*	*S_4 var 2/3*	*S_4 var 2/4*	*S_4 var 3/4*
X + DT	*0.9235*	*0.9825*	*0.8658*	*0.9391*	*0.8868*
Y + DT	***0.0466***	***0.0737***	***0.0945***	***0.0857***	*0.3283*
Z + DT	*0.5922*	*0.7823*	*0.6709*	*0.7733*	*0.91*
MÓDULO + DT	*0.7614*	*0.6315*	*0.6291*	*0.7614*	*0.676*
X + DF	0.8313	0.9966	0.9454	0.9559	0.6817
Y + DF	0.9453	0.7802	0.9957	0.8693	0.9349
Z + DF	0.4081	0.2559	0.3066	0.3085	0.337
MÓDULO + DF	0.4359	0.7381	0.9262	0.6448	0.5867

(b) Homocedasticidad $S_4^{ruidoAutomático}$

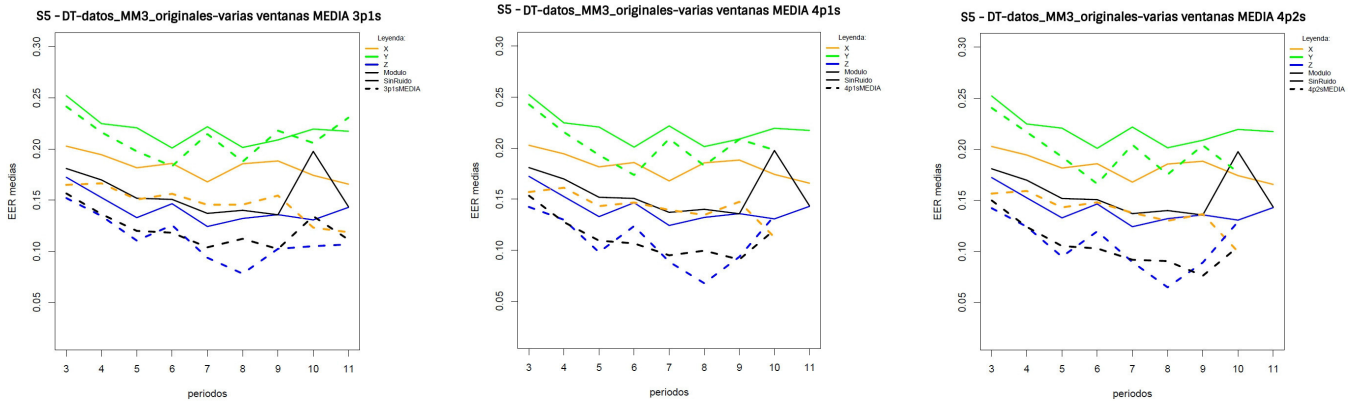
Tabla 6.6: Resultados test Wilcoxon - Eliminar ruido manualmente y datos originales eliminando baja autocorrelación.



(a) usuarios auténticos (b) usuarios impostores

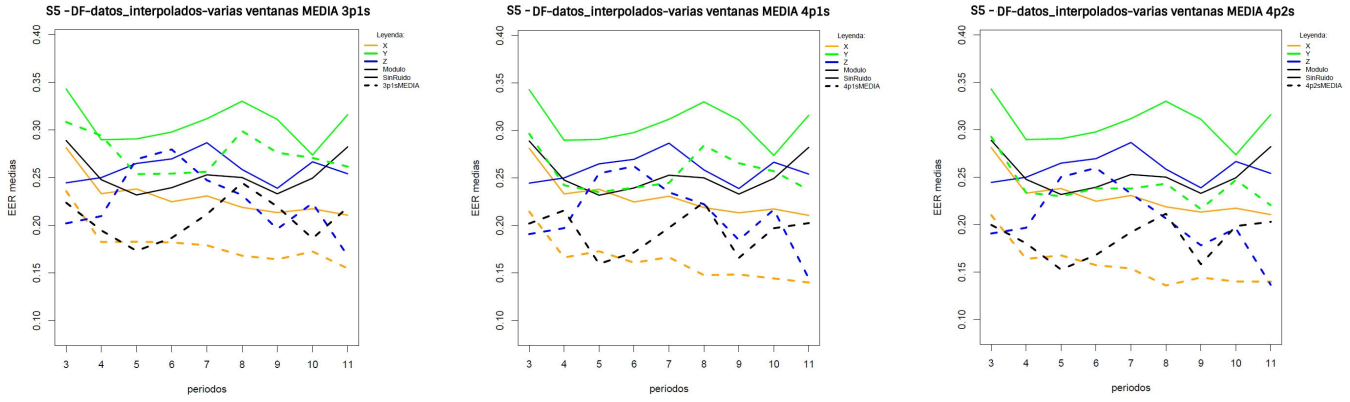
Figura 6.8: Funcionamiento de aplicar varias ventanas.

El *test de Wilcoxon* aplicado a rangos de ventanas entre 6 y 10, con tamaño de ventana para la fusión 3 y solapamiento 1, indica que en el dominio del tiempo no hay diferencias estadísticamente significativas, pero en el dominio de la frecuencia, en la componente X y el módulo sí las hay, siendo precisamente éstas, las componentes que mejor funcionan en este dominio. Para las componentes Y y Z no se obtiene p-valor alto, pero es razonable porque se tienen pocos datos y el test no es capaz de detectar las diferencias, se puede ver en la tabla 6.7 (a). Por otro lado, para tamaño de ventana para la fusión 4 y solapamiento 2, cuya tablas es la 6.8 (a), se ve como en todos los casos los p-valores son más pequeños, prácticamente la mitad, aunque las conclusiones se mantienen, pero es una manera de ver que estos parámetros de la fusión funcionan mejor y que ésta es probablemente una buena manera de mejorar los resultados.



(a) 3ventanas/1solapamiento (b) 4ventanas/1solapamiento (c) 4ventanas/2solapamiento

Figura 6.9: Dominio del Tiempo - Aplicar varias ventanas con la media.



(a) 3ventanas/1solapamiento (b) 4ventanas/1solapamiento (c) 4ventanas/2solapamiento

Figura 6.10: Dominio de la Frecuencia - Aplicar varias ventanas con la media.

	P-valor $S_5^{FusionMedia}$ 3p1s
X + DT	*0.4483*
Y + DT	*0.9268*
Z + DT	***0.1251***
MÓDULO + DT	*0.3462*
X + DF	**0.0497**
Y + DF	0.2852
Z + DF	0.4013
MÓDULO + DF	**0.0395**

(a) p-valor test Wilcoxon $S_5^{FusionMedia3p1s}$

	Homocedasticidad $S_5^{FusionMedia}$ *3p1s*
X + DT	*0.6404*
Y + DT	*0.4729*
Z + DT	*0.7367*
MÓDULO + DT	*0.7615*
X + DF	0.4444
Y + DF	0.3967
Z + DF	0.2022
MÓDULO + DF	0.6987

(b) Homocedasticidad $S_5^{FusionMedia3p1s}$

Tabla 6.7: Resultados test Wilcoxon - Aplicar varias ventanas con la media, tamaño 3, solapamiento 1.

	P-valor $S_5^{FusionMedia}$ 4p2s
X + DT	*0.26*
Y + DT	*0.6623*
Z + DT	***0.0847***
MÓDULO + DT	*0.1609*
X + DF	**0.0274**
Y + DF	0.1371
Z + DF	0.21
MÓDULO + DF	**0.0141**

(a) p-valor test Wilcoxon $S_5^{FusionMedia4p2s}$

	Homocedasticidad $S_5^{FusionMedia}$ *4p2s*
X + DT	*0.7428*
Y + DT	*0.6002*
Z + DT	*0.8631*
MÓDULO + DT	*0.5148*
X + DF	0.3131
Y + DF	0.5032
Z + DF	0.2655
MÓDULO + DF	0.4854

(b) Homocedasticidad $S_5^{FusionMedia4p2s}$

Tabla 6.8: Resultados test Wilcoxon - Aplicar varias ventanas con la media, tamaño 4, solapamiento 2.

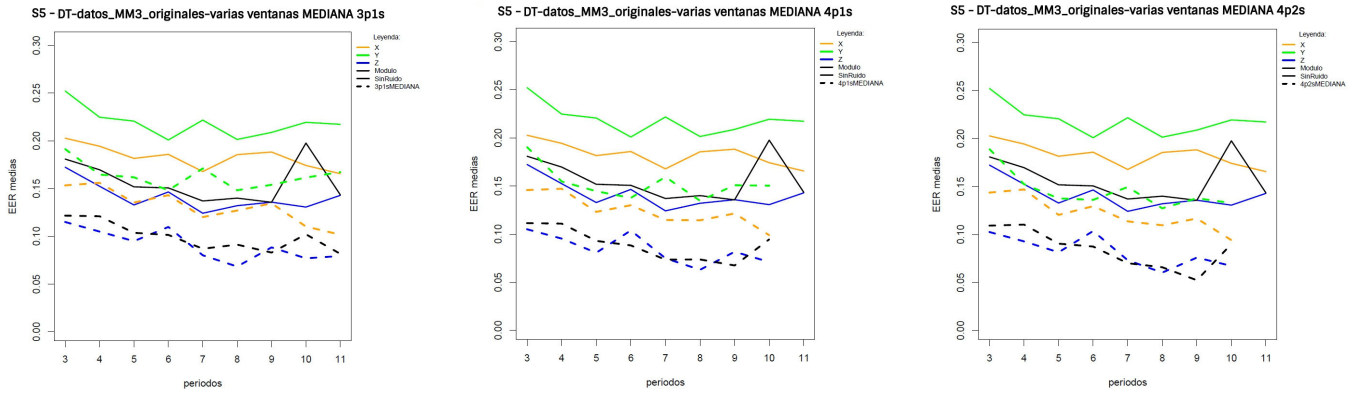
6.3.5.2. Aplicar varias ventanas a los scores con el estadístico de la mediana vs no modificar los scores

Como complemento al subapartado anterior y dado el buen funcionamiento de fusionar varias ventanas aplicando el estadístico de la media, se han probado otros estadísticos, que son la mediana, el mínimo y el máximo y se ha decidido incluir el estudio de la mediana, por ser el que mejor rendimiento mostró.

El funcionamiento es el mismo, lo único que cambia es el estadístico, usando ahora la **mediana** y probando los mismos tamaños de ventana y solapamiento que antes.

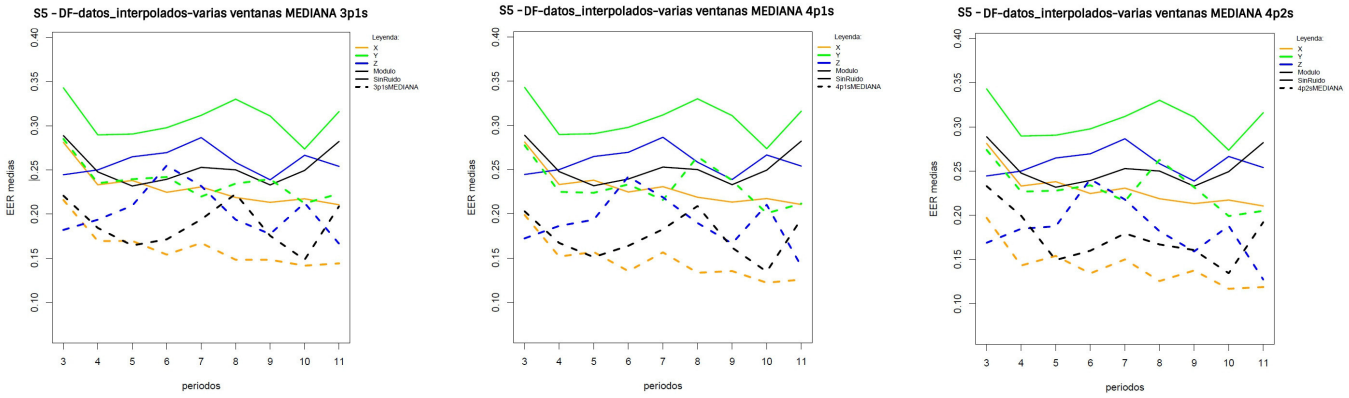
De manera visual, en los gráficos de las figuras 6.11 y 6.12 para el dominio del tiempo y de la frecuencia respectivamente, se muestra la evolución de los tamaños de ventana de 3 a 11 utilizando la fusión de la mediana en la línea discontinua y con los datos eliminando únicamente el ruido de manera manual en la línea continua. De nuevo, los resultados son los mismos, en ambos dominios, todas las componentes mejoran, incrementando la mejora con el tamaño de la ventana y el solapamiento, resultando mejor tamaño 4 y solapamiento 2.

El test de Wilcoxon aplicado a rangos de ventana entre 6 y 10 en el mejor caso, es decir, con la fusión de tamaño 4 y solapamiento 2, consigue en el dominio del tiempo p-valores más pequeños que con el estadístico de la media, sobre todo en la componente Z y en el módulo, pero también en la componente X e Y, donde se encuentran próximos al 10 %. En el dominio de la frecuencia consigue p-valores pequeños en las componentes X, Y y el módulo, una componente más que antes, demostrando al igual que se puede ver visualmente en la tabla 6.9 (a) que, parece funcionar mejor el estadístico de la mediana que el de la media, detectando más diferencias estadísticamente significativas para poder decir que funciona bien la fusión y mejor con el estadístico de la mediana.



(a) 3ventanas/1solapamiento (b) 4ventanas/1solapamiento (c) 4ventanas/2solapamiento

Figura 6.11: Dominio del Tiempo - Aplicar varias ventanas con la mediana.



(a) 3ventanas/1solapamiento (b) 4ventanas/1solapamiento (c) 4ventanas/2solapamiento

Figura 6.12: Dominio de la Frecuencia - Aplicar varias ventanas con la mediana.

