DETECCIÓN DE PÉRDIDAS DE ESPESOR EN LAS PAREDES DE TUBERÍAS DE TRANSPORTE DE HIDROCARBUROS UTILIZANDO TÉCNICAS DE PROCESAMIENTO DE SEÑALES Y MINERÍA DE DATOS

Presentado por:
ALDAIR BARAJAS ALDANA

Asesorado por:
Ph.D. CARLOS JULIO ARIZMENDI PEREIRA

UNIVERSIDAD AUTÓNOMA DE BUCARAMANGA
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
AUTOMATIZACIÓN Y CONTROL
BUCARAMANGA
2015

Tabla de contenido

Lista de Figuras

1. OBJETIVO GENERAL

- Detectar e identificar segmentos con pérdida de espesor en la pared de tuberías de transporte de hidrocarburos y gas por medio de técnicas de procesamiento de señales (Transformada *Wavelet (TW)*) e inteligencia artificial (Máquinas de Soporte Vectorial (MSV) con diferentes *Kernels*, Análisis Discriminante Lineal (ADL)), utilizando señales de flujo magnético.

2. OBJETIVOS ESPECÍFICOS

- Implementar algoritmos estadísticos para la caracterización de los datos provenientes de tuberías para la detección de disminución en el grosor de las paredes.

- Desarrollar algoritmos para realizar la selección y/o extracción de características, para la obtención de las variables más relevantes del sistema.

- Aplicar técnicas de procesamiento de datos (TW) e inteligencia artificial (MSV con diferentes *Kernels* y ADL) para identificar segmentos con pérdidas del espesor en la tubería.

- Reducir la dimensionalidad de los datos a procesar para aumentar la eficiencia y disminuir el tiempo de ejecución de las técnicas aplicadas.

- Identificar patrones de defectos en tuberías de transporte de hidrocarburos y gas.

3. INTRODUCCIÓN

Las empresas de transporte de hidrocarburos, cuentan con una longitud de tubería considerable para transportar el material hasta su destino, debido a la gran distancia que se debe recorrer, es probable que el material de la tubería sufra un desgaste considerable, lo cual podría provocar una pérdida del material transportado. Realizar una inspección manual de la integridad de la tubería no es posible en algunos casos, debido a la ubicación geográfica de las mismas.

Actualmente la Corporación de Investigación en Corrosión (CIC) está desarrollando diferentes proyectos tecnológicos de inspección y monitoreo de fenómenos de corrosión para solucionar este problema, como es el caso del proyecto ITION, el cual es un dispositivo con diferentes tipos de sensores, que viajará por la tubería con el fin de inspeccionar anomalías en ésta.

A su vez éste dispositivos generará un elevado volumen de información y el proceso de análisis de datos para la generación de información valiosa es una tarea particular para cada técnica en cada equipo o tecnología desarrollada.

Debido a esto es necesario utilizar técnicas de análisis e interpretación de señales (TW, MSV con diferentes *Kernels* y ADL) para clasificar e identificar patrones que determinen algún tipo de deterioro (pérdida del grosor de las paredes) en las tuberías de transporte de hidrocarburos y gas.

DISEÑO METODOLÓGICO

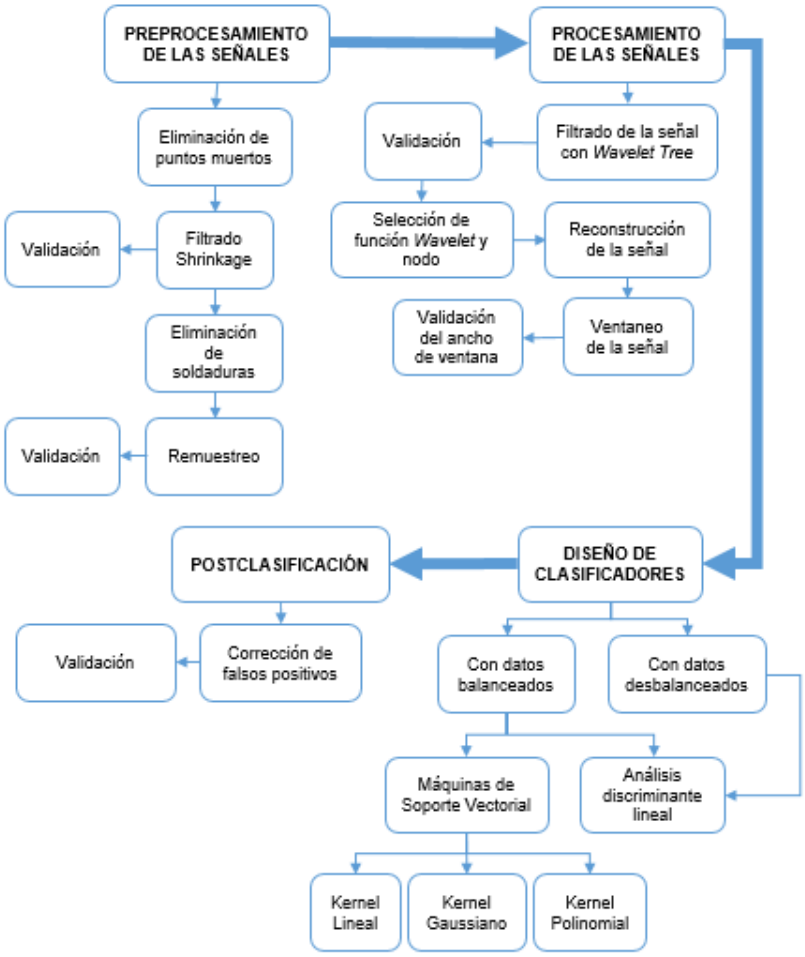

Figura 1. Diseño metodológico del proyecto.

4. FUGA DE FLUJO MAGNÉTICO

La Fuga de Flujo Magnético (FFM) es un método magnético de ensayos no destructivos que se utiliza para detectar la corrosión y picaduras en las estructuras de acero, más comúnmente tuberías y tanques de almacenamiento.

La herramienta de pérdida de flujo magnético utiliza imanes permanentes que inducen un campo magnético a niveles de saturación dentro de la pared de la tubería. Los sensores de efecto *Hall* detectan los cambios en el campo magnético provocados por la pérdida metálica, la corrosión, y otras anomalías, como se observa en la Figura 2. Estos cambios (comúnmente denominados pérdidas de flujo) y su relación con la pérdida volumétrica de la pared de la tubería ayudan a determinar la condición de ésta última.

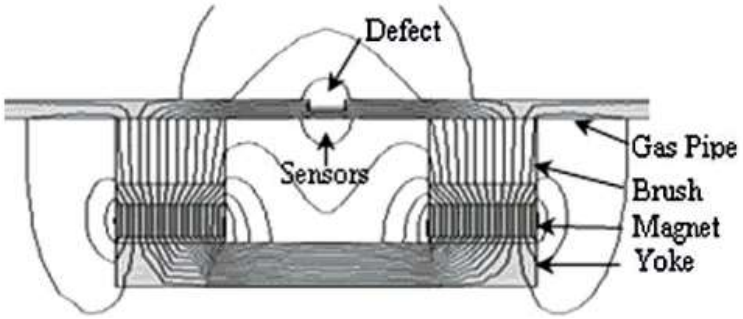

Figura 2. Inducción de campo magnético a la pared de la tubería. [1]

4.1. Señales de FFM

Cuando un campo magnético en una tubería encuentra una pérdida de material, el flujo magnético se desvía o se fuga. Los sensores (de efecto *Hall* en este caso) miden parte de la fuga de campo. En algunos casos, la fuga de campo alrededor de un defecto puede asemejarse a la forma del mismo, pero por lo general no tienen la misma forma. Por lo tanto, la forma de la fuga de campo no es necesariamente un buen indicador de la forma del defecto. Además, la ubicación del defecto, por

ejemplo en la pared interior del tubo contra la pared exterior del tubo, afecta la forma de la fuga [2].

La cantidad de fuga de campo magnético depende de la geometría del defecto, las principales variables que afectan la fuga de campo, son las que definen el volumen de la pérdida de material [3]. Las relaciones de la profundidad, longitud y ancho del defecto con la cantidad de fuga de campo magnético se pueden observar en las Figura 3, Figura 4 y Figura 5.

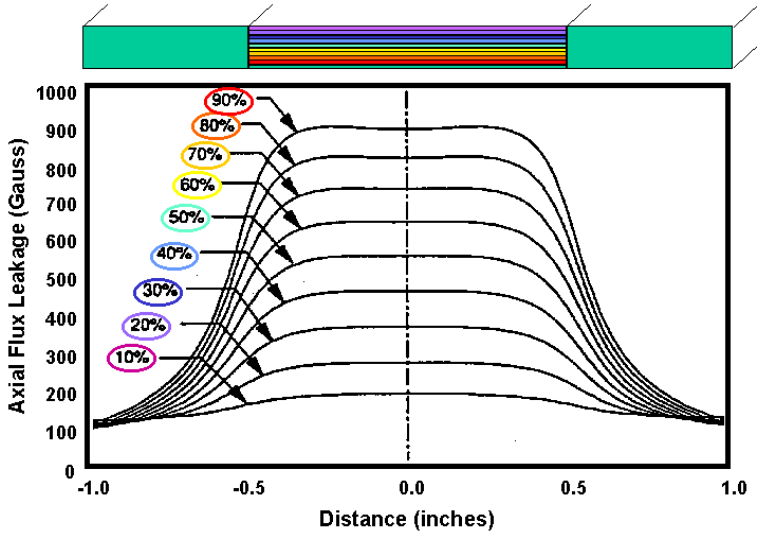

Figura 3. Relación Profundidad del defecto-FFM

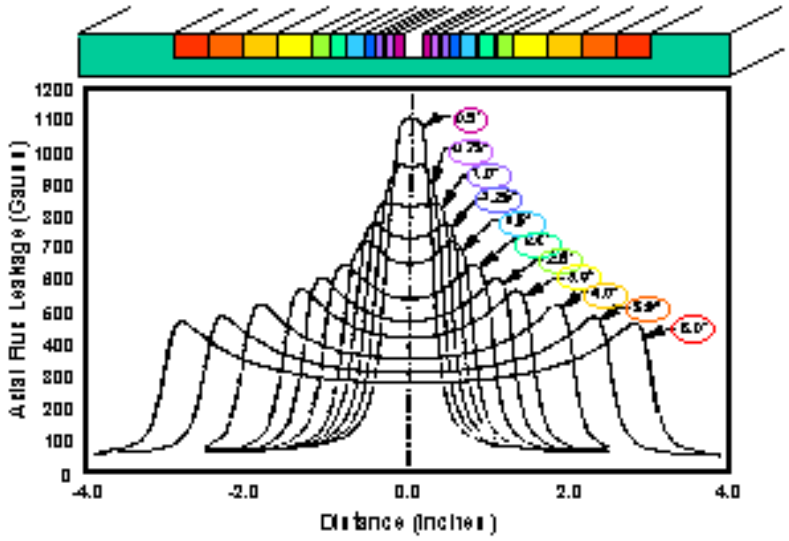

Figura 4. Relación Longitud del defecto-FFM

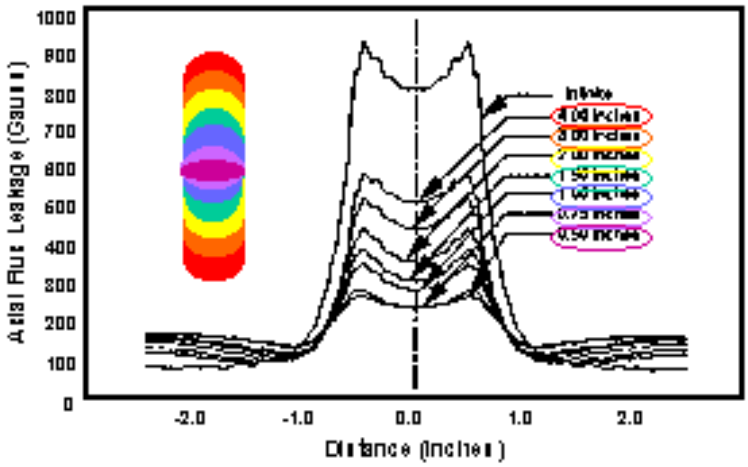

Figura 5. Relación Ancho del defecto-FFM

El flujo magnético es una cantidad vectorial y por esta razón, puede ser medida en cada una de sus componentes. Las componentes axial y radial son las más usadas comúnmente. Estas componentes se pueden observar en la Figura 6.

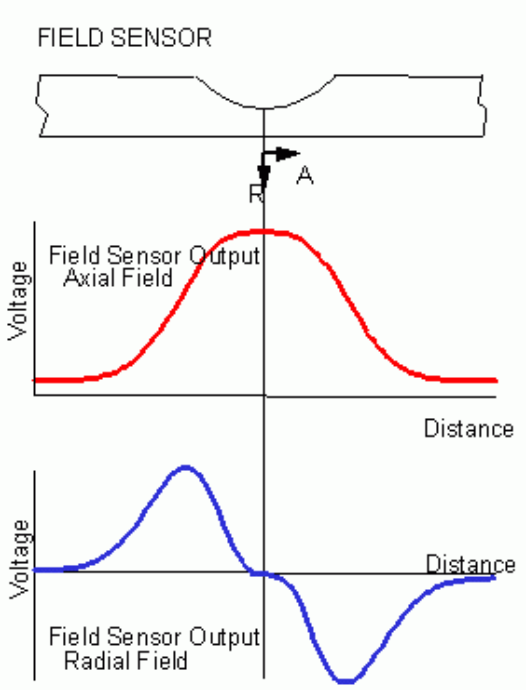

Figura 6. Componentes de FFM más usadas.

4.2. Base de datos suministradas por la CIC

La base de datos que fue suministrada por la CIC consta de 19 señales de 14.449.944 muestras cada una, el muestreo de estas señales fue hecho a 300 Hz, las señales son: 3 Acelerómetros (X, Y, Z), 3 Giroscópios (X, Y, Z), 2 Magnetómetros (X, Y), 2 Brazos Cáliper y 8 FFM; éstas corresponden a lo registrado por el dispositivo ITION en un tramo de tubería de aproximadamente 23 Km. Las marcas o etiquetas suministradas por la CIC para la Base de Datos se observan en la Tabla 1 .

MARCA	POSICIÓN APROXIMADA (M)
'VA'	0
'TE'	1,43
'GR'	2,98
'TA'	3,46
'TA'	3,81
'TA'	4,16
'TA'	4,53
'VA'	5,43
'TA'	6,17
'TA'	6,56
'TA'	6,89
'TA'	7,27
'TA'	8,25
'BR'	9,37
'GR'	10,6
'FL'	11,26
'MA'	17,37
'GR'	30,01
'PM'	600,72
'MA'	1867,41
'CC'	1983,1
'FC'	2001,17
'CC'	2769,13
'FC'	2799,33
'CC'	3686,21
'FC'	3715,85
'MA'	3954,01
'MM'	3971,26
'MM'	5938,94
'MA'	7930,63
'MA'	8002,73
'PM'	8676,37
'CC'	9027,54
'FC'	9045,64
'CC'	10169,42
'FC'	10211,9
'MA'	10300,36
'MM'	10301,52
'CC'	10302,2
'FC'	10332,28
'PM'	11291,64
'TA'	12582,44
'CC'	13095,37
'FC'	13119,5
'MM'	14047,04
'MA'	18095,39
'MM'	18163,51
'MA'	19490,07
'TA'	20265,49
'MM'	21742,03
'MA'	22051,6
'CI'	23855,35
'GR'	23855,97
'TA'	23856,97
'TA'	23857,32
'TA'	23857,66
'TA'	23858,01
'VA'	23859,01

Tabla 1. Marcas suministradas por la CIC para la base de datos.

Como el propósito del proyecto es detectar pérdidas de material, se considerarán solamente las 8 señales de FFM, la amplitud de estas señales no es conocida, las marcas “PM” (Pérdida de Material) se encuentran ubicadas en: 600.72 m, 8676.37 m y 11291.64 m. La CIC definió un intervalo de 14 m con respecto a cada marca de

pérdida de material, esto quiere decir que los defectos podrían estar a +/- 7 m de cada marca. Las secciones de la señal donde se encuentran los defectos se observan en las Figura 7, Figura 8 y Figura 9.

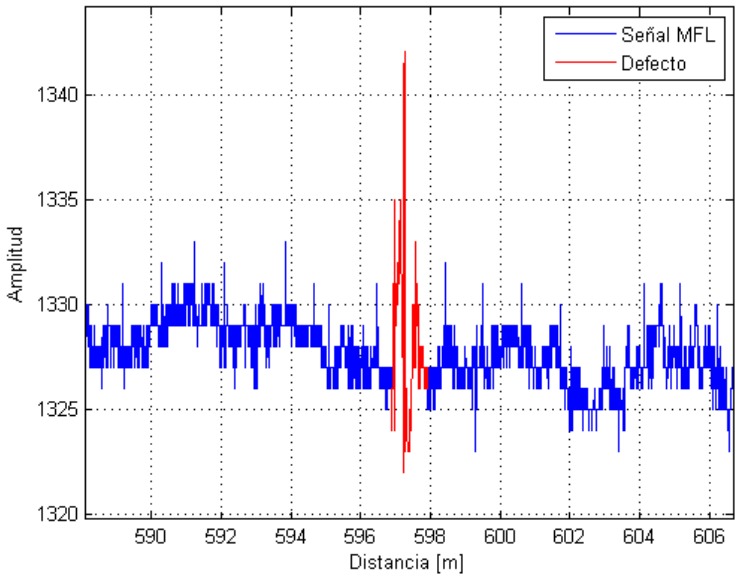

Figura 7. Primer defecto ubicado a +/- 600.72 m.

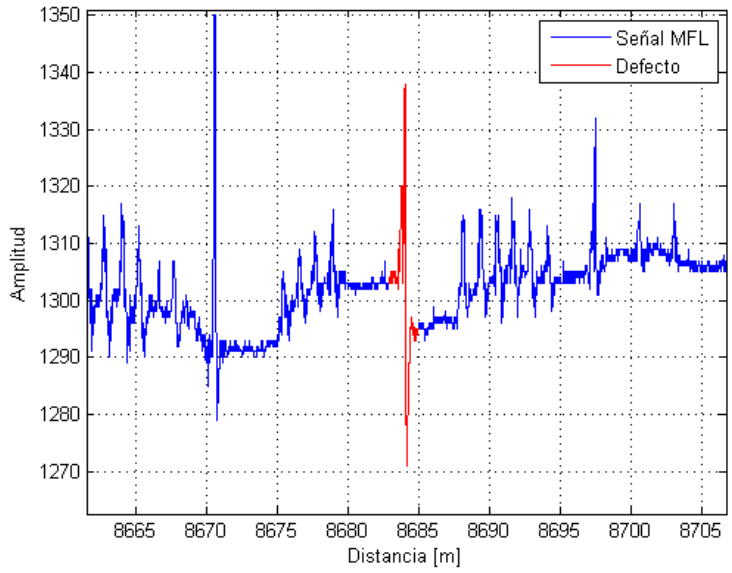

Figura 8. Segundo defecto ubicado a +/- 8676.37 m.

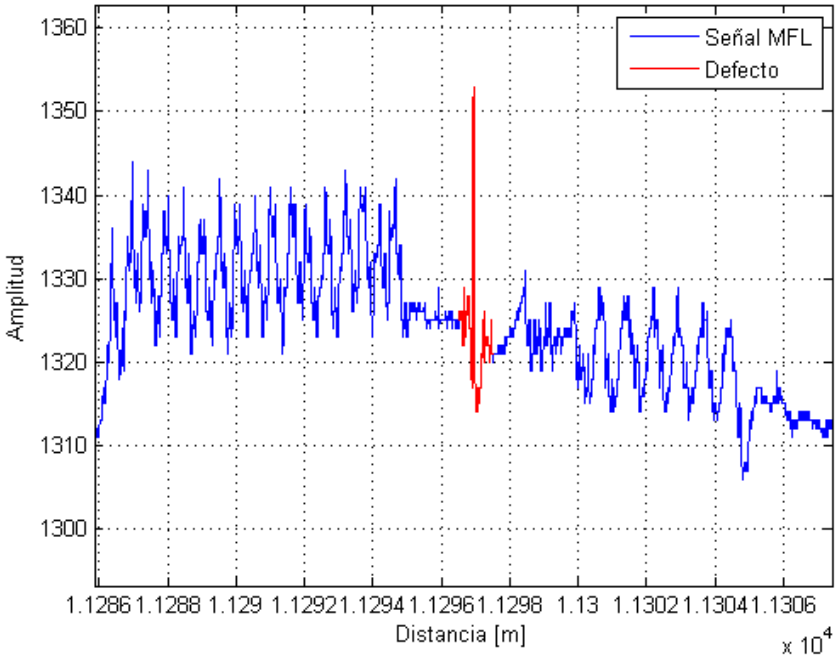

Figura 9. Tercer defecto ubicado a +/- 11291.64 m.

5. PREPROCESAMIENTO DE LOS DATOS

El procesamiento de los datos es una tarea necesaria para la preparación de los datos para su posterior análisis. La justificación de este proceso preliminar al análisis de datos, generalmente, radica en que los datos vienen con una serie de defectos, entre los cuales se encuentran:

- Datos incompletos
- Errores en los datos
- Ruido

5.1. Eliminación de puntos muertos

Las señales fueron recortadas de 14’449.944 muestras (Figura 10) a 10’139.445 muestras (Figura 11), debido a que al inicio de la toma de datos, el dispositivo ITION no se mueve por un periodo de tiempo.

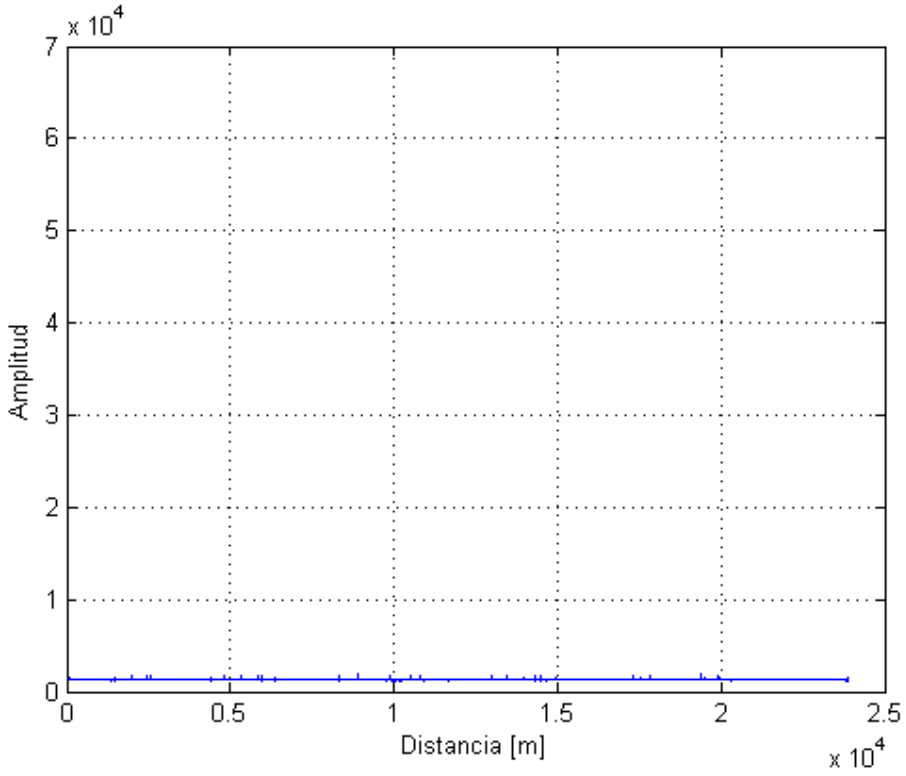

Figura 10. Señal 11 de FFM original sin recortar

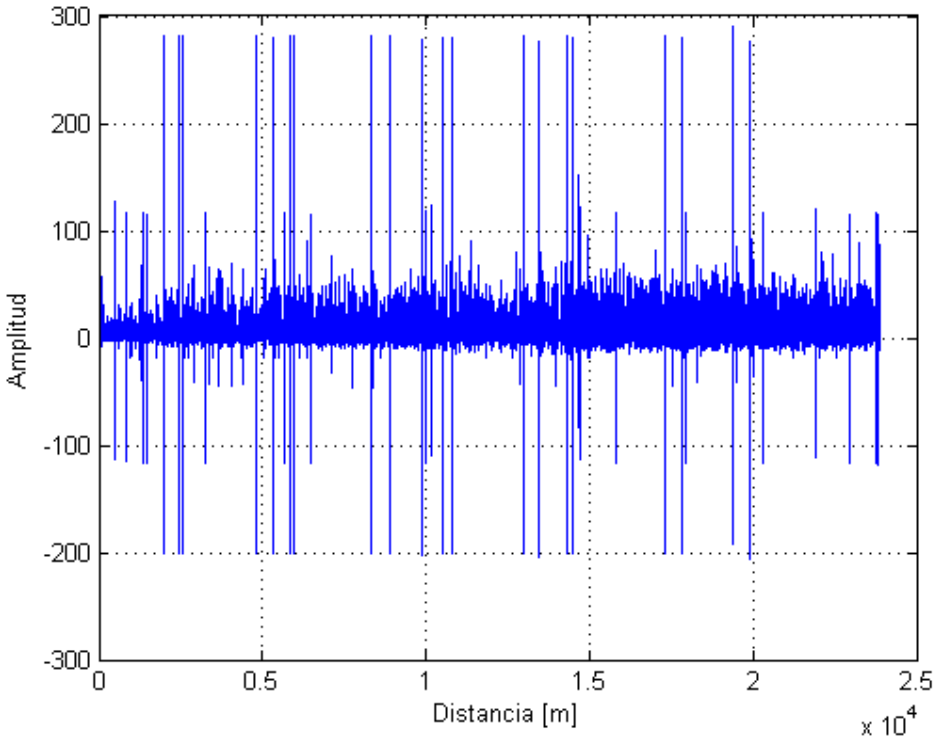

Figura 11. Señal 11 de FFM recortada.

5.2. Transformada wavelet (TW)

La TW está dada por la Transformada Wavelet Continua (TWC) y la Transformada Wavelet Discreta (TWD). Son dos herramientas que permiten el análisis de señales de manera similar a la Transformada de Fourier (TF) con la diferencia que la TW puede entregar información temporal y frecuencial en forma cuasi-simultánea, mientras que la TF sólo da una representación frecuencial. De acuerdo al principio de incertidumbre de Heisenberg, existen limitaciones con la resolución en el tiempo y frecuencia, pero es posible realizar un análisis usando la TW, que permite examinar la señal a distintas frecuencias y con diferentes resoluciones. La TW da una buena resolución temporal y baja resolución en frecuencia para eventos de altas frecuencias y da una buena resolución frecuencial pero poca resolución temporal en eventos de bajas frecuencias [4].

5.2.1. Descomposición multinivel

El proceso de descomposición multinivel es un proceso iterativo con descomposiciones sucesivas que se descomponen a su vez, de modo que una señal se divide en muchos componentes de menor resolución. Esto es el árbol de descomposición *Wavelet,* éste se puede observar en la Figura 12.

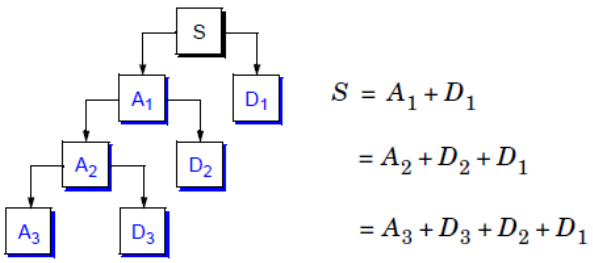

Figura 12. Árbol de descomposición Wavelet multinivel.

En la Figura 13 se observa un ejemplo de descomposición *Wavelet* multinivel, donde “S” es la señal original, cA_x y cD_x son los componentes de aproximación y detalle en el nivel “x”.

