Detección de pérdidas de espesor en las paredes de tuberías de transporte de hidrocarburos utilizando técnicas de procesamiento de señales y minería de datos

ALDAIR BARAJAS ALDANA

INGENIERÍA MECATRÓNICA

2015

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.

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.

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.

MARCA	POSICIÓN APROXIMADA (M)
'PM'	600,72
'PM'	8676,37
'PM'	11291,64

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.

PREPROCESAMIENTO DE LOS DATOS

- ***Eliminación de puntos muertos.***

Las señales fueron recortadas de 14’449,944 muestras (izquierda) a 10’139,445 muestras (derecha)

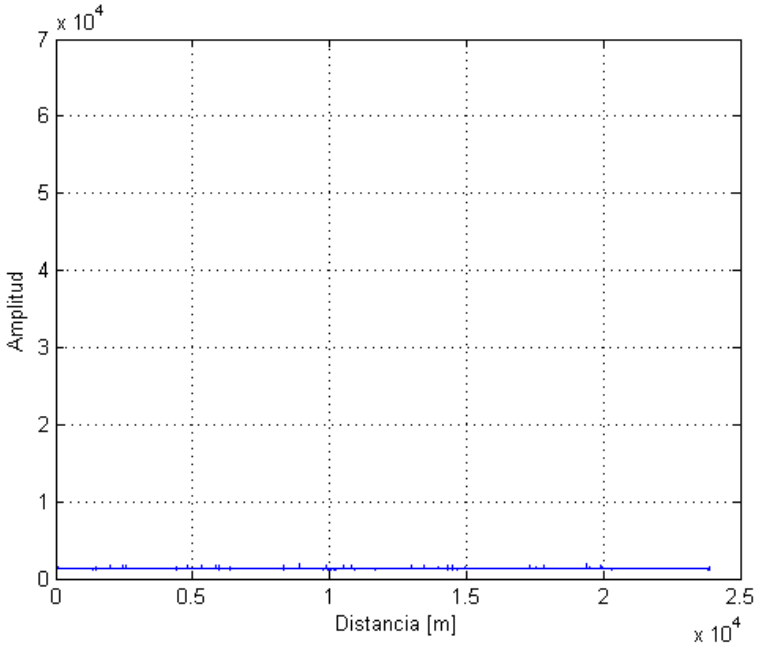


Figura 1. Señal 11 de FFM original sin recortar

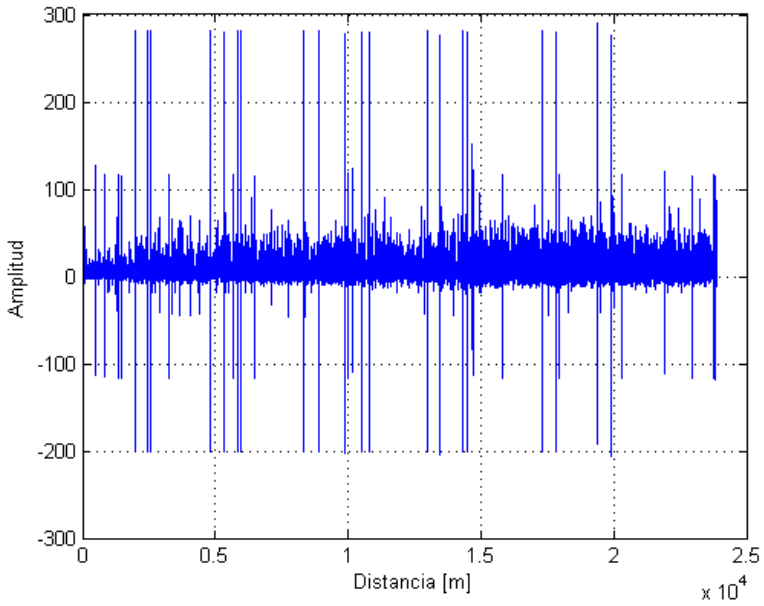


Figura 2. Señal 11 de FFM recortada

PREPROCESAMIENTO DE LOS DATOS

- **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, cuyo proceso se observa en la figura.

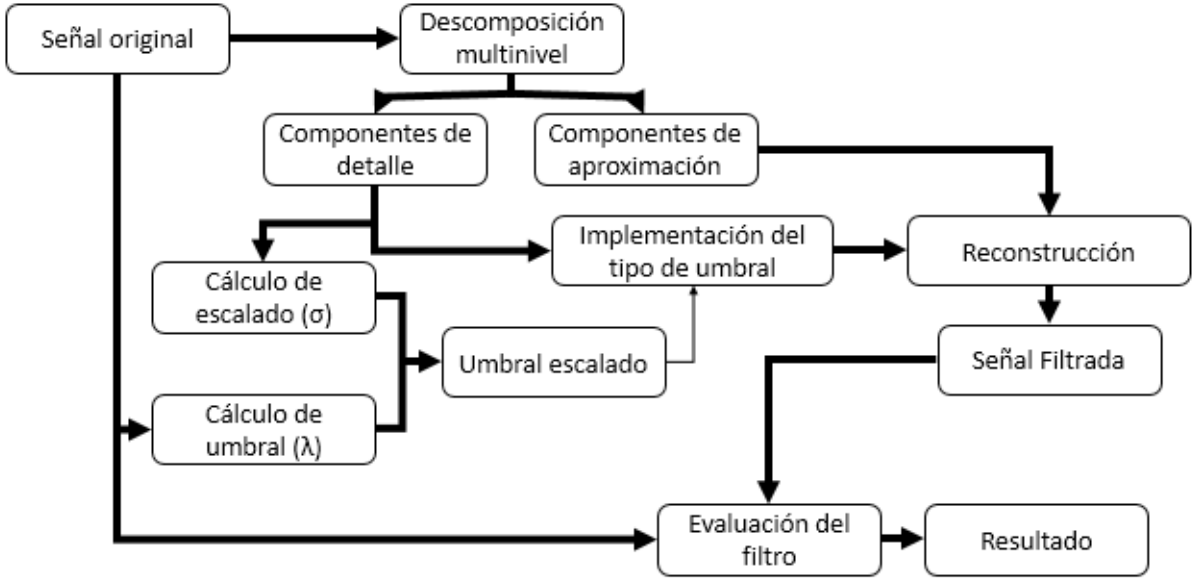


Figura 3. Diagrama Filtrado Shrinkage

- Descomponer la señal.
- 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.
- Reconstruir la señal con cada nivel de aproximación y detalles filtrados.

PREPROCESAMIENTO DE LOS DATOS

- **Filtrado Shrinkage**

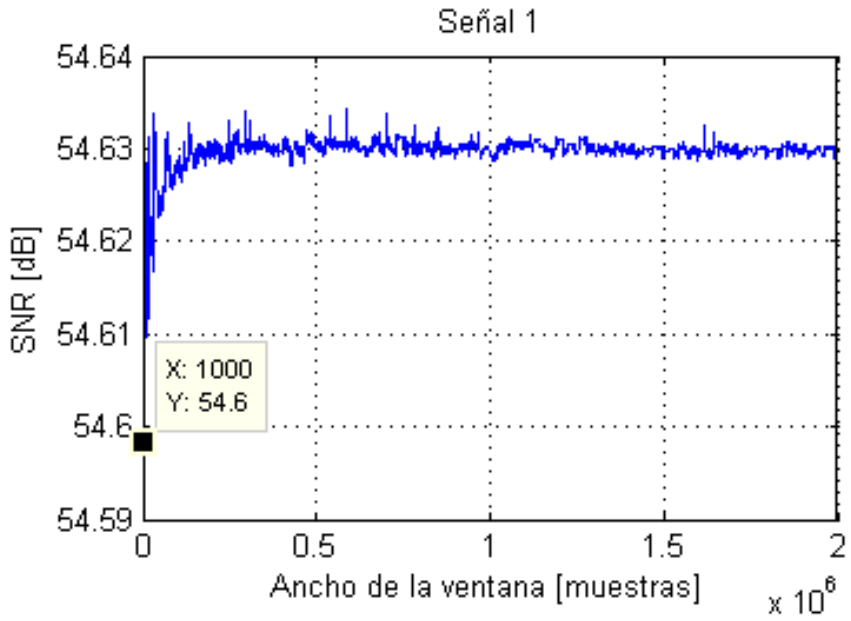


Figura 4. SNR para diferentes anchos de ventana, Señal 11 de FFM.

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 usadas

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. Mejor función Wavelet y mejor nivel de descomposición