	P-valor $S_5^{FusionMediana}$ 4p2s
X + DT	***0.1027***
Y + DT	***0.1127***
Z + DT	***0.0363***
MÓDULO + DT	***0.0365***
X + DF	0.0091
Y + DF	**0.062**
Z + DF	0.1499
MÓDULO + DF	0.0049

(a) p-valor test Wilcoxon $S_5^{FusionMediana3p1s}$

	Homocedasticidad $S_5^{FusionMediana}$ *4p2s*
X + DT	*0.679*
Y + DT	*0.8781*
Z + DT	*0.3538*
MÓDULO + DT	***0.1044***
X + DF	0.7501
Y + DF	0.4942
Z + DF	0.3064
MÓDULO + DF	0.6841

(b) Homocedasticidad $S_5^{FusionMediana3p1s}$

Tabla 6.9: Resultados test Wilcoxon - Aplicar varias ventanas con la mediana.

6.4. Análisis de resultados

Utilizando los resultados obtenidos en el apartado anterior y con el incremento en el conocimiento del problema que se había conseguido tras las diversas pruebas realizadas, se tomaron la siguiente serie de decisiones para reducir el problema y seguir buscando mejoras sobre un problema más concreto.

Para el primer parámetro del tamaño de la ventana, se decidió reducir la prueba a un rango de tamaños entre 6 y 10, cuando hasta ahora se habían probado rangos más amplios, concretamente entre 2 y 15, reduciéndolo después a 3 y 11 ciclos.

Respecto al segundo parámetro de aplicar alguna técnica de suavizado a los datos, se deja como una cuestión abierta que se debe probar. No se han tenido suficientes datos como para poder extraer buenas conclusiones. En este proyecto y dado que hay que continuar por un único camino, se ha decidido aplicar el filtro de la media móvil de orden 3 al dominio del tiempo, pero no al dominio de la frecuencia.

En la tercera y cuarta suposición de eliminar el ruido de manera manual o automática, dado las ventajas que genera se ha decidido que merece la pena hacer una segmentación automática basada en la autocorrelación con el criterio de eliminar sobre los datos originales, todas aquellas ventanas con autocorrelación inferior a $\frac{3}{4}$ del valor de la autocorrelación máxima, en valor absoluto, de cada usuario/sesión/muestra y componente. Hay que contemplar la posibilidad de que $\frac{3}{4}$ sea un criterio muy fuerte y se pierdan las muestras de algunos usuarios. Como no es posible perder usuarios, ya que ni los resultados serían comparables ni podemos permitirlo con los pocos que se tienen se ha decidido ir comprobando el número de ventanas resultantes y si no es suficiente, ir reduciendo el valor del criterio de la autocorrelación en 5 centésimas hasta que sí se tengan suficientes ventanas.

Por último, en la quinta suposición de aplicar varias ventanas, se han conseguido resultados suficientemente buenos como para considerar la técnica de la fusión de los scores una cuestión abierta y necesaria de probar en esta biometría. Para continuar, en nuestro caso, se va a utilizar el estadístico de la mediana con tamaño de ventana en la fusión 4 y solapamiento 2.

6.5. Sistema final

En los apartados anteriores, se ha estudiado el comportamiento de los diversos parámetros, hasta obtener el mejor valor de cada uno. Ahora se pretende dar un rendimiento final que nos permita seguir el estudio y comparar los rendimientos en los distintos escenarios planteados:

- Procedimiento experimental: Monosesión-Monomuestra, Multisesión-Monomuestra, Multisesión-Multimuestra.
- Dispositivo: Motorola, Microsoft.

- Tipo de sensor en el dispositivo: Acelerómetro, Giroscopio.

Para ello, sigue habiendo una cuestión abierta, *¿qué tamaño de ventana utilizar?*. Se ha reducido el rango de posibles valores entre 6 y 10, pero no se conoce el valor óptimo, dependiendo del dominio y la componente varía entre los valores posibles. Aprovechando el buen funcionamiento evaluado en la técnica de fusión, se va a probar a obtenidos sobre ese sistema los scores fusionados de cada tamaño de ventana, entre 6 y 10 en este caso, volverlos a fusionar, obteniendo una única secuencia de valores fruto de los scores fusionados en todos los posibles tamaños de ventana. Evidentemente, no se tiene el mismo número de scores en cada tamaño, pero se ha solucionado seleccionando el tamaño de la longitud más pequeña y para cada tamaño de longitud más grande, eligiendo una secuencia equiespaciada de dichos valores. Para elegir la secuencia equiespaciada se han seleccionado el primer y último elemento junto con valores intermedios espaciados. Se puede ver un ejemplo de su funcionamiento en la figura 6.13, donde la longitud más pequeña es 5 y para la primera secuencia de longitud 10, se seleccionan los elementos marcados a color. Una vez conseguida la misma longitud en todos los tamaños, se ha aplicado la fusión de todos ellos con un nuevo estadístico, media, mediana, mínimo y máximo, para obtener la secuencia de scores final.

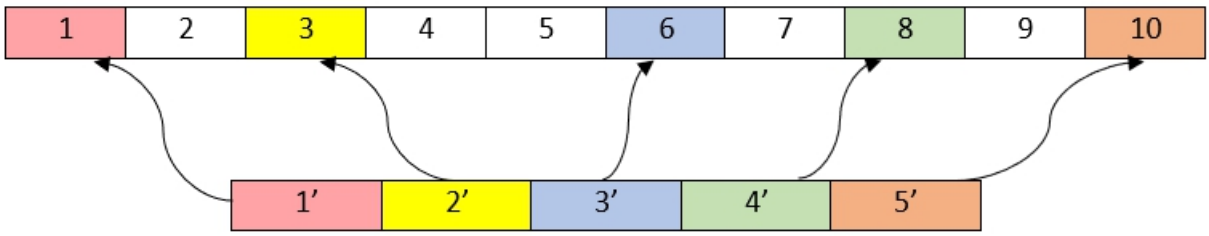


Figura 6.13: Secuencia equiespaciada de valores.

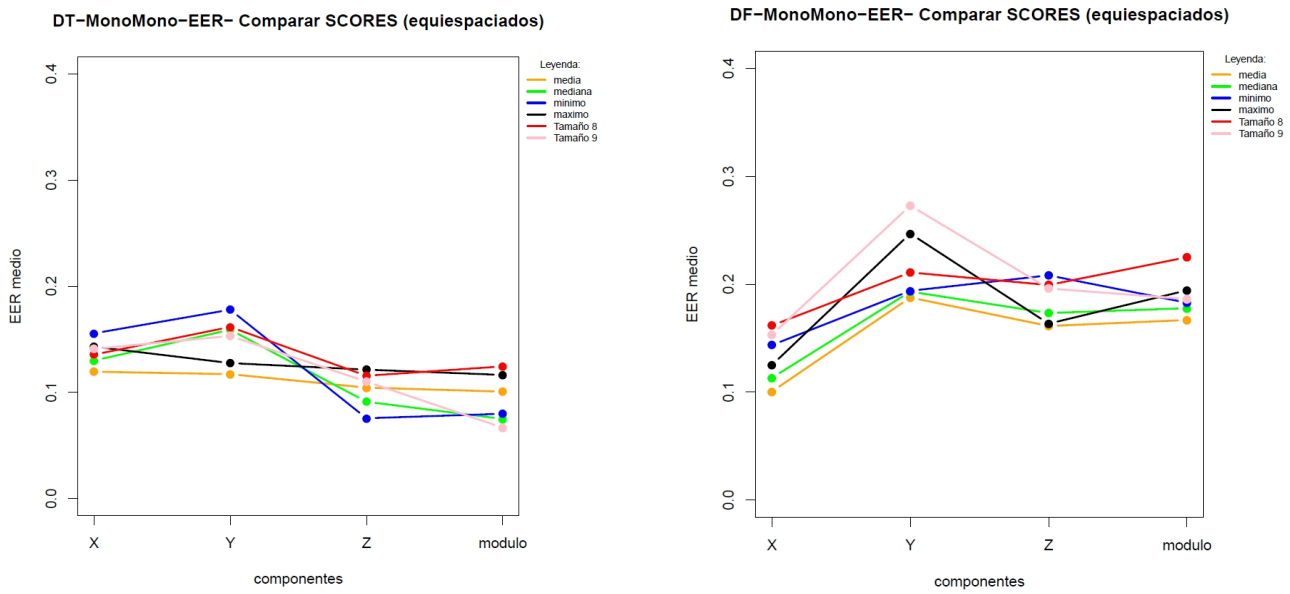
El objetivo es comparar está nueva fusión con los resultados obtenidos con un tamaño fijo de ventana, donde no siempre hay un óptimo. Utilizando en ambos casos, la técnica de fusión ya seleccionada con el estadístico de la mediana, tamaño de ventana para la fusión 4 y solapamiento 2, que ha sido el que parecía obtener mejores resultados en el apartado anterior.

En los gráficos de la figura 6.14 se puede ver el efecto de esta doble fusión, tras fusionar los scores a los que ya se les había aplicado la técnica de fusión en tamaños de ventana entre 6 y 10 aplicando los diferentes estadísticas mencionados y utilizando un tamaño fijo de ventana de 8 y 9. Se han elegido 8 y 9 porque de manera global eran los que mejor funcionaban. El eje de abscisas muestra las distintas componentes estudiadas y el eje de ordenadas la tasa de equierror media en todos los usuarios disponibles.

En el dominio de la frecuencia, los resultados escogiendo la fusión funcionan mejor en todas las componentes, obteniendo resultados bastante similares con la media y la mediana, pero resultando mejor la media. En el dominio del tiempo existen pocas diferencias entre utilizar tamaño fijo o la fusión, funcionando mejor el tamaño de 9 para el módulo, pero la fusión con la media en la

componente X. Lo que sí se puede ver es que la fusión con la media funciona de manera más estable en todas las componentes.

Como decisión, se va a construir el sistema final con la técnica mencionada en este apartado, de fusionar los rangos entre 6 y 10 con el estadístico final de la media, utilizando la técnica de fusión con el estadístico de la mediana cada 4 ventanas con 2 de solapamiento.



(a) Dominio del Tiempo

(b) Dominio de la frecuencia

Figura 6.14: Fusionar scores en rango de tamaños de 6 a 10 vs coger un tamaño fijo.

Capítulo 7

Experimentos: Selección de características

Todo el trabajo realizado en el estudio prospectivo donde se han analizado las técnicas de preprocesamiento en el capítulo 5 y se ha hecho un análisis de los parámetros en el capítulo 6 con los datos de la pulsera *Microsoft* por ser un entorno más cerrado que el del reloj *Motorola*, permitiendo la adquisición de los datos a través de su SDK, con mayor fiabilidad y mejor conexión para tomar las mejores decisiones. Respecto al sensor, el acelerómetro, por ser el más habitualmente utilizado en la bibliografía. Con las conclusiones obtenidas se ha construido un sistema de reconocimiento final que se ha aplicado al otro sensor de la pulsera *Microsoft*, giroscopio, por resultar muy parecido en autocorrelación y forma al acelerómetro. De la misma manera se ha aplicado a los dos sensores del reloj *Motorola* y se han utilizado todos los procedimientos experimentales diseñados, en los cuales se prueba la variabilidad del rasgo biométrico con el tiempo. En el **caso experimental intermedio** en el que se utilizan muestras de entrenamiento de distintas sesiones, las mejores tasas de equierror han sido las que se muestran en la tabla 7.1. Se puede ver que los resultados son muy similares en los dos dispositivos, pero algo diferentes al cambiar de sensor. Si se observa la información de cada usuario, se puede decir que existe una especie de "animalario", término ya conocido en biometría y que, dentro de esa variabilidad, existen usuarios malos que mejoran en alguna de sus componentes o al cambiar de sensor, por lo que podría parecer apropiada una combinación de ambos sensores.

Dispositivo	**DT-ACC**	**DT-GYR**	**DF-ACC**	**DF-GYR**
MICRO	0.1241	0.1026	0.1175	0.1681
MOTO	0.1142	0.124	0.1375	0.1531

Tabla 7.1: Resumen Mejores Resultados en Multisesión-Multimuestra.

Con el módulo de los datos, que es el que obtiene los mejores resultados globalmente y continuando con el **procedimiento experimental intermedio**, se ha seleccionado el mejor sensor de cada usuario, consiguiendo las tasas de equierror de la tabla 7.2. Estos serían los resultados si fuésemos capaces de predecir que sensor funciona mejor en cada usuario, pero son suficientes para

mostrar que una combinación de ellos podría ser apropiada, ya que ambos sensores, aunque son parecidos en autocorrelación y forma, no recogen la misma información.

Dispositivo	DT	DF
MICRO	0.0604	0.0887
MOTO	0.0761	0.1233

Tabla 7.2: Resumen Mejores Resultados - Combinando ACC/GYR - módulo - Multisesión-Multimuestra.

Este estudio se ha realizado en profundidad en [6], categorizando a los usuarios según se comportamiento y viendo su funcionamiento al cambiar de dispositivo y sensor.

Un número excesivo de características puede provocar sobreajuste. Una forma robusta de controlarlo es haciendo una selección de características, que lleva a sistemas más simples y con menor coste computacional, siempre que se controle que no empeore la calidad del procedimiento.

En este problema se están utilizando las características que se explicaron en el apartado 4.1 y que se muestran en la tabla 7.3, siendo un total de 12 en el dominio del tiempo y 13 en el dominio de la frecuencia. Coincide el mismo número de características en cada componente X/Y/Z y en el módulo, ya que se extraen los estadísticos para la componente individual utilizada o para el módulo, resultando el mismo número total.

No obstante, aunque en nuestro problema, no existan muchas características y no afecte al coste computacional, tener menos facilita la comprensión del problema.