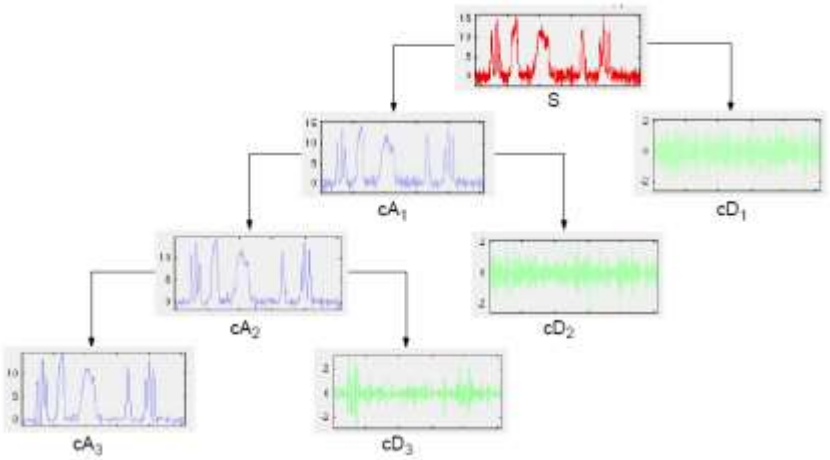

Figura 13. Ejemplo de descomposición Wavelet multinivel.

Los componentes de aproximación y detalle equivaldrían al resultado de aplicar a la señal un filtro pasa bajos y uno pasa altos, respectivamente [5], como se observa en la Figura 14.

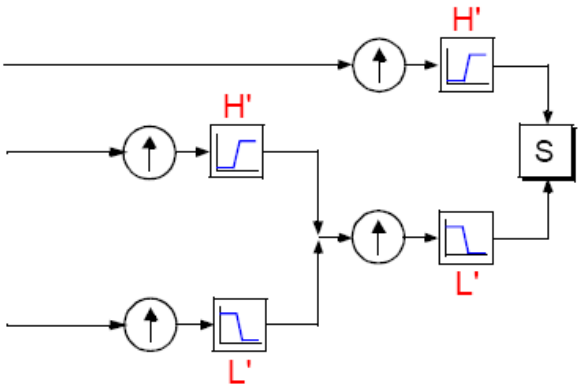

Figura 14. Similitud de la descomposición Wavelet con un banco de filtros.

5.3. Filtrado Shrinkage

El objetivo del filtrado de una señal es eliminar, en lo posible, el ruido producido durante la adquisición de la misma. En este proyecto se plantea el uso de un filtrado Shrinkage, el cual consiste en cinco pasos: descomponer la señal en múltiples niveles aplicando la TWD, determinar un umbral (λ) que esté en función del ruido de la señal, escalar (σ) dicho umbral en función de los coeficientes de detalle hallados anteriormente, implementar el umbral ya escalado en cada uno de los coeficientes

y finalmente, reconstruir la señal con cada nivel de aproximación y detalles filtrados [6]. Un diagrama de esta metodología se puede observar en la Figura 15.

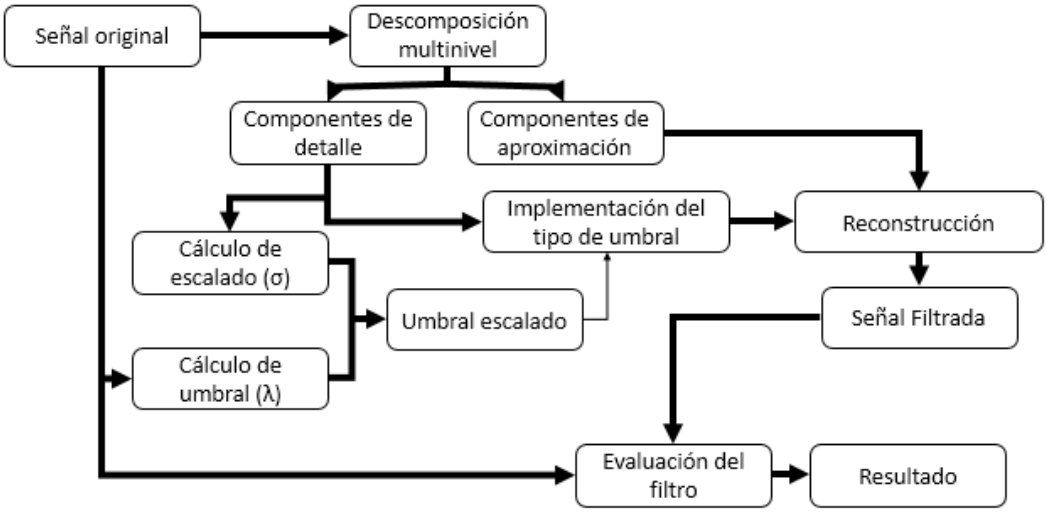

Figura 15. Diagrama Filtrado Shrinkage.

Para realizar este tipo de filtrado es necesario usar una función *Wavelet*, la cual se debe escoger de tal forma que la señal filtrada no difiera tanto de la señal original. Para esto se usa como índice de validación el Error Medio Cuadrático (EMC, Ecuación (1)), lo cual implica que se debe buscar el menor EMC entre la señal original y la señal filtrada, éste último fue propuesto por David L. Donoho.

$$EMC = \frac{1}{N}\sum_{i=1}^{N}\left(x(i) - \hat{x}(i)\right) \quad (1)$$

Donde:

x es la señal original

$\hat{x}$ es la señal filtrada

N es el número de muestra de la señal

Las funciones *Wavelet* utilizadas para realizar el filtrado Shrinkage, así como su orden se ilustran en la Tabla 2.

ACRÓNIMO DE LA FUNCIÓN WAVELET	*NOMBRE DE LA FAMILIA WAVELET*	*ORDEN*
'DB'	***Daubechies***	***Db1:Db20***
'SYM'	***Symlets***	***Sym1:Sym20***
'BIOR'	***Biorthogonal***	***Bior1.3:Bior6.8***

Tabla 2. Funciones Wavelet utilizadas en el filtro.

Para determinar el máximo nivel de descomposición que se utilizará en el proceso de filtrado, se aplicó el criterio de la Entropía de Shannon, cuyo autor es Claude E. Shannon, el cual mide el nivel de incertidumbre en un sistema.

Siguiendo los criterios anteriormente descritos, la funcion *Wavelet* y el máximo nivel de descomposición para cada señal se muestran en la Tabla 3.

SEÑAL	**FUNCIÓN WAVELET ADECUADA**	**MÁXIMO NIVEL DE DESCOMPOSICIÓN**
SEÑAL 11	Symlets 9	**5**
SEÑAL 12	Symlets 9	**5**
SEÑAL 13	Symlets 9	**5**
SEÑAL 14	Symlets 9	**5**
SEÑAL 15	Symlets 9	**5**
SEÑAL 16	Symlets 9	**5**
SEÑAL 17	Symlets 9	**5**
SEÑAL 18	Symlets 9	**5**

Tabla 3. Función Wavelet y nivel de descomposición máximo para cada señal.

Una vez se tiene la función *Wavelet* y el nivel de descomposición máximo para cada señal, se procede a realizar el cálculo del umbral (λ) y su escalado (σ).

Para calcular el umbral (λ) adecuado se compararon cuatro métodos: umbral universal (“*sqtwolog*”), umbral de aplicación del principio de riesgo imparcial de Stein (“*rigsure*”), umbral minmax (“*minmax*”), y umbral por combinación (“*heursure*”), los cuales se encuentran en la “*Wavelet Toolbox*” de *Matlab*®

Para escoger el escalado (σ) del umbral se compararon tres métodos

- **Escalado unitario:** en este caso el escalado es igual a uno.

$$\sigma = 1 \tag{2}$$

- **Escalado Sln**: el escalado "Sln" del umbral está definido por:

$$\sigma = \frac{\overline{|d_1|}}{0.6745} \tag{3}$$

Donde:

- $\overline{|d_1|}$ es la media de los coeficientes del primer nivel de detalle.

El escalado del umbral está en función del primer nivel de detalle.

- **Escalado Mln**: el escalado "Mln" del umbral está definido por:

$$\sigma = \frac{\overline{|d_\iota|}}{0.6745} \tag{4}$$

Donde:

- $\overline{|d_\iota|}$ es la media de los coeficientes del nivel "i" de detalle.

En este caso habrá un escalado del umbral para cada nivel de detalle.

Para determinar el método de implementación se tuvieron en cuenta dos métodos:

- **Hard:** este método está descrito por la Ecuación 5

$$D_h(Y,\lambda) = \begin{Bmatrix} Y, |Y| \geq \lambda \\ 0, |Y| < \lambda \end{Bmatrix} \tag{5}$$

Donde:

- D_h son los coeficientes de detalle obtenidos al implementar el método *Hard.*
- Y son los coeficientes de detalle originales.

- λ es el umbral aplicado.
- sgn es la función signo.

- **Soft:** este método está descrito por la Ecuación 6

$$D_s(Y,\lambda) = \begin{cases} sgn(Y)(Y-\lambda), |Y| \geq \lambda \\ 0, |Y| < \lambda \end{cases} \quad (6)$$

Donde:

- D_s son los coeficientes de detalle obtenidos al implementar el método *Soft.*
- Y son los coeficientes de detalle originales.
- λ es el umbral aplicado.
- sgn es la función signo.

Para identificar la combinación adecuada de ésta metodología, se filtraron las ocho señales con todas las combinaciones posibles y se calculó el EMC entre la señal original y la señal filtrada, en cada una de las combinaciones.
Las combinaciones de umbral, escalado del umbra y método de implementación con menor EMC se muestran en la Tabla 4.

SEÑAL	**CÁLCULO DEL UMBRAL**	**ESCALADO DEL UMBRAL**	**MÉTODO DE IMPLEMENTACIÓN**
SEÑAL 11	*minmax*	Unitario	*Soft*
SEÑAL 12	*minmax*	Unitario	*Soft*
SEÑAL 13	*minmax*	Unitario	*Soft*
SEÑAL 14	*minmax*	Unitario	*Soft*
SEÑAL 15	*minmax*	Sln	*Soft*
SEÑAL 16	*minmax*	Sln	*Soft*
SEÑAL 17	*minmax*	Sln	*Soft*
SEÑAL 18	*minmax*	Sln	*Soft*

Tabla 4. Mejor metodología de filtrado para cada señal.

El cálculo del umbral y el método de implementación fueron los mismos para todas las señales, pero las señales 11 a 14 se obtuvieron los menores EMC escalando el umbral de forma unitaria, mientras que las señales 15 a 18, se obtuvieron los menores EMC con el escalado del umbral Sln.

Finalmente se determina el mejor ancho de ventana para realizar el filtrado de las señales, esto con el fin de aumentar el rendimiento y la velocidad del filtrado. Para este caso se utiliza el criterio de buscar la mayor relación señal a ruido (SNR por sus siglas en inglés, Ecuación 7), entre la señal filtrada con una ventana y un segmento la señal del mismo tamaño que la ventana.

$$SNR = 10log\left[\frac{\sum_{i=1}^{N}[x(i)]^2}{\sum_{i=1}^{N}[x(i) - \hat{x}(i)]}\right] \quad (7)$$

Donde

- x es la señal original.
- $\hat{x}$ es la señal filtrada.
- N es el número de muestras de la señal.

Los anchos de la ventana variaron desde 1.000 muestras hasta 2’000.000 de muestras, con incrementos de 1.000 muestras, el desplazamiento de la ventana se realizó cada $N - 1$ muestras, y las muestras solapadas se promediaron, además, si el último segmento de señal restante era menor que el tamaño de la ventana, se toma como tamaño de ventana final el total de muestras restantes. La Figura 16 muestra un ejemplo de este ventaneo.

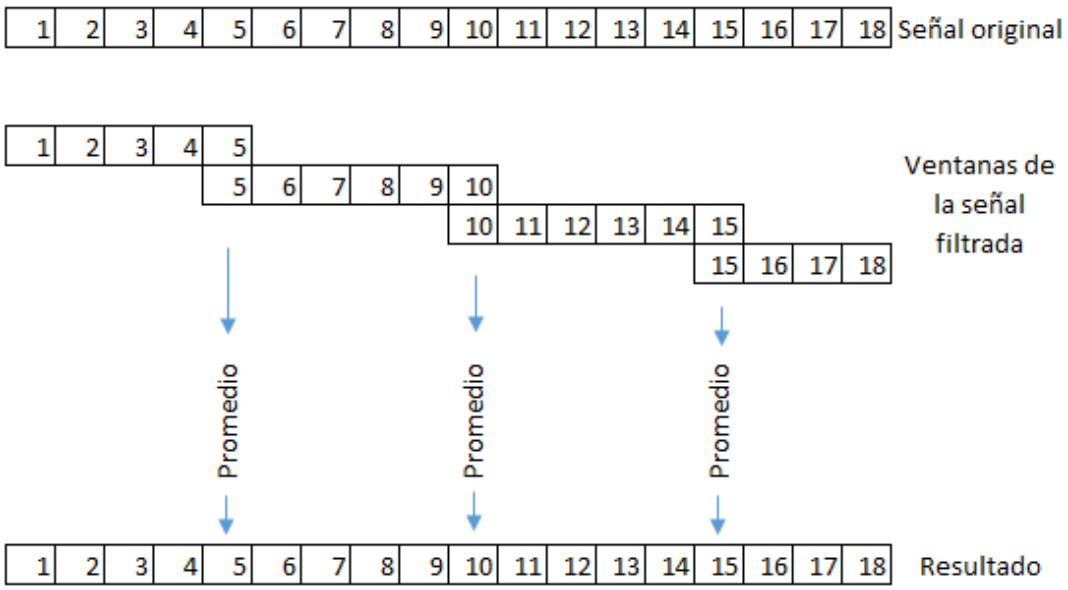

Figura 16. Ejemplo de filtrado con ventaneo.

La SNR para la Señal 11 de FFM se muestra en la Figura 17. La variación de la SNR va de 54.6 hasta 54.63

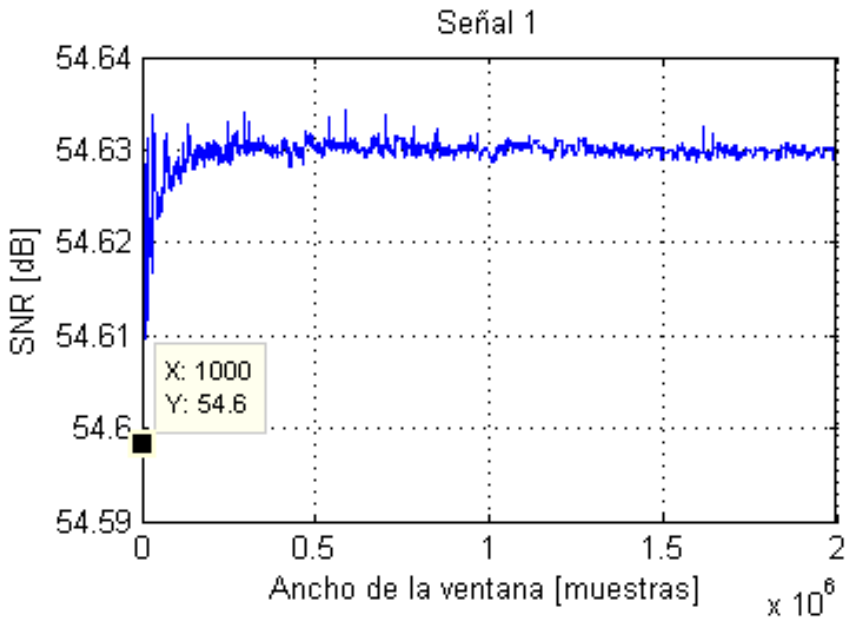

Figura 17. SNR para diferentes anchos de ventana, Señal 11 de FFM.

El resultado fue similar para todas las señales. La variación de la SNR entre una ventana de 1.000 muestras y otra de 2'000.000 de muestras fue mínima.

Por lo mencionado anteriormente se selecciona un ancho de ventana de 2^{20} muestras (1'048.576), ya que la eficiencia de los algoritmos aumenta al utilizar ventanas cuyos anchos de ventana sean potencia de 2.

Teniendo en cuenta que en el momento de realizar el filtrado se contaba con un computador con 4 GB de RAM y un procesador de 2.4 GHz, no se podía seleccionar una ventana tan grande.

El resultado de este filtro se puede observar en la Figura 18, Figura 19 y Figura 20.

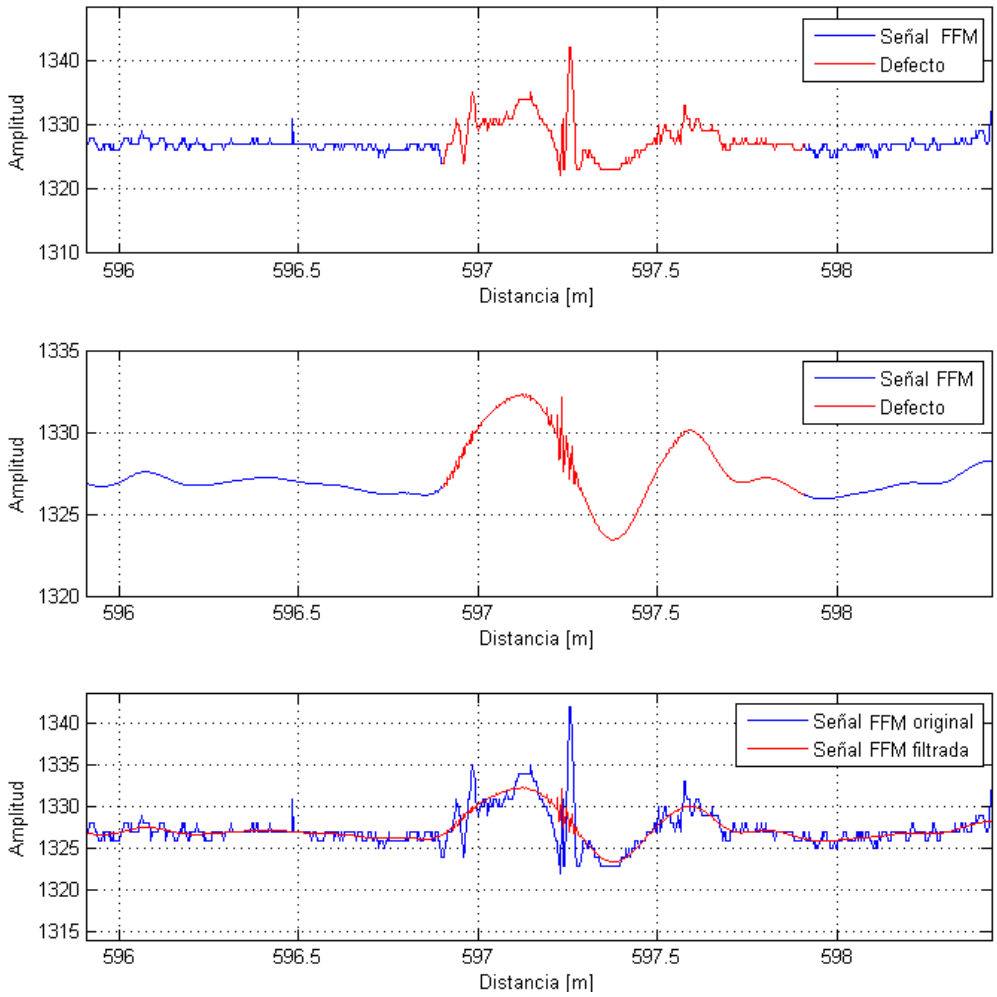

Figura 18. Comparación entre la señal original y la señal filtrada (Defecto 1).

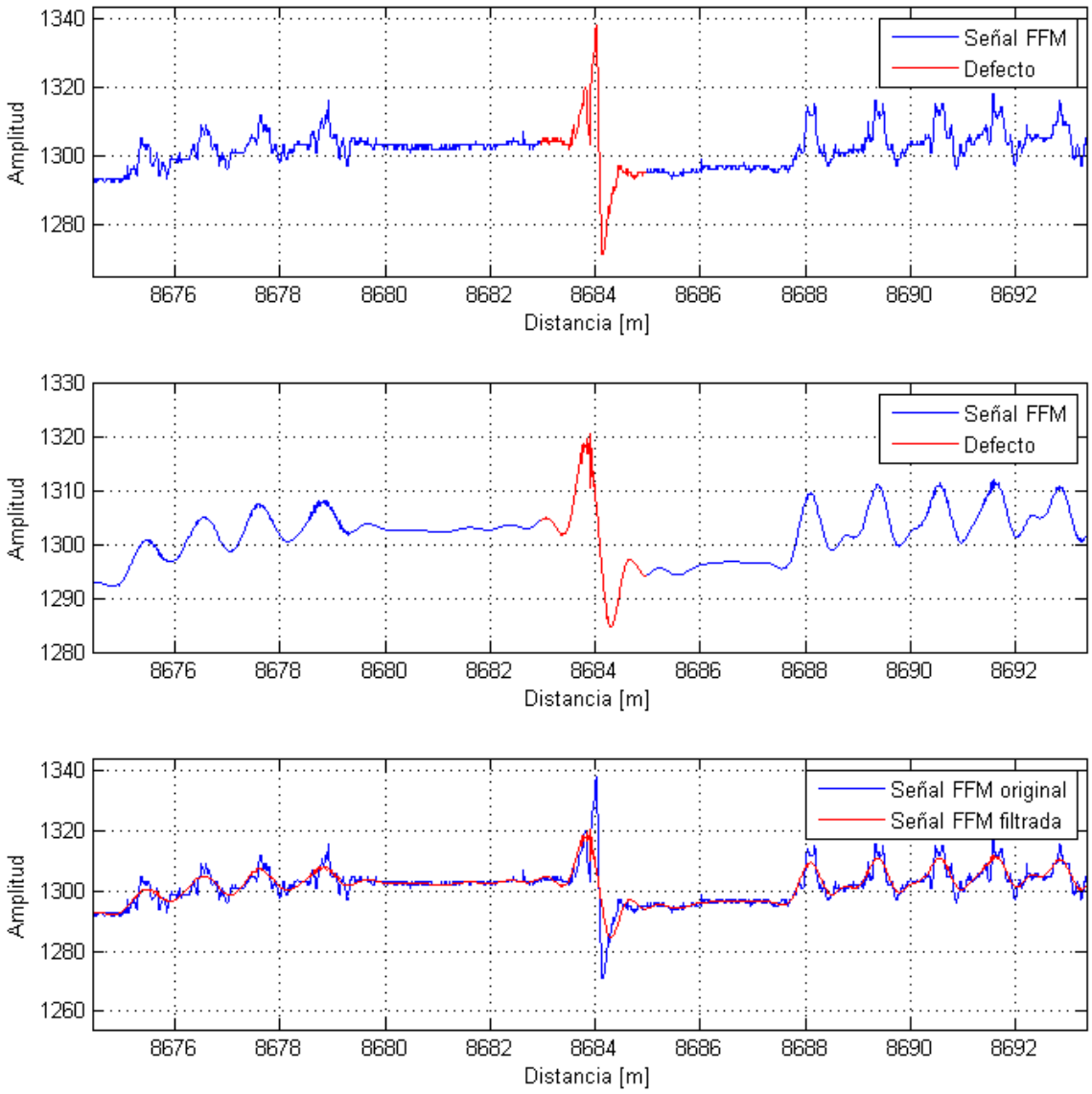

Figura 19. Comparación entre la señal original y la señal filtrada (Defecto 2).

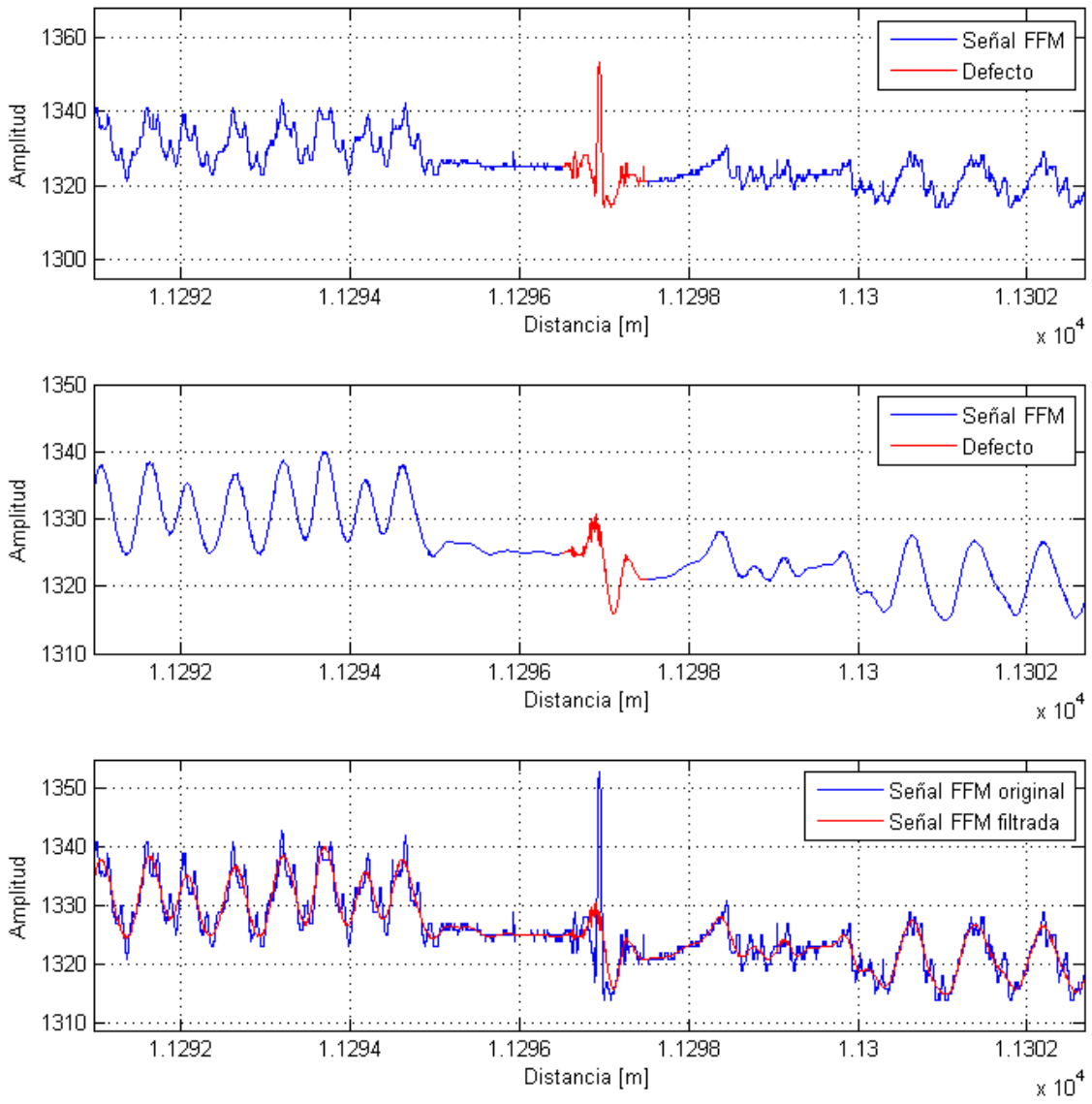

Figura 20. Comparación entre la señal original y la señal filtrada (Defecto 3).

El filtro eliminó casi todo el ruido existente conservando el la forma de onda de los defectos y de la señal en general.

5.4. Corrección de Línea Base

Es un tipo de preprocesamiento que intenta corregir determinadas tendencias que aportan ruido a la señal. Existen varios tipos de corrección de línea base según el efecto que se desea corregir.

Un tipo de corrección es el que ajusta la señal original a una función cuadrática, sustrayéndola posteriormente de la señal original, eliminando la característica curvatura ascendente a medida que aumenta la longitud de onda [7].

De esta forma, el conjunto de señales corregidas tiende a tener una media de valor 0 y una varianza diferente de cero dependiendo de la señal. La Figura 21 muestra un ejemplo de corrección de línea base a una señal médica.