PREPROCESAMIENTO DE LOS DATOS

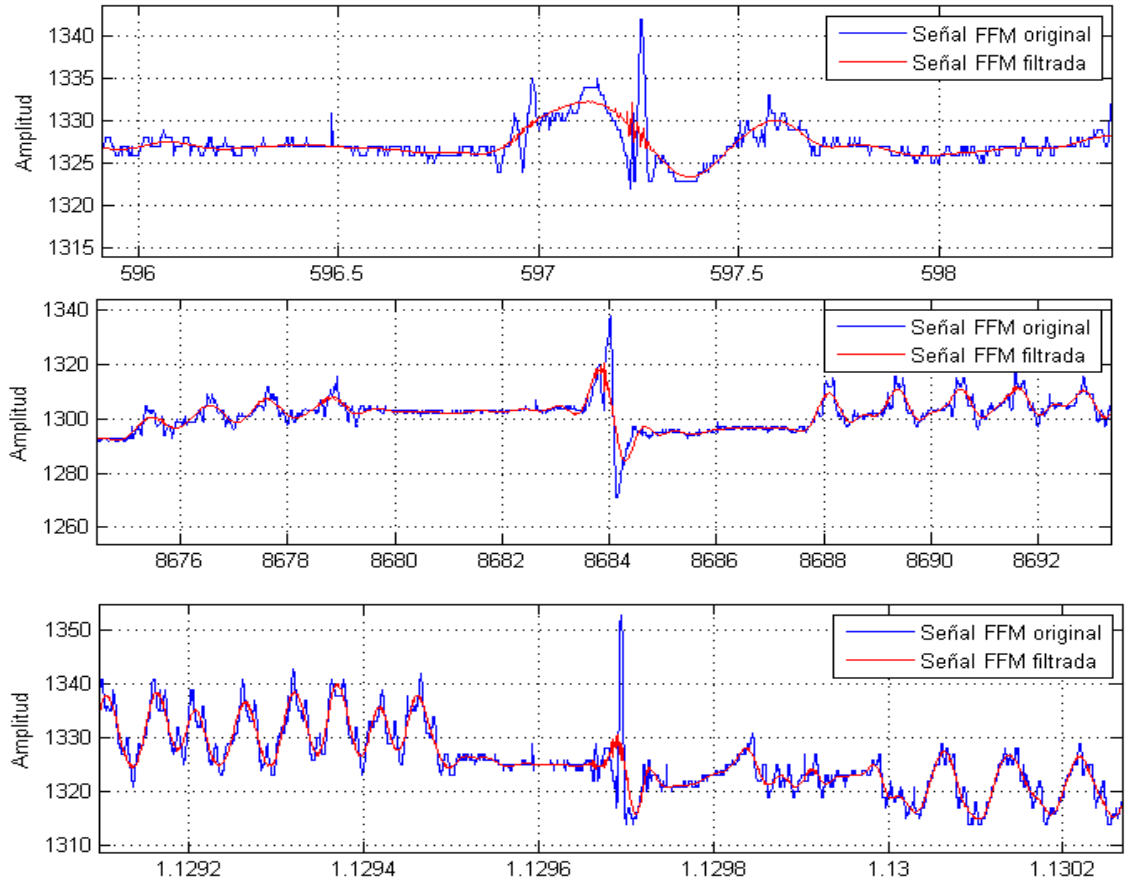


Figura 5. Comparación entre la señal original y la señal filtrada.

PREPROCESAMIENTO DE LOS DATOS

- **Corrección de Línea Base**

Es un tipo de preprocesamiento que intenta corregir determinadas tendencias que aportan ruido a la señal. Un tipo de corrección es el que ajusta la señal original a una función cuadrática, sustrayéndola posteriormente de la señal original.

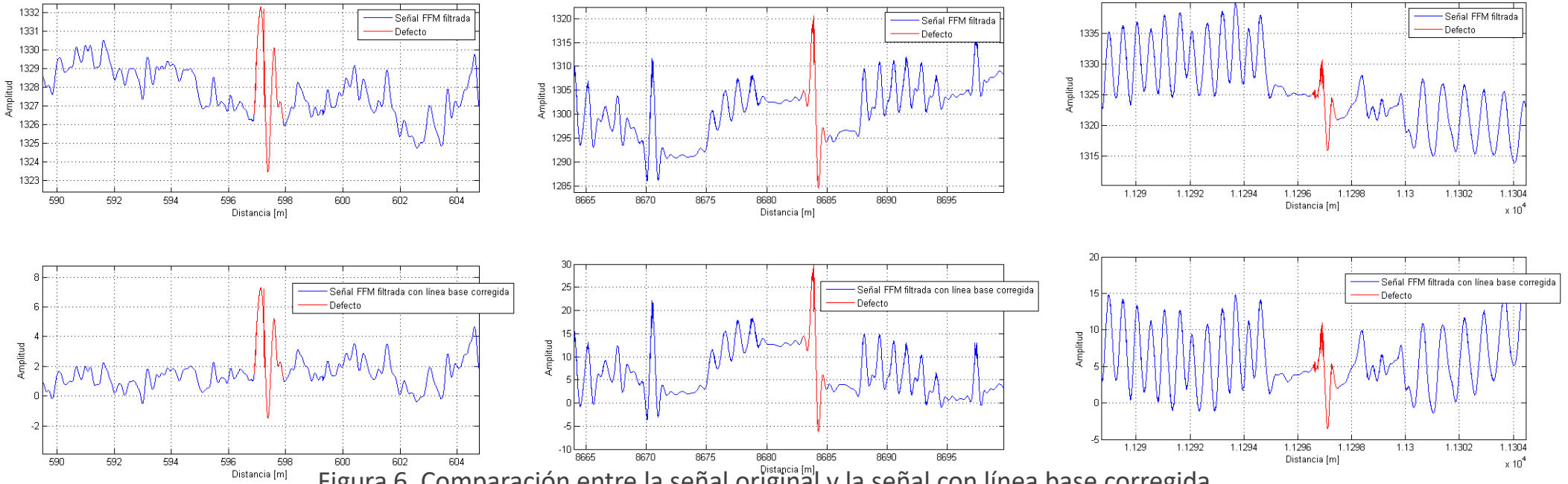


Figura 6. Comparación entre la señal original y la señal con línea base corregida

PREPROCESAMIENTO DE LOS DATOS

- **Eliminación de las soldaduras**

Tomando las etiquetas de soldaduras suministradas por la CIC y en conjunto con las etiquetas que se encontraron en [1] se procede a eliminar las soldaduras de la señal

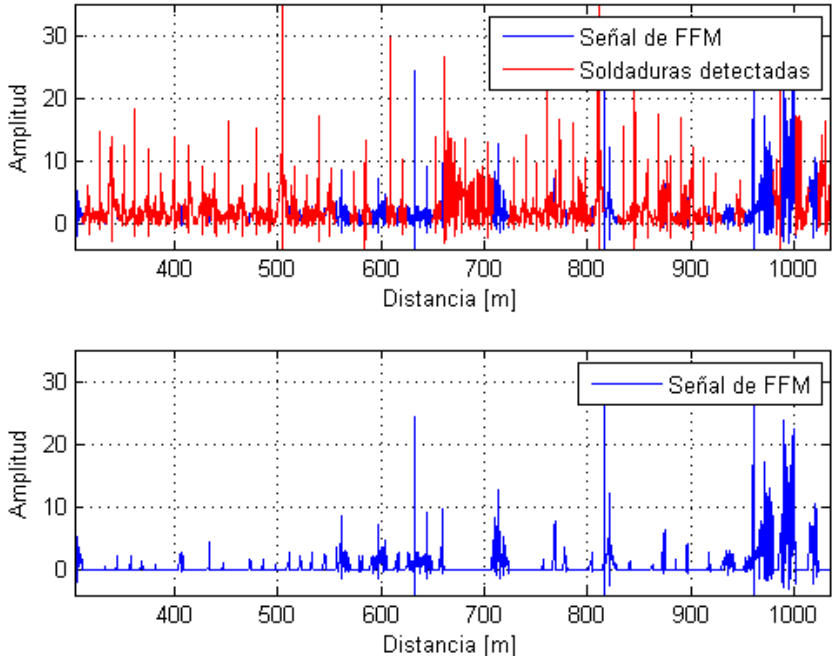


Figura 7.Comparación entre la señal original y la señal sin soldaduras

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

PREPROCESAMIENTO DE LOS DATOS

- **Remuestreo de la señal**

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.

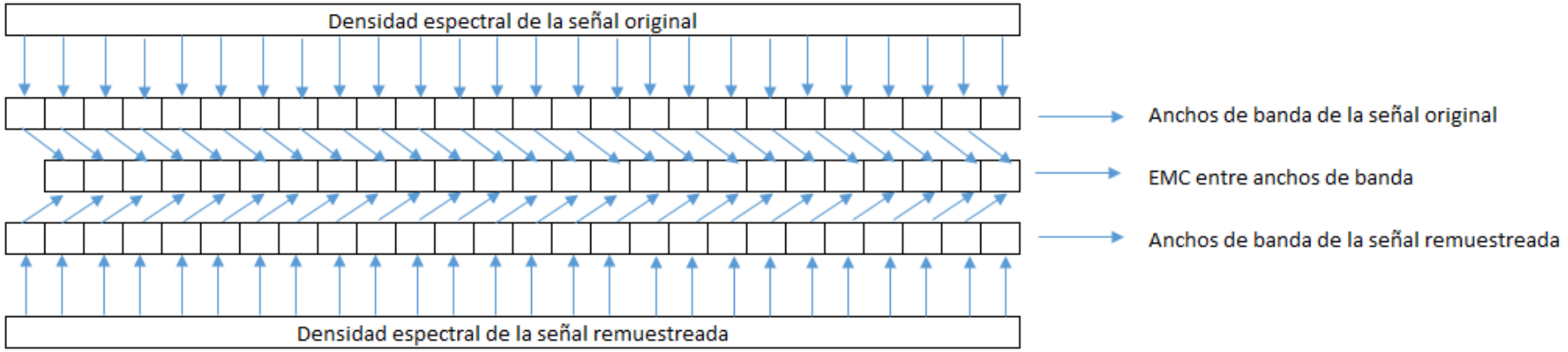


Figura 8. Ejemplo de la comparación de los anchos de banda entre la densidad espectral de frecuencia de la señal original y de la señal remuestreada

PREPROCESAMIENTO DE LOS DATOS

- **Remuestreo de la señal**

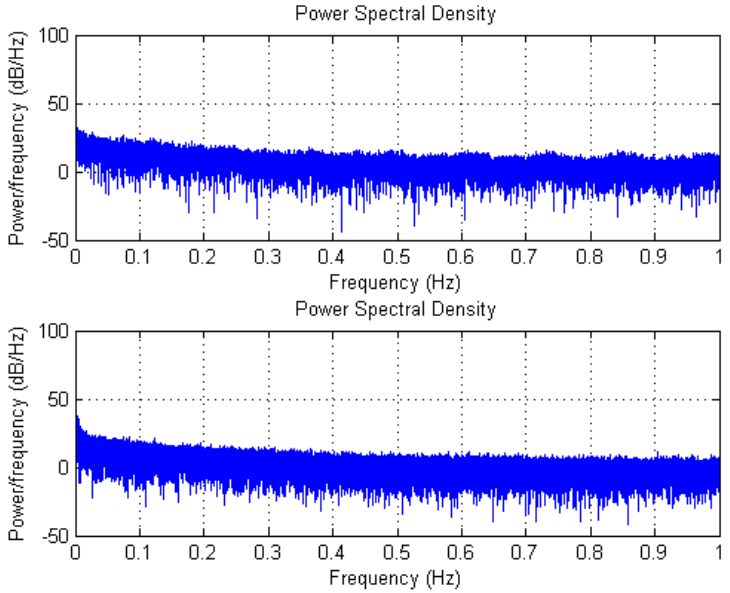


Figura 9. Densidad espectral de potencia de un ancho de banda de 0.5 Hz de la señal remuestreada a 5 mm (arriba) y la señal original (abajo).