Dominio	Características	Nºtotal
Dominio del Tiempo		
X/Y/Z/ módulo	Periodo, autocorrelación, media, mediana, máximo, mínimo, desviación estándar, rango, curtosis, cuantil 25 % y 75 %, coeficiente de asimetría	12/12/12/12
Dominio de la Frecuencia		
X/Y/Z/ módulo	Primera y segunda frecuencia dominante, primera y segunda amplitud dominante, área bajo la curva, estadísticos para la amplitud (media, mediana, desviación estándar, rango, curtosis, cuantil 25 % y 75 %, coeficiente de asimetría)	13/13/13/13

Tabla 7.3: Resumen de las características utilizadas en el problema.

7.1. Técnica de selección utilizada

Al aplicar la técnica de selección de características cambia ligeramente el problema, tal y como se había visto. Hasta ahora, se ha trabajado con muestras de la misma clase, ya que el clasificador

solamente necesitaba muestras del usuario para crear su patrón, pero para este capítulo se va a realizar una especie de estudio a posteriori, donde es necesario conocer la clase. Aunque el conjunto de entrenamiento, prueba auténtica y prueba impostor de cada usuario va a seguir siendo el mismo, es decir, el explicado en el apartado 4.3, ahora se va a tener una variable respuesta binaria que indicará si la instancia de prueba pertenece a un usuario auténtico (prueba auténtico) o a un usuario impostor (prueba impostor), ya que va a ser necesario juntar los dos conjuntos de prueba mencionados.

El nombre de la técnica es **Relief** y se categoriza como un *aprendizaje basado en instancias*. Su objetivo es ponderar la relevancia de cada atributo utilizando la distancia entre las instancias. Va muestreando instancias y comprobando la vecindad, su funcionamiento para cada usuario se resume en el algoritmo 4 [63].

Algorithm 4: Algoritmo de la técnica de selección Relief

Input Conjunto de atributos A
Número de instancias a muestrear r
Tabla de datos de prueba del usuario correspondiente X
Output Scores de cada atributo w

for *i* ***in*** *1* ***to*** r **do**
  $x \leftarrow instancia\_aleatoria\_de\_X$
  Encuentra:
  $h \leftarrow near\_hit$
  $m \leftarrow near\_miss$

  for *each attribute* A **do**
    $w_A \leftarrow w_A + \frac{1}{r} \cdot (d_A(m, x) - d_A(h, x))$
  end
end

Donde:

- **Aciertos cercanos *(near hits)*:** vecinos más próximos de la misma clase.
- **Fallos cercanos *(near miss)*:** vecinos más próximos de distinta clase.
- $\boldsymbol{d_A(x, y)}$ es la distancia del atributo A entre las muestras x e y.
- **Número de instancias a muestrear**, $\boldsymbol{r}$, se ha fijado a 12, igual que la frecuencia de muestreo que se ha utilizado en la interpolación lineal.
- El **número de vecinos** a encontrar para cada instancia muestreada, se ha fijado a 1, igual que la técnica que se ha estado utilizando de 1-vecino más próximo.

Se ha utilizado está técnica por estar relacionado su funcionamiento con el del clasificador utilizado, k-vecinos más próximos.

La manera de actuar va a ser, utilizando un tamaño fijo de ventana (el usado es 8 por ser uno de los que mejores resultados han conseguido de manera global) obtener el orden de las características más importantes para cada usuario junto con el peso asignado a cada una de ellas. Como se va a realizar la selección de características en cada posibilidad de dominio, se van a utilizar las decisiones que se tienen hasta el momento, es decir, extrayendo las características en los datos originales con el filtro de la media móvil de orden 3 en el Dominio del Tiempo y los datos interpolados en el Dominio de la Frecuencia. Una vez conseguidos los pesos de las características de cada usuario, se van a probar dos cosas diferentes:

- Considerar **cada usuario por separado**: probar las características (1, 3, 5, 7, 9, 11, máximo) más importantes de cada usuario y con cada posibilidad de número de características, ejecutar el modelo final construido, obtener y guardar la tasa de equierror de dicho usuario con ese número de características. En cada usuario se tendrá un orden de características diferente. Máximo es 12 en el Dominio del Tiempo y 13 en el Dominio de la Frecuencia.
- Considerar **todos los usuarios juntos**: teniendo la matriz que contiene por columnas las características en orden de cada usuario, junto con su peso, utilizar un valor umbral. Como en cada dominio, los pesos máximos son diferentes, se ha utilizado como valor umbral $\frac{1}{2}$ del máximo peso. Teniendo en cuenta todas las características que contienen en al menos un usuario, un peso mayor a dicho valor umbral. Ahora todos los usuarios tienen las mismas características con el mismo orden, pero diferente número total entre los distintos dominios. El peso de las características se ha asignado utilizando el máximo peso entre todos los usuarios que superasen el valor umbral. Probando todas las posibilidades de número de características entre 1 y el número total que superaban el valor umbral en dicho dominio.

Considerar cada usuario por separado es una manera de actuar menos frecuente en *clasificación*, pero muy frecuente en biometría, donde cada usuario, en definitiva, es un problema de clasificación independiente. Se puede ver considerando el campo de la *medicina* donde a cada paciente le funcionará mejor una terapia personalizada que una terapia general. Aquí la terapia personalizada es considerar cada usuario por separado y la terapia general considerar a todos los usuarios juntos.

7.2. Resultados obtenidos

Para las dos situaciones comentadas en el apartado anterior se van a generar como resultados tanto el número de características seleccionado como la tasa de equierror obtenida con los mismos. La mejor manera de mostrar los resultados es con un gráfico de líneas de cada usuario en cada dominio, donde el eje de abscisas va a ser el número de características y el eje de ordenadas la tasa de equierror. También se va a mostrar el EER medio de todos los usuarios para diferente número de características totales.

Estos resultados son una prueba inicial de lo que se puede conseguir realizando un análisis de las características, por lo que se va a utilizar como dispositivo, la pulsera de *Microsoft*, como sensor el acelerómetro y como procedimiento experimental *Monosesión-Monomuestra*, tal y como se ha utilizado al realizar el análisis de los parámetros en el capítulo 6.

7.2.1. Considerando cada usuario

Se está tratando a cada usuario como un problema de clasificación independiente, teniendo su propia selección de características y utilizando las (1, 3, 5, 7, 9, 11 y máximo) características más importantes para estudiar el efecto de tener menos características. En este caso, al estar utilizando el máximo, significa estar usando todas las características.

En las tablas referenciadas como 7.4 se muestra el EER medio de todos los usuarios utilizando sus k características más importantes entre 1 y el máximo. Las filas indican el número de características y las columnas, el dominio. En el dominio del tiempo, tabla de la izquierda, la componente X funciona mejor con todas las características. En la componente Y igual, pero utilizar una característica menos funciona de manera similar con diferencias inferiores al 1 %. La componente Z y el módulo funcionan mejor con 3 características menos, en el caso de la componente Z la diferencia con utilizar todas es inferior al 1 %, mientras que en el módulo mejora el resultado final en casi un 4 %. Por otro lado, en el Dominio de la Frecuencia, tabla de la derecha, se puede ver cómo utilizar menos características es positivo, consiguiéndose en las componentes Y y Z con 6 y 8 características menos, mejoras del 5 y 1.5 % en el resultado final, respectivamente. En la componente X y el módulo no se consiguen mejoras, pero sí resultados muy similares utilizando 6 características menos en ambos casos y empeorando la tasa de equierror media en menos de un 0.5 %.

En los gráficos de las figuras 7.1 y 7.2 se puede ver el comportamiento de cada usuario al variar el número de características, utilizando el módulo. En el Dominio del Tiempo, hay usuarios que tienen un comportamiento normal, con resultados muy similares para cualquier número de características, ejemplo de ello son el usuario 1 o el 15. Otros, como el 2, 8, 9, 10 o 13, con 1 característica consiguen el peor resultado y después van variando hasta conseguir el mejor con el máximo número de características. Y otros como el usuario 5 que consiguen su mejor valor en un número intermedio de 7 características e incrementan su error al aumentar el número de características. Pero sin duda, cada usuario se comporta de una manera diferente consiguiendo mejores o peores resultados. En el dominio de la frecuencia, hay usuarios con los mismos comportamientos, pero también otros, como el usuario 8 que son muy extraños, ya que parece que 1 característica es lo que mejor funciona, consiguiendo un EER de 0.25, mientras que con cualquier otro número de características consigue el peor EER posible de 1. La característica que está usando es la segunda frecuencia dominante, cualquier otra parece estar perjudicando a dicho usuario. Los mismos gráficos para las componentes X, Y y Z de ambos dominios se pueden encontrar en el anexo A.1, pero las conclusiones son las mismas, cada usuario se comporta de una manera diferente.

Realizando un análisis visual de las tablas que contienen las características por orden de impor-

	DT_X	DT_Y	DT_Z	modulo_DT
1	0.3867	0.2398	0.1993	0.2212
3	0.3234	0.169	0.1305	0.1012
5	0.3216	0.1674	0.1232	0.0751
7	0.2863	0.168	0.113	0.0719
9	0.2151	0.1601	**0.1009**	**0.0663**
11	0.177	0.1206	0.1058	0.0874
12	**0.1196**	**0.1173**	0.1046	0.1008

(a) Dominio del Tiempo

	DF_X	DF_Y	DF_Z	modulo_DF
1	0.1504	0.1816	0.1943	0.1799
3	0.165	0.1498	0.1552	0.1752
5	0.1377	0.1398	**0.1475**	0.1735
7	0.1068	**0.1377**	0.1552	0.1678
9	0.1038	0.1837	0.1655	0.1679
11	0.1008	0.1754	0.164	**0.1665**
13	**0.1006**	0.1877	0.1614	0.1669

(b) Dominio de la Frecuencia

Tabla 7.4: Selección de características considerando cada usuario, EER medio.

tancia, se puede decir que destacan las siguientes sobre cada posibilidad de dominio. Se consideran las primeras dos posiciones de todos los usuarios y se marcan en **negrita** aquellas que se repiten entre las distintas posibilidades del mismo dominio.

- **Dominio Del Tiempo:**
  - Componente X: **media**, **cuantil 25 y 75 %**, mínimo, máximo, mediana, autocorrelación, **periodo**.
  - Componente Y: **cuantil 75 y 25 %**, **media**, mediana, **periodo**, desviación estándar, autocorrelación, coeficiente de asimetría.
  - Componente Z: **cuantil 25 y 75 %**, **media**, mediana, **periodo**, desviación estándar.
  - Módulo: **media**, coeficiente de asimetría, **cuantil 25 y 75 %**, autocorrelación, **periodo**, desviación estándar, máximo.
- **Dominio De la Frecuencia:**
  - Componente X: **frecuencias dominantes 1 y 2**, **curtosis**, mediana, **desviación estándar**, **coeficiente de asimetría**.
  - Componente Y: **frecuencias dominantes 1 y 2**, **curtosis**, **desviación estándar**, AUC, media, **coeficiente de asimetría**, amplitud dominante 1 y 2, rango.
  - Componente Z: **frecuencias dominantes 1 y 2**, **curtosis**, **coeficiente de asimetría**, cuantil 25 %, amplitud dominante 1 y 2, **desviación estándar**, AUC, rango, media.
  - Módulo: **frecuencias dominantes 1 y 2**, amplitud dominante 1, AUC, media, **desviación estándar**, rango, **curtosis**, **coeficiente de asimetría**, mediana, cuantil 25 y 75 %.

En cualquier posibilidad del Dominio del Tiempo aparece la media, los cuantiles 25 y 75 % y el periodo, mientras que en el Dominio de la Frecuencia aparecen las frecuencias dominantes 1 y 2, la curtosis, la desviación estándar y el coeficiente de asimetría, pareciendo las 4 y 5 características más importantes en cada dominio.

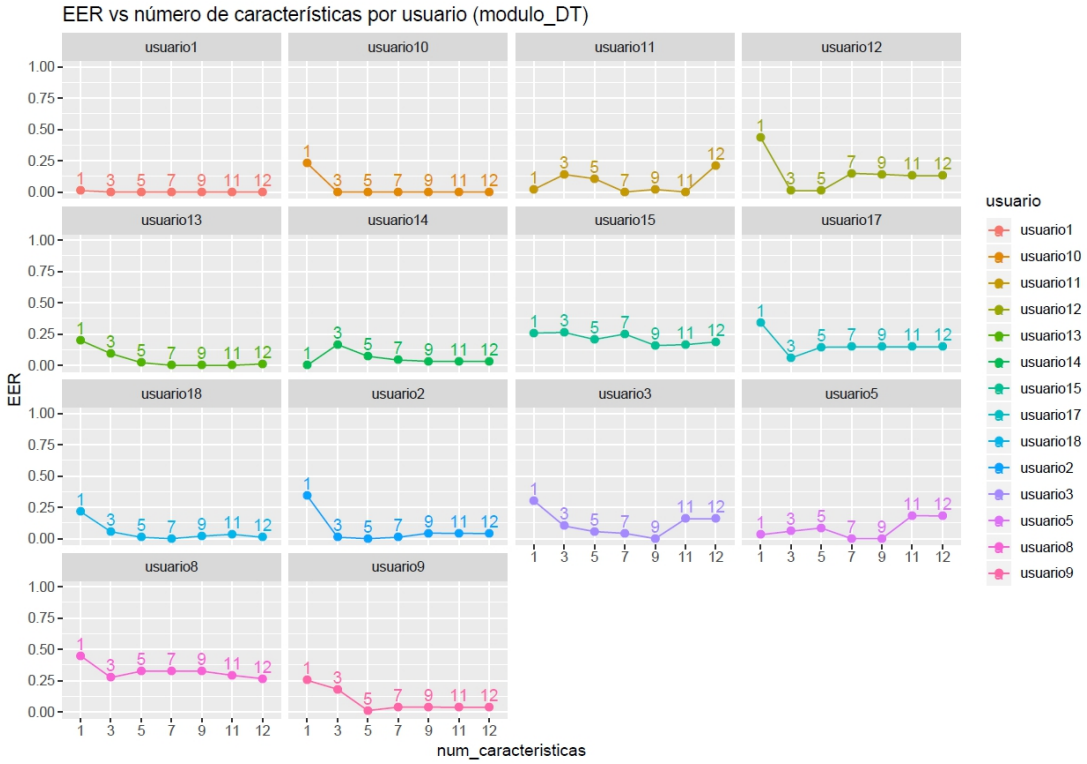


Figura 7.1: Resultados de la selección de características considerando cada usuario. Gráficos por usuarios. Módulo - Dominio del Tiempo.