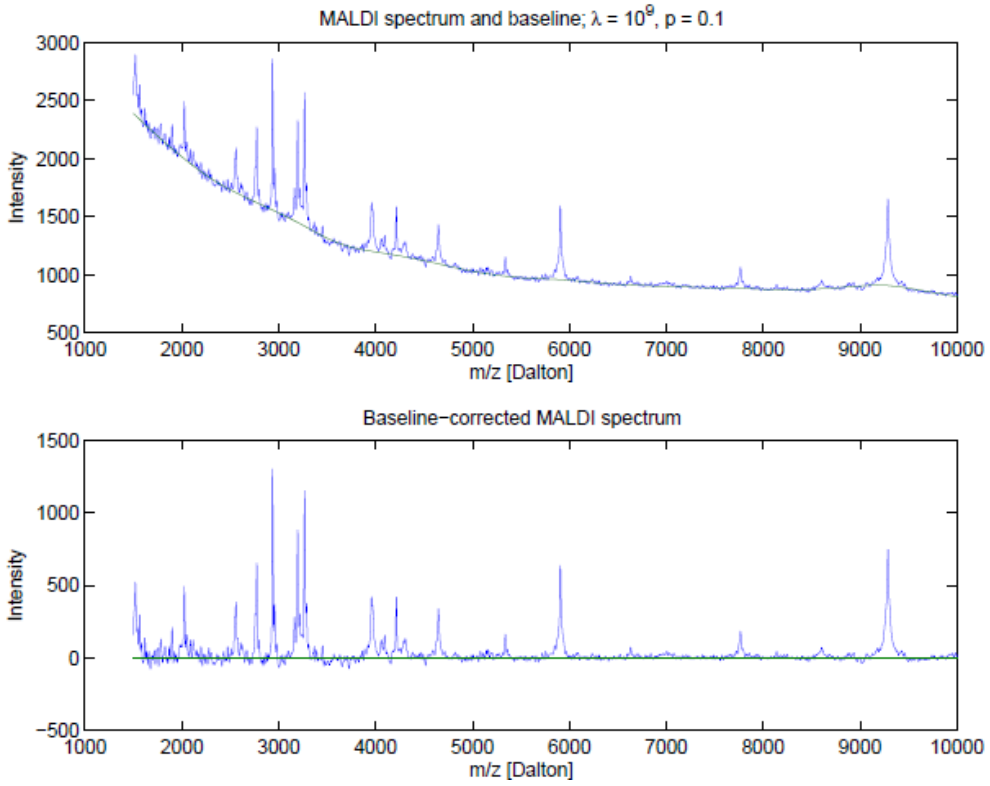

Figura 21. Corrección de línea de base de un cromatograma GC. Arriba: cromatograma GC y línea base estimada. Abajo: cromatograma con línea base corregida [8].

La corrección de línea base se hizo a cada señal de manera independiente, y el resultado se puede observar en las Figura 22, Figura 23 y Figura 24. Estas correcciones de línea base corresponden a la señal 11 de FFM.

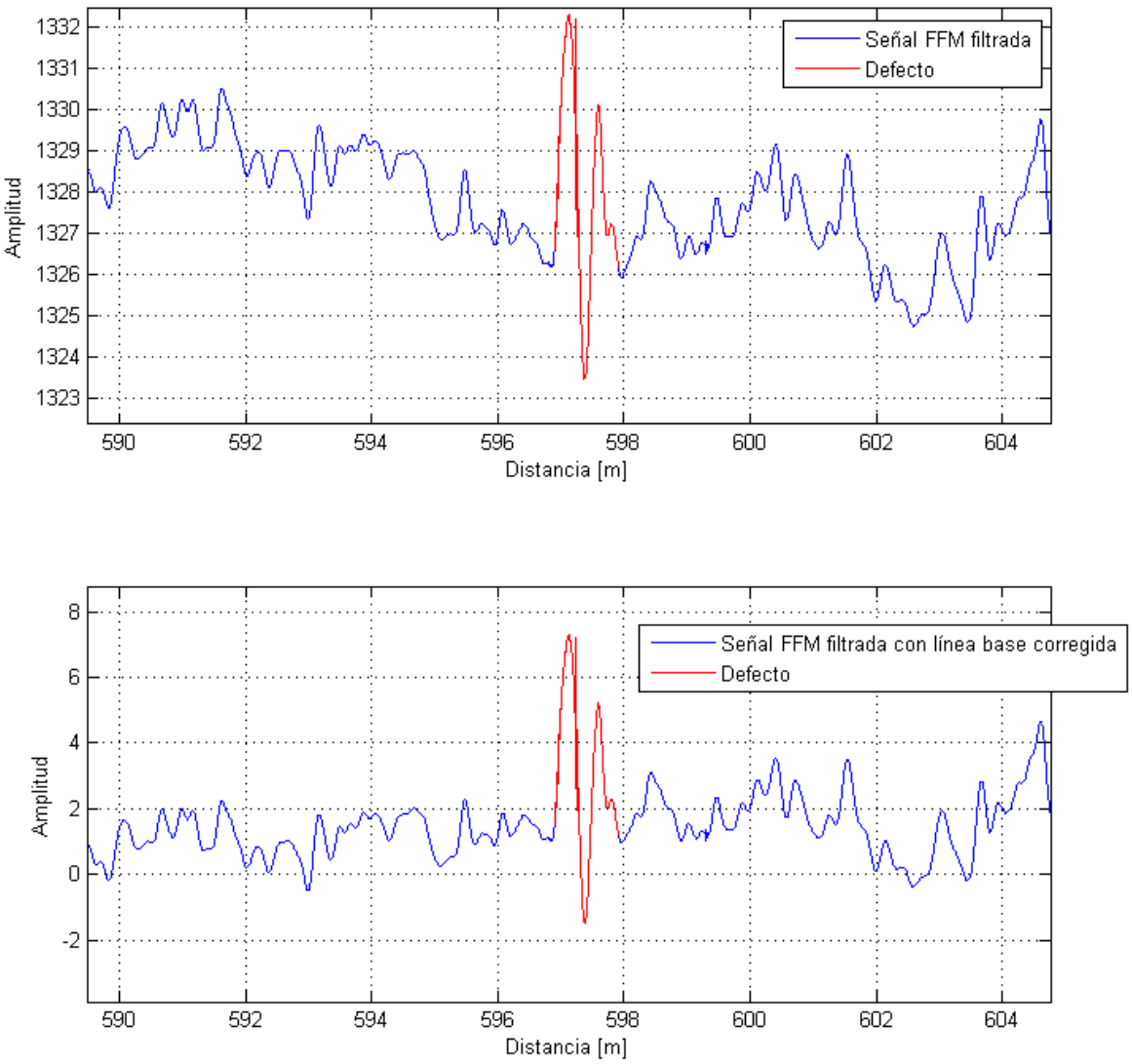

Figura 22. Comparación entre la señal original y la señal con línea base corregida (Defecto 1).

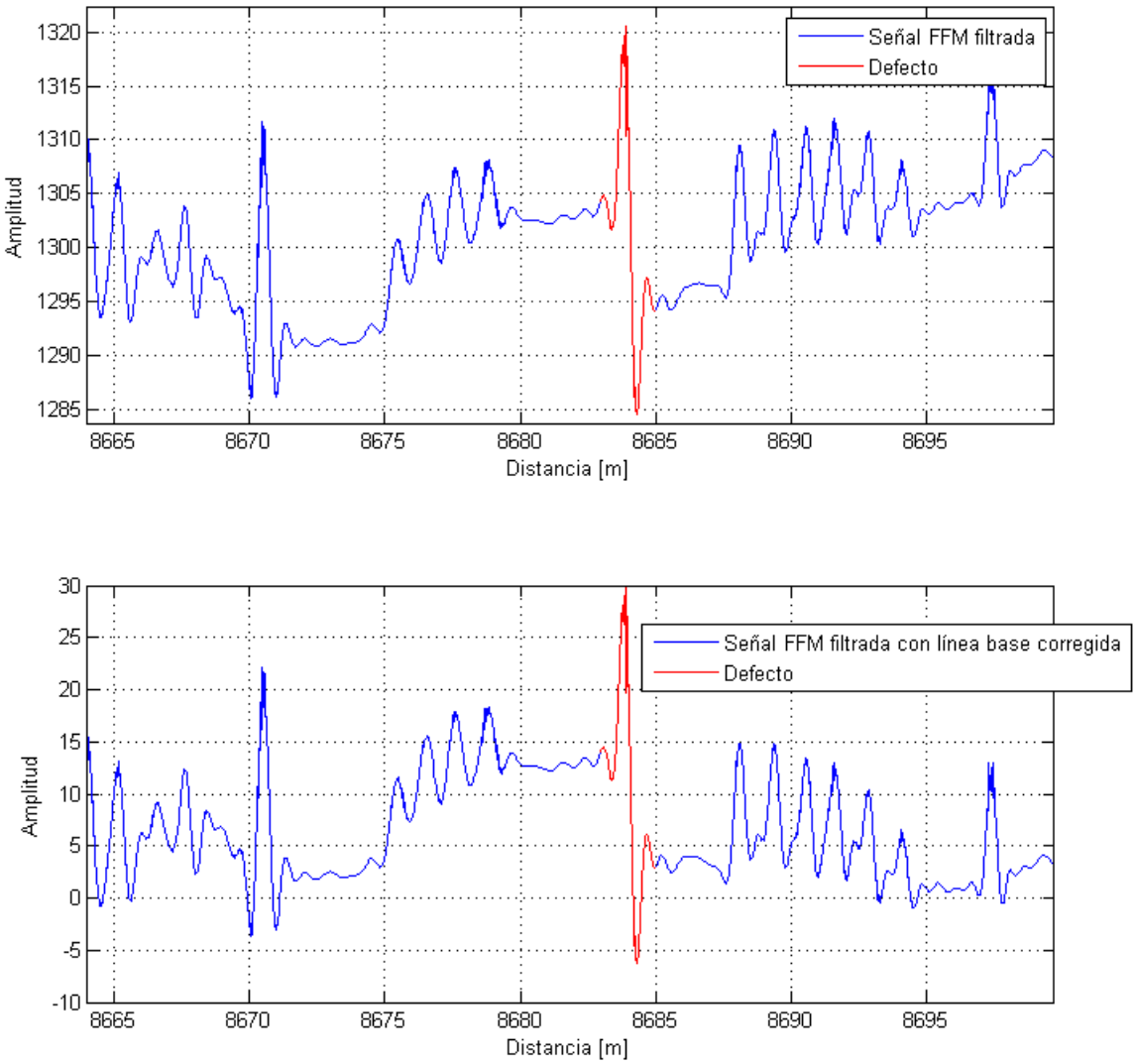

Figura 23. Comparación entre la señal original y la señal con línea base corregida (Defecto 2).

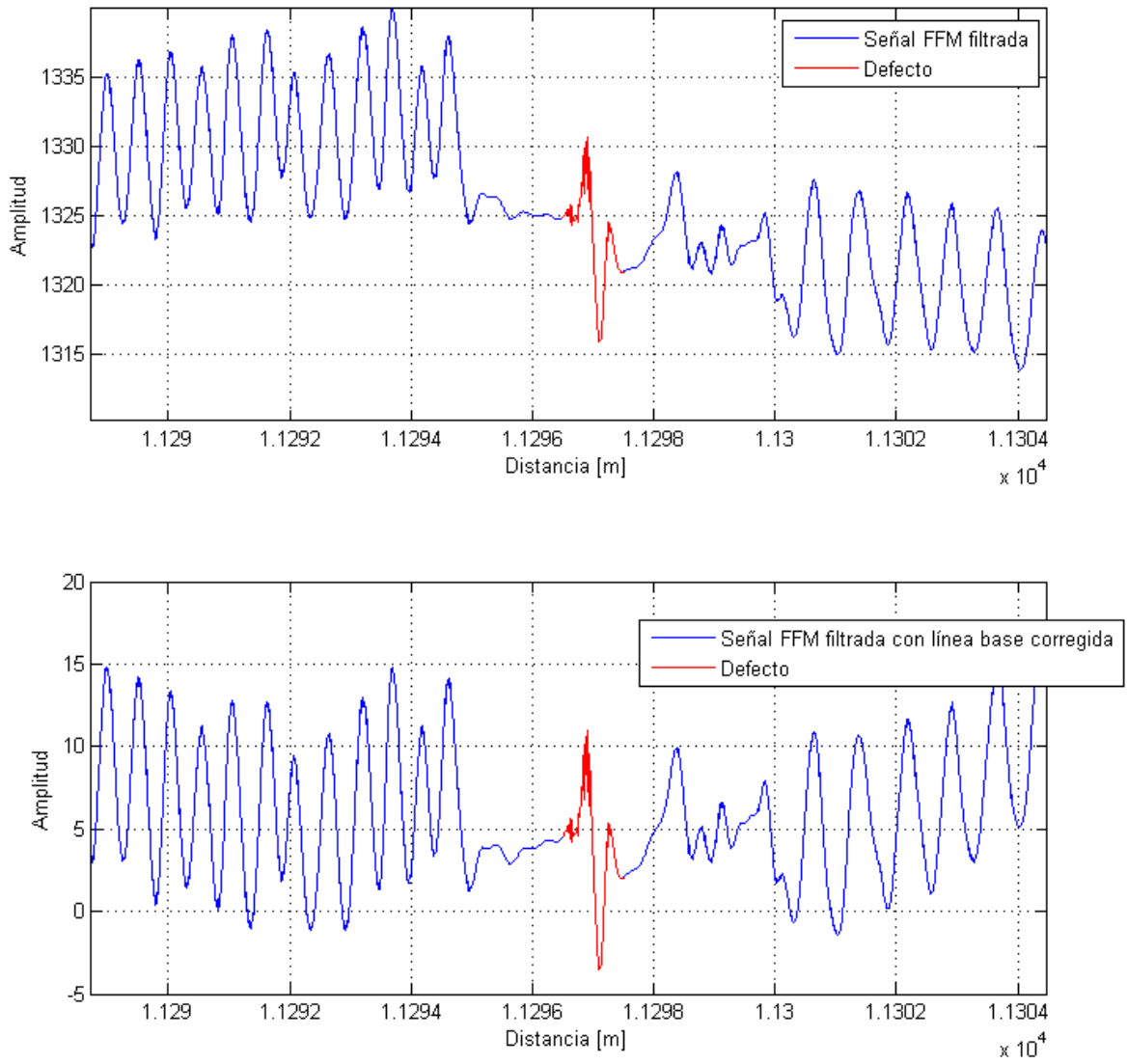

Figura 24. Comparación entre la señal original y la señal con línea base corregida (Defecto 3).

5.5. Eliminación de las soldaduras

Tomando las etiquetas de soldaduras suministradas por la CIC y en conjunto con las etiquetas que se encontraron en [9] se procede a eliminar las soldaduras de la señal.

Esta eliminación se muestra en la Figura 25, en total, fueron eliminadas 1430 soldaduras.

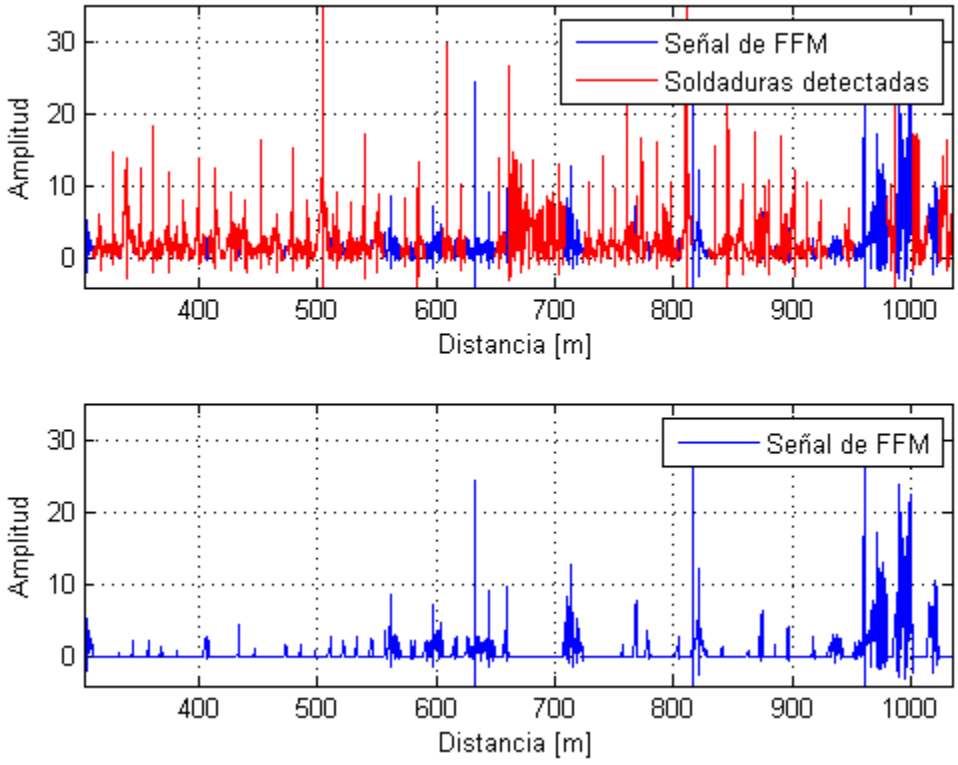

Figura 25. Muestra de eliminación de soldaduras.

5.6. Remuestreo de la señal

En procesamiento de digital de señales, es frecuente recurrir al remuestreo de las mismas. En este caso las señales de FFM son visualizadas teniendo como abscisa la señal de odómetro. Para realizar un procesamiento con transformada *Wavelet* es necesario que la distancia entre muestras sea fija. En este punto se propone realizar un remuestreo en distancia, desde 5 mm hasta 300 mm con incremento de 5 mm.

Para seleccionar la mejor tasa de remuestreo, se propone comparar la densidad espectral de potencia de la señal original con la de la señal remuestreada, a diferentes bandas de frecuencia, y se calcula el EMC entre estas, esto con el fin de conservar la forma de onda de toda la señal original.

Teniendo en cuenta que la señal original fue tomada a 300 Hz, se proponen 5 anchos de banda a comparar, 0.5 Hz, 1 Hz, 1.5 Hz, 2 Hz, 2.5 Hz. Estos anchos de banda recorren toda la densidad espectral de potencia de las dos señales comparadas, en total serán 600, 300, 200, 150 y 120 comparaciones con cada ancho de banda respectivamente, este proceso se repetirá para cada señal de FFM y para cada tasa de remuestreo propuesta. En la Figura 26 se muestra un ejemplo de este proceso, independiente del ancho de banda utilizado.

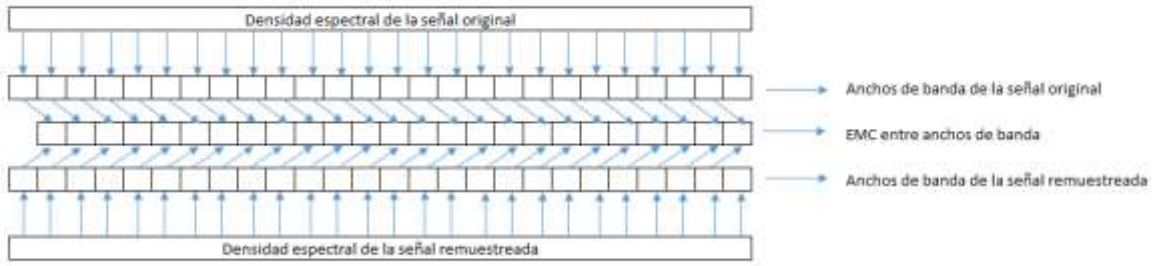

Figura 26. Ejemplo de la comparación de los anchos de banda entre la densidad espectral de frecuencia de la señal original y de la señal remuestreada.

La Figura 27, Figura 28, Figura 29, Figura 30 y Figura 31 muestran los anchos de banda utilizados (0.5 Hz, 1 Hz, 1.5 Hz, 2 Hz, 2.5 Hz respectivamente) para calcular el EMC.

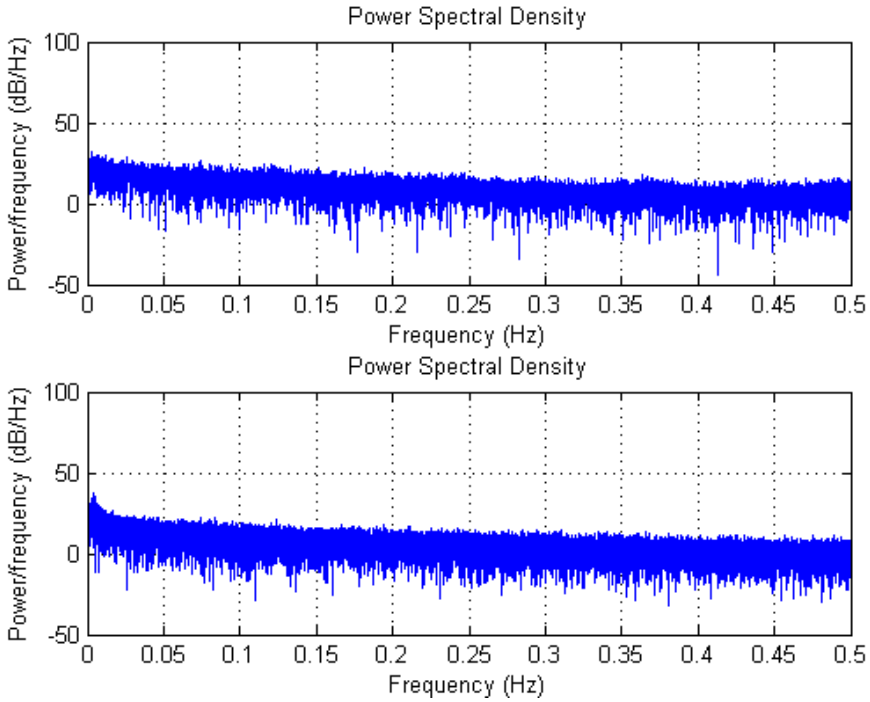

Figura 27. Densidad espectral de potencia de un ancho de banda de 0.5 Hz de la señal remuestreada a 5 mm (arriba) y la señal original (abajo).

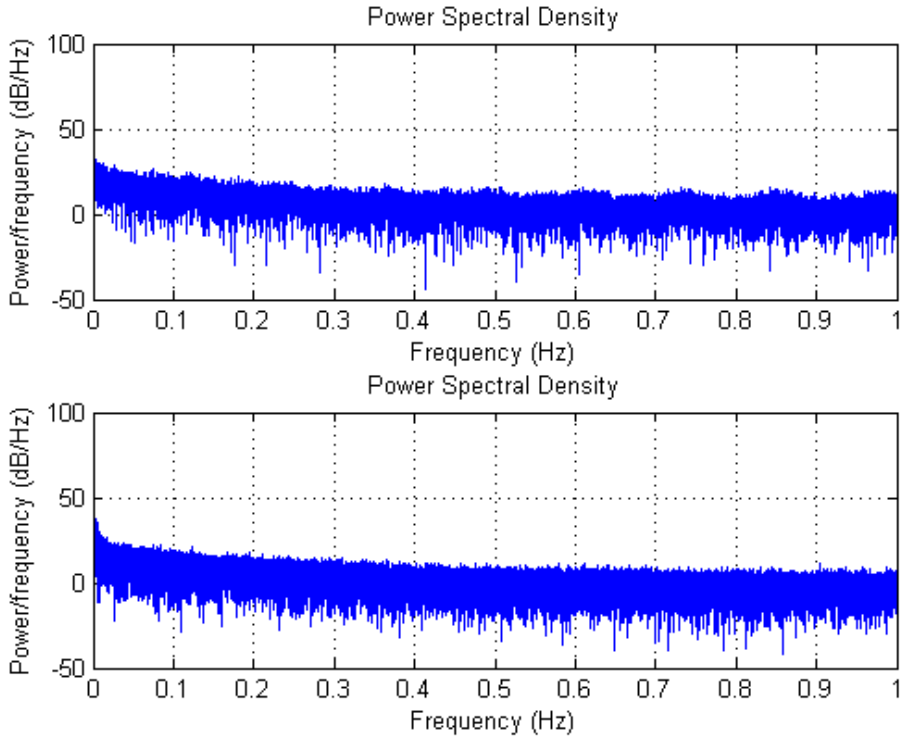

Figura 28. Densidad espectral de potencia de un ancho de banda de 1 Hz de la señal remuestreada a 5 mm (arriba) y la señal original (abajo).

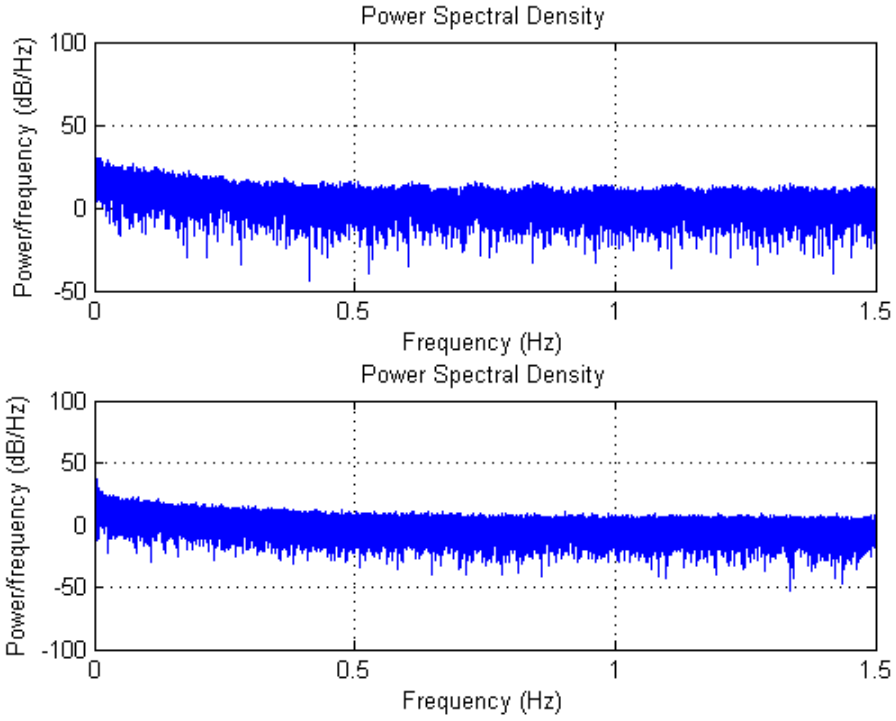

Figura 29. Densidad espectral de potencia de un ancho de banda de 1.5 Hz de la señal remuestreada a 5 mm (arriba) y la señal original (abajo).

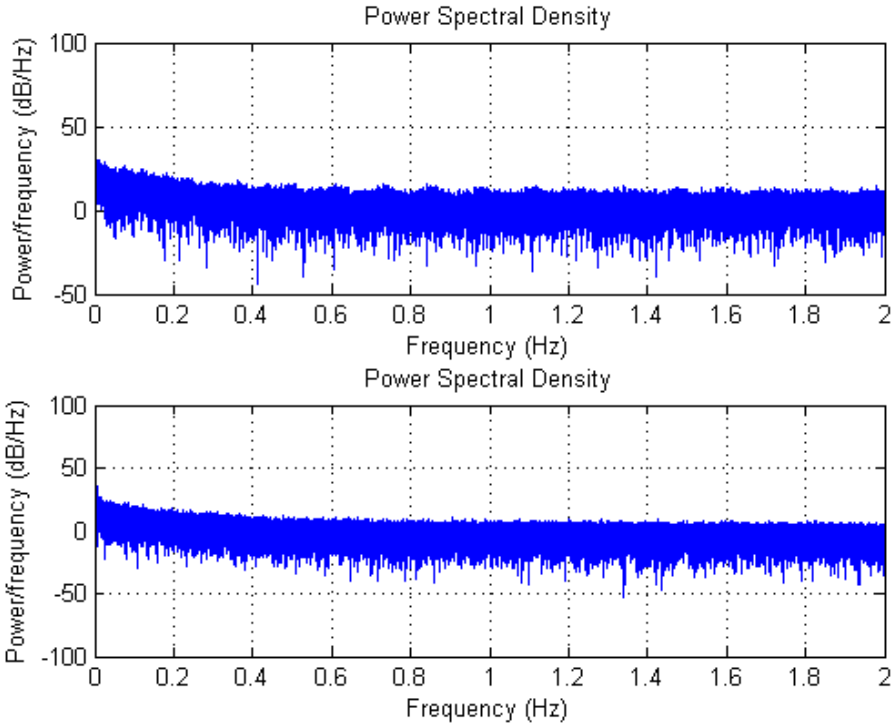

Figura 30. Densidad espectral de potencia de un ancho de banda de 2 Hz de la señal remuestreada a 5 mm (arriba) y la señal original (abajo).