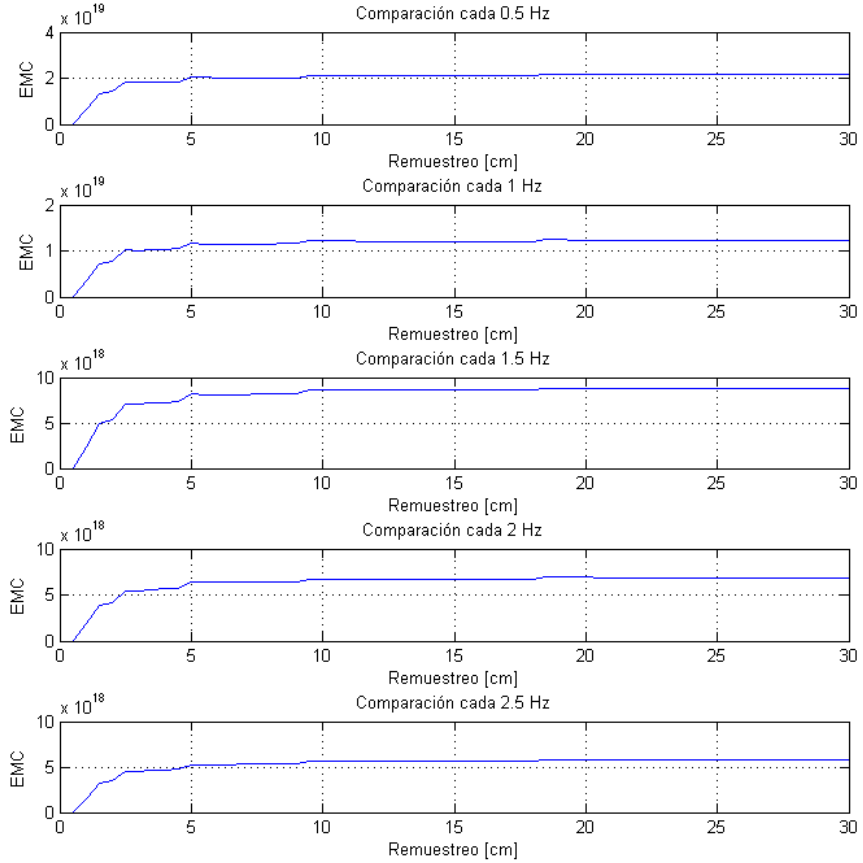


Figura 10. EMC de los diferentes anchos de banda y para todas las tasas de remuestreo.

PROCESAMIENTO DE LOS DATOS

- **Corrección de ruido con Wavelet Tree**

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.

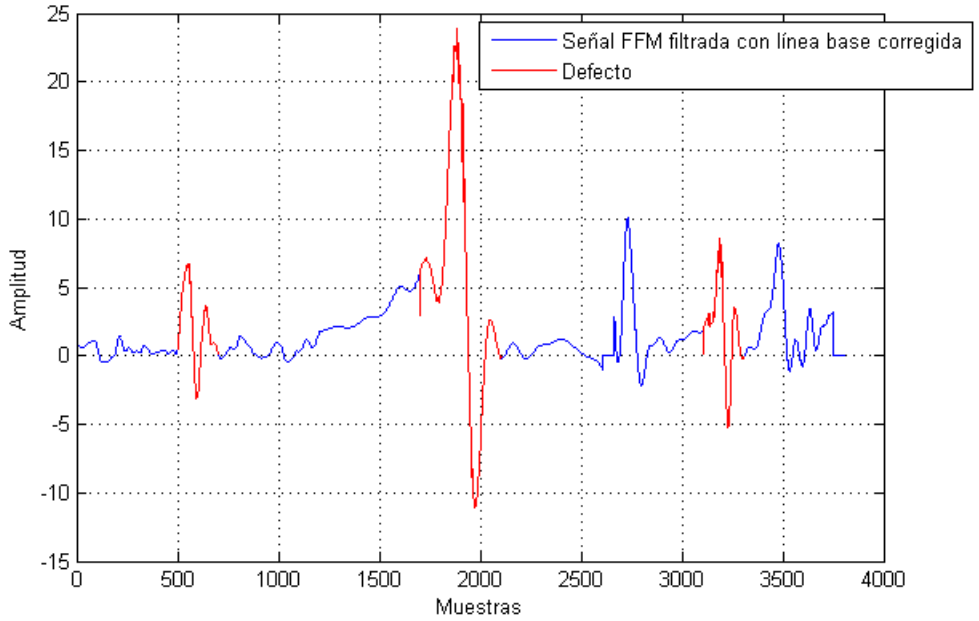


Figura 11. Señal utilizada para realizar la descomposición

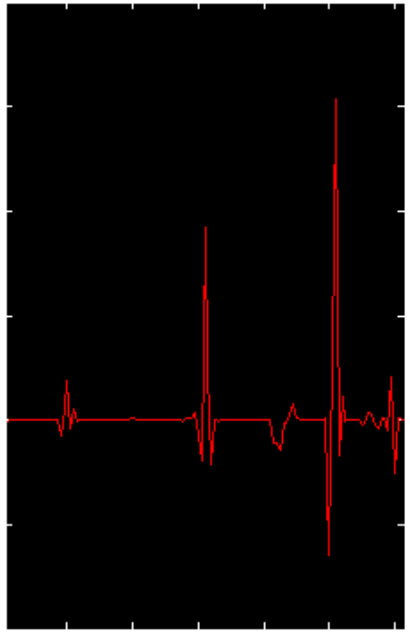


Figura 12. Mejor nodo para los datos de la prueba en el Nodo 58

PROCESAMIENTO DE LOS DATOS

- **Corrección de ruido con Wavelet Tree**

El resultado de todas las distancias euclideas para los 64 nodos y las 36 funciones Wavelet para la señal 11 de FFM se observa en la

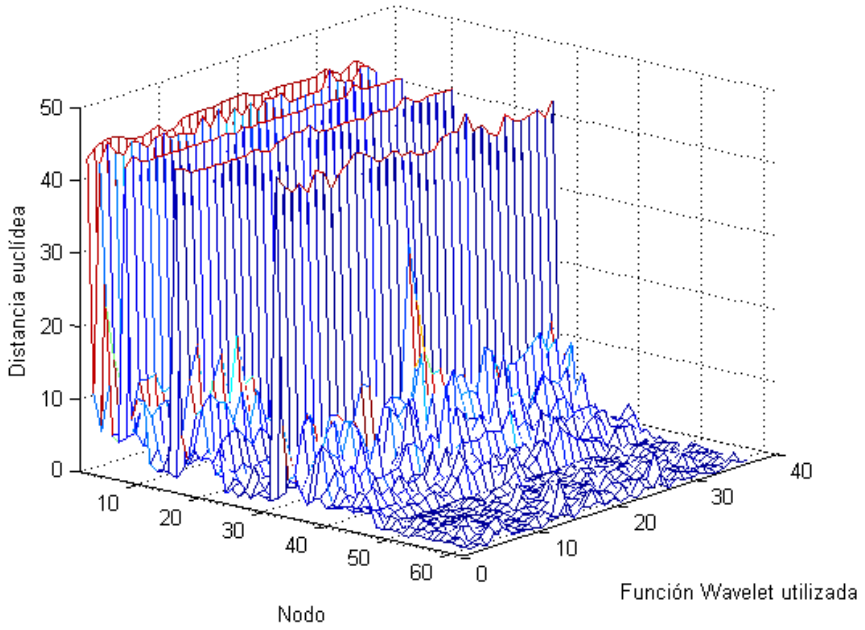


Figura 13. Distancias euclídeas para la señal 11 de FFM.

PROCESAMIENTO DE LOS DATOS

- **Selección del nodo y función *Wavelet* adecuados**

Para realizar una selección de manera adecuada, se deben normalizar las distancias euclídeas. Cada señal de FFM cuenta con una matriz de tamaño 36x64, 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.

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 4. Nodo y función Wavelet óptima para cada señal de FFM

PROCESAMIENTO DE LOS DATOS

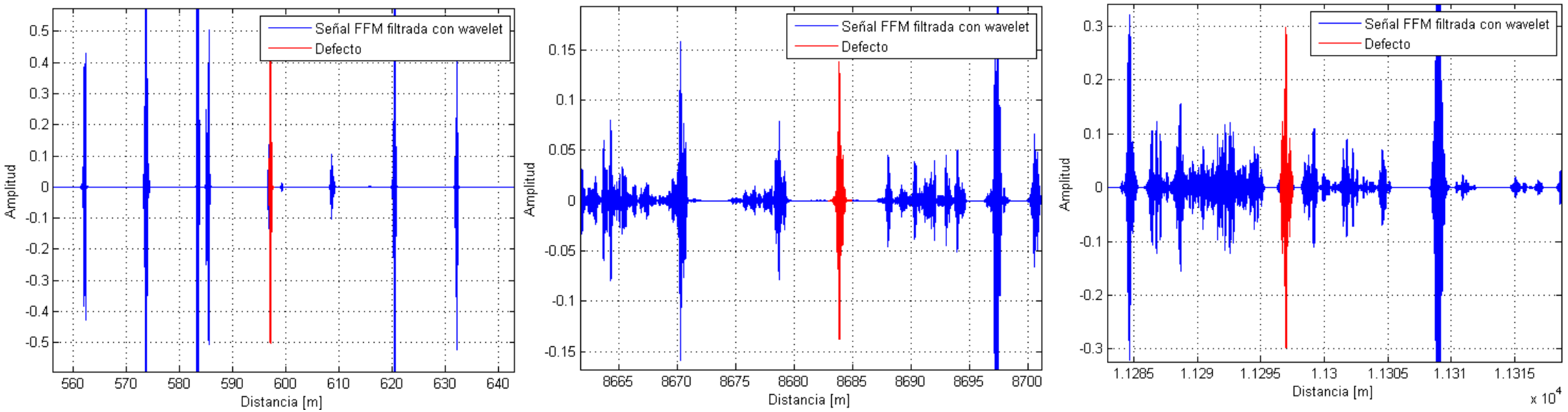


Figura 14. Señal filtrada con Wavelet

PROCESAMIENTO DE LOS DATOS

- **Ventaneo con estadísticos**

Para este punto se propone ventanear las señales con 8 estadísticos: media, mediana, rango intercuartil, desviación media absoluta, rango, desviación estándar, energía y potencia.