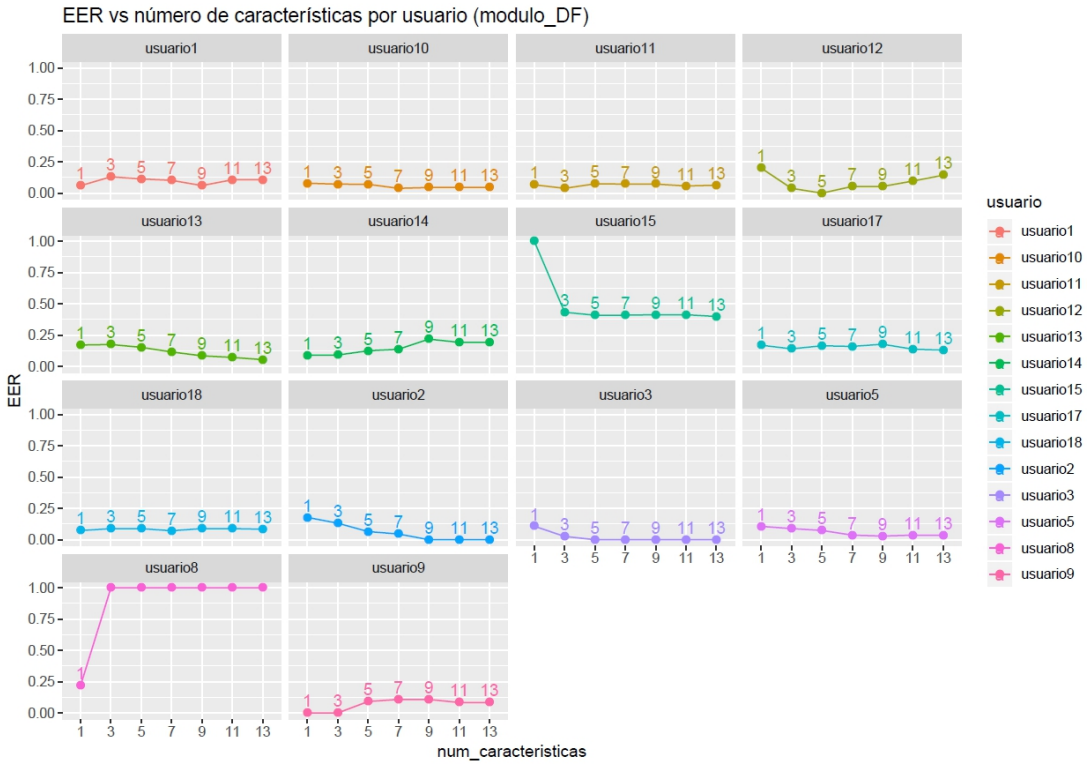


Figura 7.2: Resultados de la selección de características considerando cada usuario. Gráficos por usuarios. Módulo - Dominio de la Frecuencia

7.2.2. Considerando todos los usuarios

En este segundo caso, en que se trata a todos los usuarios juntos, se busca un orden de las características común para todos ellos con el criterio del valor umbral 0.5 del máximo peso. Todas las características que tengan un peso superior a dicho valor, en al menos un usuario, serán seleccionadas y ordenadas considerando su máximo peso. Como el número de características utilizadas en cada dominio era desconocido a priori, se han probado todos los posibles valores entre 1 y el número seleccionado. En la tabla 7.5 se puede ver el número de características seleccionado en cada dominio junto con el valor aproximado del máximo peso. En el Dominio del Tiempo los pesos son más altos en las componentes que en el módulo, a pesar de funcionar bien, aunque parece que el máximo peso es 0.25 y todos los atributos tienen valores similares. En el Dominio de la Frecuencia los pesos son más pequeños en las componentes que en el módulo, además parece que en el módulo existe una única característica con peso alto, ya que el quedarse con todos los atributos con peso superior a 0.2, selecciona sólo 2.

Componente	**Nºcaracterísticas**	**Peso máximo**
DT_X	8	0.6
DT_Y	6	0.6
DT_Z	10	0.3
módulo_DT	11	0.25
DF_X	10	0.2
DF_Y	13	0.25
DF_Z	11	0.25
módulo_DT	2	0.4

Tabla 7.5: Resumen de la combinación de sensores en el dispositivo MICRO.

Igual que antes, en las tablas referenciadas como 7.6, se muestra el EER medio de todos los usuarios utilizando sus k características comunes más importantes entre 1 y el máximo seleccionado, que no tiene por qué ser el mismo en todos los casos. En el Dominio del Tiempo, la componente X y el módulo consiguen el mejor resultado con el máximo de las características seleccionadas, siendo en la componente X la mejor opción, pero el módulo consigue un resultado muy similar utilizando 5 o incluso 6 características menos. La componente Y consigue el mejor resultado con 4 características menos que las seleccionadas, pero con una diferencia muy pequeña, por lo que el número seleccionado podría ser apropiado, y la componente Z consigue el mejor resultado con 3 características menos, pero muy similar al máximo número de características, que también parece apropiado.

En los gráficos de las figuras 7.3 y 7.4 se puede ver el comportamiento de cada usuario al variar el número de características, utilizando el módulo. En el Dominio del Tiempo se distinguen tres comportamientos: usuarios más normales que tienen mayor EER con 1 característica y van disminuyendo la tasa de equierror al aumentar el número de características consiguiendo su mejor

valor con el máximo o un número intermedio apropiado, como ocurre en los usuarios 1, 2, 6, 10, 12, 13, 14, 15 o 18. Después usuarios como el 11 en los que ocurre justo lo contrario y consiguen su mejor EER con 1 característica. Y por último usuarios estables que consiguen los mismos resultados con los distintos números de características probados, como ocurre con los usuarios 9 o 17. Por otro lado, en el Dominio de la Frecuencia, donde únicamente 2 características superan el valor umbral, se tienen usuarios como el 15 o el 17, que tienen el mismo comportamiento que aquellos que se han llamado "usuarios normales" en el Dominio del Tiempo. De nuevo, usuarios en los que ocurre justo lo contrario, que sería el caso del usuario 2, 8 o 11, pudiendo decir que el resto son usuarios estables, aunque con 2 características no se pueden extraer muchas conclusiones. Igual que antes, los mismos gráficos para las componentes X, Y y Z de ambos dominios se pueden encontrar en el anexo A.2.

De nuevo, se va a realizar un análisis visual observando todas las características seleccionadas con el criterio del valor umbral 0.5 en cada una de las posibilidades de dominio, marcando con **negrita** aquellas comunes.

- **Dominio Del Tiempo:**
  - Componente X: **media**, **cuantil 25 y 75 %**, **mediana**, **mínimo**, **máximo**, autocorrelación, periodo.
  - Componente Y: **cuantil 25 y 75 %**, **media**, **mediana**, **máximo**, **mínimo**.
  - Componente Z: **mediana**, **cuantil 25 y 75 %**, **media**, desviación estándar, autocorrelación, rango, periodo, **máximo**, **mínimo**.
  - Módulo: desviación estándar, **media**, **máximo**, **cuantil 25 y 75 %**, periodo, coeficiente de asimetría, autocorrelación, rango, **mediana**, **mínimo**.
- **Dominio De la Frecuencia:**
  - Componente X: **frecuencias dominantes 1** y 2, curtosis, cuantil 25 %, coeficiente de asimetría, **mediana**, AUC, amplitud dominante 1, rango, media.
  - Componente Y: AUC, media, **frecuencias dominantes 1** y 2, **mediana**, cuantil 25 y 75 %, amplitud dominante 1 y 2, desviación estándar, rango, coeficiente de asimetría, curtosis (todas)
  - Componente Z: **frecuencias dominantes 1** y 2, AUC, media, amplitud dominante 1, rango, cuantil 25 y 75 %, desviación estándar, **mediana**, curtosis.
  - Módulo: **frecuencia dominante 1**, **mediana**.

De manera que ahora, las características comunes en el Dominio del Tiempo son la media, los cuantiles 25 y 75 %, la mediana, el mínimo y el máximo y en el Dominio de la Frecuencia aquellas dos que se encuentran en el módulo que son la frecuencia dominante 1 y la mediana.

	DT_X	DT_Y	DT_Z	modulo_DT
1	0.3277	0.2287	0.2146	0.1813
2	0.3212	**0.2198**	0.1756	0.1429
3	0.2802	0.2228	0.1685	0.1388
4	0.2823	0.2311	0.1564	0.1287
5	0.3198	0.2248	0.146	0.1041
6	0.3109	0.2239	0.1247	0.0959
7	0.3032		**0.0985**	0.0929
8	**0.1147**		0.1144	0.0935
9			0.1124	0.0923
10			0.1138	0.0918
11				**0.0906**

(a) Dominio del Tiempo

	DF_X	DF_Y	DF_Z	modulo_DF
1	0.1619	0.1401	0.1352	0.1667
2	0.2176	0.1398	0.1336	**0.1241**
3	0.1978	0.1204	0.1336	
4	0.1965	0.1155	**0.1312**	
5	0.1956	0.1185	0.1364	
6	0.1916	0.1052	0.1376	
7	0.1262	0.1125	0.1376	
8	0.1125	0.1106	0.1371	
9	0.1066	**0.1038**	0.1362	
10	**0.1055**	0.1042	0.142	
11		0.111	0.1681	
12		0.1193		
13		0.1877		

(b) Dominio de la Frecuencia

Tabla 7.6: Selección de características considerando todos los usuarios juntos, EER medio.

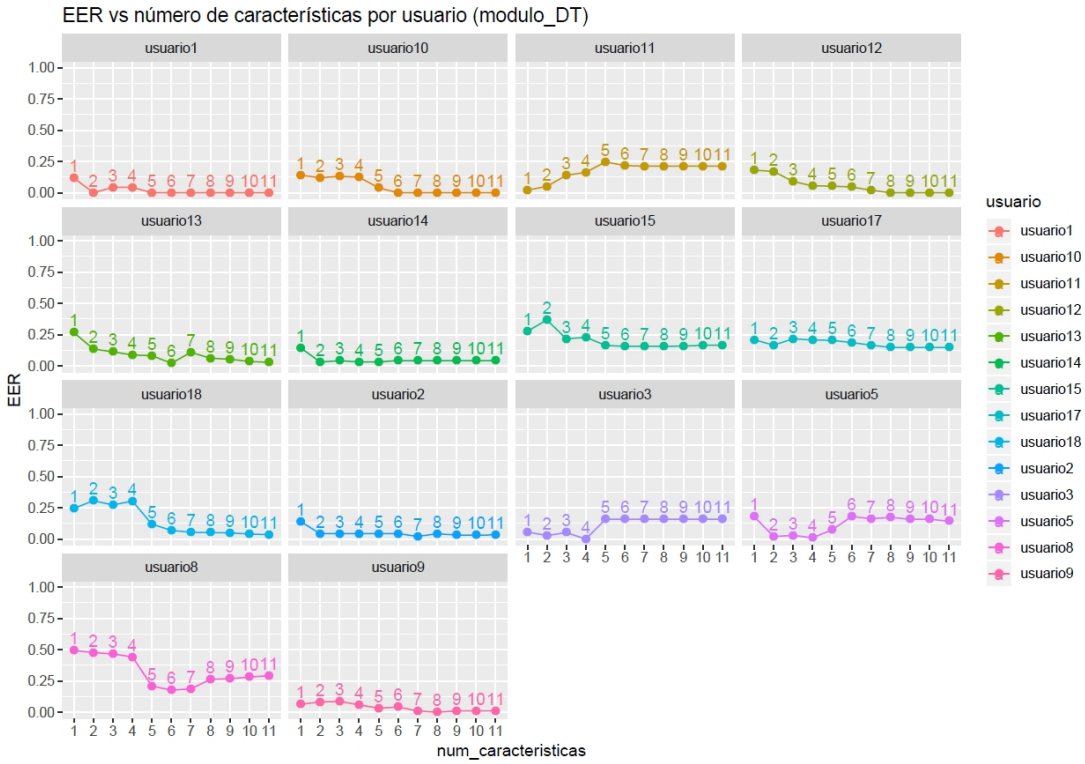


Figura 7.3: Resultados de la selección de características considerando todos los usuarios juntos. Gráficos por usuarios. Módulo - Dominio del Tiempo.