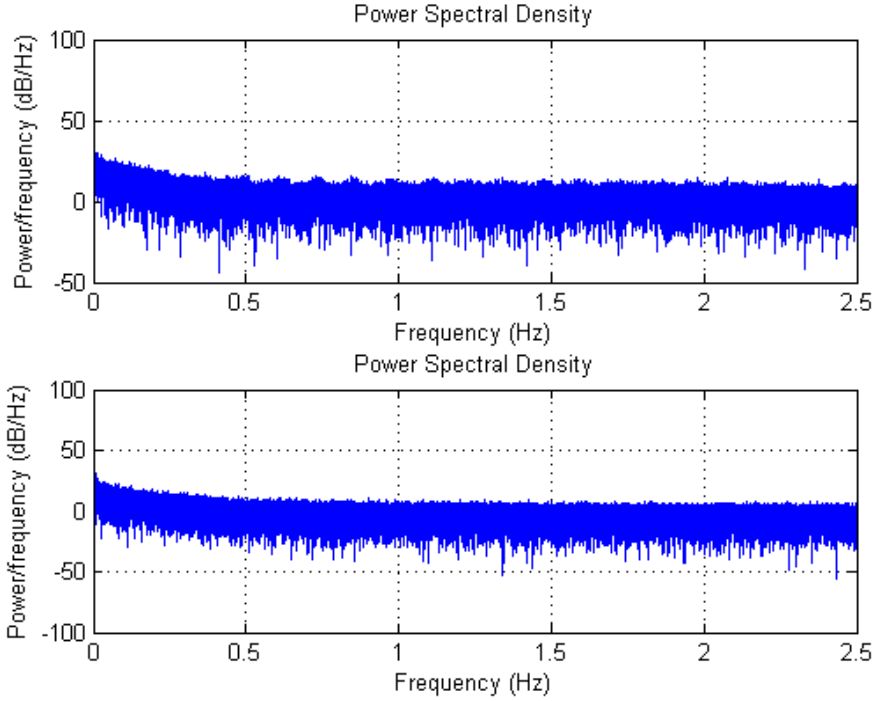

Figura 31. Densidad espectral de potencia de un ancho de banda de 2.5 Hz de la señal remuestreada a 5 mm (arriba) y la señal original (abajo).

El resultado del EMC para cada ancho de banda y con todas las tasas de remuestreo de la señal 1 de FFM se puede observar en la Figura 32. Se puede observar que mientras más pequeña es la tasa de remuestreo, menor es el EMC, por esta razón las señales se remuestrearon a 5 mm. Las señales de FFM pasaron de tener 10’139.445 de muestras a tener 4’778.915 de muestras. Como la distancia entre muestras es de 5 mm, la señal sería de 23’894.575 mm, lo cual equivale aproximadamente a 23.9 Km, que es la distancia recorrida por el dispositivo ITION.

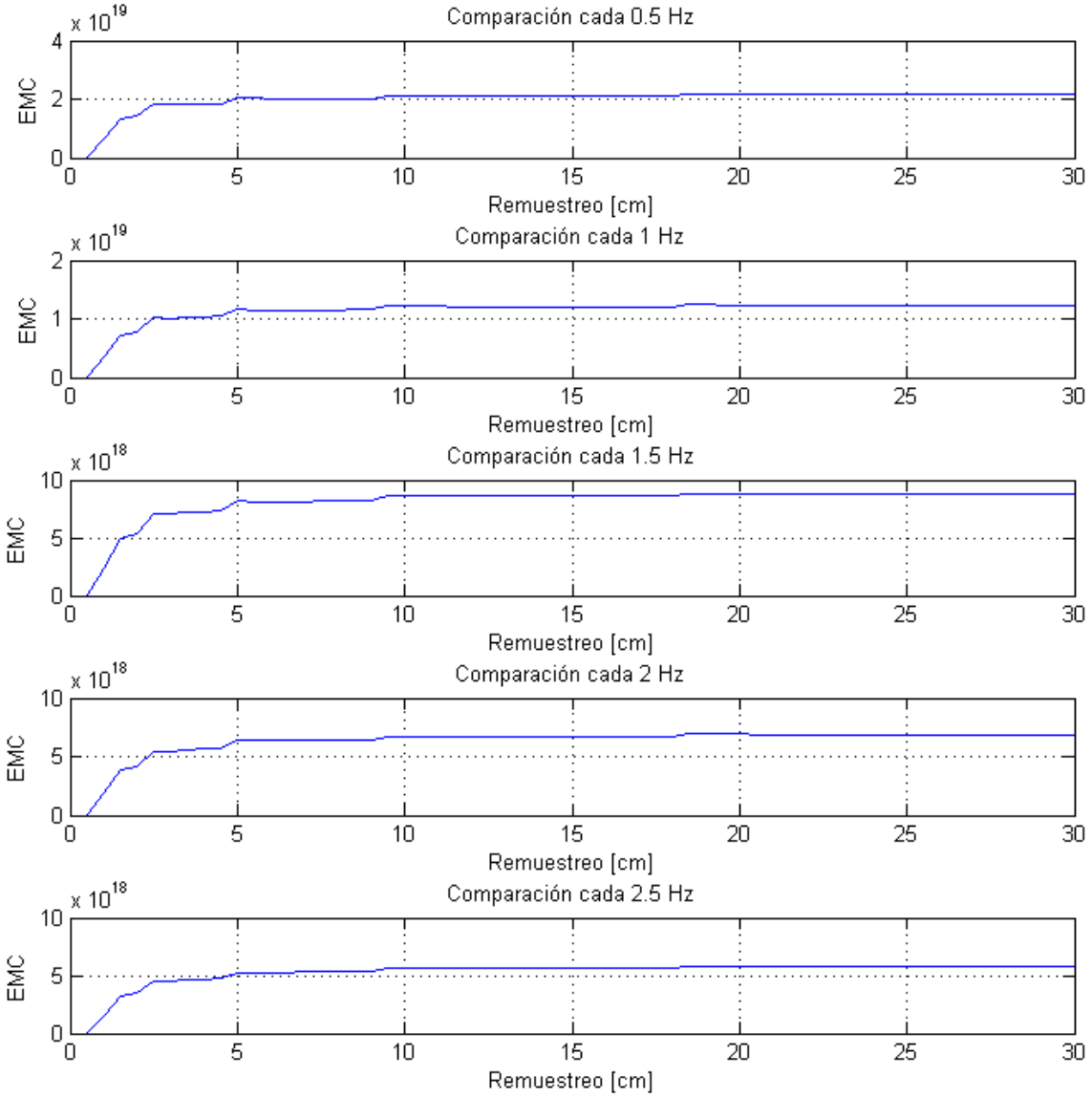

Figura 32. EMC de los diferentes anchos de banda y para todas las tasas de remuestreo.

La Figura 33 muestra una superposición entre la señal filtrada sin soldaduras y la señal remuestreada, se observa que la forma de onda de a señal se mantuvo en casi su totalidad.

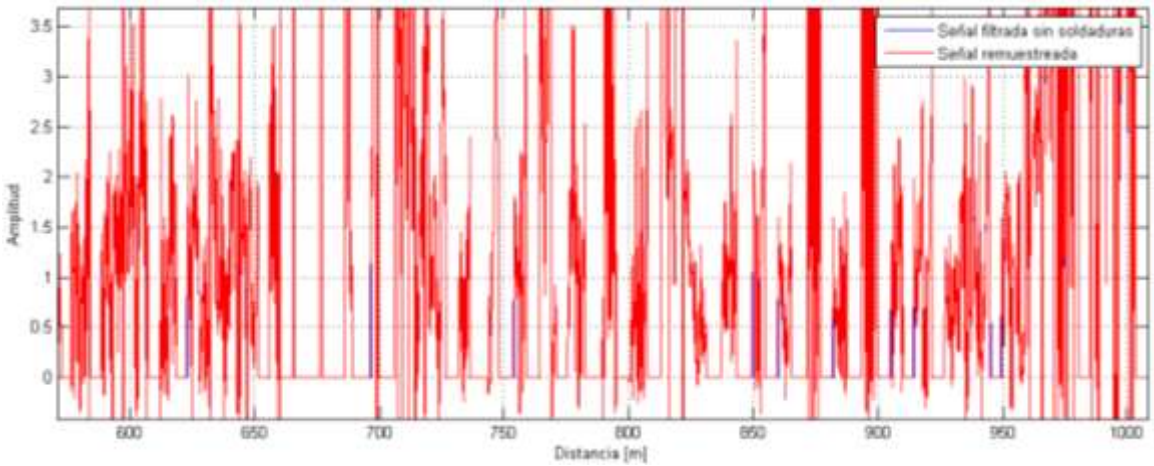

Figura 33. Comparación Señal filtrada y Señal remuestreada.

6. PROCESAMIENTO DE LOS DATOS

El procesamiento de los datos es la aplicación de una serie de operaciones lógicas y matemáticas a un conjunto de datos provenientes de una señal.

Los objetivos de este procesamiento pueden ser diversos, por ejemplo, añadir una información a una señal y que se pueda recuperar más tarde en otro lugar o en otro momento o modificar la señal de tal forma que satisfaga ciertas necesidades o requerimientos.

6.1. Corrección de ruido con Wavelet Tree

A continuación se realiza una descomposición *Wavelet* para las señales de FFM con el fin de encontrar una señal reconstruida que pueda diferenciar más fácilmente el ruido y los defectos en toda la señal.

Se utilizó la herramienta *Wavelet packet* de Matlab® para hallar el árbol de descomposición, y se utilizó la distancia euclídea para encontrar el nodo que mejor diferencie el ruido de los defectos, mientras más grande sea dicha distancia, la separación entre el ruido y el defecto será mayor. Para realizar esta prueba se

utilizan las funciones *Wavelet* que se muestran en la Tabla 2, y el nivel de descomposición que se muestra en la Tabla 3, de modo que son 36 funciones *Wavelet* y 64 nodos.

Para realizar este proceso, se toman las muestras correspondientes a los 3 defectos, y muestras aleatorias que corresponden no defecto. La Figura 34 muestra un ejemplo de los datos que se utilizarán para esta prueba.

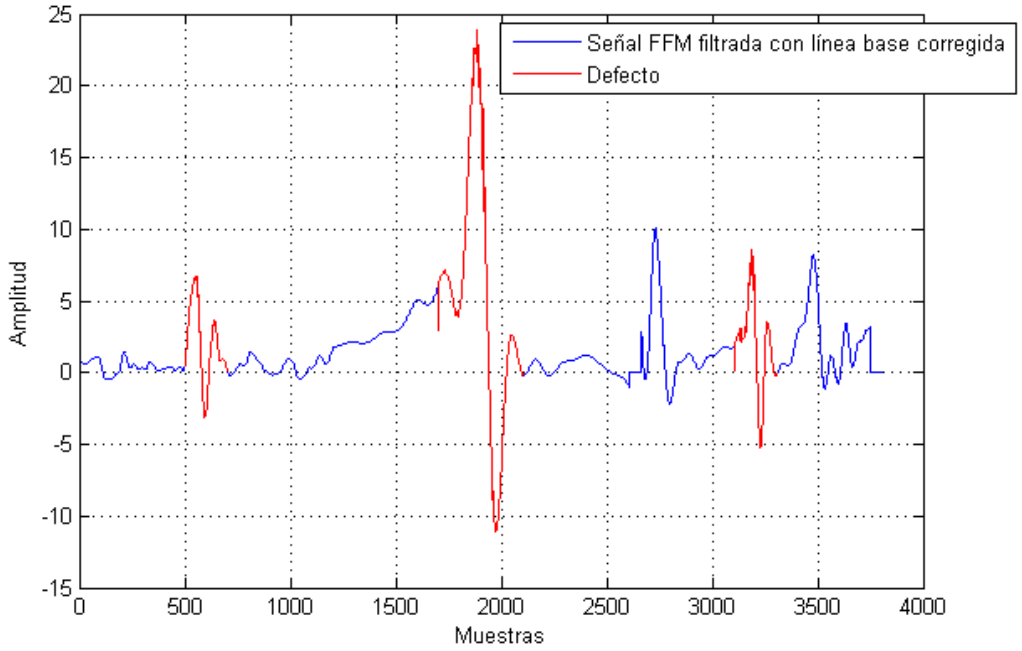

Figura 34. Señal utilizada para realizar la descomposición.

La Figura 35 muestra el árbol de descomposición *Wavelet* con los datos de la Figura 34 y el nodo y la función *Wavelet* que maximizan la distancia euclídea entre el ruido y el defecto se observa en la Figura 36.

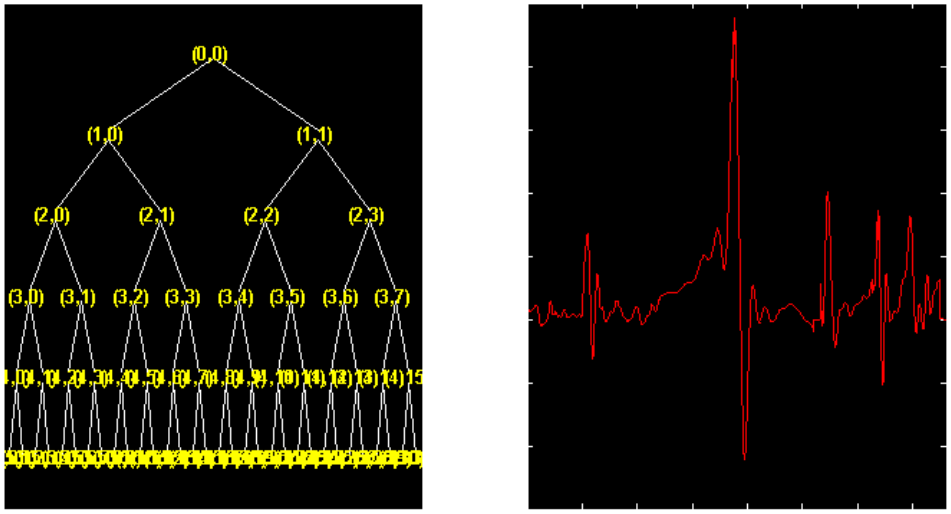

Figura 35. Árbol de descomposición Wavelet.

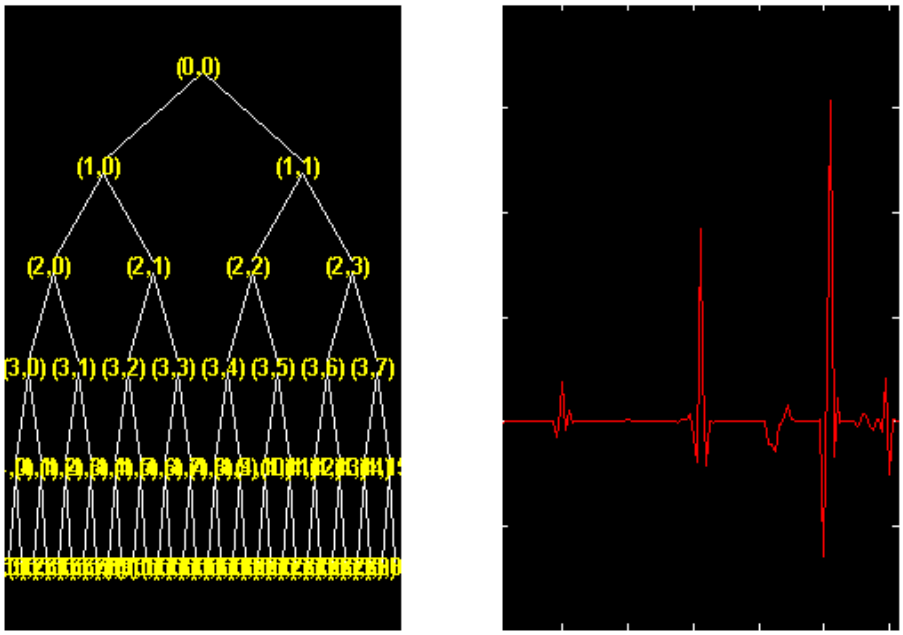

Figura 36. Mejor nodo para los datos de la prueba en el Nodo 58.

El resultado de todas las distancias euclideas para los 62 nodos y las 36 funciones Wavelet para la señal 11 de FFM se observa en la Figura 37.

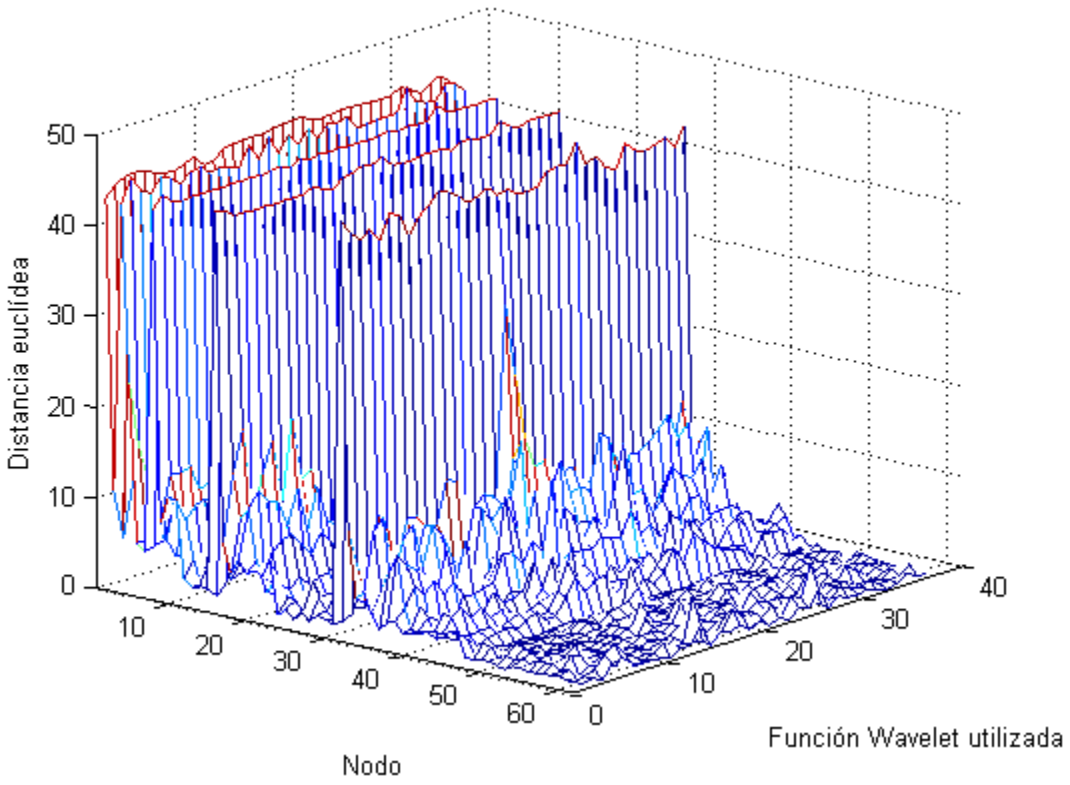

Figura 37. Distancias euclídeas para la señal 11 de FFM.

6.1.1. Selección del nodo y función *Wavelet* adecuada

Para realizar una selección de manera adecuada, se deben normalizar las distancias euclídeas. Cada señal de FFM cuenta con una matriz de tamaño 36x62, las filas corresponden a las funciones *Wavelet* y las columnas a los nodos. Se calcula el promedio de las distancias en cada nodo, luego cada columna de la matriz se divide entre el promedio del nodo correspondiente hallado en el paso anterior, esto se hace con el fin de normalizar cada grupo de datos.

Luego de esto se selecciona el que tenga el mayor índice normalizado para cada señal. El nodo y la función *Wavelet* óptimos para cada señal de FFM se muestran en la Tabla 5.

SEÑAL	NODO	FUNCIÓN WAVELET
SEÑAL 11	48	bior3.1
SEÑAL 12	53	bior1.5
SEÑAL 13	26	bior1.5
SEÑAL 14	58	coif1
SEÑAL 15	48	bior3.1
SEÑAL 16	48	bior3.1
SEÑAL 17	52	coif2
SEÑAL 18	44	bior2.2

Tabla 5. Nodo y función Wavelet óptima para cada señal de FFM.

El resultado del filtro se consigue reconstruyendo la señal con la función *Wavelet* y el nodo seleccionados. La Figura 38, Figura 39 y Figura 40 muestran el segmento donde se encuentran los defectos, después de haber aplicado este filtrado en la señal 11 de FFM.

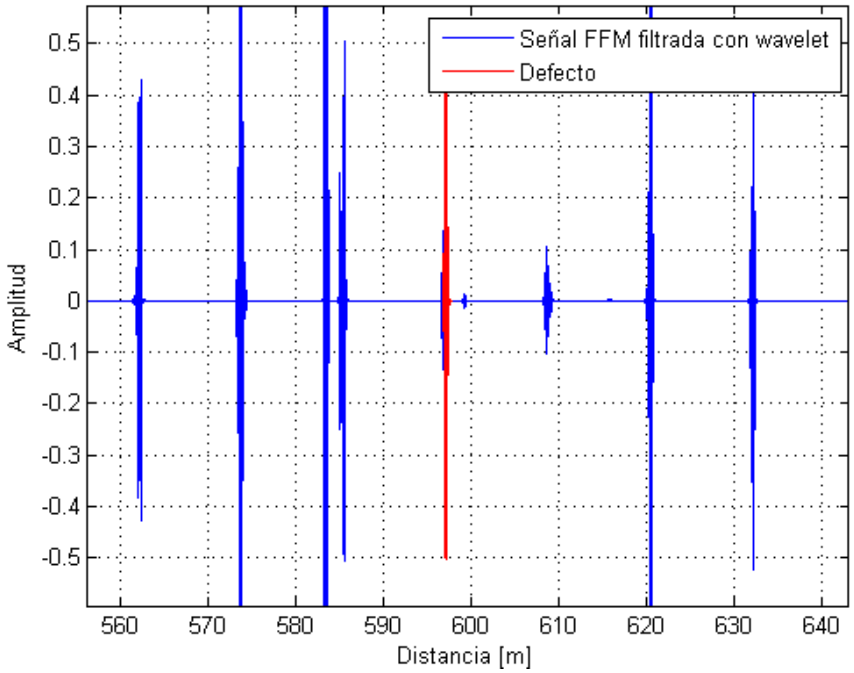

Figura 38. Señal filtrada con Wavelet (Defecto 1).

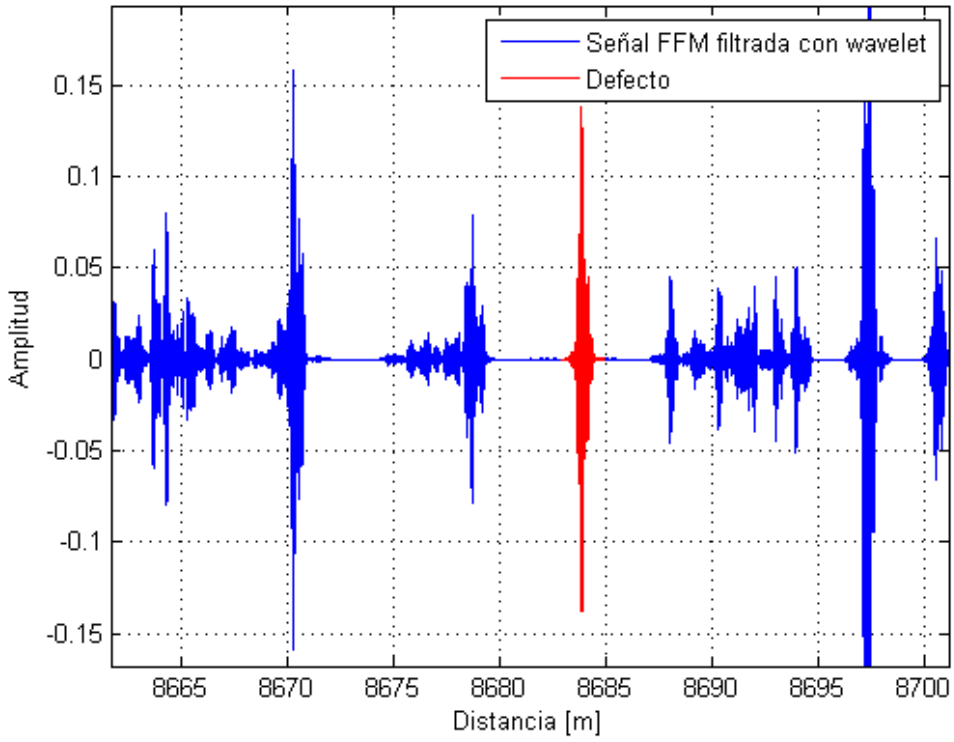

Figura 39. Señal filtrada con Wavelet (Defecto 2).

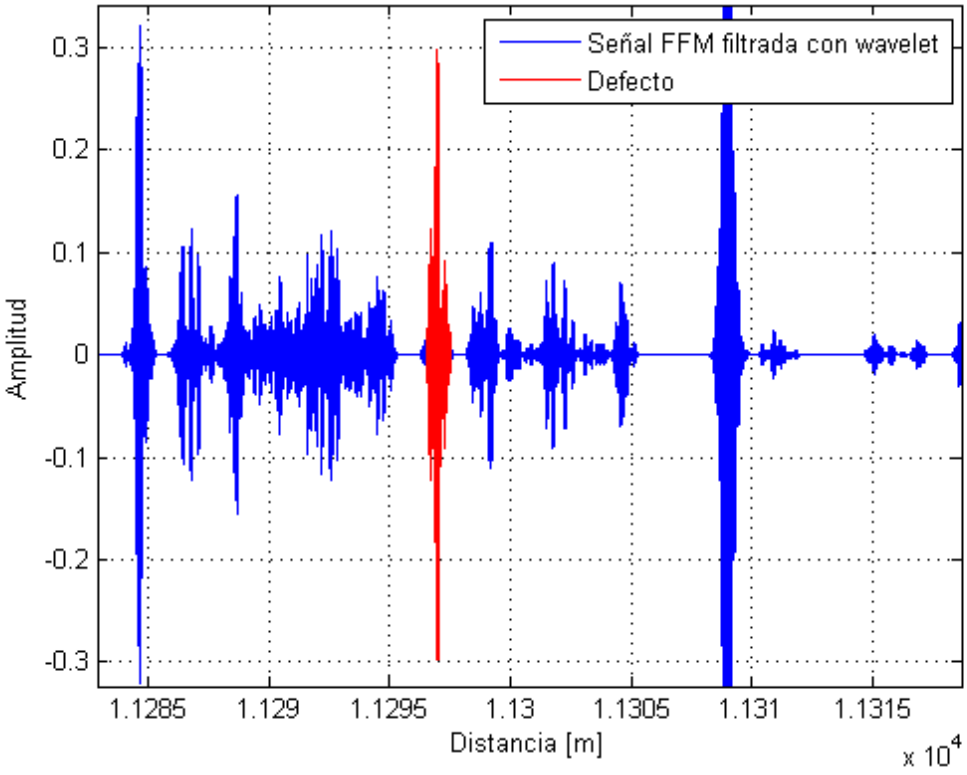

Figura 40. Señal filtrada con Wavelet (Defecto 3).

6.2. Ventaneo con estadísticos

Para caracterizar series temporales, se han usado estadísticos como la media, la desviación estándar y el rango intercuartil de las señales, para este punto se propone ventanear las señales con 8 estadísticos: media, mediana, rango intercuartil, desviación media absoluta, rango, desviación estándar, energía y potencia.

Para seleccionar el ancho de ventana óptimo se procede a hacer la prueba de U de Mann-Whitney, la cual es una prueba no paramétrica para verificar si las medianas de dos muestras son independientes. Cuando el p-Value (resultado de la prueba) es menor que 0.05, se dice que existe separación entre las medianas de cada clase.