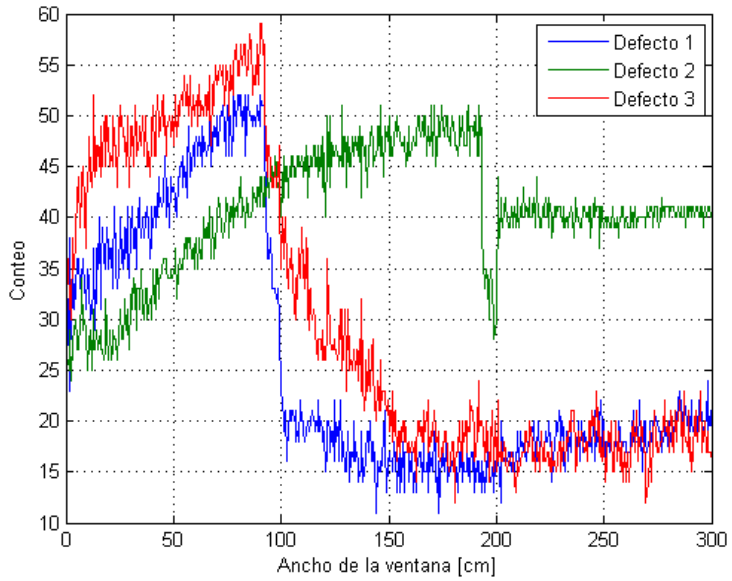


Figura 15. Conteo de los p-Value menores a 0.05 para cada ancho de ventana en todos los defectos.

- Para seleccionar el ancho de ventana óptimo se procede a hacer la prueba de U de Mann-Whitney.
- No realizó el ventaneo a toda la señal, se hizo en 3 segmentos de prueba para cada señal.
- Luego de que todos los descriptores han sido calculados, se calcula la sumatoria de los p-Value menores a 0.05.

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.

DISEÑO DE CLASIFICADORES

- **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.

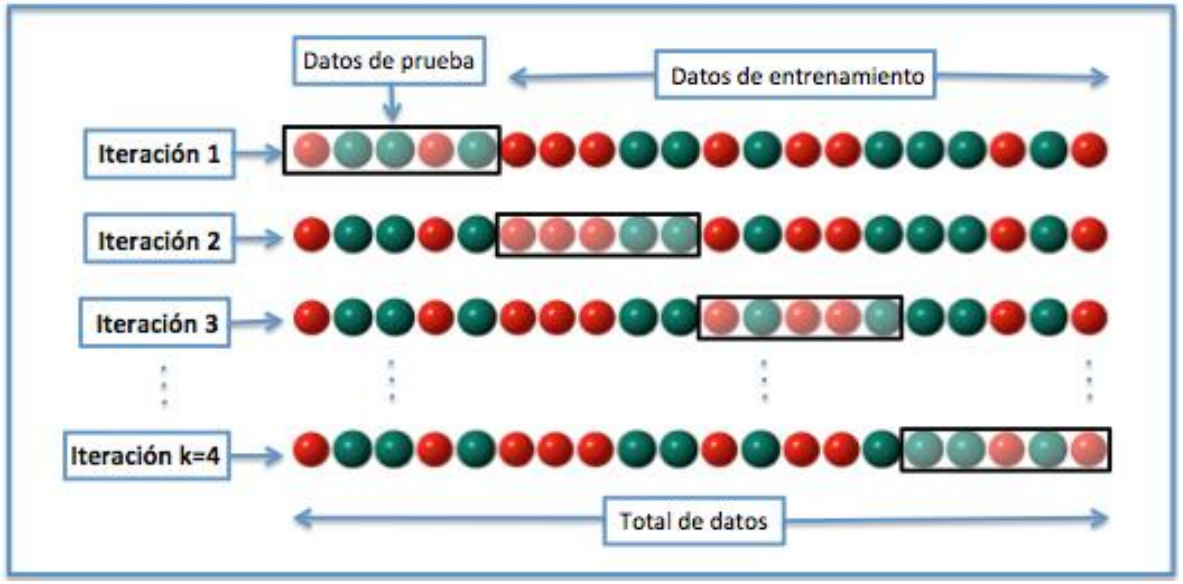


Figura 16. Ejemplo de validación cruzada.

$$PB = \frac{1}{2} * \frac{Cp}{Cp + Fp} + \frac{1}{2} * \frac{Cn}{Cn + Fn}$$

Donde:

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

DISEÑO DE CLASIFICADORES

- **Análisis Discriminante Lineal**

El objetivo de ADL es encontrar una función discriminante la cual maximice la separación entre ambas medias de las distribuciones gaussianas supuestas.

$$Y = a0 + a1 * x1 + a2 * x2 + \cdots + ap * xp$$

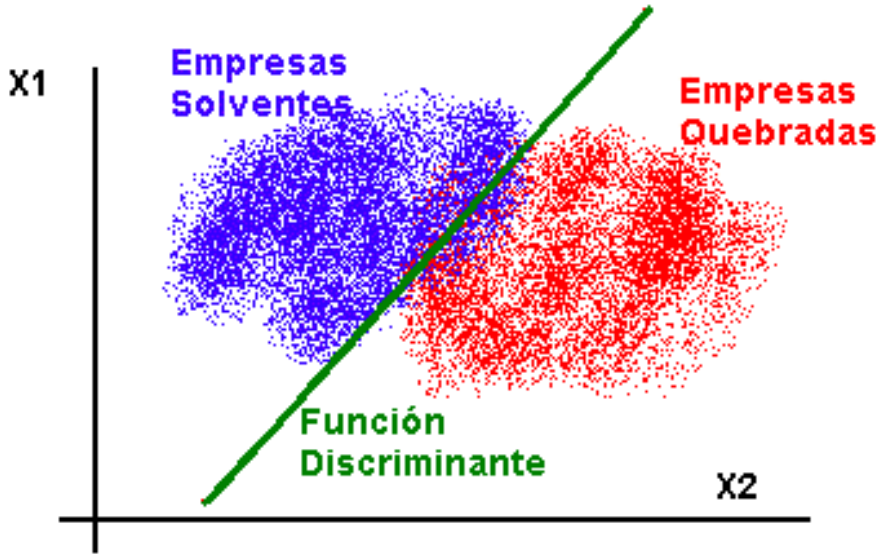
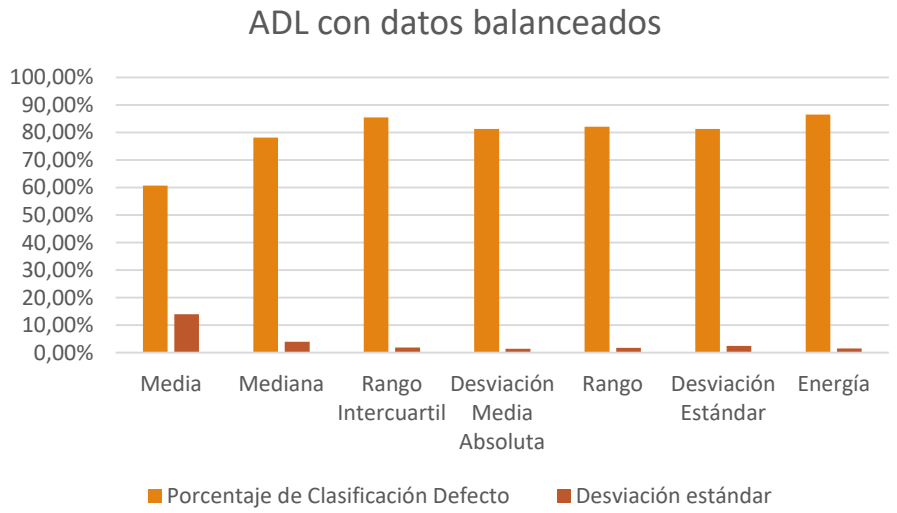


Figura 17. Ejemplo de Análisis discriminante lineal.