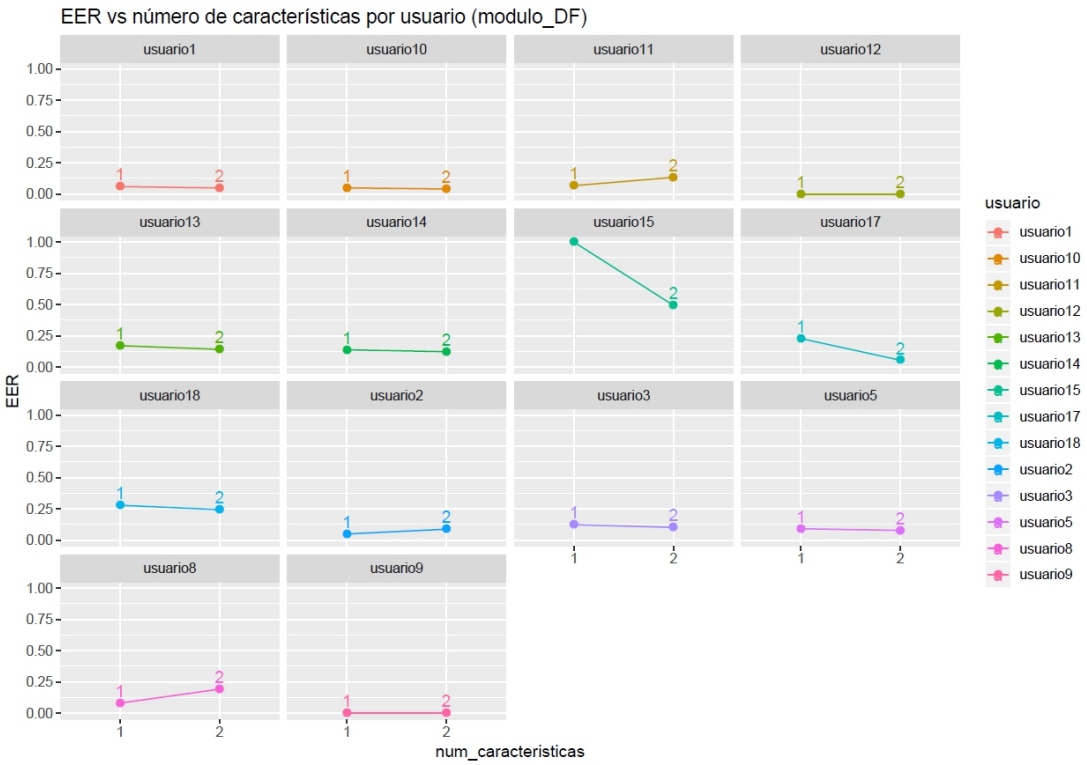


Figura 7.4: Resultados de la selección de características considerando todos los usuarios juntos. Gráficos por usuarios. Módulo - Dominio de la Frecuencia

7.3. Análisis de los resultados

A continuación, se va a realizar una comparación de los resultados obtenidos en el Dominio del Tiempo y en el Dominio de la Frecuencia. La tabla 7.7 contiene la tasa de equierror media de todos los usuarios sin hacer la selección de características, quedándose con todas, y la mejor opción conseguida realizando la selección de características en cada manera de actuar: considerando cada usuario por separado en la columna "Mejor selección 1" y todos los usuarios juntos en la columna "Mejor selección 2". Entre paréntesis se indica el número de características utilizadas. Respecto al Dominio del Tiempo, en las componentes X y Z funciona mejor realizar la selección de todos los usuarios juntos, pero con diferencias inferiores al 0.5 %, mientras que en la componente Y y el módulo funciona mejor realizar la selección de características considerando cada usuario, con diferencias más grandes del 10 % e inferior al 3 % respectivamente. En el Dominio de la Frecuencia, en la componente X hay pocas diferencias, funcionando ligeramente mejor la selección de características considerando cada usuario, pero en el resto de las posibilidades funciona mejor la selección de características de todos los usuarios juntos, con resultados mejores en aproximadamente el 3, 1.5 y 4 % para las componentes Y, Z y el módulo, respectivamente. Como conclusión, puede parecer que en el Dominio del Tiempo funciona mejor la selección de características considerando cada usuario por separado, ya que cuando consigue mejoras, estás son más grandes, mientras que en el Dominio de la Frecuencia funciona mejor la selección de características considerando todos los usuarios juntos. No obstante, estas conclusiones se obtienen considerando pocos usuarios y el EER medio

en la mejor opción conseguida, es decir, siendo capaces de predecir el número de características óptimo en cada posibilidad de dominio y tipo de selección. Pero si nos fijamos en los resultados de manera global, el mejor resultado del Dominio del Tiempo es 0.0663 con el módulo y 0.1006 con la componente X del Dominio de la Frecuencia, es decir, como hemos visto en todo momento, el Dominio del Tiempo parece funcionar mejor, aunque las diferencias entre dominios son inferiores al 4 %, siendo ambos resultados obtenidos considerando la selección de características de cada usuario por separado.

Por otro lado, en la tabla 7.8, se muestran las características que parecían más importantes en cada dominio por tener más peso y ser comunes en las 3 componentes y en el módulo, siendo en ambos tipos de selección, características muy similares.

Componente	**Sin selección**	**Mejor selección 1**	**Mejor selección 2**
DT_X	0.1196 (12)	0.1196 (12)	**0.1147 (8)**
DT_Y	**0.1173 (12)**	**0.1173 (12)**	0.2198 (2)
DT_Z	0.1046 (12)	0.1009 (9)	**0.0985 (7)**
módulo_DT	0.1008 (12)	**0.0663 (9)**	0.0906 (11)
DF_X	**0.1006 (13)**	**0.1006 (13)**	0.1055 (10)
DF_Y	0.1877 (13)	0.1377 (7)	**0.1038 (9)**
DF_Z	0.1614 (13)	0.1475 (5)	**0.1312 (4)**
módulo_DF	0.1669 (13)	0.1665 (11)	**0.1241 (2)**

Tabla 7.7: Resumen Resultados - Selección de Características (EER medio)

Tipo de selección (usuarios)	**Nombres de las características**
Dominio del Tiempo	
Independientes	media, cuantil 25 y 75 %, periodo.
Juntos	media, cuantil 25 y 75 %, mediana, mínimo, máximo.
Dominio de la Frecuencia	
Independientes	Frecuencias dominantes 1 y 2, kurtosis, desviación estándar, coeficiente de asimetría.
Juntos	Frecuencia dominante 1, mediana.

Tabla 7.8: Resumen Resultados - Selección de Características (nombres de las características)

Capítulo 8

Conclusiones y trabajo futuro

8.1. Conclusiones

Tras el trabajo expuesto se puede concluir que se han cumplido todos los objetivos inicialmente planteados.

Se ha realizado un análisis de la señal original, para comprobar que es factible usarla en biometría. Una de las razones en que se basa está afirmación es en el análisis de la autocorrelación, el cual demostraba la clara existencia de periodicidad en los datos.

Con la señal limpia, se han analizado distintas opciones de preprocesamiento, para ir optimizando y reduciendo el problema. Se han considerado las técnicas de normalización e interpolación, ya que se disponía de datos capturados con tiempo entre muestras variable, viéndose la necesidad de hacer la prueba de pasarlos a datos con tiempo entre muestras fijo.

Teniendo la señal preprocesada, se ha hecho un estudio prospectivo para valorar los parámetros que existían en el experimento y tratar de buscar sus mejores valores. Se ha considerado el tamaño de la ventana en que se dividían las muestras de los usuarios con los que se trabajaba, la posibilidad de aplicar un método de suavizado a los datos y eliminar ventanas con autocorrelación baja. Esto nos ha llevado a crear una segmentación automática basada en la autocorrelación con pérdidas no significativas en los resultados y mucho ahorro de tiempo, consiguiendo quedarnos con las zonas de las muestras donde sí hay periodicidad. También se ha probado una técnica de fusión a los scores, que produce mejoras, incluso en los usuarios malos y llega a resultados finales realmente buenos del 5.24 % en el dominio del tiempo y 11.68 % en el dominio de la frecuencia.

Una vez tomadas todas las decisiones se ha construido un sistema de reconocimiento final cuya prueba en diversos sensores y dispositivos se ha muestra en [6].

Para continuar con el estudio, se ha realizado una selección de características que se ha abordado considerando dos maneras diferentes, produciendo mejoras significativas y pudiéndose decir que es una posible alternativa a tratar.

Se puede concluir que los resultados obtenidos son prometedores y muestran que el uso de dispositivos ponibles puede ser una alternativa muy interesante, aunque con muchas cuestiones todavía abiertas. Para poder abordar este estudio es imprescindible tener una base de datos más completa.

Respecto a las conclusiones personales, este trabajo me ha permitido poner en práctica los conocimientos adquiridos en diversas asignaturas a lo largo de mis estudios, aprender nuevos conceptos, el mundo la biometría, el conocimiento y funcionamiento de problemas de clasificación trabajando con muestras de la misma clase o *one-class*, ampliar y reforzar mis habilidades con la herramienta de programación R, así como vivir la experiencia de un problema actual y novedoso que no tiene una solución única, sino una infinidad de posibilidades con las que poder seguir trabajando.

Por otro lado, este trabajo me ha enseñado a tomar decisiones, al ser un problema amplio, donde se pueden hacer muchas cosas, pero hay que decidir cuáles son las mejores, asumiendo la posibilidad de que nos podemos equivocar. Me ha permitido trabajar en un grupo de investigación y entender mejor cómo funciona el proceso de la investigación científica desde dentro. Habiendo conseguido, con todo ello, una experiencia muy enriquecedora, de manera personal y académica.

8.2. Líneas de trabajo futuro

En este proyecto se ha conseguido sentar unas bases sólidas sobre las que poder empezar a trabajar en esta biometría, pero aún existe bastante trabajo por delante.

Los resultados obtenidos mediante nuestro sistema de reconocimiento final construido han sido buenos y prometedores, pero se ha utilizado únicamente el clasificador de k-vecinos más próximos. Al decidir optar, como comienzo, por un sistema simple. No obstante, algo necesario a hacer, además de ser una posible manera de mejorar los resultados es a través de la aplicación de más clasificadores. Así como la prueba de más métodos de selección de características.

Idealmente, como trabajo futuro, se tendría que buscar una manera de optimizar el sistema a las características de cada individuo, ya que como se ha visto a lo largo de esta memoria, si fuésemos capaces de predecir la componente tridimensional mejor de cada individuo, los resultados mejorarían mucho. Incluso, simplemente siendo capaces de predecir el tipo de sensor que va a funcionar mejor, aunque por esta vía se deja pendiente la prueba de una combinación de ambos sensores.

Pero sin duda, a la vista de los resultados obtenidos, el presente TFG planta una semilla muy interesante para futuros trabajos al estarse verificando que la señal recogida en los dispositivos ponibles comerciales puede utilizarse para el reconocimiento biométrico de las personas.

Acrónimos y abreviaturas

ACC Acelerómetro

AI Inteligencia Artificial

AI HLEG Grupo de Expertos de Alto Nivel sobre Inteligencia Artificial

DCT Transformada del Coseno Discreta

DF Dominio de la Frecuencia

DT Dominio del Tiempo

DTree Decision Trees

DTW Dynamic Time Warping

ECTS European Credit Transfer and Accumulation System

EER tasa de equierror

EU Unión Europea

FNR tasa de falsos negativos

FPR tasa de falsos positivos

GMM Gaussian Mixture Model

GPS Sistema de Posicionamiento Global

GYR Giroscopio

H precisión

HMM Hidden Markov Model

I identificación

ICA Independent Component Analysis

ID Identificador

IMEI International Mobile Equipment Identity

KNN K-Nearest Neighbors

LDA Linear Discriminant Analysis

MICRO Microsoft

ML Machine Learning

MOTO Motorola

MonoMono Monosesión-Monomuestra

MultiMono Multisesión-Monomuestra

MultiMulti Multisesión-Multimuestra

PCA Principal Component Analysis

PPG sensores fotopletismográficos

ROC Receiver Operating Characteristic

SVM Support Vector Machines

TFG Trabajo Fin de Grado

UVa Universidad de Valladolid

V verificación

Índice alfabético

Anexos

Apéndice A

Selección de características

A.1. Considerando cada usuario

Los gráficos de la selección de características considerando cada usuario por separado son: A.1, A.2, A.3, A.4, A.5, A.6.

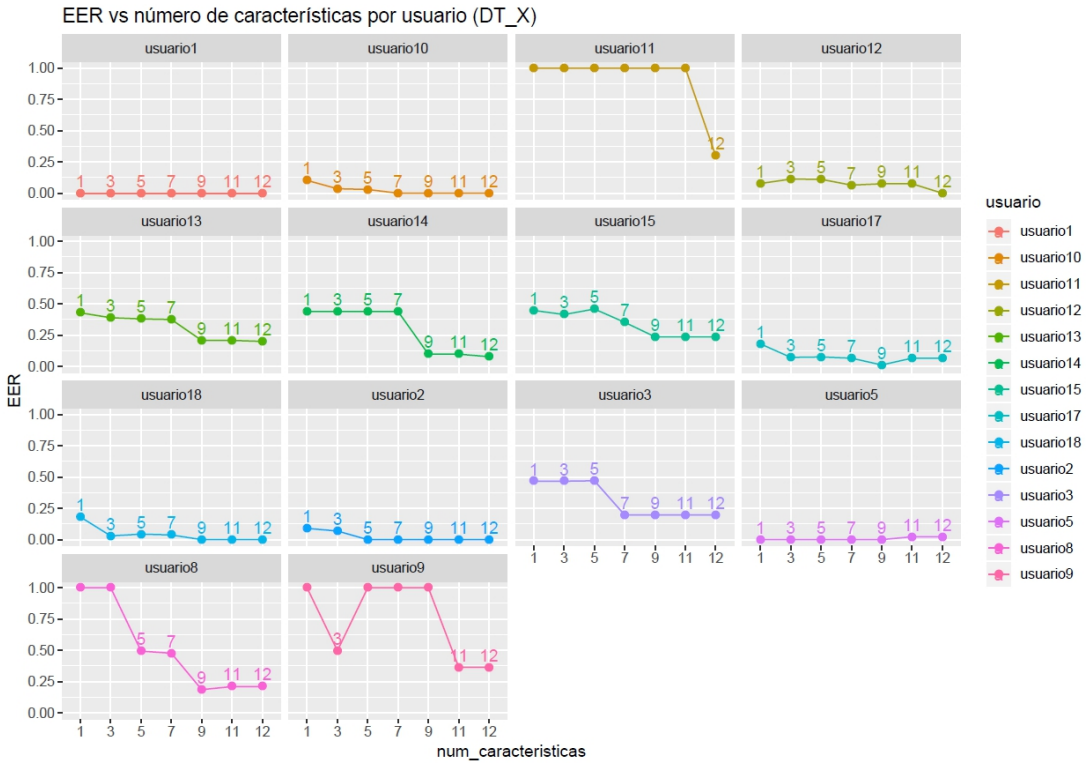


Figura A.1: Selección de características considerando cada usuario. Gráficos por usuarios. ComponenteX - Dominio del Tiempo.