Para este punto se utilizó una ventana cuyo ancho variaba desde 2 hasta 600 muestras, en este caso no realizó el ventaneo a toda la señal, se hizo en 3 segmentos de prueba para cada señal, los cuales correspondían a cada defecto más un número muestras que correspondían a ruido; el número de muestras del segmento era 600 (entre muestras de defecto y muestras de ruido) que era el doble del número de muestras del defecto más grande, esto se hizo con el fin de mostrar qué pasaría si se selecciona una ventana demasiado grande.

A medida que la ventana recorría los segmentos de prueba, se calculan los descriptores antes mencionados, cuando un ancho de ventana termina de recorrer toda la señal, se realiza la prueba de U de Mann-Whitney y se comienza a calcular cada descriptor para todas las señales con el ancho de ventana siguiente.

Luego de que todos los descriptores han sido calculados con todos los anchos de ventana para todas las señales, se procede a realizar la sumatoria o conteo de los p-Value menores a 0.05 para cada ancho de ventana en todas las señales y con cada defecto, es decir, se obtuvo una gráfica para cada defecto en la cual muestra el conteo de los p-Value del mismo a diferentes anchos de ventana.

Como son 8 estadísticos y también 8 señales de FFM, el máximo valor del conteo para cada defecto será de 64. La Figura 41 muestra el conteo de cada defecto con todos los anchos de ventana utilizados, teniendo en cuenta que 2 muestras son 1 cm. Como se puede observar, mientras el ancho de ventana se acerca al ancho de cada defecto, el conteo tiende a aumentar.

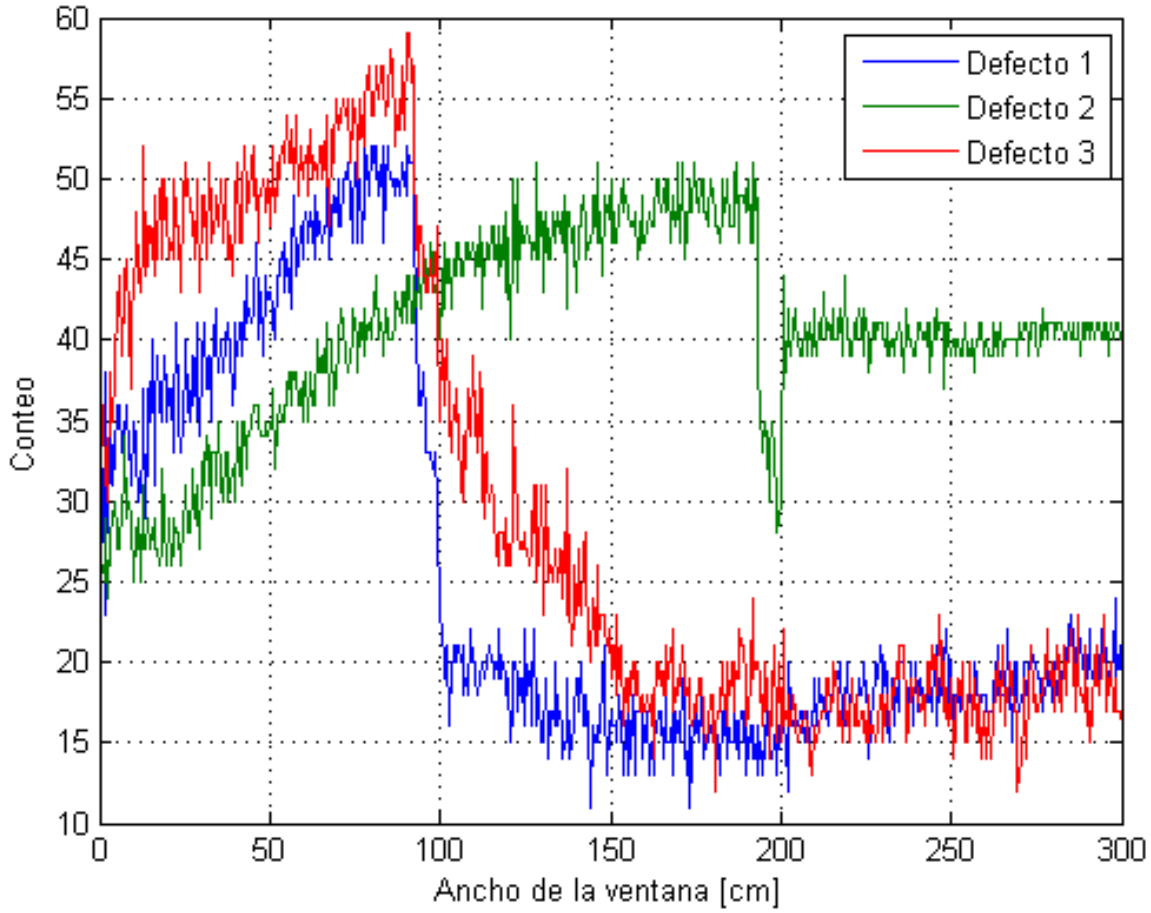

Figura 41. Conteo de los p-Value menores a 0.05 para cada ancho de ventana en todos los defectos.

Por lo tanto el valor seleccionado para realizar el ventaneo es de 181 muestras (90.5 cm). El ventaneo con los estadísticos y el ancho de ventana seleccionado mostrando cada defecto (en rojo) para la señal 11 de FFM se muestra en la Figura 42, Figura 43 y Figura 44.

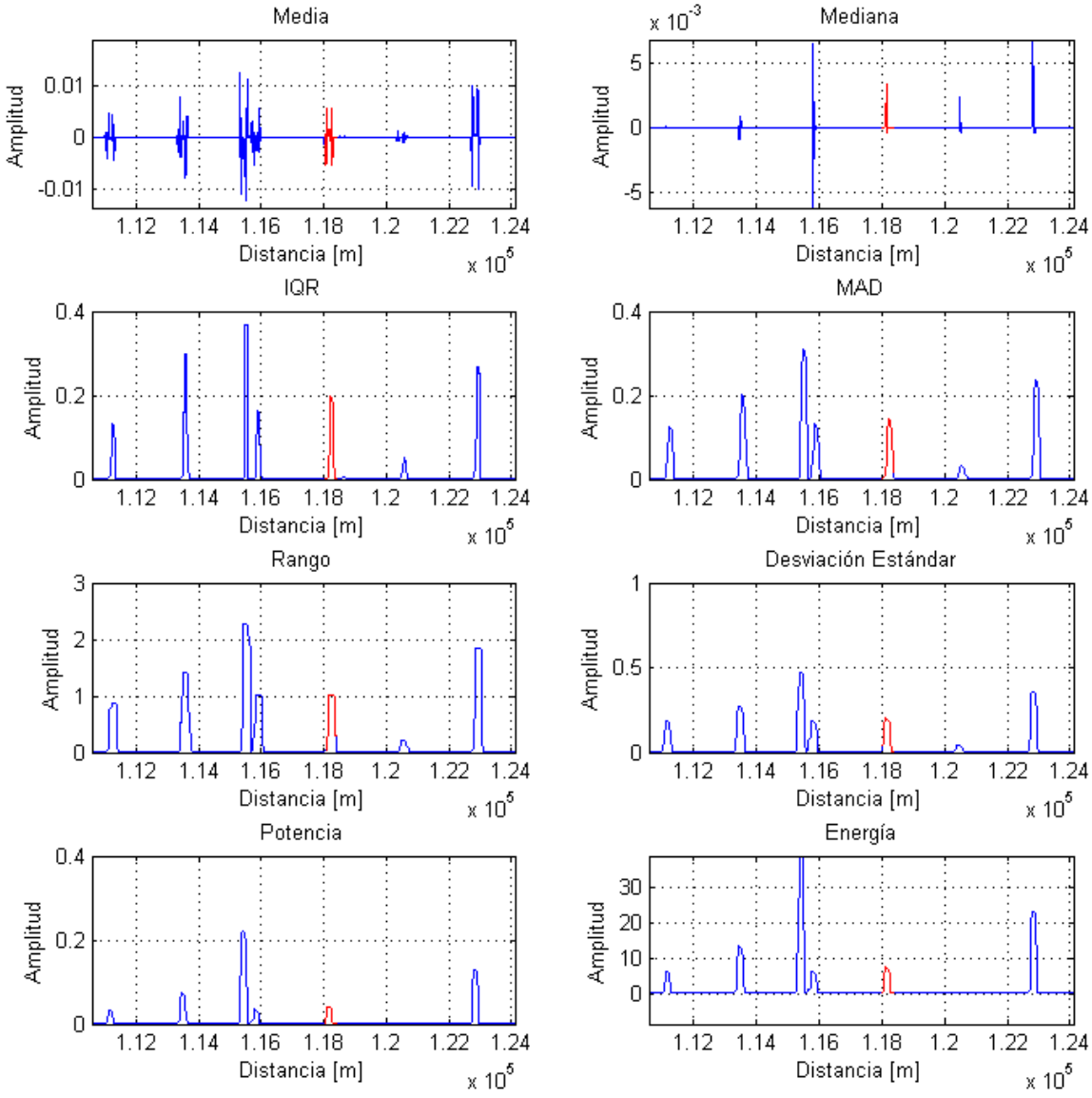

Figura 42. Ventaneo calculando los estadísticos (Defecto 1).

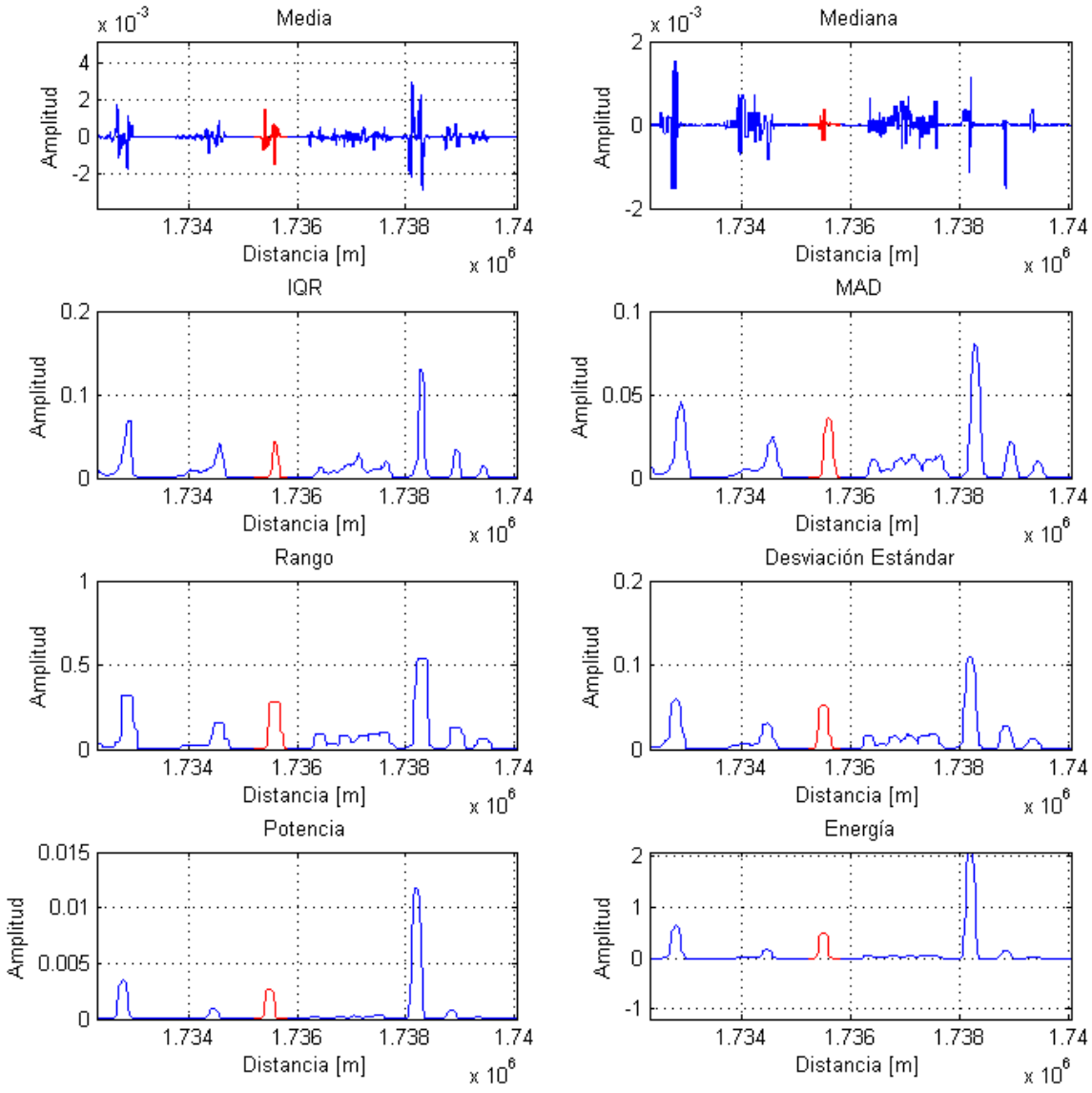

Figura 43. Ventaneo calculando los estadísticos (Defecto 2).

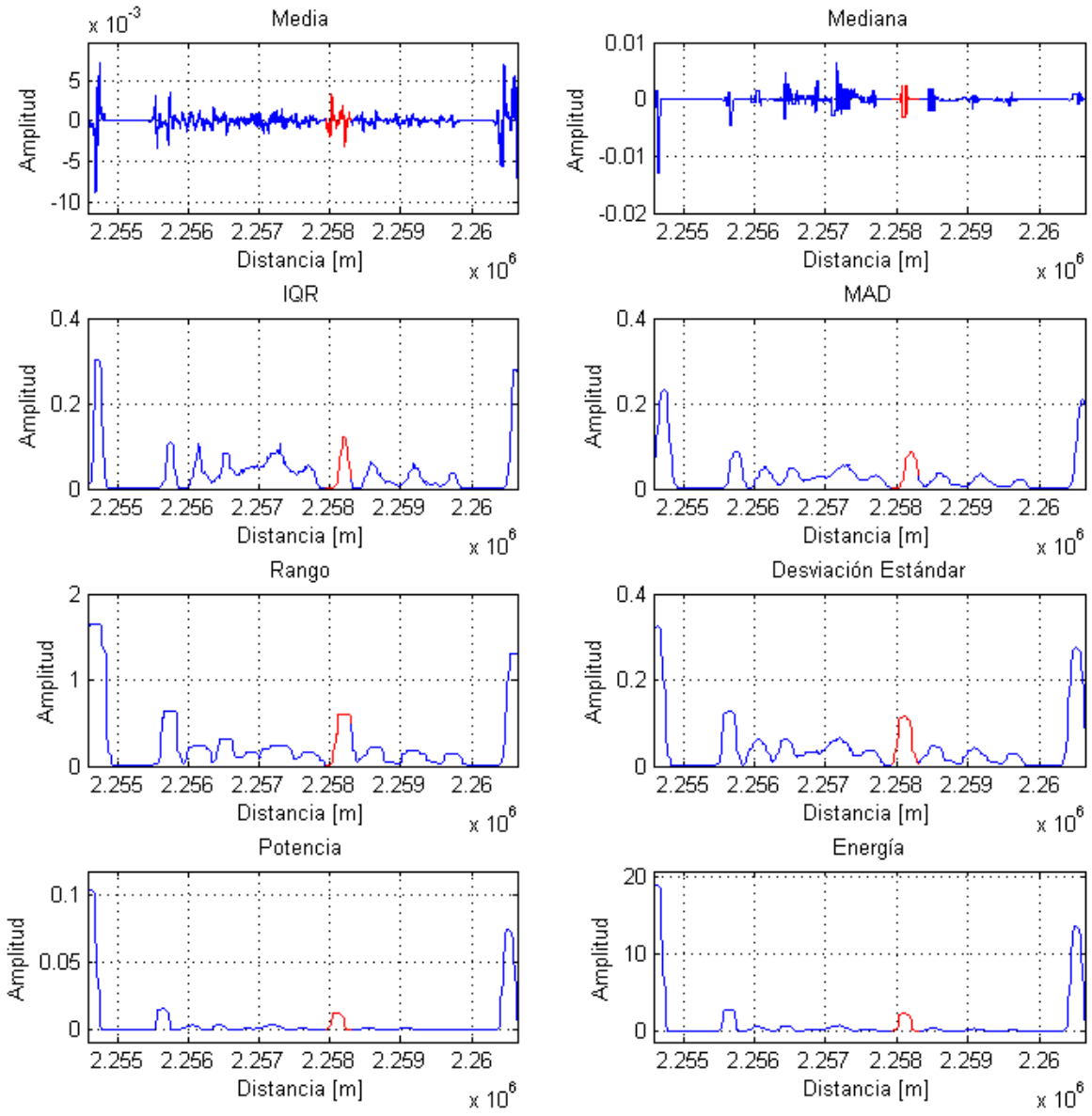

Figura 44. Ventaneo calculando los estadísticos (Defecto 3).

7. DISEÑO DE CLASIFICADORES

En este proyecto se planteó el uso de dos tipos de clasificadores diferentes: ADL y MSV (con diferentes *kernels*), y con dos metodologías de clasificación: balanceada y desbalanceada. . Para los clasificadores con datos balanceados se entrenó cada clasificador con las muestras que correspondían a defectos (marcados como “1”) y un randómico de datos que correspondían a no defecto (marcados como “0”), en total el número de muestras usadas para entrenar los clasificadores fue de 1938.

En total se repitió el proceso 100 veces para cada clasificador. Para el clasificador ADL con datos desbalanceados se entrenó cada clasificador con todos los datos de cada estadístico, y se clasificó cada estadístico de cada señal de manera independiente. Cada clasificador tiene como salida un ”1” si determina que dicho dato es un defecto o un “0” si considera que no es defecto.

7.1. Metodología de Validación

Se utilizó el método de Validación Cruzada con K=4 iteraciones para asegurar que el modelo posea un nivel de generalización aceptable y así garantizar el porcentaje de clasificación.

En se realizó una aleatorización de los datos de las cuatro señales tomadas a diferentes velocidades para luego dividirlos en 4 grupos (cajas) diferentes. Como se puede observar en la Figura 45.

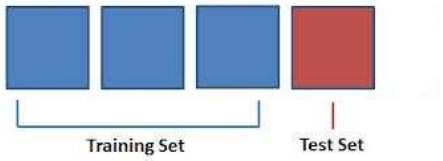

Figura 45. Grupos de datos para entrenamiento y test.

Luego se determinaron los grupos de datos que se utilizaran para el entrenamiento, prueba de los clasificadores. En este caso, las primeras tres cajas se utilizan para el entrenamiento, la caja número cuatro se utiliza para probar el clasificador y obtener el porcentaje de clasificación. Cuando todas las iteraciones terminen, todas las cajas habrán sido para entrenamiento y prueba.

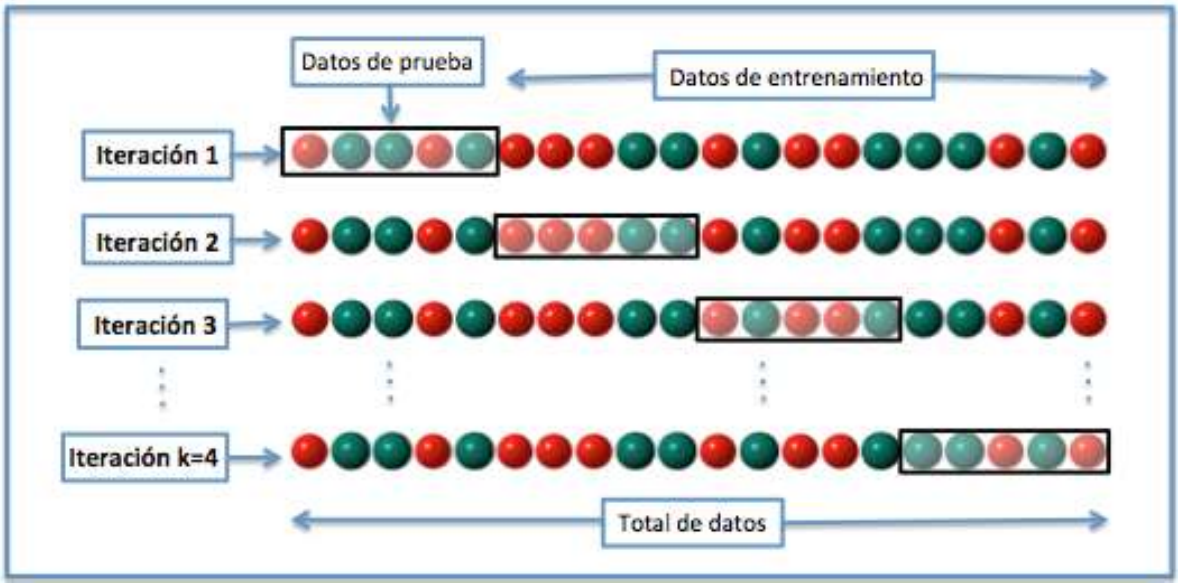

Figura 46. Ejemplo de validación cruzada.

Luego de realizar este proceso, se debe volver a aleatorizar los datos para repetir el procedimiento nuevamente. Se repitió 100 veces, obteniendo 100 porcentajes de clasificación utilizando diferentes datos de prueba.

El porcentaje de clasificación se obtuvo a partir de la ecuación de precisión balanceada (Ecuación 8) [10].

$$Precision\ balanceada = \frac{1}{2} * \frac{Cp}{Cp + Fp} + \frac{1}{2} * \frac{Cn}{Cn + Fn} \quad (8)$$

Donde:

- Cp correcto positivo
- Fp falso positivo
- Cn Correcto negativo
- Fn Falso negativo

7.2. Análisis Discriminante Lineal

Es una técnica lineal de clasificación de datos en la que se supone la existencia de dos o más clases, las cuales siguen una distribución gaussiana, donde la mayoría de los datos se concentran alrededor de la media. El objetivo de ADL es encontrar una función discriminante la cual maximice la separación entre ambas medias de las distribuciones gaussianas supuestas [11]. La función de ADL es de la forma:

$$Y = a0 + a1 * x1 + a2 * x2 + \cdots + ap * xp \quad (9)$$

Se entrenó un clasificador independiente para cada estadístico calculado, esto con el fin de realizar una reducción de la dimensionalidad más adelante.

Como se dijo anteriormente, se entrenaron 2 tipos de clasificadores basados en ADL, uno con los datos balanceados, y otro con los datos desbalanceados. Para realizar el entrenamiento con el método balanceado, se utilizó el método de validación descrito en 7.1.

7.2.1. ADL Balanceado

Como se dijo anteriormente, se entrenaron dos tipos de clasificadores, uno con los datos balanceados y otro con los datos desbalanceados. Para realizar el entrenamiento de los clasificadores, se utilizó la función de Matlab® R2014a "`fitcdiscr`" para entrenar el clasificador y la función "`predict`" para predecir las clases de las nuevas muestras. Las funciones discriminantes para cada estadístico se encuentran a continuación:

$$Y_{media} = -12.5213 * x_1 + 1.8899 * 10^{-4} \quad (10)$$

$$Y_{mediana} = 278.6866 * x_2 + 0.0135 \quad (11)$$

$$Y_{IQR} = -21.8083 * x_3 + 0.3991 \quad (12)$$

$$Y_{DMA} = -19.5763 * x_4 + 0.4668 \quad (13)$$

$$Y_{rango} = -0.9985 * x_5 + 0.2653 \quad (14)$$

$$Y_{std} = -1.6864 * x_6 + 0.0932 \quad (15)$$

$$Y_{energía} = -96.5964 * x_7 + 0.3852 \quad (16)$$

Donde:

- $Y_{media}, Y_{mediana}, Y_{IQR}, Y_{DMA}, Y_{rango}, Y_{std}$ y $Y_{energía}$ son la salida de los clasificadores con datos balanceados.
- $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ son las entradas de los clasificadores.

Los porcentajes de clasificación de los diferentes estadísticos para todas las señales junto con su desviación estándar se muestran en la Figura 47 (Defectos) y en la Figura 48 (No Defectos).

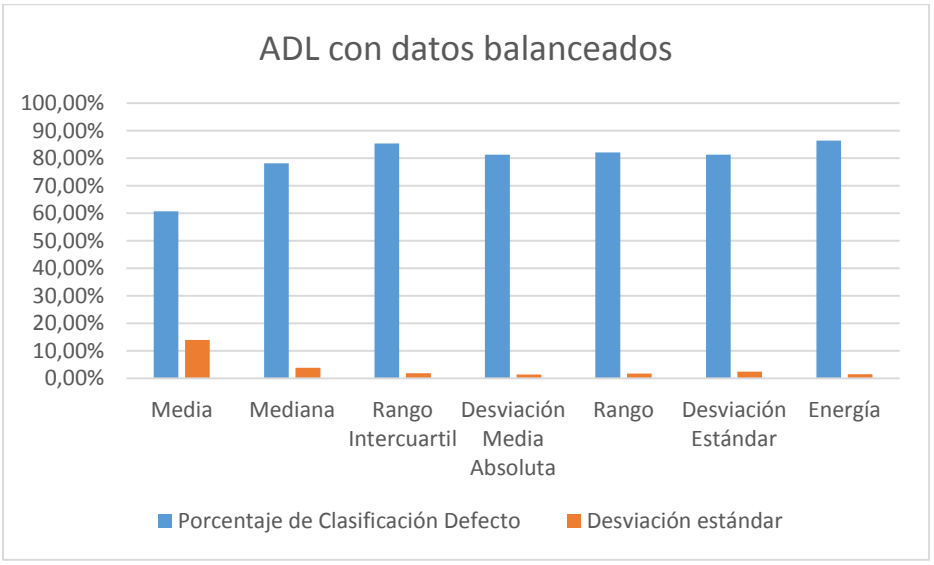

Figura 47. Porcentaje de clasificación ADL con datos balanceados (Defectos).

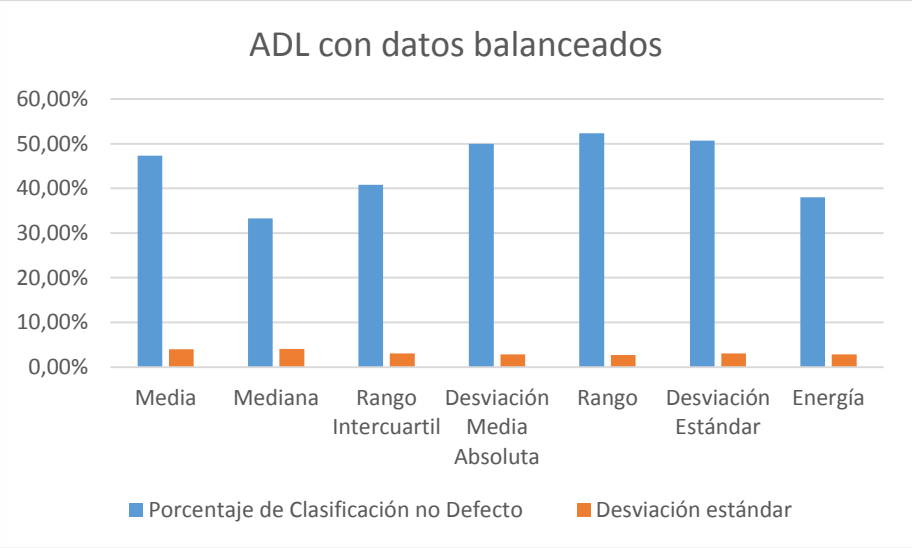

Figura 48. Porcentaje de clasificación ADL con datos balanceados (No Defectos).

El resultado del clasificador, visto en la Señal 11 de FFM se puede observar en la Figura 49 (Defecto 1), la Figura 50 (Defecto 2) y la Figura 51 (Defecto 3).