DISEÑO DE CLASIFICADORES

- **Análisis Discriminante Lineal con datos balanceados**



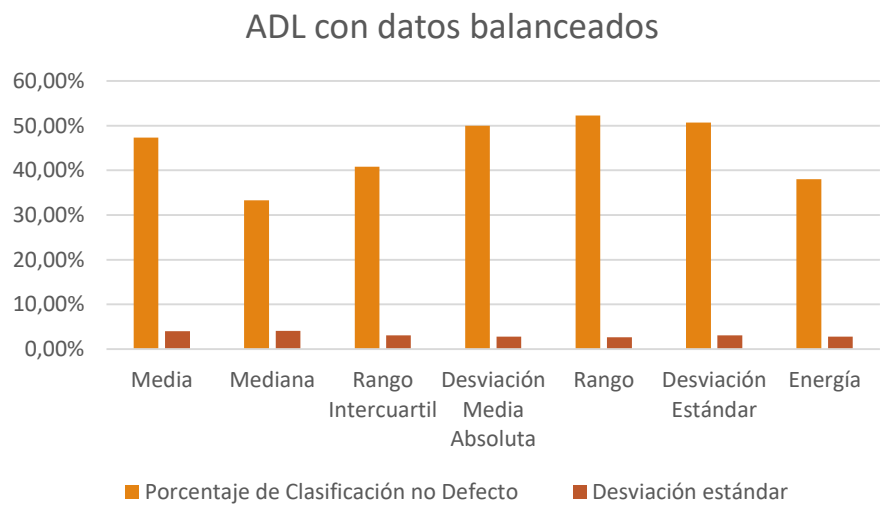


Figura 18. Porcentaje de clasificación ADL con datos balanceados

DISEÑO DE CLASIFICADORES

- **Análisis Discriminante Lineal con datos balanceados**

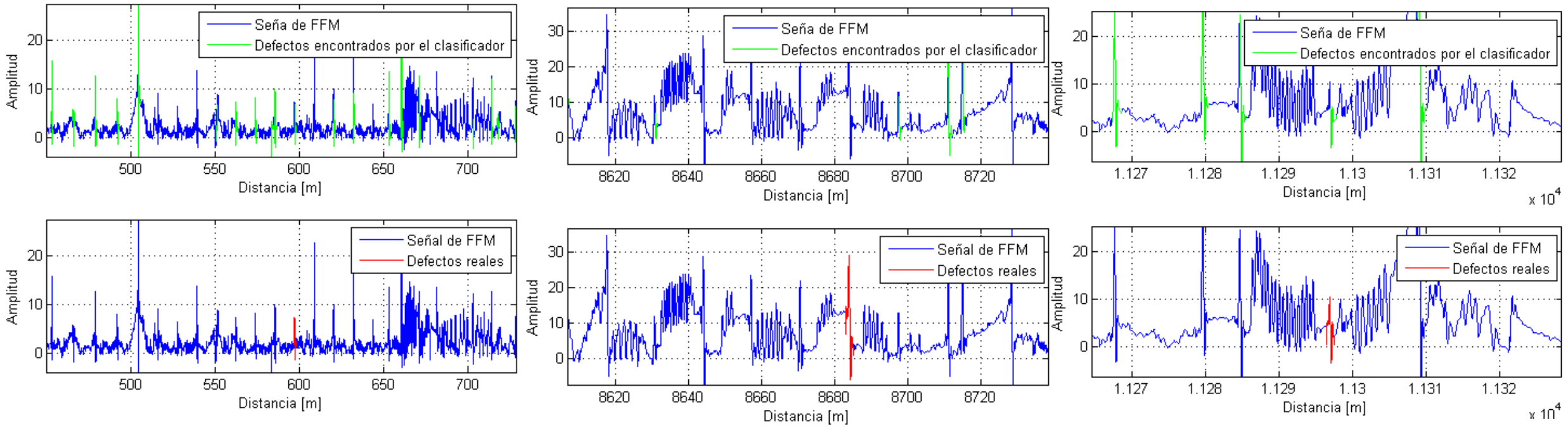
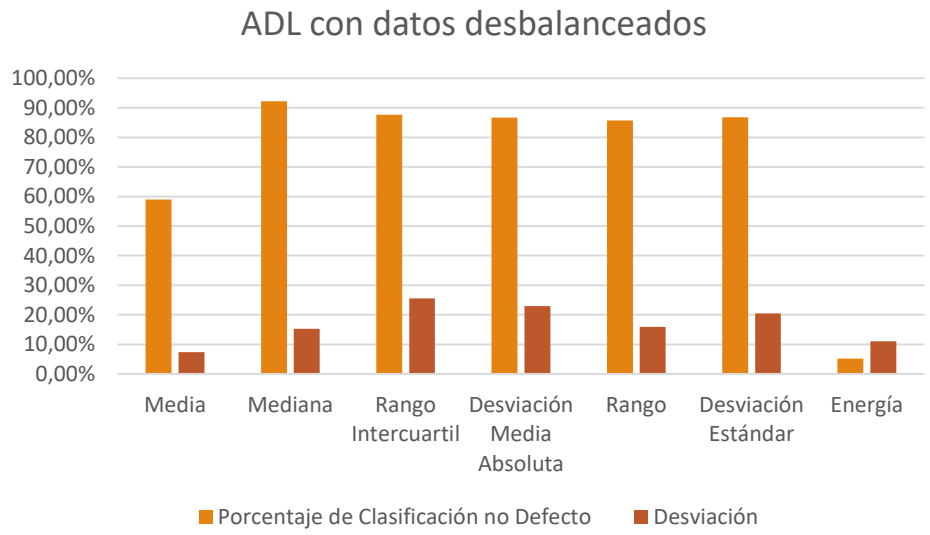


Figura 19. Resultado del clasificador en la Señal 11 de FFM - ADL con datos balanceados

DISEÑO DE CLASIFICADORES

- **Análisis Discriminante Lineal con datos desbalanceados**



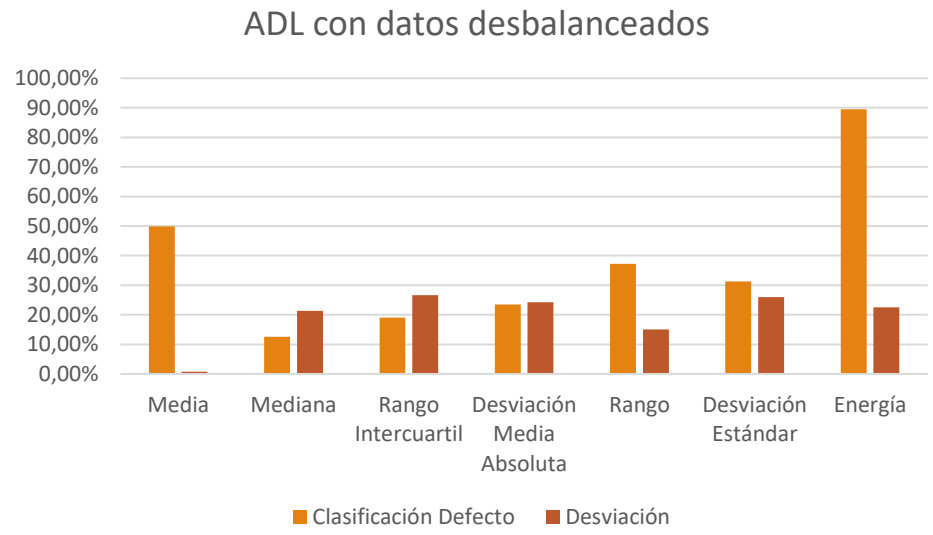


Figura 20. Porcentaje de clasificación ADL con datos desbalanceados

DISEÑO DE CLASIFICADORES

- **Análisis Discriminante Lineal con datos desbalanceados**

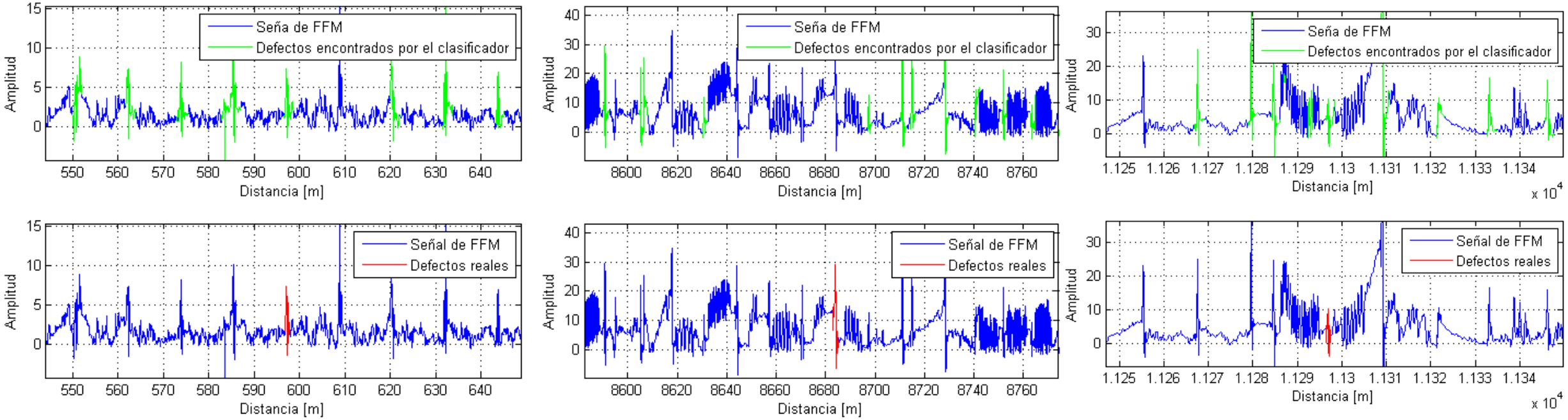


Figura 21. Resultado del clasificador en la Señal 11 de FFM - ADL con datos desbalanceados

DISEÑO DE CLASIFICADORES

- **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.