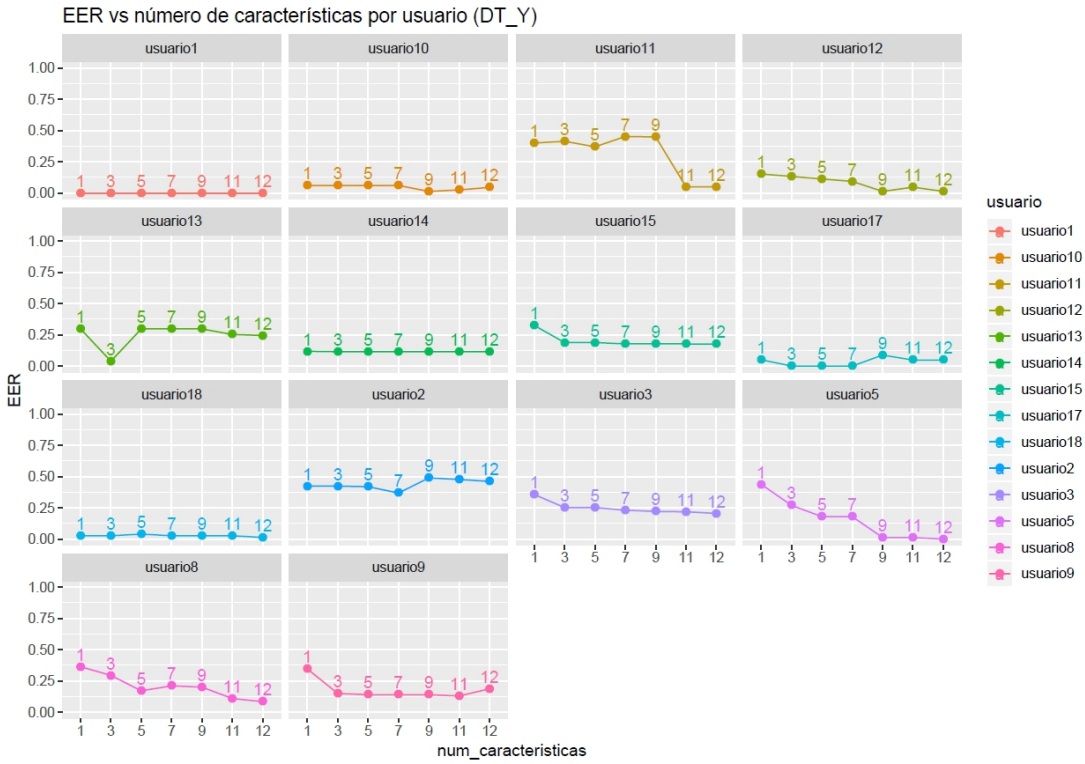


Figura A.2: Selección de características considerando cada usuario. Gráficos por usuarios. ComponenteY - Dominio del Tiempo.

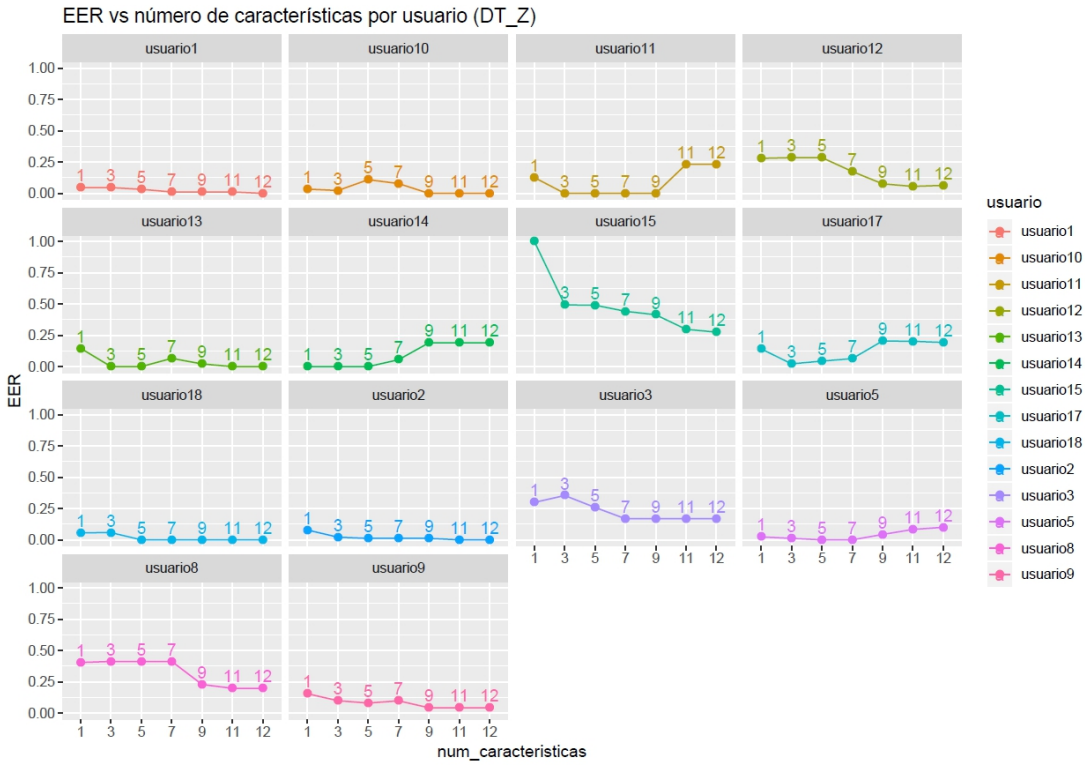


Figura A.3: Selección de características considerando cada usuario. Gráficos por usuarios. ComponenteZ - Dominio del Tiempo.

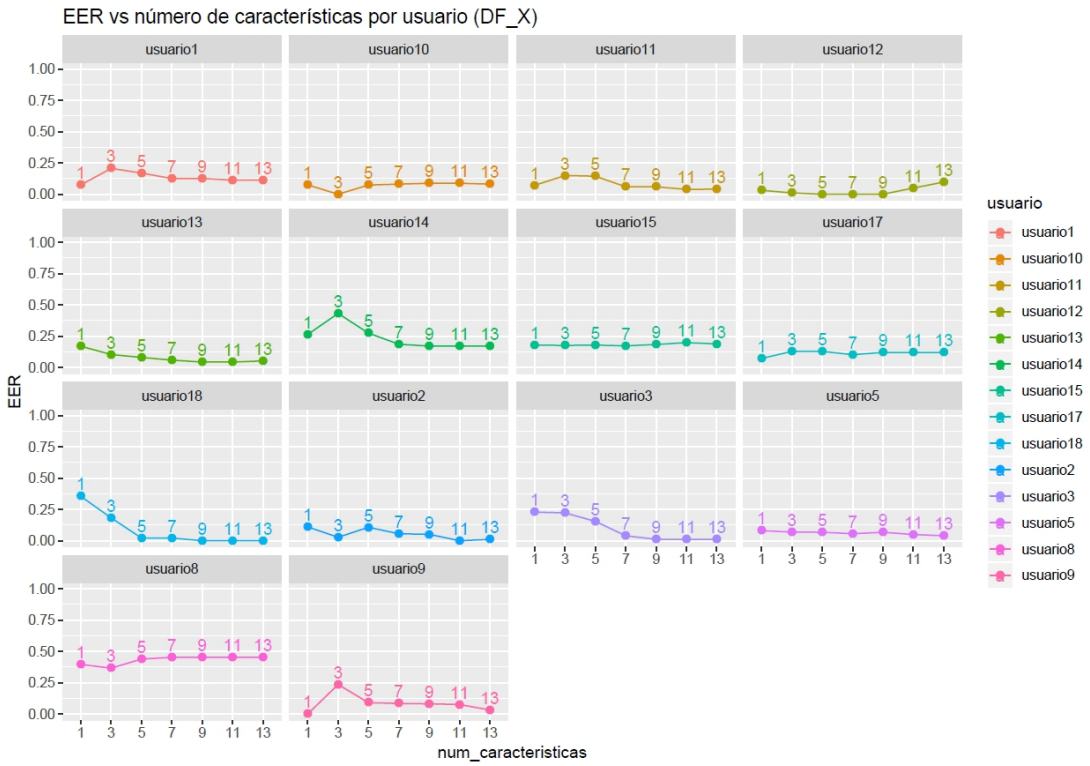


Figura A.4: Selección de características considerando cada usuario. Gráficos por usuarios. ComponenteX - Dominio de la Frecuencia

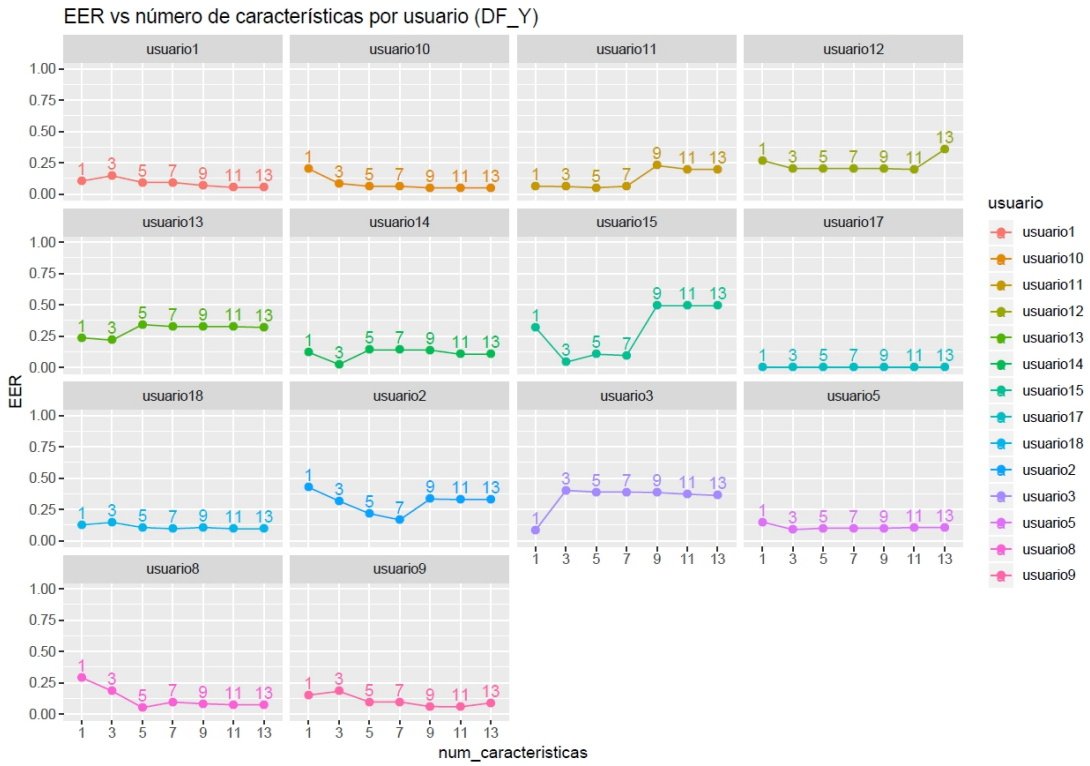


Figura A.5: Selección de características considerando cada usuario. Gráficos por usuarios. ComponenteY - Dominio de la Frecuencia

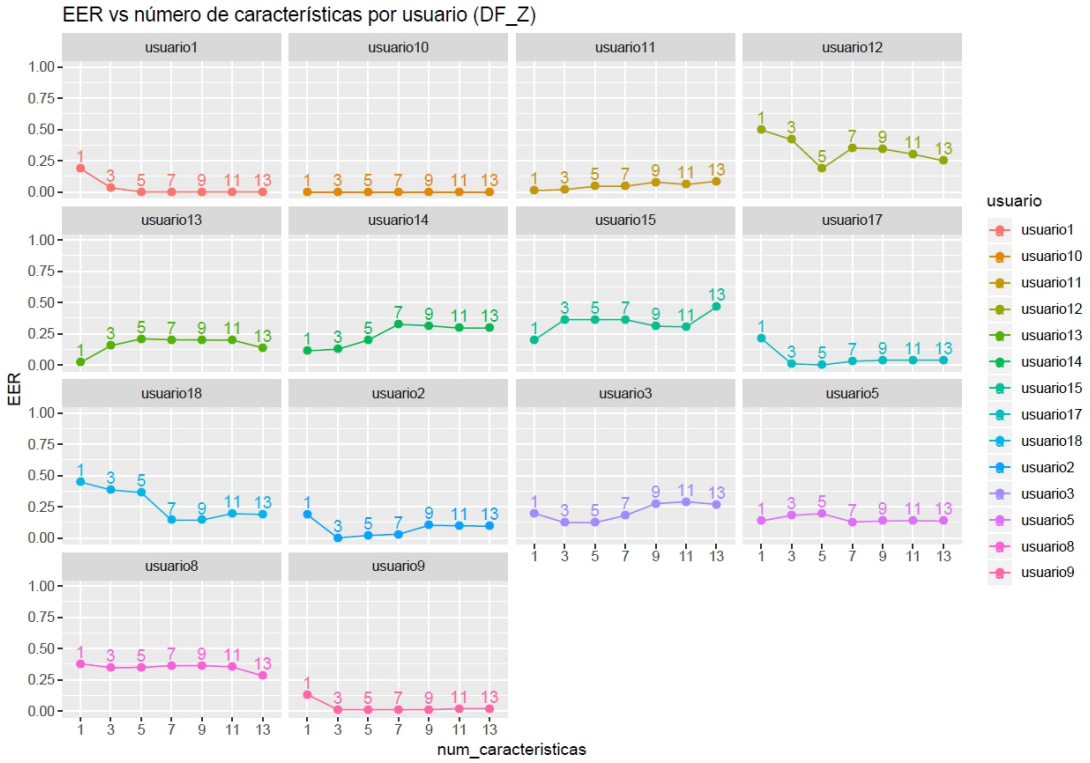


Figura A.6: Selección de características considerando cada usuario. Gráficos por usuarios. ComponenteZ - Dominio de la Frecuencia

A.2. Considerando todos los usuarios juntos

Los gráficos de la selección de características considerando todos los usuarios juntos son: A.7, A.8, A.9, A.10, A.11, A.12.