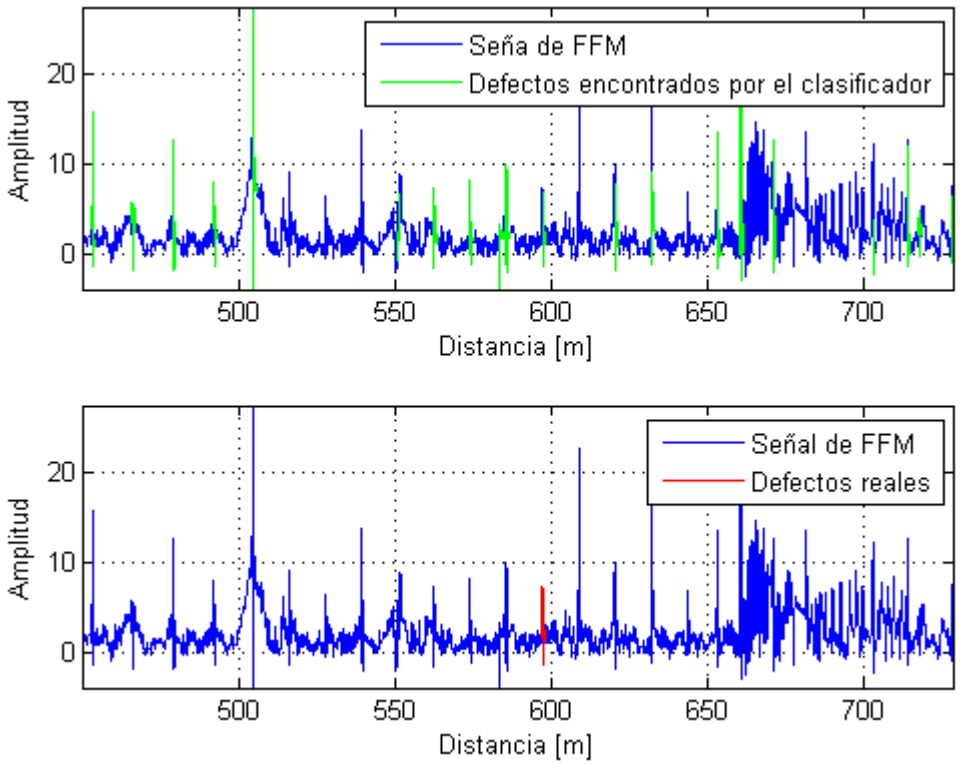

Figura 49. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados (Defecto 1).

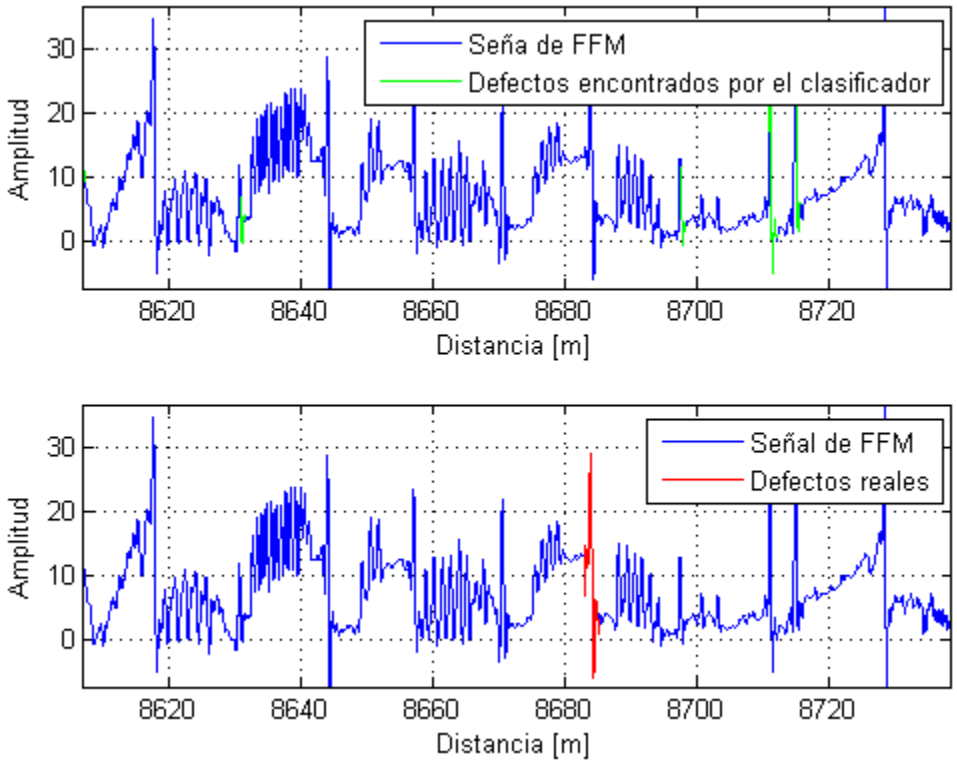

Figura 50. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados (Defecto 2).

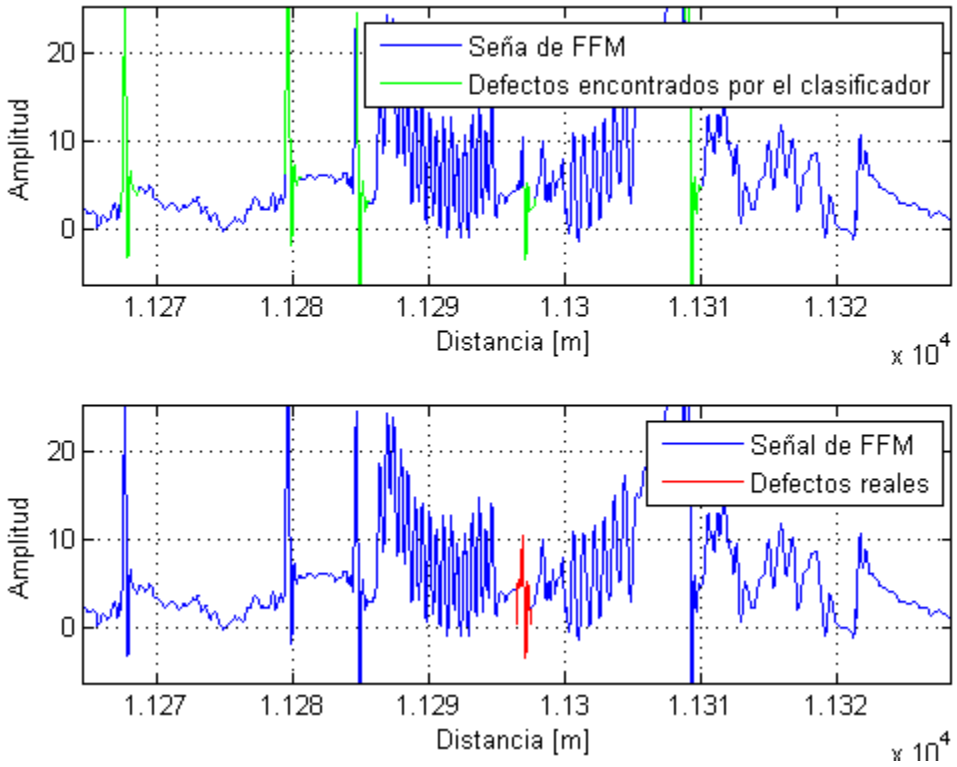

Figura 51. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados (Defecto 3).

7.2.2. ADL desbalanceado

Las funciones discriminantes para cada estadístico se encuentran a continuación:

$$Y_{media} = 0.0142 * x_1 + 3.5727 * 10^{-9} \quad (17)$$

$$Y_{mediana} = -2.2829 * x_2 + 0.0025 \quad (18)$$

$$Y_{IQR} = -0.4752 * x_3 + 0.0422 \quad (19)$$

$$Y_{DMA} = -0.8464 * x_4 + 0.0527 \quad (20)$$

$$Y_{rango} = -0.1099 * x_5 + 0.0501 \quad (21)$$

$$Y_{std} = -0.6463 * x_6 + 0.0584 \quad (22)$$

$$Y_{energía} = 0.0027 * x_7 - 1.9532 * 10^{-4} \quad (23)$$

Donde:

- $Y_{media}, Y_{mediana}, Y_{IQR}, Y_{DMA}, Y_{rango}, Y_{std}$ y $Y_{energía}$ son la salida de los clasificadores con datos balanceados.
- $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ son las entradas de los clasificadores.

Los porcentajes de clasificación de los diferentes estadísticos para todas las señales junto con su desviación estándar se muestran en la Figura 52 (Defectos) y en la Figura 53 (No Defectos).

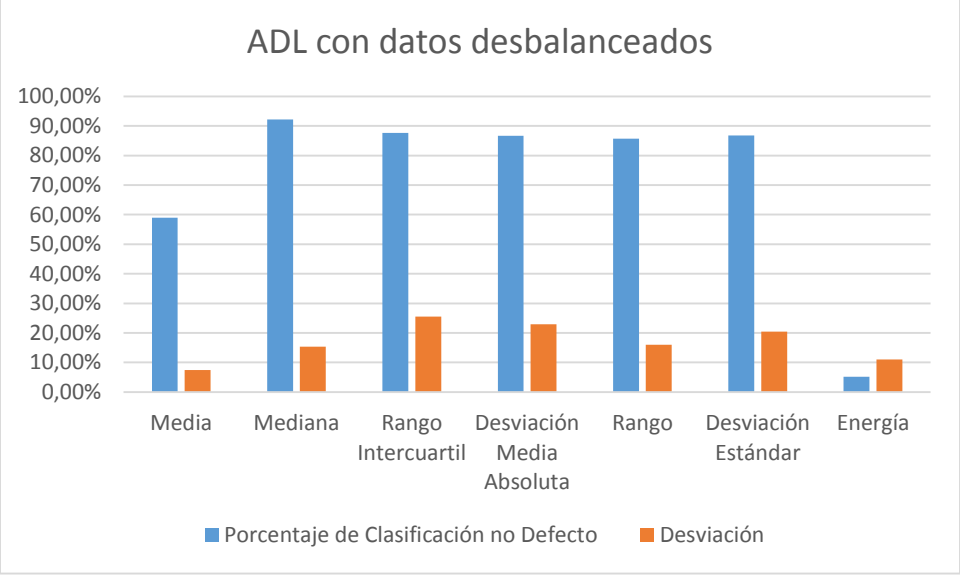

Figura 52. Porcentaje de clasificación ADL con datos desbalanceados (Defectos).

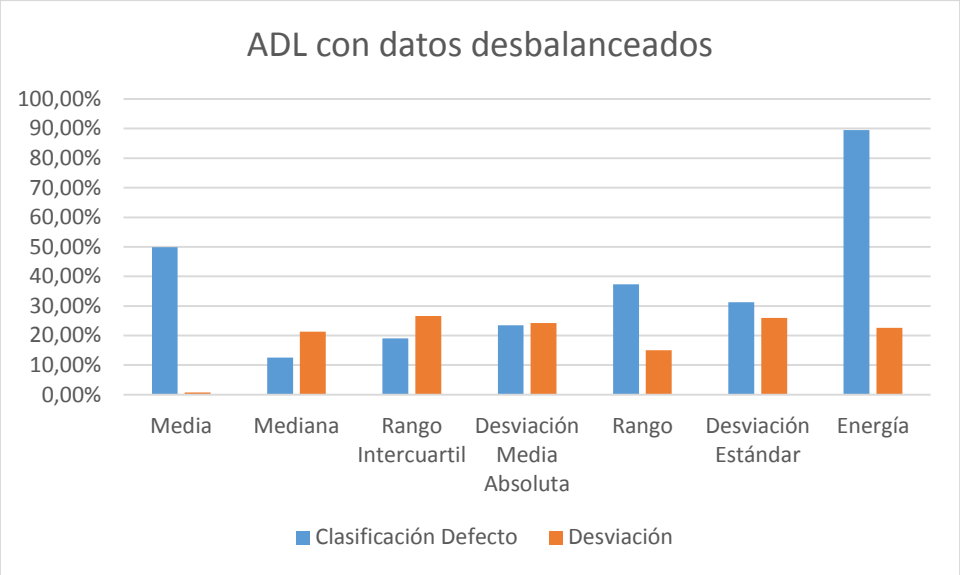

Figura 53. Porcentaje de clasificación ADL con datos desbalanceados (No Defectos).

El resultado del clasificador, visto en la Señal 11 de FFM se puede observar en la Figura 54 (Defecto 1), la Figura 55 (Defecto 2) y la Figura 56 (Defecto 3).

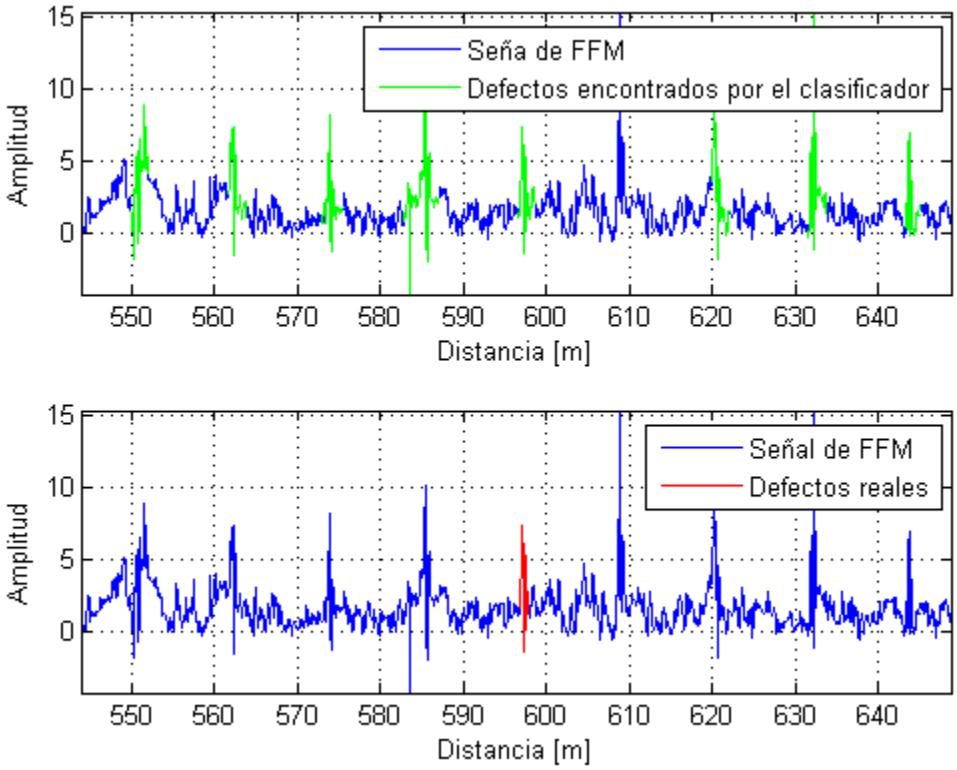

Figura 54. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados (Defecto 1).

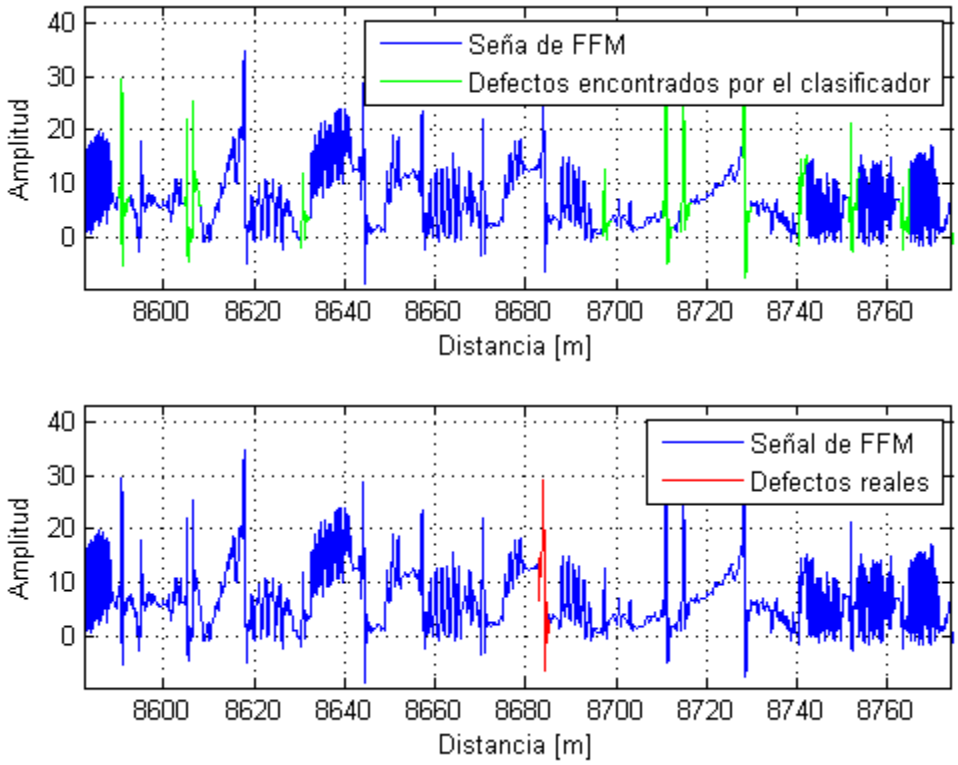

Figura 55. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados (Defecto 2).

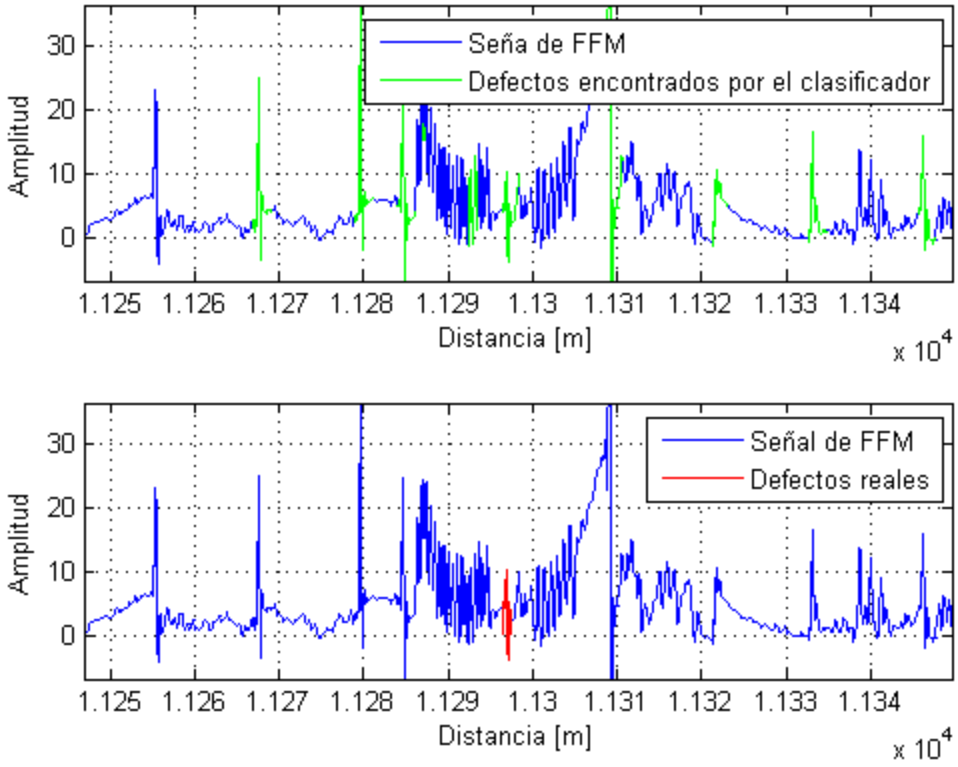

Figura 56. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados (Defecto 3).

7.3. Máquinas de Soporte Vectorial

Una MSV construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación correcta [12].

Dado un conjunto de ejemplos de entrenamiento (de muestras) podemos etiquetar las clases y entrenar una MSV para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una MSV es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase [13].

7.3.1. Funciones Kernel

La manera más simple de realizar la separación es mediante una línea recta, un plano recto o un hiperplano N-dimensional. Desafortunadamente los universos a estudiar no se suelen presentar en casos ideales, sino que una MSV debe tratar con:

- Más de dos variables predictoras.
- Curvas no lineales de separación.
- Casos donde los conjuntos de datos no pueden ser completamente separados.
- Clasificaciones en más de dos categorías.

Debido a las limitaciones computacionales de las máquinas de aprendizaje lineal, éstas no pueden ser utilizadas en la mayoría de las aplicaciones del mundo real. La representación por medio de funciones Kernel ofrece una solución a este problema, proyectando la información a un espacio de características de mayor dimensión el cual aumenta la capacidad computacional de las máquinas de aprendizaje lineal.

Es decir, mapearemos el espacio de entradas a un nuevo espacio de características de mayor dimensionalidad [14].

7.3.1.1. Función Polinomial

El *Kernel* es una función comúnmente utilizada con MSV y otros modelos que utilicen *Kernels*, que representa la similitud de los vectores (muestras de entrenamiento) en un espacio de características sobre polinomios de las variables originales, lo que permite el aprendizaje de modelos no lineales, está dado por la siguiente ecuación:

$$K(xi, xj) = (xi \cdot xj)^n \quad (10)$$

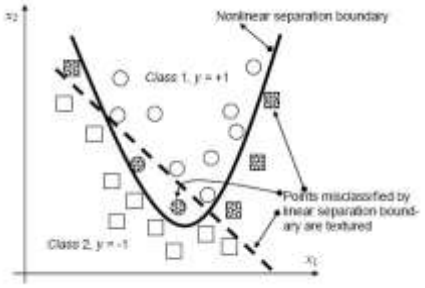

Figura 57. Separación polinomial de dos grupos de datos.

7.3.1.2. Función de base radial Gaussiana

Este *Kernel* toma la forma de una función que mapea un único vector a un vector de mayor dimensionalidad, que se aproxima al *Kernel*:

$$K(xi, xj) = \exp\left(-\frac{(xi-xj)^2}{2(sigma)^2}\right) \quad (11)$$

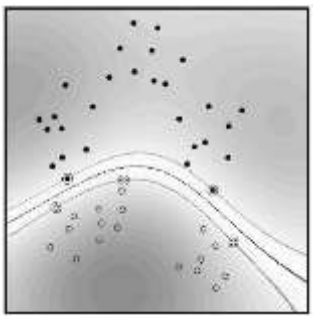

Figura 58. Separación gaussiana de dos grupos de datos.

7.3.2. MSV con Kernel Gaussiano

Para realizar el entrenamiento de los clasificadores, se utilizó la función de Matlab® R2014a "`fitcsvm`" con `'KernelFunction' = 'gaussian'` para entrenar el clasificador y la función "`predict`" para predecir las clases de las nuevas muestras. Los porcentajes de clasificación de los diferentes estadísticos para todas las señales junto con su desviación estándar se muestran en la Figura 59 (Defectos) y en la Figura 60 (No Defectos).

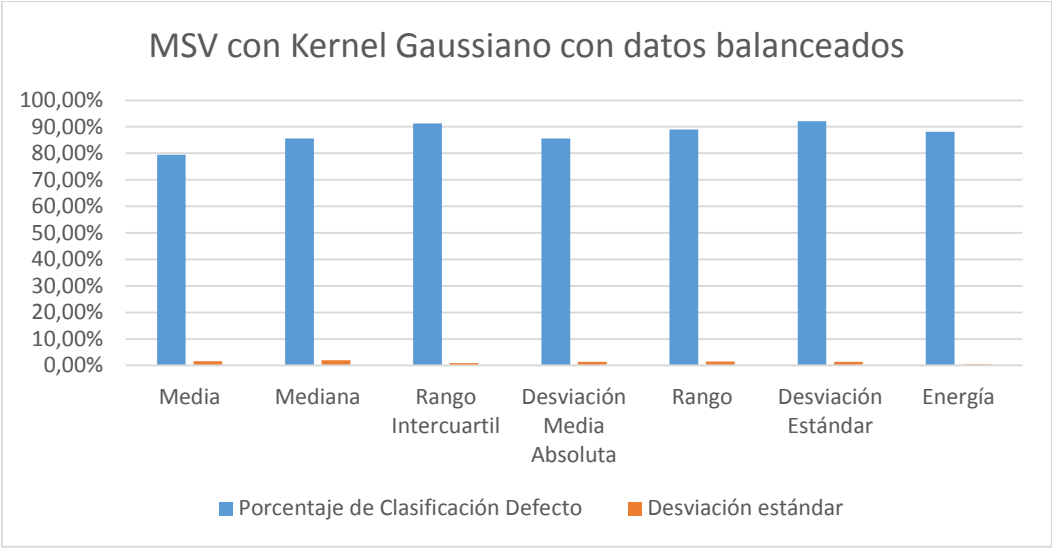

Figura 59. Porcentajes de Clasificación MSV con Kernel Gaussiano (Defecto).

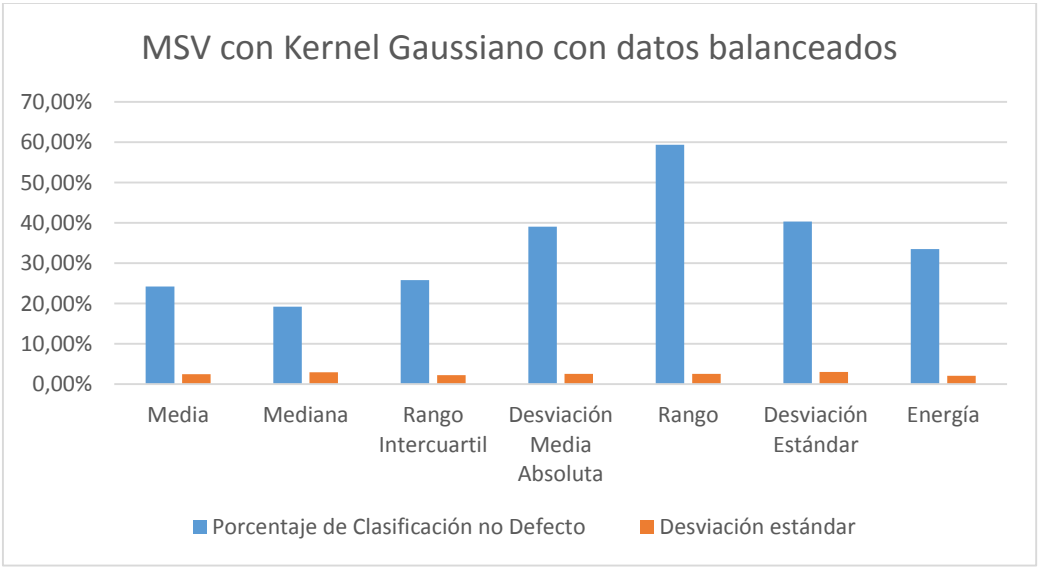

Figura 60. Porcentajes de Clasificación MSV con Kernel Gaussiano (No Defecto).

El resultado del clasificador, visto en la Señal 11 de FFM se puede observar en la Figura 61 (Defecto 1), la Figura 62 (Defecto 2) y la Figura 63 (Defecto 3).

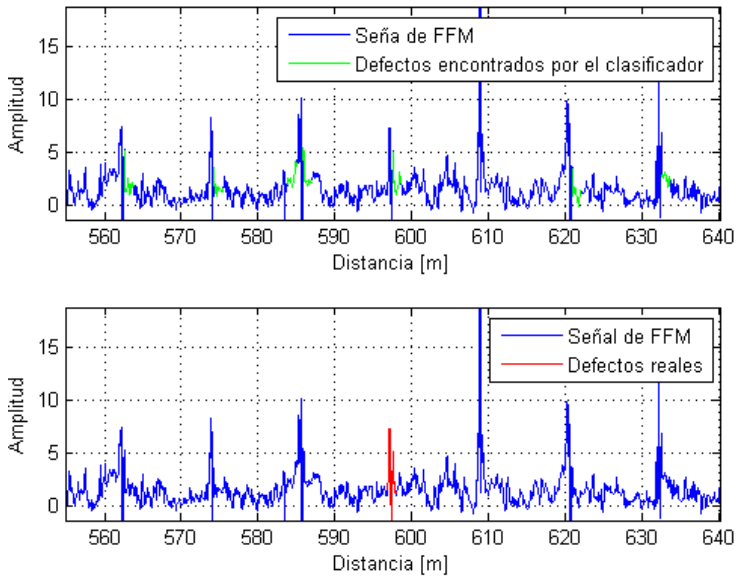

Figura 61. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Gaussiano (Defecto 1).