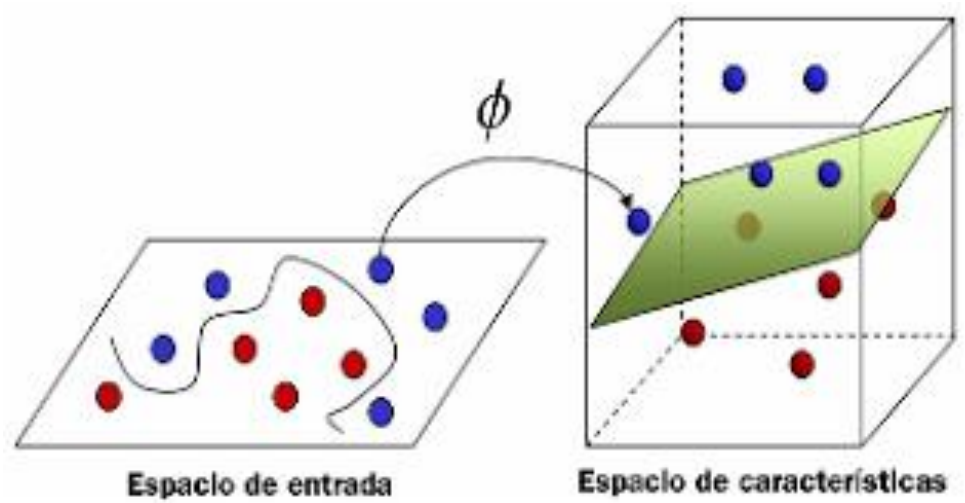
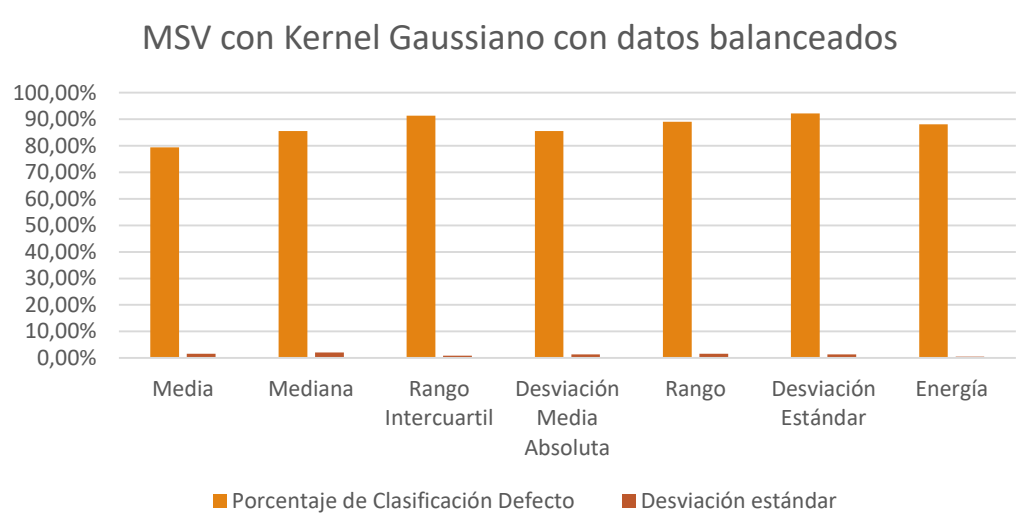


Figura 22. Ejemplo de Máquinas de soporte vectorial

DISEÑO DE CLASIFICADORES

- **MSV con Kernel Gaussiano**



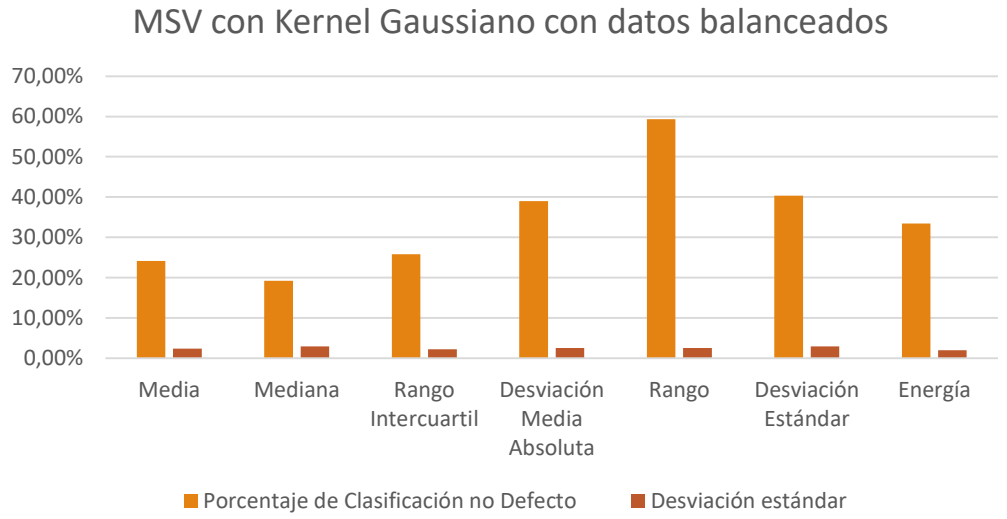


Figura 23. Porcentajes de Clasificación MSV con Kernel Gaussiano.

DISEÑO DE CLASIFICADORES

- **MSV con Kernel Gaussiano**

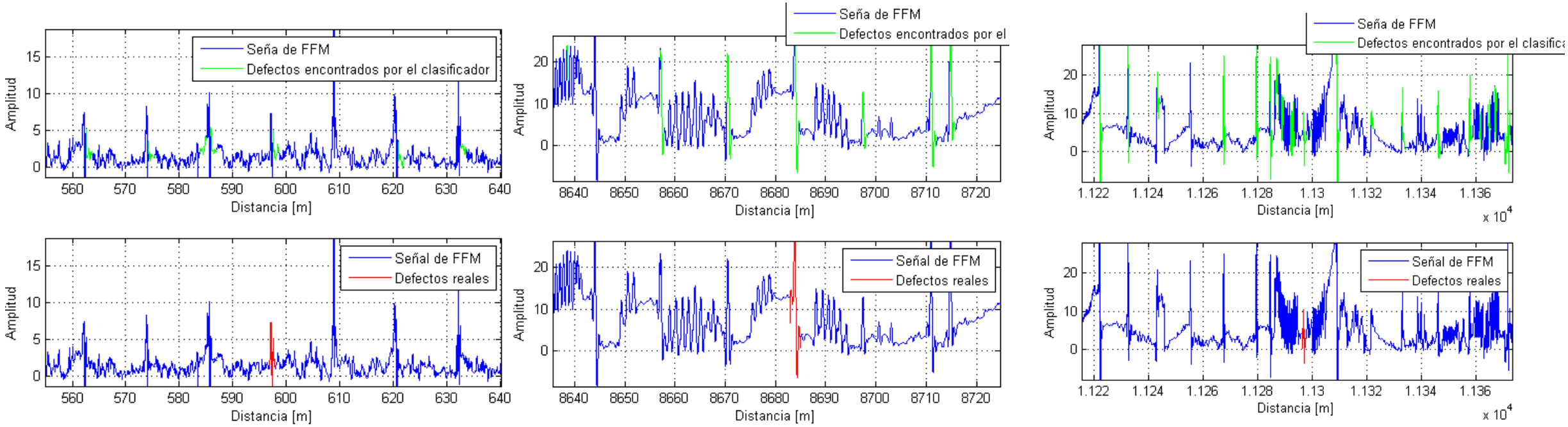
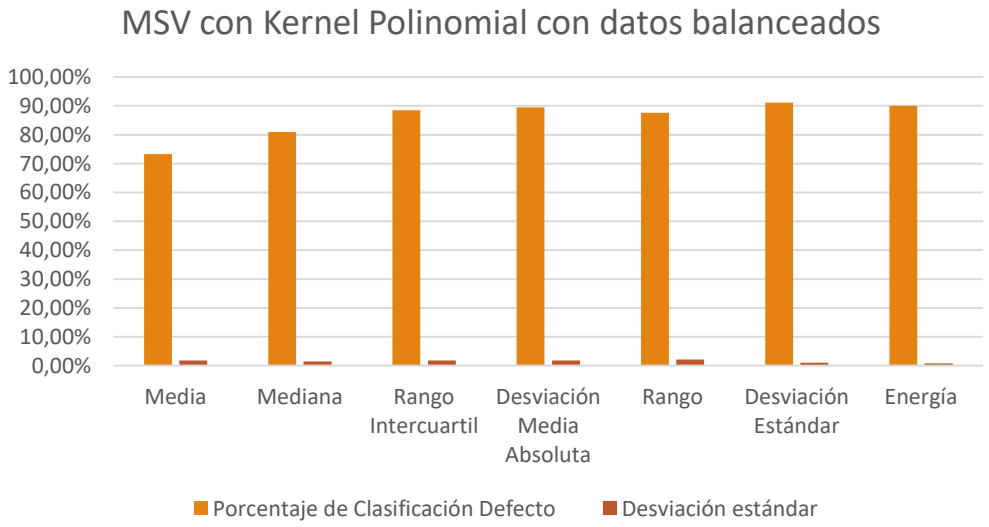


Figura 24. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Gaussiano

DISEÑO DE CLASIFICADORES

- **MSV con Kernel Polinomial**



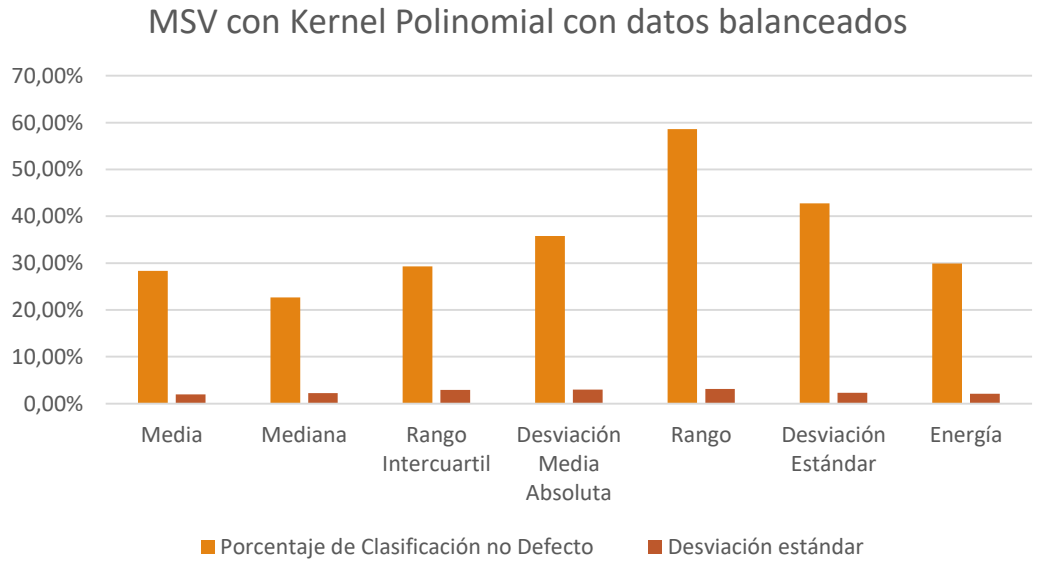


Figura 25. Porcentajes de Clasificación MSV con Kernel Lineal.