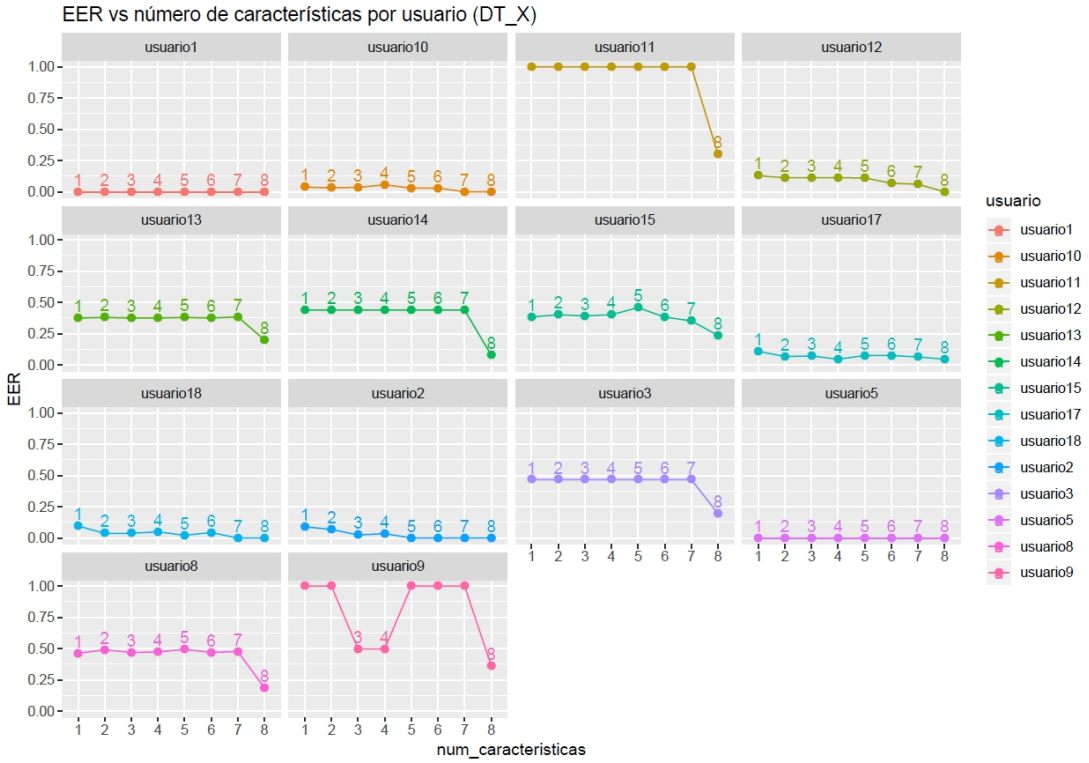


Figura A.7: Selección de características considerando todos los usuarios juntos. Gráficos por usuarios. ComponenteX - Dominio del Tiempo.

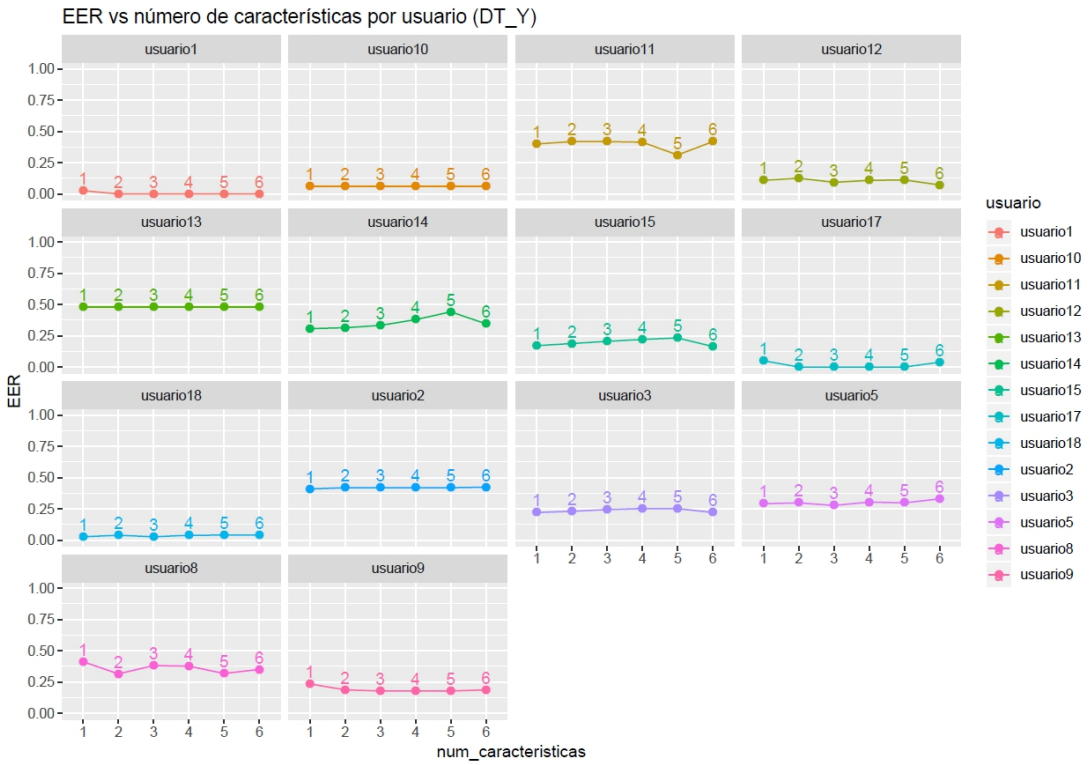


Figura A.8: Selección de características considerando todos los usuarios juntos. Gráficos por usuarios. ComponenteY - Dominio del Tiempo.

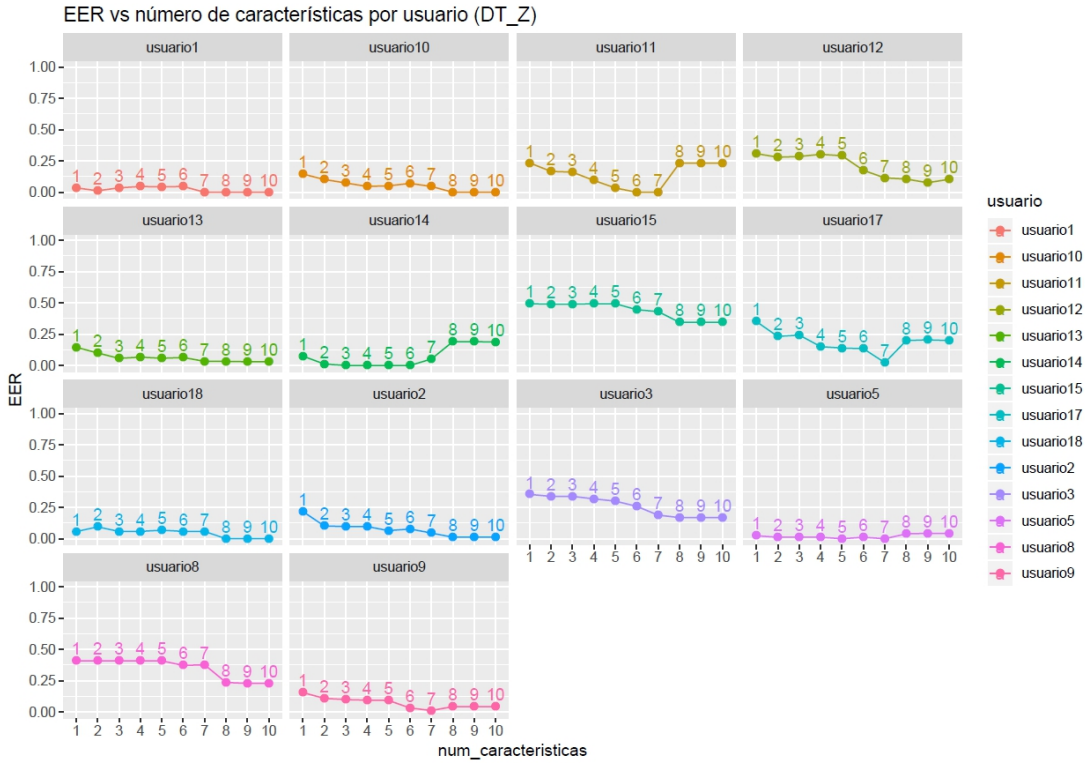


Figura A.9: Selección de características considerando todos los usuarios juntos. Gráficos por usuarios. ComponenteZ - Dominio del Tiempo.

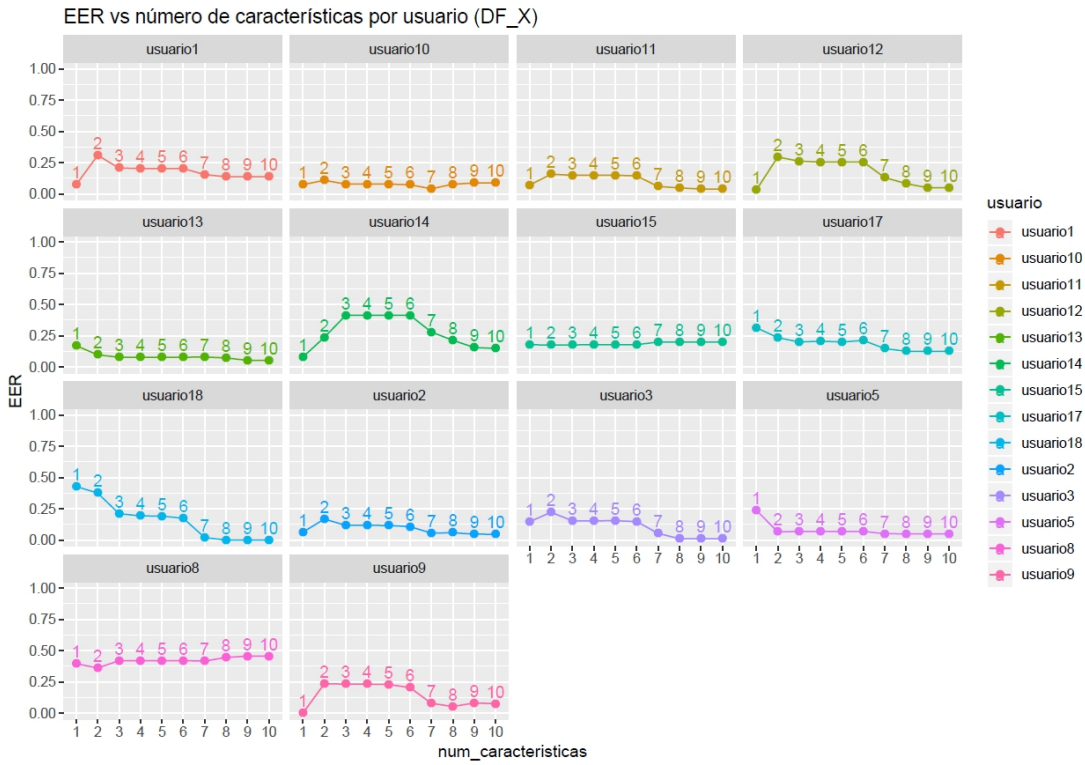


Figura A.10: Selección de características considerando todos los usuarios juntos. Gráficos por usuarios. ComponenteX - Dominio de la Frecuencia

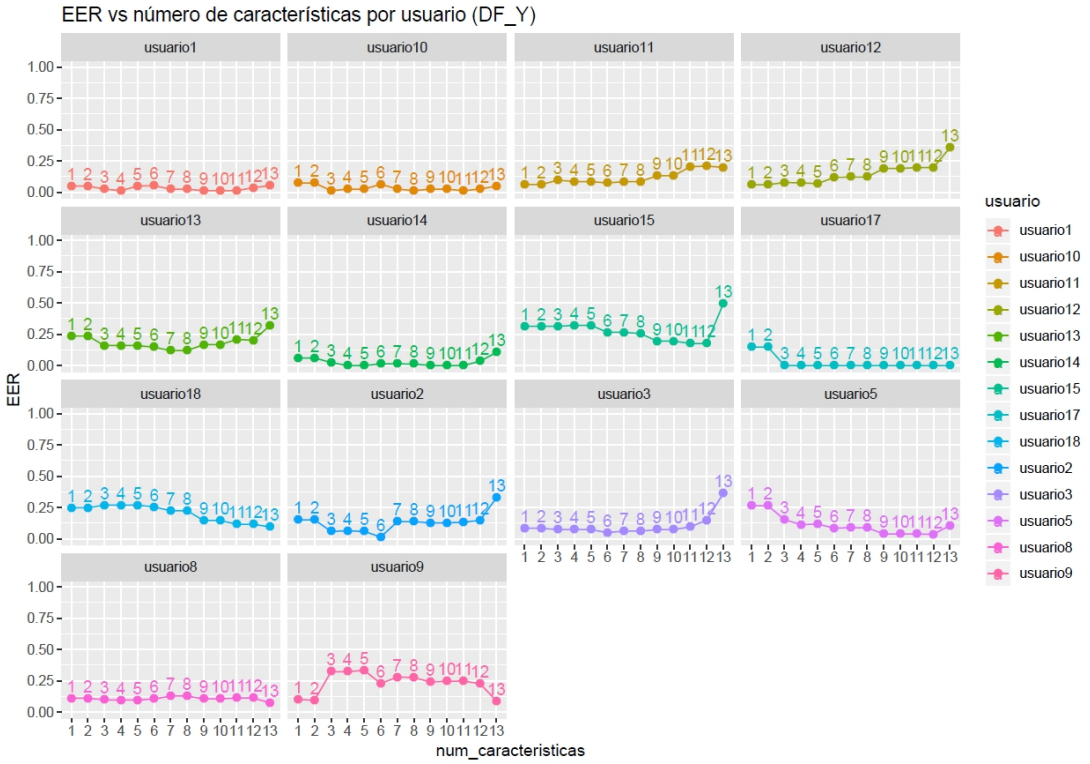


Figura A.11: Selección de características considerando todos los usuarios juntos. Gráficos por usuarios. ComponenteY - Dominio de la Frecuencia