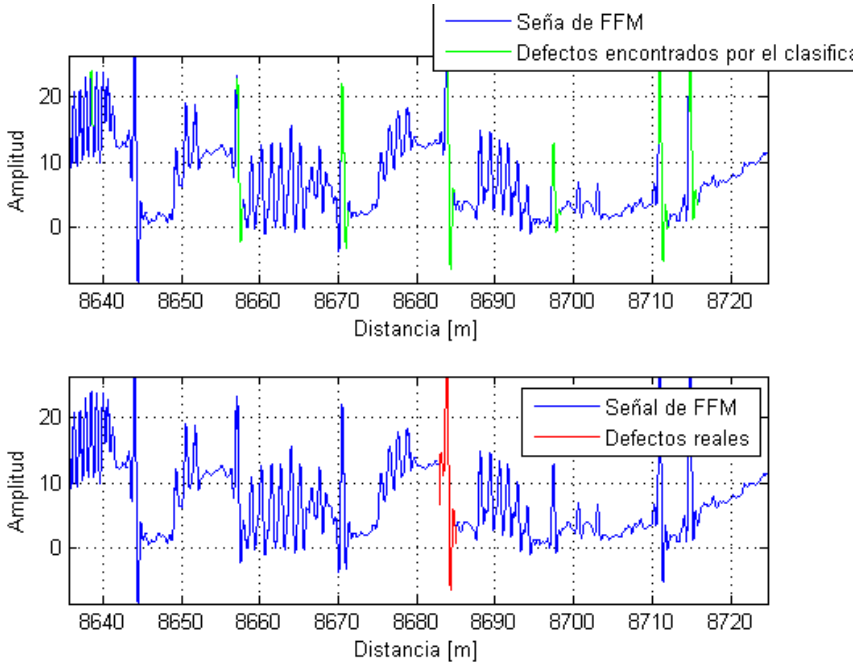

Figura 62. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Gaussiano (Defecto 2).

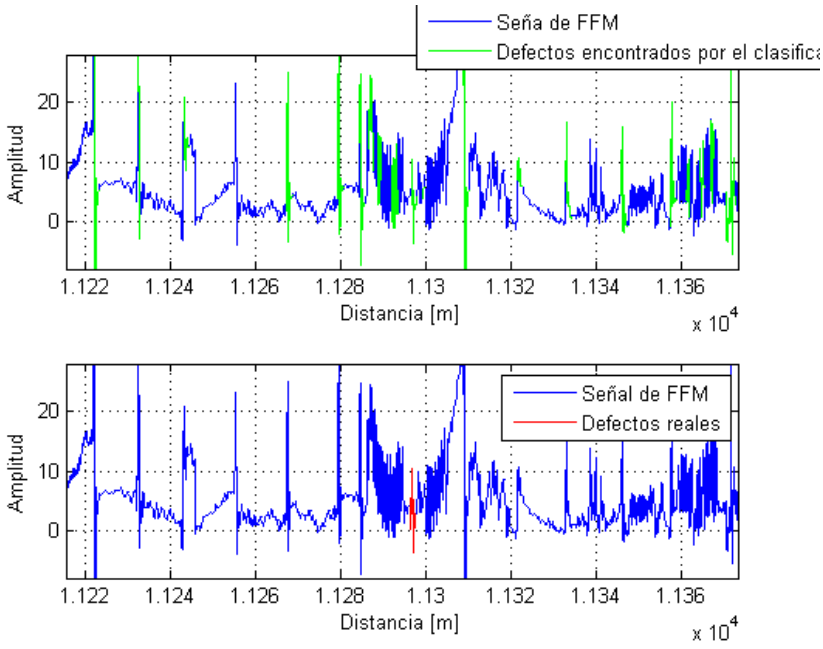

Figura 63. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Gaussiano (Defecto 3).

7.3.3. MSV con Kernel Polinomial

Para realizar el entrenamiento de los clasificadores, se utilizó la función de Matlab® R2014a "`fitcsvm`" con `'KernelFunction' = 'polynomial'` para entrenar el clasificador y la función "`predict`" para predecir las clases de las nuevas muestras. Los porcentajes de clasificación de los diferentes estadísticos para todas las señales junto con su desviación estándar se muestran en la Figura 64Figura 47 (Defectos) y en la Figura 65 (No Defectos).

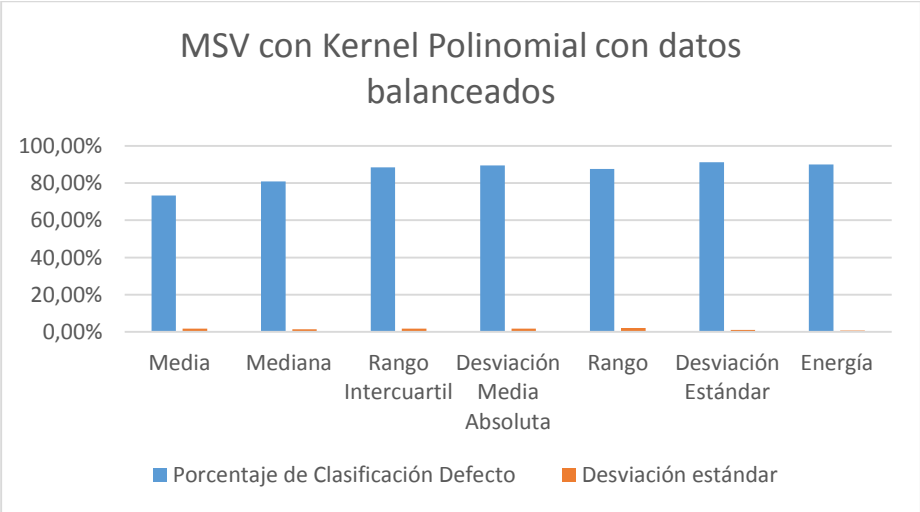

Figura 64. Porcentajes de Clasificación MSV con Kernel Polinomial (Defecto).

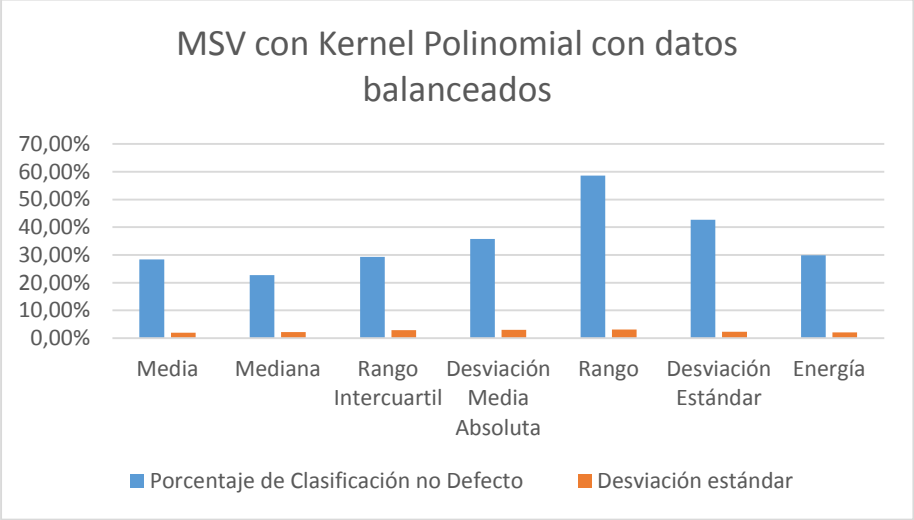

Figura 65. Porcentajes de Clasificación MSV con Kernel Polinomial (No Defecto).

El resultado del clasificador, visto en la Señal 11 de FFM se puede observar en la Figura 66 (Defecto 1), la Figura 67 (Defecto 2) y la Figura 68 (Defecto 3).

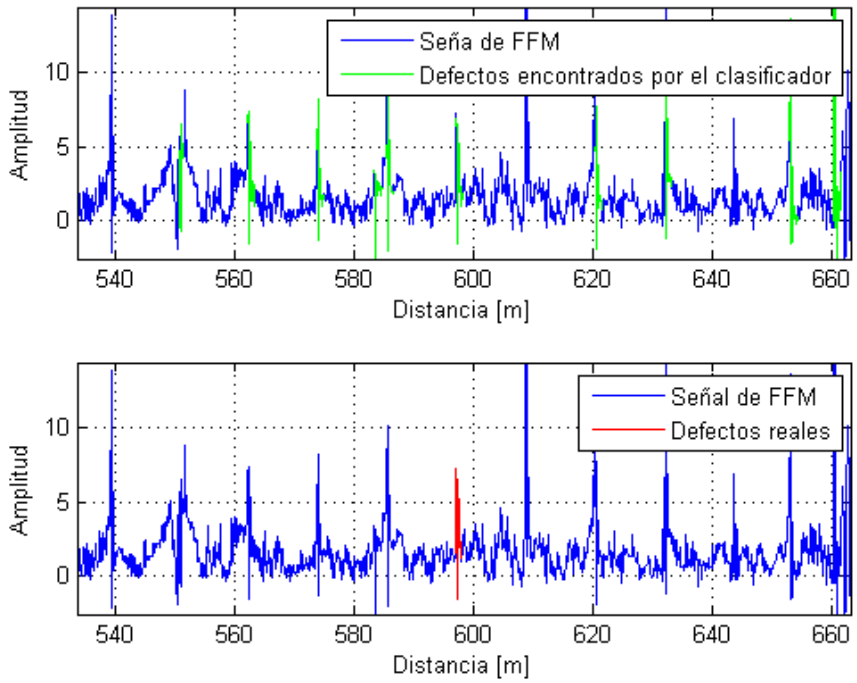

Figura 66. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Polinomial (Defecto 1).

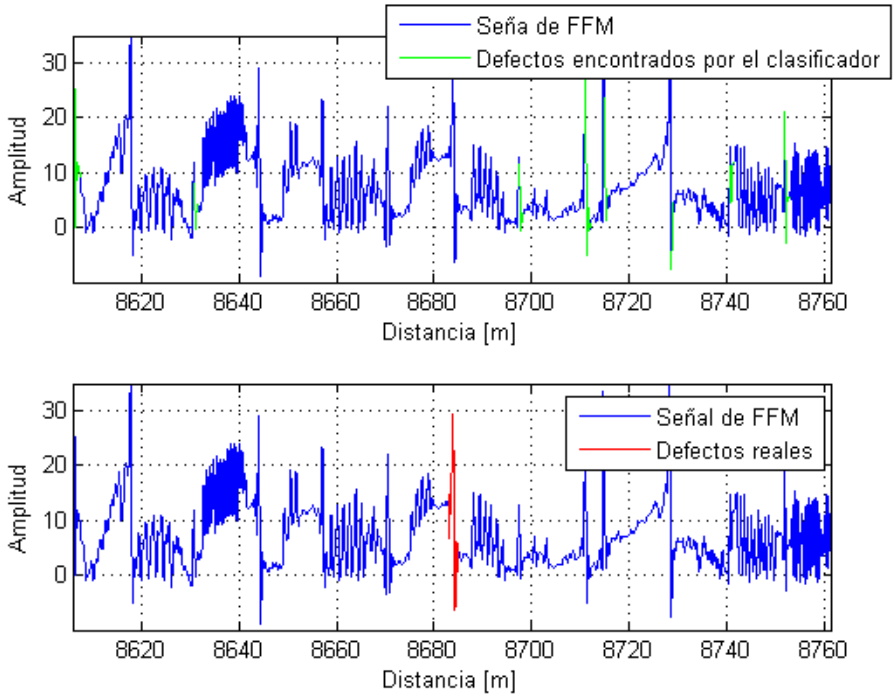

Figura 67. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Polinomial (Defecto 2).

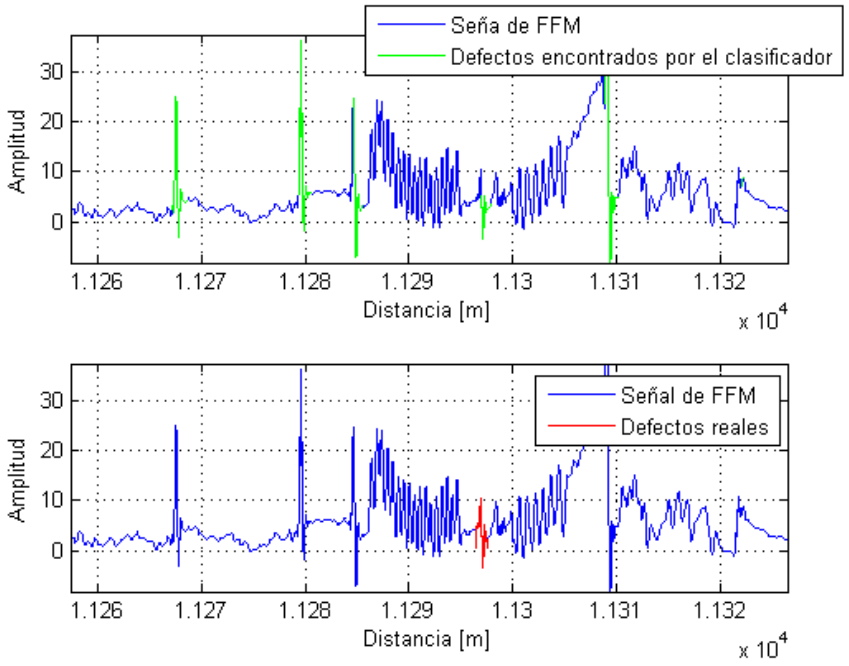

Figura 68. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Polinomial (Defecto 3).

7.3.4. MSV con Kernel Lineal

Para realizar el entrenamiento de los clasificadores, se utilizó la función de Matlab® R2014a "`fitcsvm`" con `'KernelFunction' = 'linear'` para entrenar el clasificador y la función "`predict`" para predecir las clases de las nuevas muestras. Los porcentajes de clasificación de los diferentes estadísticos para todas las señales junto con su desviación estándar se muestran en la Figura 69Figura 47 (Defectos) y en la Figura 70 (No Defectos).

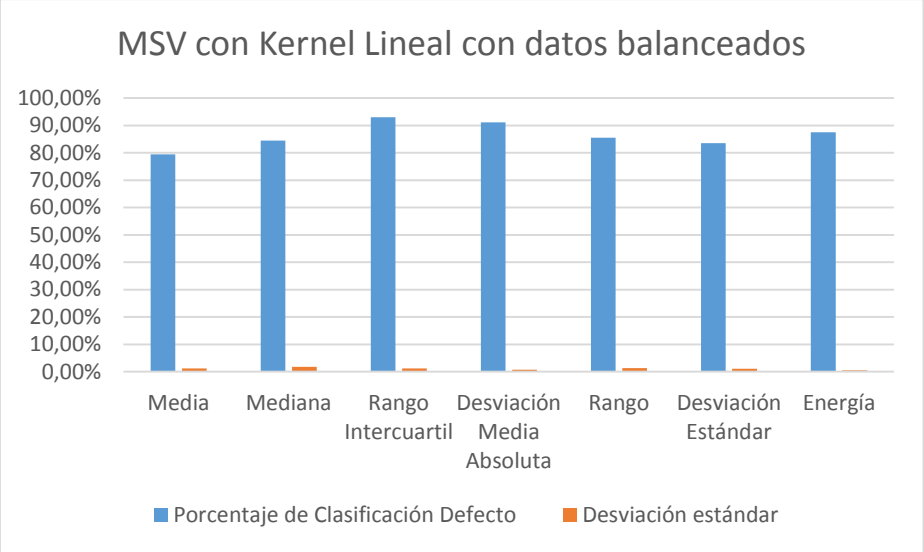

Figura 69. Porcentajes de Clasificación MSV con Kernel Lineal (Defecto).

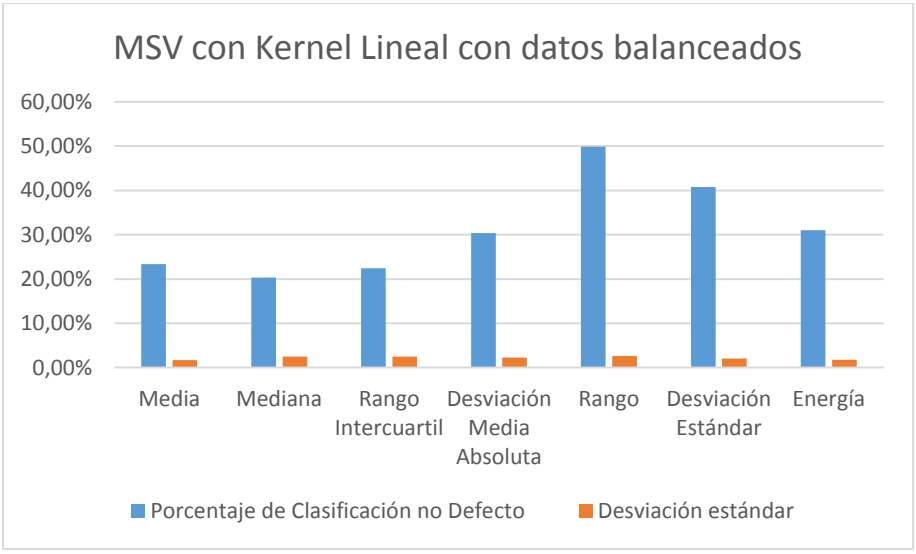

Figura 70. Porcentajes de Clasificación MSV con Kernel Lineal (No Defecto).

El resultado del clasificador, visto en la Señal 11 de FFM se puede observar en la Figura 71 (Defecto 1), la Figura 72 (Defecto 2) y la Figura 73 (Defecto 3).

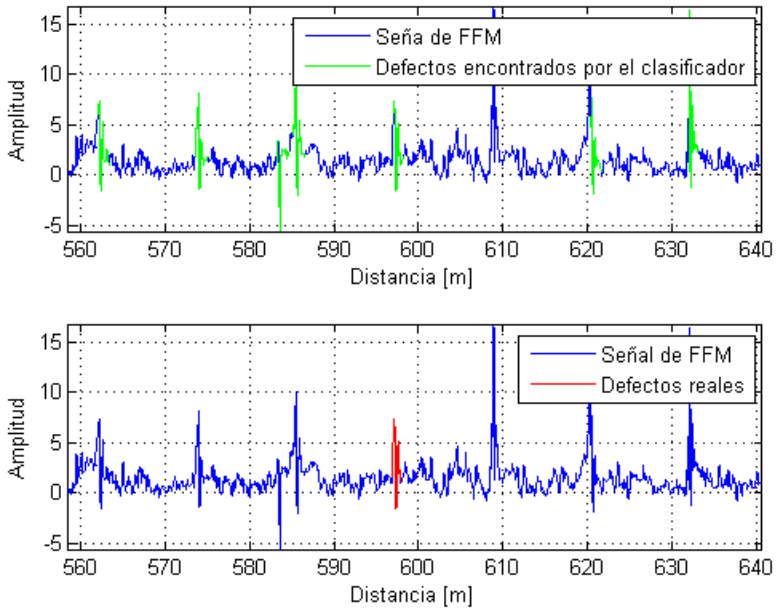

Figura 71. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Lineal (Defecto 1).

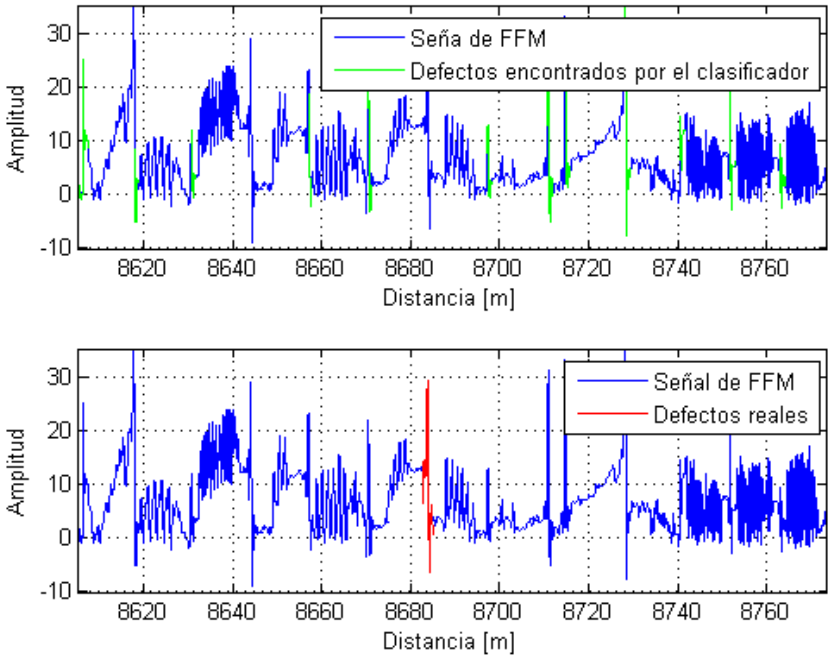

Figura 72. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Lineal (Defecto 2).

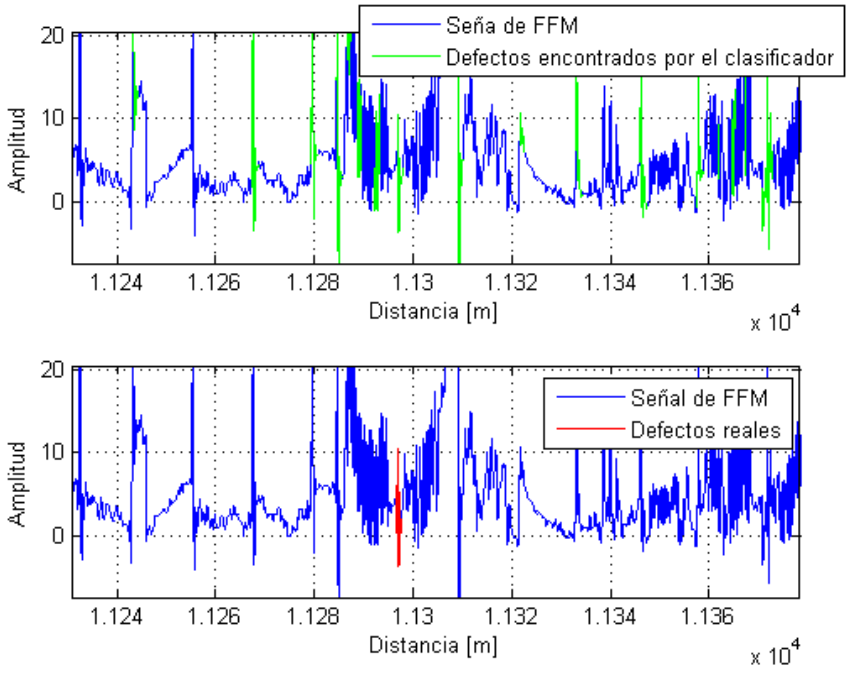

Figura 73. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Lineal (Defecto 3).

8. CORRECCIÓN DE FALSOS POSITIVOS

Los resultados de los clasificadores incluyeron “no defectos” en su clasificación como “defecto”, la razón por la que esto ocurre es que existen patrones similares a los defectos que se repiten a lo largo de la señal. Para eliminar los falsos positivos y aumentar la eficacia de los clasificadores, se propone realizar un ventaneo por los defectos, con un ancho variable (T) de 2 muestras al número total de muestras de cada defecto, para determinar qué porcentajes de “1” se encuentran en cada defecto variando el ancho de la ventana, (teniendo en cuenta que en algunos casos los defectos fueron clasificados de manera intermitente). Luego se selecciona un ancho de ventana óptimo la cual irá recorriendo toda la señal calculando dicho porcentaje, y a este ventaneo se le aplicará un criterio de selección, (también se hará ventaneado (σ)), el cual decidirá si es un falso positivo o un posible defecto.

8.1. Ventaneo T del resultado de los clasificadores

En este punto se recorre cada defecto en las 8 señales de FFM, con el fin de determinar el porcentaje de “1” que existe en los mismos, y ver, además, cómo varía dicho porcentaje cuando varía el ancho de la ventana.

Se puede observar que en los 3 casos, cuando el ancho de ventana es igual a la mitad de las muestras totales del defecto, la mayoría de los estadísticos tienden a tener un mayor porcentaje de “1”. La Figura 74, la Figura 75 y la Figura 76, muestran los 3 porcentajes de “1” de los 3 defectos.

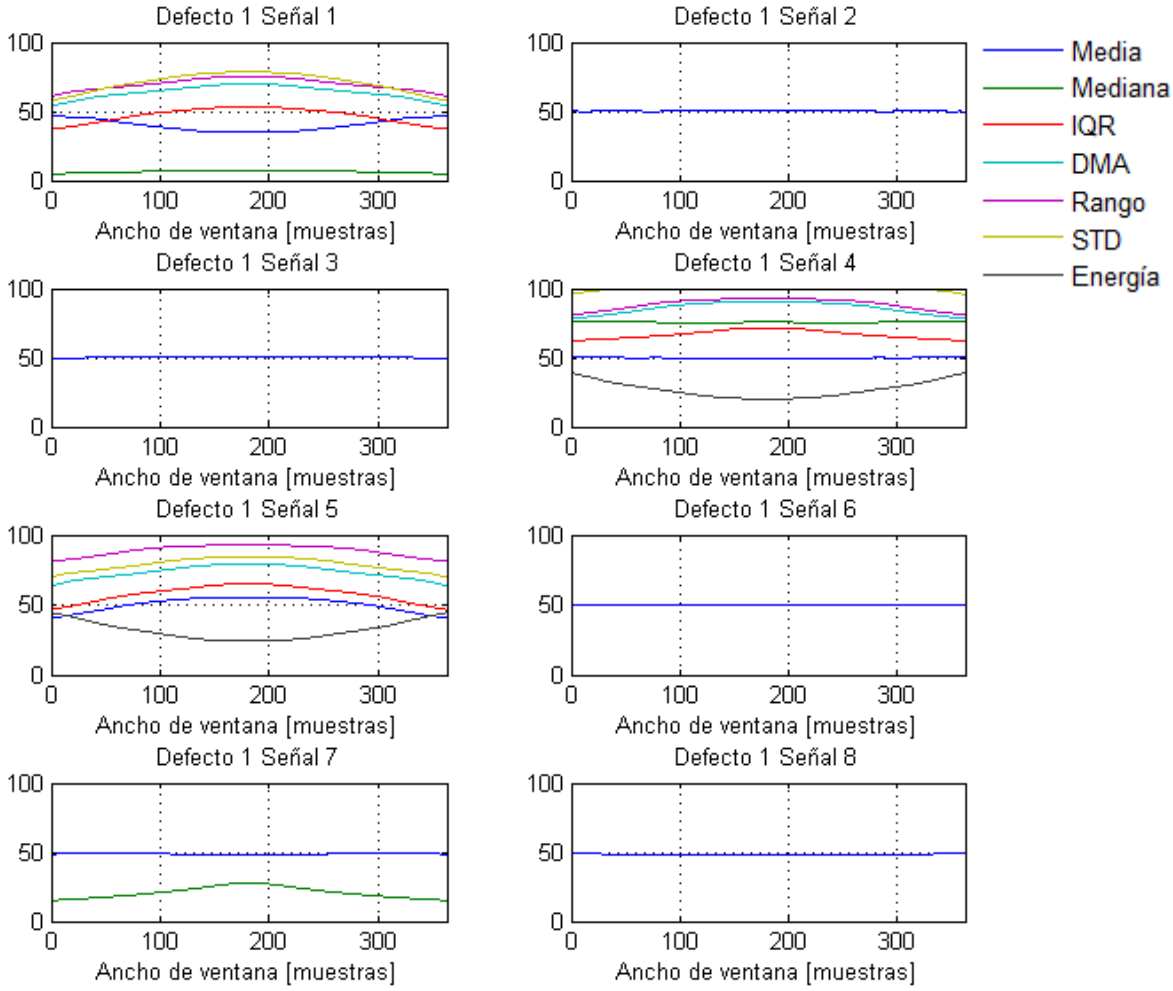

Figura 74. Porcentaje de "1" para el primer defecto, todas las señales (ADL con datos desbalanceados).