DISEÑO DE CLASIFICADORES

- **MSV con Kernel Polinomial**

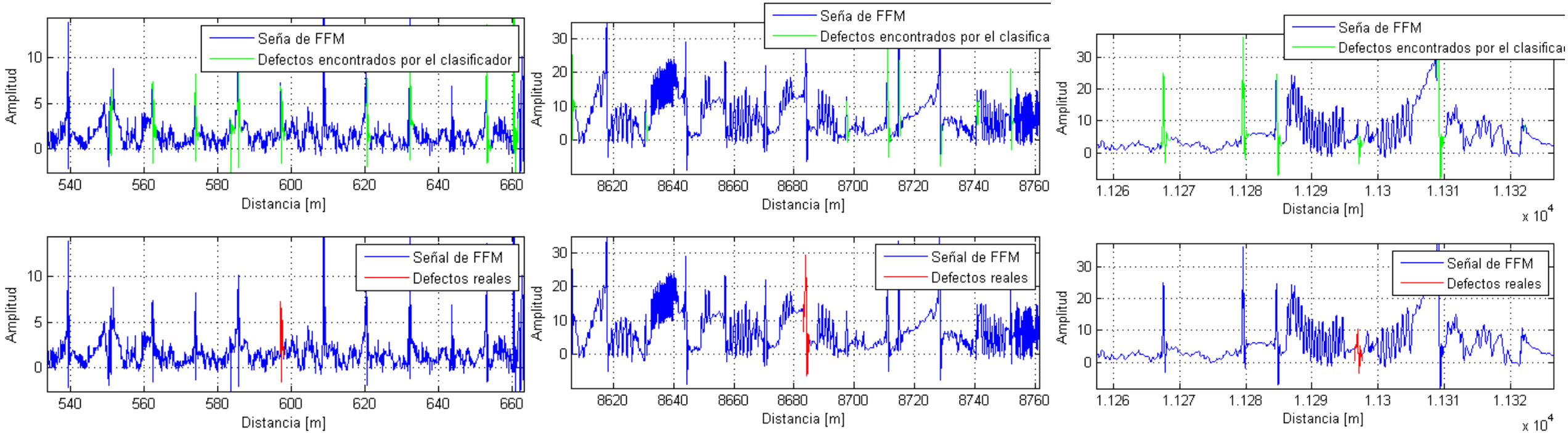
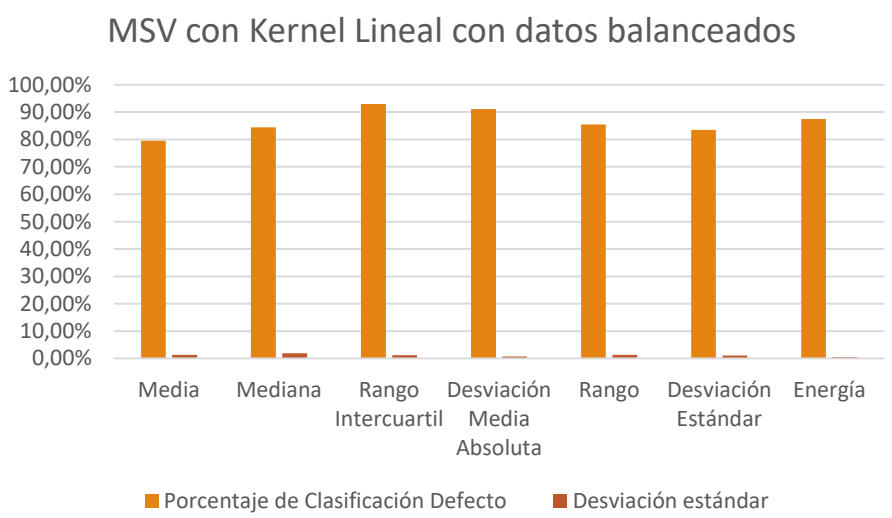


Figura 26. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Polinomial

DISEÑO DE CLASIFICADORES

- **MSV con Kernel Lineal**



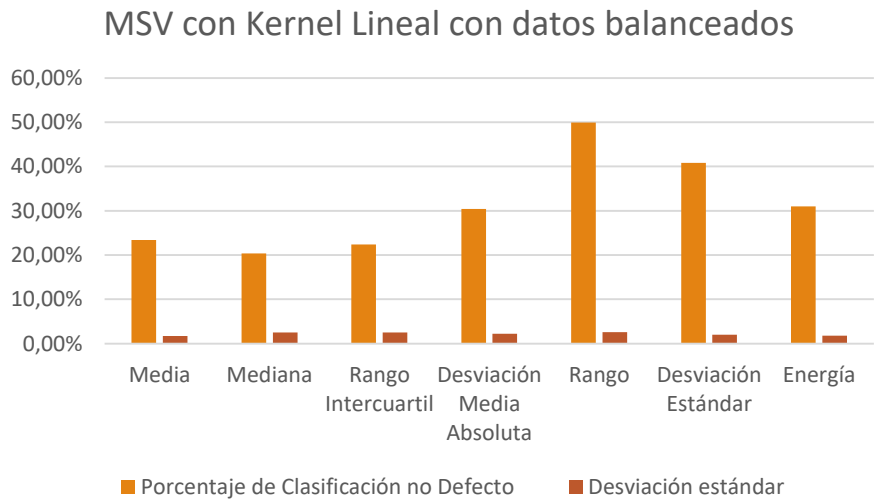


Figura 27. Porcentajes de Clasificación MSV con Kernel Lineal.

DISEÑO DE CLASIFICADORES

- **MSV con Kernel Lineal**

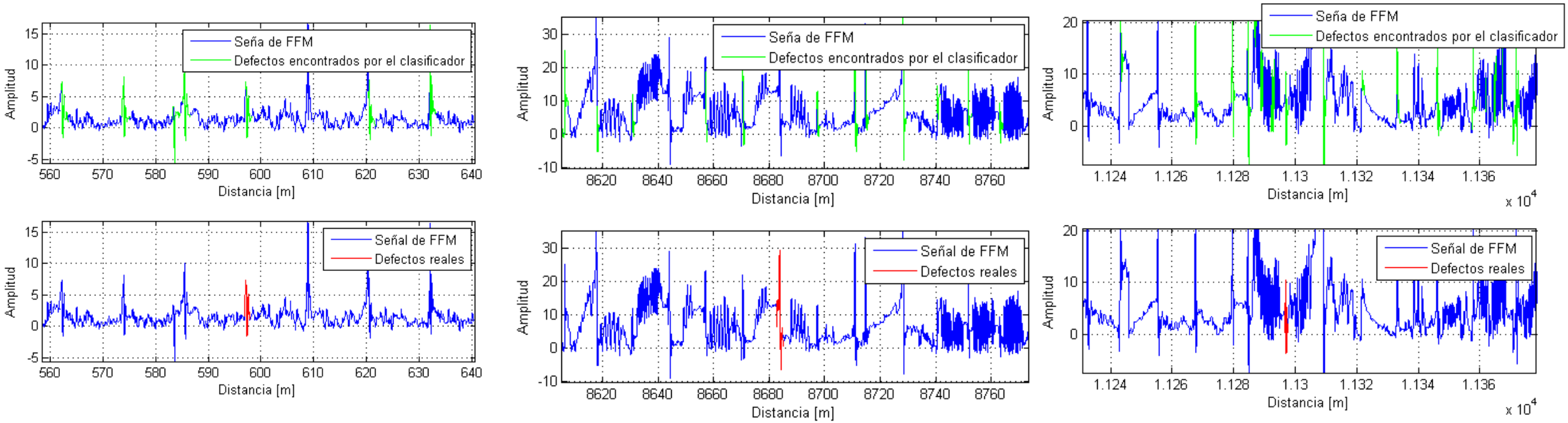


Figura 28. Resultado del clasificador en la Señal 11 de FFM - MSV con Kernel Lineal

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.
- 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.
- 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.