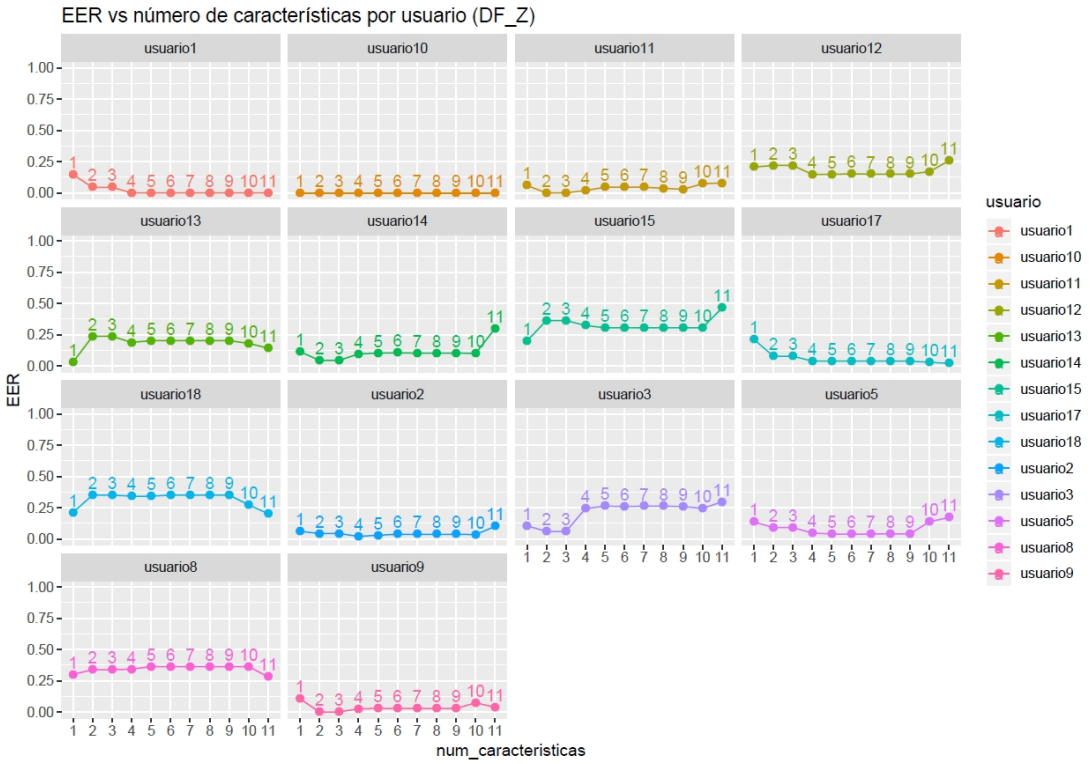


Figura A.12: Selección de características considerando todos los usuarios juntos. Gráficos por usuarios. ComponenteZ - Dominio de la Frecuencia

Apéndice B

Contenido del CD

- /
  - titulo_memoria.pdf ..... Memoria del Trabajo de Fin de Grado de Estadística.
  - Programas
    - construyeTablasCaracteristicas.R ..... Script utilizado para construir las tablas de características.
    - SeleccionCaracteristicas_RELIEF.R ..... Script para obtener los ordenes de características de todos los usuarios junto con el score de cada una de ellas.
    - Tabla_Xcaract_UsuariosIndep.R ..... Script que obtiene la métrica para cada usuario y posibilidad de número de características, considerando cada usuario.
    - Tabla_Xcaract_UsuariosJuntos.R ..... Script que obtiene la métrica para cada usuario y posibilidad de número de características, considerando todos los usuarios juntos.
  - Imágenes ..... Carpeta con todas las imágenes utilizadas en este documento. Por cada capítulo, se tiene una carpeta con su nombre y las imágenes correspondientes.

Bibliografía

[1] Darina Lynkova, 22 de abril de 2019, "The Surprising Reality of How Many Emails Are Sent Per Day", https://techjury.net/stats-about/how-many-emails-are-sent-per-day/

[2] Alimarket Alimentación, 28 de enero de 2019, "Minsait lanza una solución para la compra online de frescos", https://www.alimarket.es/alimentacion/noticia/292022/minsait-lanza-una-solucion-para-la-compra-online-de-frescos

[3] Gartner, 11 de abril de 2019, "Tendencias de Inteligencia Artificial que están arrasando en 2019", https://decidesoluciones.es/tendencias-de-inteligencia-artificial-2019/

[4] J. M. Galán, TFG de la Escuela de Ingeniería Informática de la Universidad de Valladolid curso 2015/2016, "Wearables: Análisis de dispositivos y recogida de datos en Android para estudios biométricos"

[5] Daniel González Alonso, TFG de la Escuela de Ingeniería Informática de la Universidad de Valladolid curso 2016/2017, "Estudio preliminar del uso de Wearables en reconocimiento biométrico de personas"

[6] Irene Salvador Ortega, 26 de junio de 2019, "Investigación y Desarrollo en Técnicas de Reconocimiento Biométrico mediante Dispositivos Ponibles (wearables)"

[7] Jorge Blasco, septiembre de 2016, "A Survey of Wearable Biometric Recognition Systems"

[8] Anil K. Jain, 2004, "Multibiometric systems. Communications", ACM

[9] Roman V. Yampolskiy, 2010, "Taxonomy of behavioral biometrics", Behavioral Biometrics for Human Identification

[10] Salil Prabhakar, 2003, "Biometric recognition: Security and privacy concerns", IEEE Security & Privacy

[11] European Commission's High-Level Expert Group on Artificial Intelligence, 18 de diciembre de 2018, "Draft Ethics guidelines for trustworthy AI", https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai

[12] Mohammad Omar Derawi, 2010, "Accelerometer-Based Gait Analysis, A survey"

[13] Joe Belfiore, 2015, "Making Windows 10 More Personal and More Secure with Windows Helloy", https://blogs.windows.com/windowsexperience/2015/03/17/making-windows-10-more-personal-and-more-secure-with-windows-hello/

[14] Toshiyo Tamura, 2014, "Wearable photoplethysmographic sensors—Past and present", Electronics.

[15] Davide Maltoni, 2009, "Handbook of Fingerprint Recognition", Springer.

[16] Lucas Introna and Helen Nissenbaum, 2010, "Facial recognition technology a survey of policy and implementation issues. Technical Report", The Department of Organisation, Work and Technology, Lancaster University.

[17] Pallavi Meharia and Dharma P. Agrawal, 2015, "Unobtrusive gait verification for mobile phones", Journal of Information Privacy & Security

[18] Hong Lu, 2014, "Unobtrusive gait verification for mobile phones", ACM

[19] Mohammad Derawi, 2015, "Wireless chest-based ECG biometrics", Springer

[20] Hindra Kurniawan, 2013, "Stress detection from speech and galvanic skin response signals", IEEE 26th International Symposium on Computer-Based Medical Systems

[21] Adam S. Venable, 2013, "Gender differences in skin and core body temperature during exercise in a hot, humid environment", Internal Journal of Exercise Science: Conference Proceedings, Vol. 2. 9.

[22] Kenneth Revett, 2008, "Behavioral Biometrics: A Remote Access Approach"

[23] Hoang Minh Thang, 2012, "Gait Identification Using Accelerometer on Mobile Phone"

[24] Oresti Banos, 9 de abril de 2014, "Window Size Impact in Human Activity Recognition"

[25] Guannan Wu, 10 de junio de 2018, "A Continuous Identity Authentication Scheme Based on Physiological and Behavioral Characteristics"

[26] Akram Bayat, 2017, "Classifying Human Walking Patterns using Accelerometer Data from Smartphone"

[27] Liu Yiyan, noviembre de 2016, "An Hidden Markov Model based Complex Walking Pattern Recognition Algorithm"

[28] Weitao Xu, 2017, "Gait-Watch: A Context-aware Authentication System for Smart Watch Based on Gait Recognition"

[29] Samer K Al Kork, 2017, "Biometric Database for Human Gait Recognition using Wearable Sensors and a Smartphone"

[30] Sherif Said, 26 de julio de 2018, "Experimental Investigation of Human Gait Recognition Database using Wearable Sensors"

[31] Davrondzhon Gafurov and Einar Snekkenes, 26 de abril de 2009, "Gait Recognition Using Wearable Motion Recording Sensors"

[32] Thomas Bernecker, (s.f), "Activity Recognition on 3D Accelerometer Data (Technical Report)"

[33] Fan Yang, 2017, "Real-Time Human Activity Classification by Accelerometer Embedded Wearable Devices"

[34] Liu Rong, 2017, "A Wearable Acceleration Sensor System for Gait Recognition"

[35] Davrondzhon Gafurov, noviembre de 2006, "Biometric Gait Authentication Using Accelerometer Sensor"

[36] Bing Sun, 2014, "Gait Characteristic Analysis and Identification Based on the iPhone's Accelerometer and Gyrometer"

[37] Miikka Ermes, 2006, "Detection of Daily Activities and Sports With Wearable Sensors in Controlled and Uncontrolled Conditions"

[38] Chiung Ching Ho, 2010, "An Unobtrusive Android Person Verification Using Accelerometer Based Gait I"

[39] Martin Reese Hestbek, 13 de abril de 2012, "Biometric Gait Recognition For Mobile Devices Using Wavelet Transform And Support Vector Machines", 19th International Conference on Systems, Signals and Image Processing

[40] Michael Fitzgerald Nowlan, 2009, "Human Identification via Gait Recognition Using Accelerometer Gyro Forces"

[41] Heikki Ailisto, 2005, "Identifying people from gait pattern with accelerometers"

[42] Pierluigi Casale, 2012, "Personalization and user verification in wearable systems using biometric walking patterns", Personal and Ubiquitous Computing.

[43] Ghina Dandachi, 2013, "A novel identification/verification model using smartphone's sensors and user behavior", 2nd International Conference on Advances in Biomedical Engineering

[44] Chiung Ching Ho, 2012, "An unobtrusive Android person verification using accelerometer based gait II", 10th International Conference on Advances in Mobile Computing & Multimedia

[45] RashaWahid, 2012, "A Gaussian mixture models approach to human heart signal verification using different feature extraction algorithms", Computer Applications for Bio-technology, Multimedia, and Ubiquitous City

[46] Zhidong Zhao and Qinqin Shen, 2011, "A human identification system based on heart sounds and Gaussian mixture models", 4th International Conference on Biomedical Engineering and Informatics

[47] Zoubin Ghahramani, 2001, "An introduction to hidden Markov models and Bayesian networks", International Journal of Pattern Recognition and Artificial Intelligence

[48] J. Ross Quinlan, 2014, "C4.5: Programs for Machine Learning", Elsevier

[49] Daisuke Sugimori, 2011, "A study about identification of pedestrian by using 3-axis accelerometer".

[50] Claudia Nickel, 2012, "Authentication of smartphone users based on the way they walk using k-NN algorithm", 8th International Conference on Intelligent Information Hiding and Multimedia Signal Processing

[51] Claudia Nickel, 2011, "Using hidden Markov models for accelerometer-based biometric gait recognition", IEEE 7th International Colloquium on Signal Processing and Its Applications

[52] Alexander J. Casson, 6 de diciembre de 2016, "Gyroscope vs. accelerometer measurements of motion from wrist PPG during physical exercise"

[53] Wikipedia, última edición el 10 de abril de 2019, "Autocorrelation", https://en.wikipedia.org/wiki/Autocorrelation

[54] Wikipedia, última edición el 20 de febrero de 2019, "Curtosis", https://es.wikipedia.org/wiki/Curtosis

[55] Wikipedia, última edición el 24 de mayo de 2018, "Asimetría estadística", https://es.wikipedia.org/wiki/Asimetr %C3 %ADa_estad %C3 %ADstica

[56] João Neto, marzo de 2013, "Fourier Transform: A R Tutorial", http://www.di.fc.ul.pt/ jpn/-r/fourier/fourier.html

[57] (s.a), extraído el 20 de febrero de 2019, "dygraphs for R", https://rstudio.github.io/dygraphs/ (https://github.com/rstudio/dygraphs/issues/29)

[58] Wikipedia, última edición el 5 de junio de 2019, "Frecuencia", https://es.wikipedia.org/wiki/Frecuencia

[59] Wikipedia, última edición el 28 de agosto de 2018, "Teorema de muestreo de Nyquist-Shannon", https://es.wikipedia.org/wiki/Teorema_de_muestreo_de_Nyquist-Shannon

[60] Santiago Morante Cendrero, 1 de noviembre de 2018, "Precauciones a la hora de normalizar datos en Data Science", https://data-speaks.luca-d3.com/2018/11/precauciones-la-hora-de-normalizar.html

[61] Wikipedia, última edición el 18 de noviembre de 2018, "Media móvil", https://es.wikipedia.org/wiki/Media_m %C3 %B3vil

[62] Joaquín Amat Rodrigo, Julio 2017, "Test de Wilcoxon-Mann-Whitney como alternativa al t-test", https://rpubs.com/Joaquin_AR/218456

[63] Carlos J. Alonso González, diapositivas para el curso 2018/2019, asignatura *Minería de Datos*, Universidad de Valladolid.