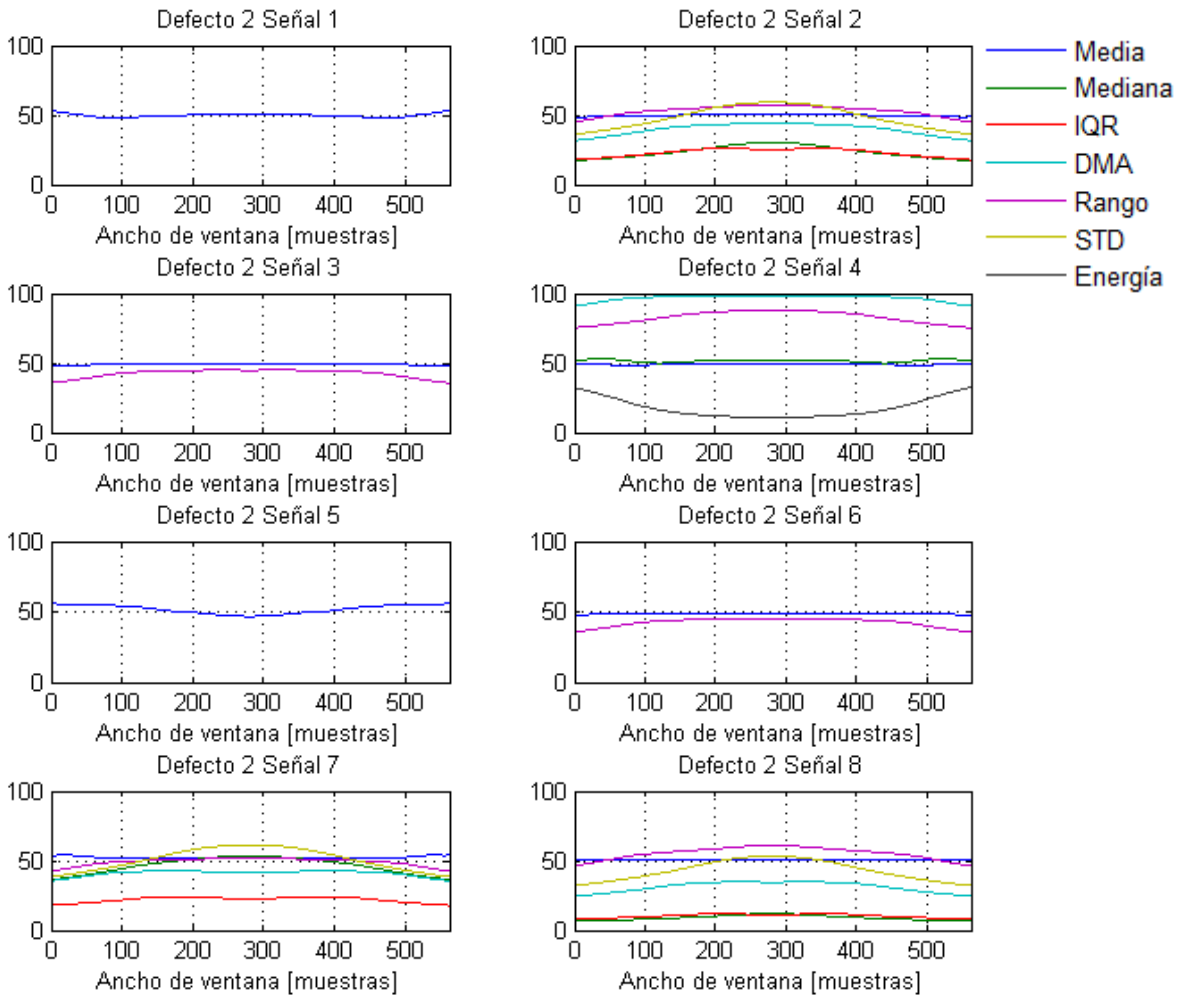

Figura 75. Porcentaje de "1" para el segundo defecto, todas las señales (ADL con datos desbalanceados).

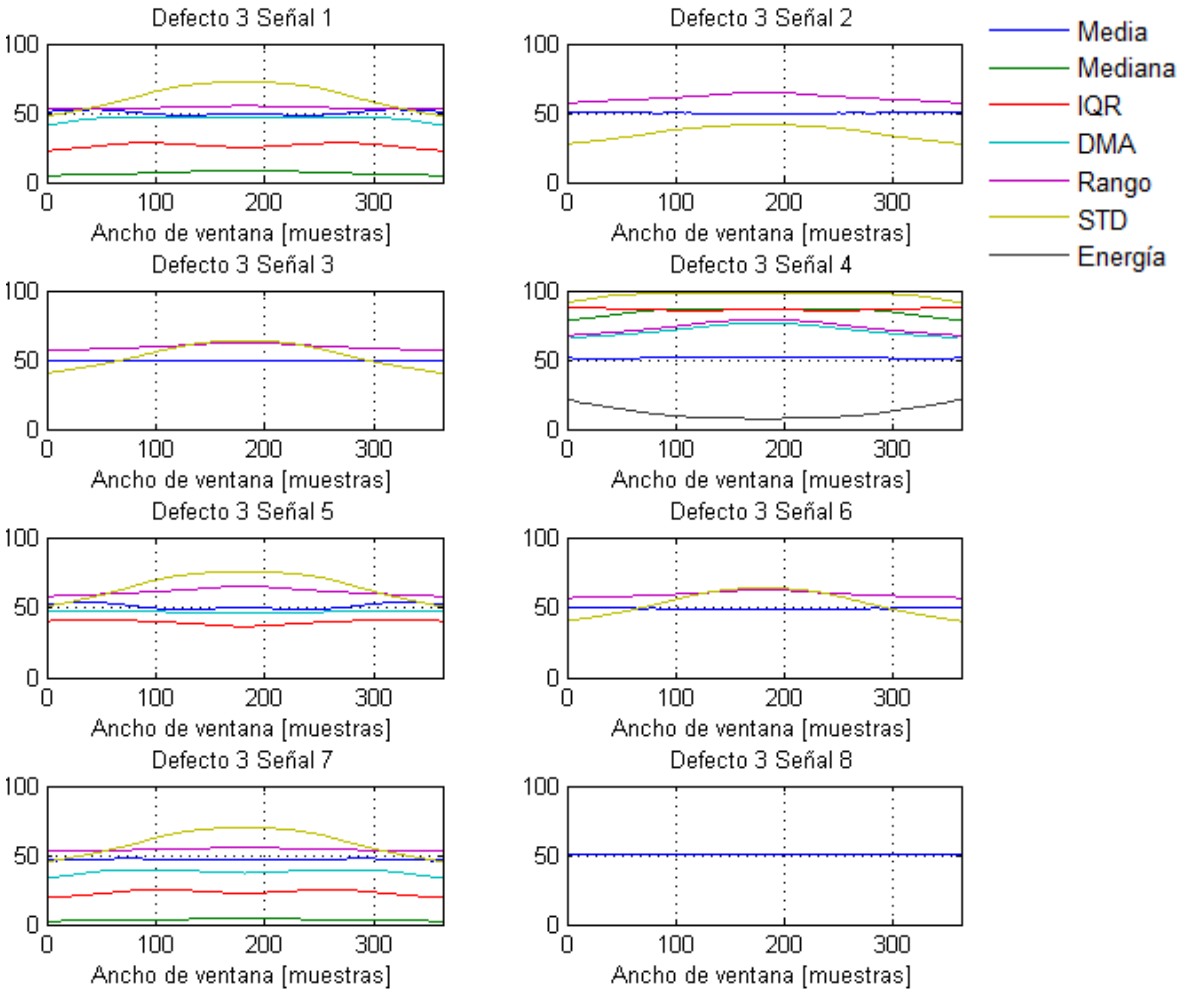

Figura 76. Porcentaje de "1" para el tercer defecto, todas las señales (ADL con datos desbalanceados).

Como se pudo observar, desde T=150 hasta T=200 el porcentaje de “1” tiende a ser estable en cada defecto, por lo cual se toma este rango para realizar el siguiente paso.

8.2. Ventaneo σ de los mejores T

Después de obtener el rango de anchos de ventana (T), se procede a realizar el ventaneo con dichos anchos de ventana, luego se calcula el promedio de los 51 ventaneos, uno de los ventaneos se muestra en la Figura 77.

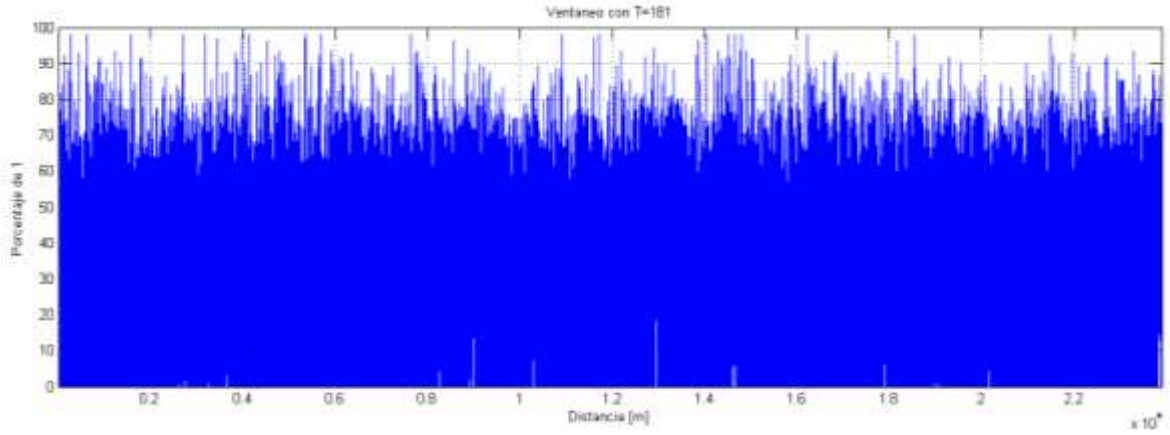

Figura 77. Promedio del ventaneo con los 51 T para el resultado del clasificador ADL, señal 11 de FFM.

Luego de esto, se procede a seleccionar el ancho de ventana (σ) con el cual se aplicará el criterio de selección, para esto se realiza la prueba de U de *Mann Whitney* con una ventana variable desde σ=2 hasta σ=364 (ancho del defecto mas pequeño), este rango se seleccionó teniendo en cuenta que era el rango que mayor porcentaje de "1" mostró en 8.1, y el principal objetivo era buscar separabilidad entre defectos y no defectos.

Nuevamente, se crea un segmento de datos, el cual corresponde a 50% defecto y 50% no defecto y se realiza la prueba, este proceso se repitió 50 veces con cada ancho de ventana, es decir, cada defecto se compara con 50 grupos aleatorios de datos que corresponden a no defectos.

El resultado de los 363 anchos de ventana de los 3 defectos se observa en la Figura 78. El rango desde σ=2 hasta σ=343, es válido para realizar el criterio, ya que el p-value es menor que 0.05. Se selecciona σ=182, ya que en la Figura 74, la Figura 75 y Figura 76 presenta un valor alto de porcentajes de "1" para la mayoría de los estadísticos.

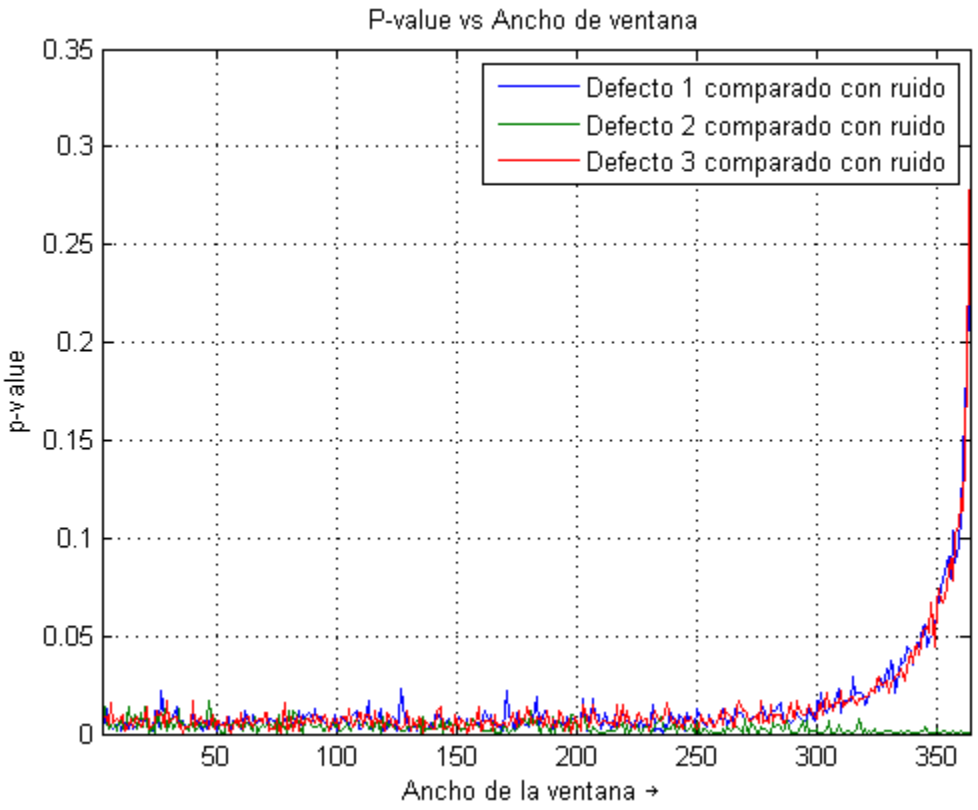

Figura 78. P-Value vs anchos de ventana para los 3 defectos.

8.3. Criterio de corrección

En este punto se decide qué es falso positivo y qué es defecto, para lograr esto, se utiliza el criterio de en una ventana el porcentaje de “1” debe ser mayor al 90% para que se afirme que existe un defecto, si dicho porcentaje es menor a 90% se considera un falso positivo y se elimina de la clasificación.

La Figura 79 muestra la señal de clasificación corregida para el clasificador ADL con datos desbalanceados en la señal 11 de FFM. La Figura 80, la Figura 81 y la Figura 82 muestran la clasificación corregida en la señal 11 de FFM, mostrando los 3 defectos.

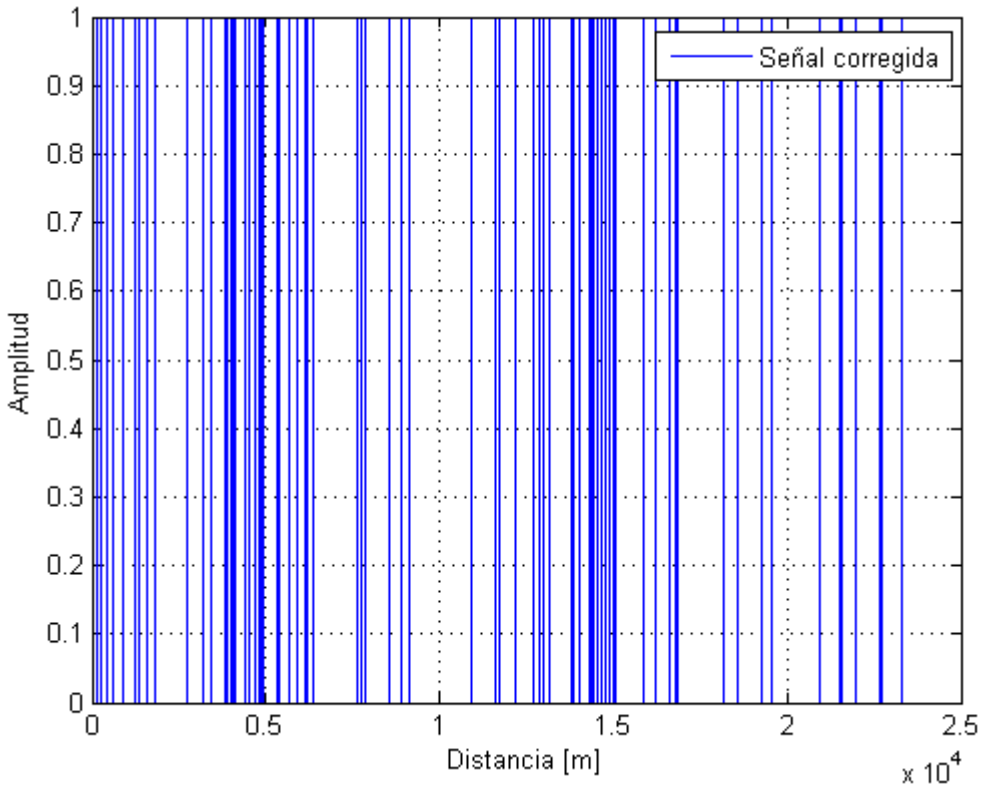

Figura 79. Señal de clasificación corregida, ADL Señal 11 FFM.

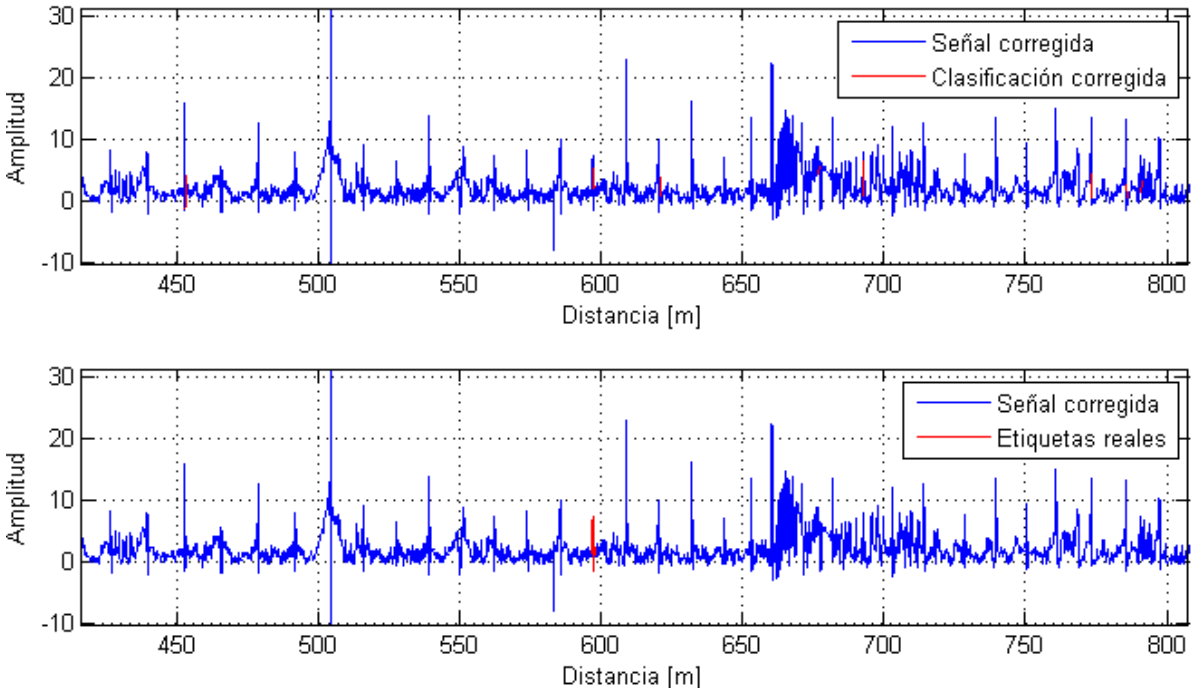

Figura 80. Clasificación corregida, ADL con datos desbalanceados Señal 11 FFM (Defecto 1).

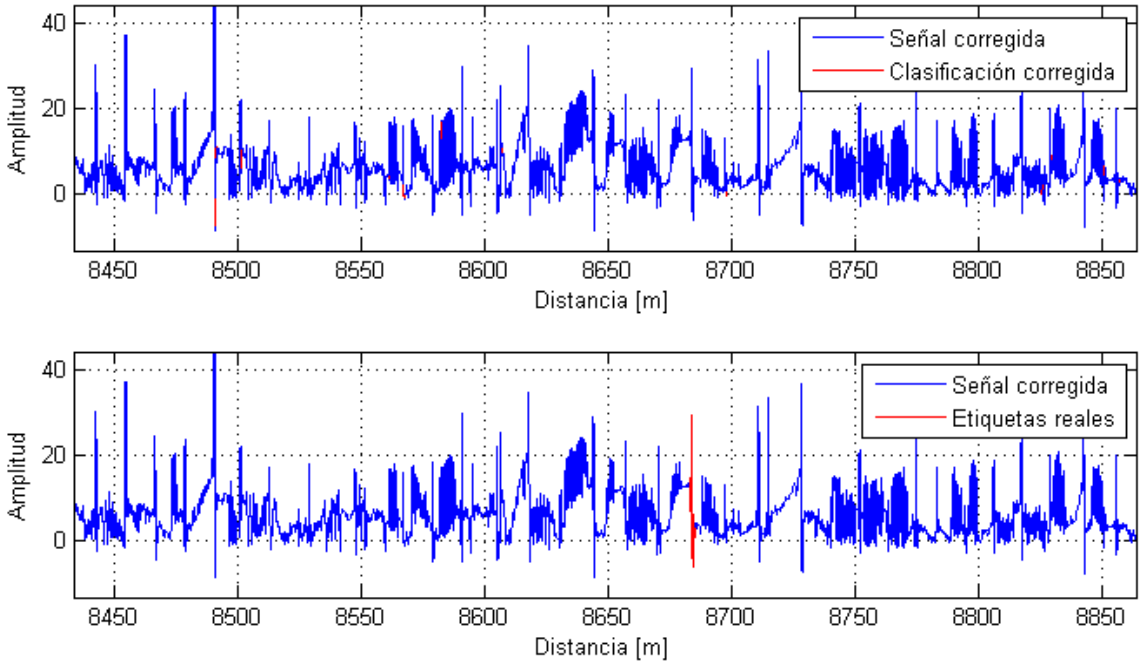

Figura 81. Clasificación corregida, ADL con datos desbalanceados Señal 11 FFM (Defecto 2).

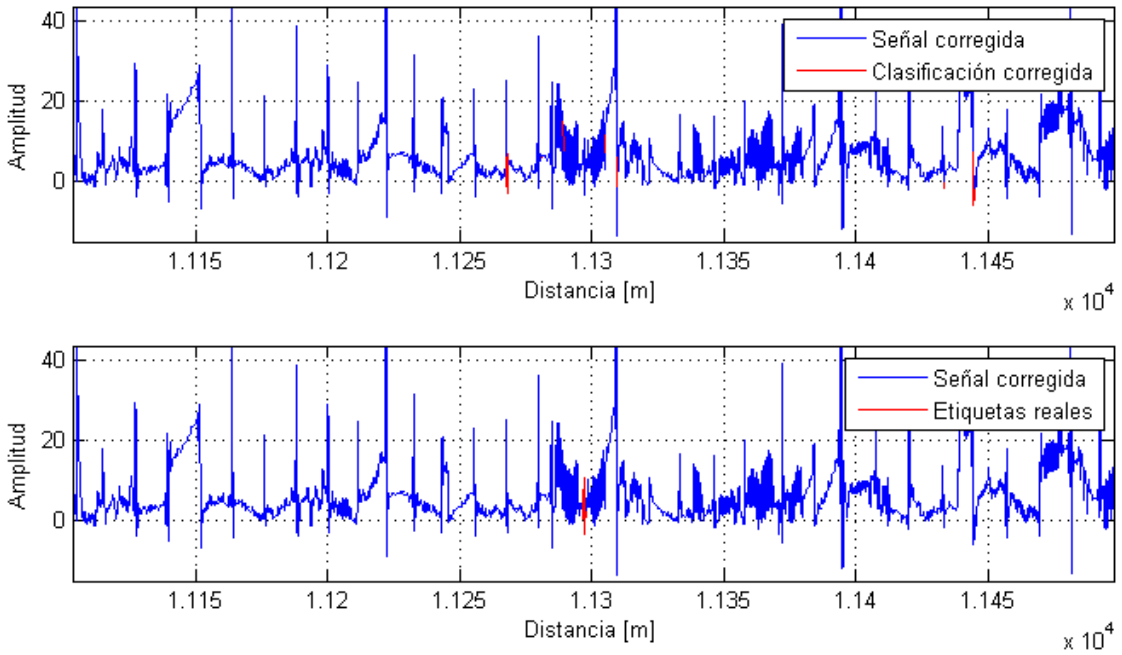

Figura 82. Clasificación corregida, ADL con datos desbalanceados Señal 11 FFM (Defecto 3).

9. COLCUSIONES

- Dada la característica no lineal y además adaptativa del filtro Shrinkage, lo hacen una buena solución para este tipo de aplicaciones, ya que los niveles de ruido son variables y dependen de diversos factores.

- Dada la naturaleza lineal del ADL y de las MSV con Kernel lineal, la corrección de línea base permite un mejor resultado de clasificación que el que se podría obtener sin realizarla dicha corrección.

- Los clasificadores no siempre pueden encontrar el número de muestras totales en cada defecto, esto se debe principalmente a que un dato con un valor “x” se repite varias veces a lo largo de la señal, y en el entrenamiento pudo tomarse dicho dato como no defecto.

- La precisión balanceada permite evaluar el desempeño de los clasificadores en este problema en específico, ya que al tener una cantidad mucho mayor de patrones de una clase con respecto a la otra es necesario utilizar un índice muestre un porcentaje válido sin importar el tamaño de cada clase.

- Es probable que existan más de los tres defectos indicados por la CIC, ya que los clasificadores coincidieron al darlos como defectos, considerando que eran clasificadores diferentes.

- El éxito de la corrección de los falsos positivos depende directamente de los clasificadores, si por algún motivo un clasificador no encontró algún defecto, dicha corrección no podrá detectarlo.

10. BIBLIOGRAFÍA

[1] AFZAL, Muhammad. Advanced signal processing of magnetic flux leakage data obtained from seamless gas pipeline. Canadá: Ndt & E International, 2002, vol. 35, no 7, p. 449-457.

[2] BUBENIK, T. A., Magnetic flux leakage (MFL) technology for natural gas pipeline inspection. Canadá: *NDT and E International*, 1997, vol. 30, no 1, p. 36-36.

[3] MUKHOPADHYAY, S.; SRIVASTAVA, G. P. Characterisation of metal loss defects from magnetic flux leakage signals with discrete wavelet transform. *Ndt & E International*, 2000, vol. 33, no 1, p. 57-65.

[4] CHUI, Charles K. *Wavelets*: a mathematical tool for signal analysis. Siam, 1997.

[5] MALLAT, Stephane G. A Theory for Multiresolution Signal Decomposition: The *Wavelet* Representation. Filadelfia. 1987.

[6] ARIZMENDI-PEREIRA, Carlos Julio. *Signal processing techniques for brain tumour diagnosis from magnetic resonance spectroscopy data*. 2012. Tesis Doctoral. Universitat Politècnica de Catalunya.

[7] ZACCARO, Jorge A., et al. Clasificación de Señales Cerebrales Relacionadas con la Imaginación de Movimientos para Aplicaciones de BCI.

[8] SUBASI, Abdulhamit; ISMAIL GURSOY, M. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, 2010, vol. 37, no 12, p. 8659-8666.

[9] BOHORQUEZ-REYES, J, C. *Interpretación de señales para la detección de obstrucciones en inspecciones instrumentadas de tuberías*, 2015.

[10] BRODERSEN, Kay Henning, et al. The balanced accuracy and its posterior distribution. En *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010. p. 3121-3124.

[11] SCHOLKOPFT, Bernhard; MULLERT, Klaus-Robert. Fisher discriminant analysis with kernels. En Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, Madison, WI, USA. 1999. p. 23-25.

[12] HEARST, Marti A. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 1998, vol. 13, no 4, p. 18-28.

[13] JOACHIMS, Thorsten. Making large scale SVM learning practical. Ithaca1999, p. 169-184.

[14] HU, Mingqing; CHEN, Yiqiang; KWOK, JT-Y. Building sparse multiple-kernel SVM classifiers. *Neural Networks, IEEE Transactions on*, 2009, vol. 20, no 5, p. 827-839.