CORRECCIÓN DE FALSOS POSITIVOS

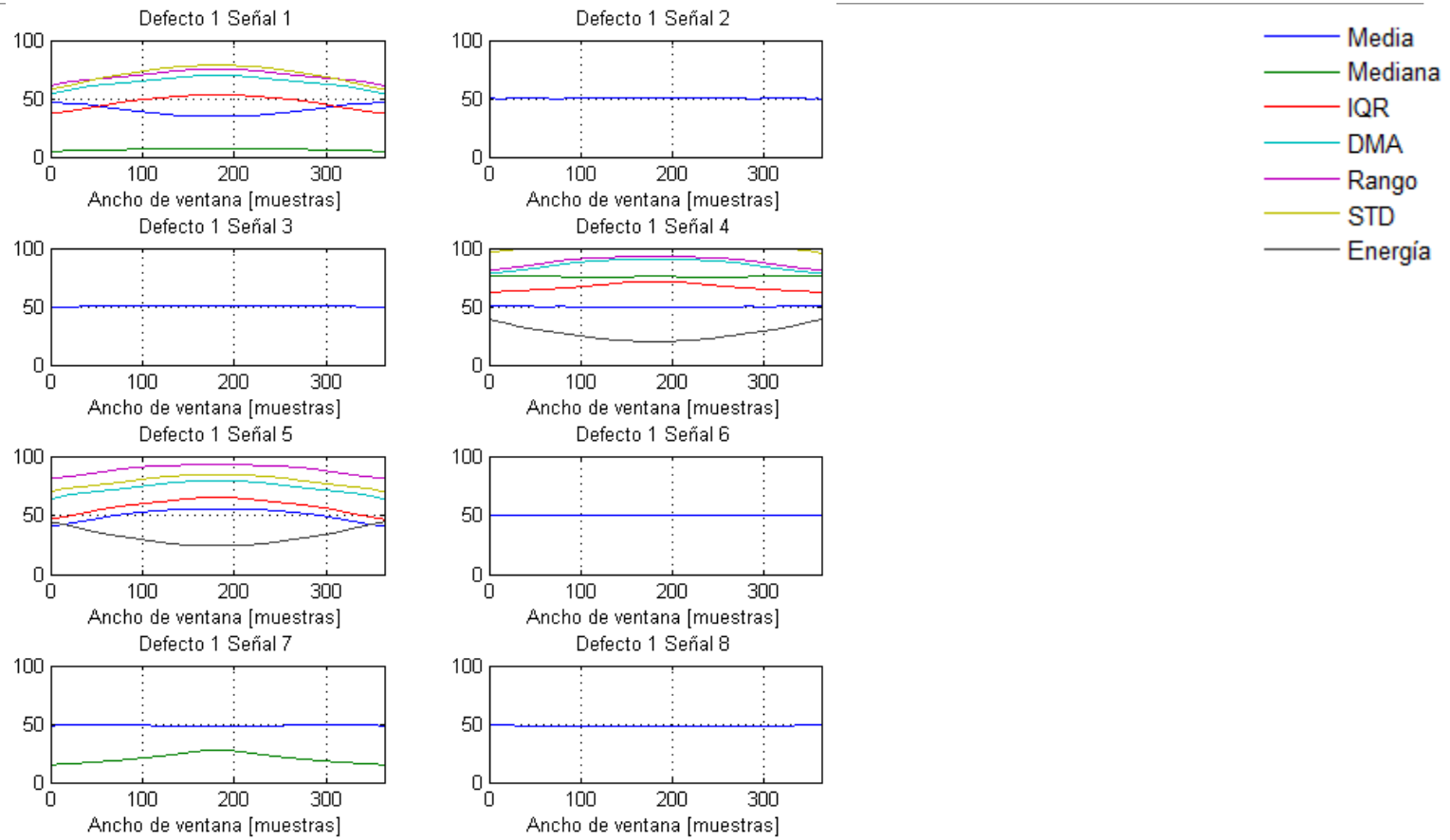


Figura 29. Porcentaje de "1" para el primer defecto, todas las señales (ADL con datos desbalanceados).

CORRECCIÓN DE FALSOS POSITIVOS

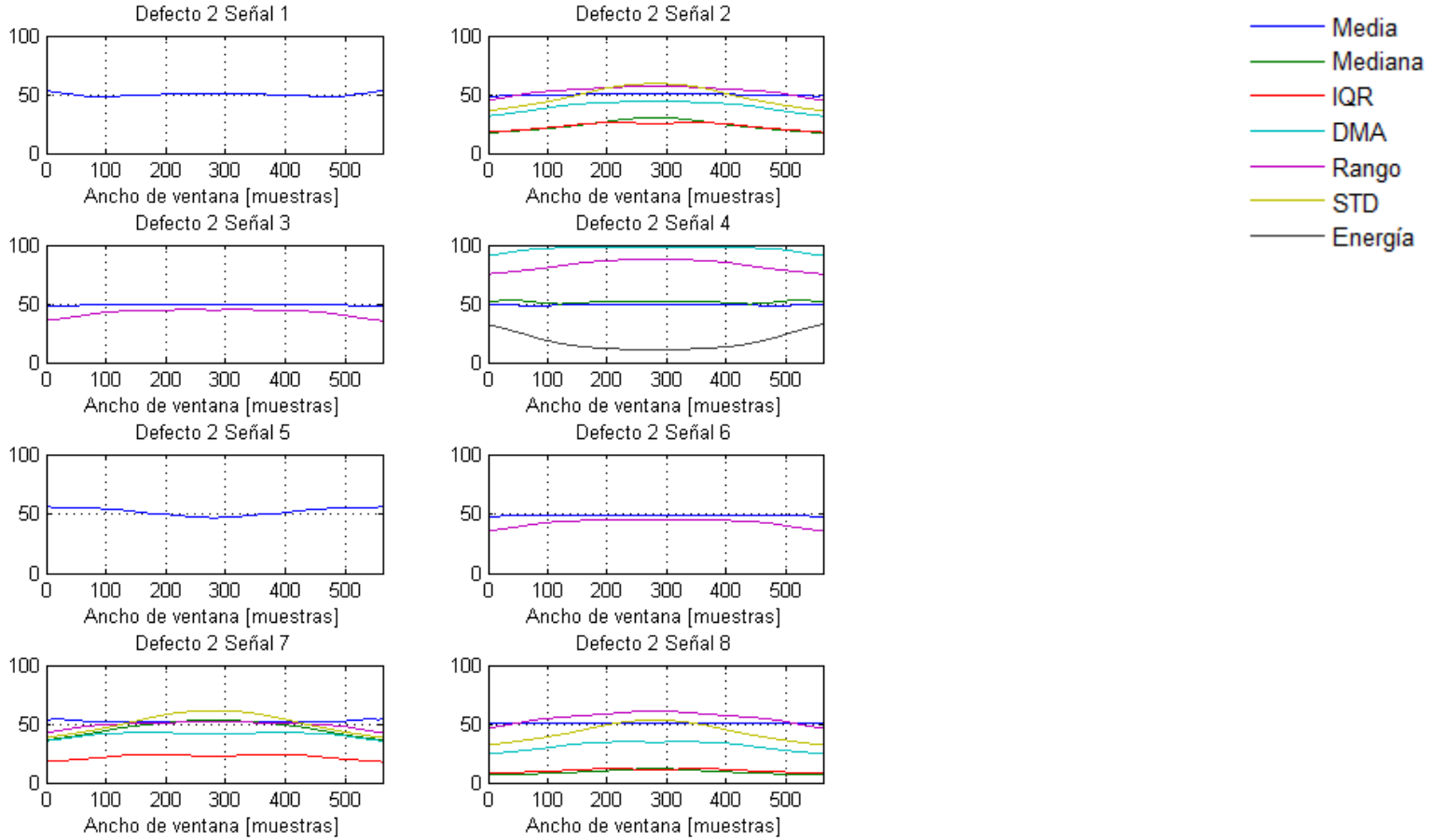


Figura 30. Porcentaje de "1" para el segundo defecto, todas las señales (ADL con datos desbalanceados).

CORRECCIÓN DE FALSOS POSITIVOS

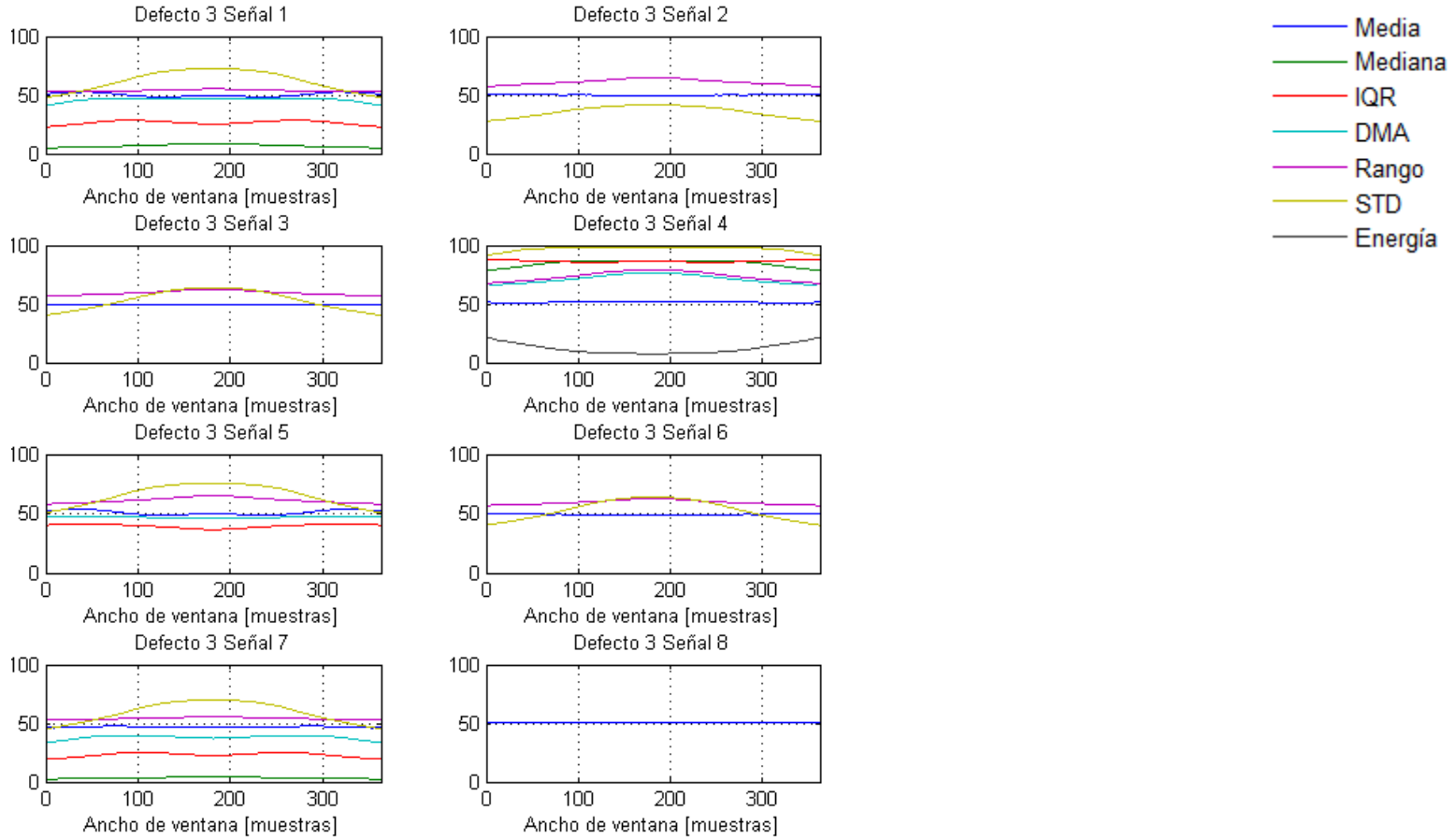


Figura 31. Porcentaje de "1" para el tercerdefecto, todas las señales (ADL con datos desbalanceados).

CORRECCIÓN DE FALSOS POSITIVOS

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.

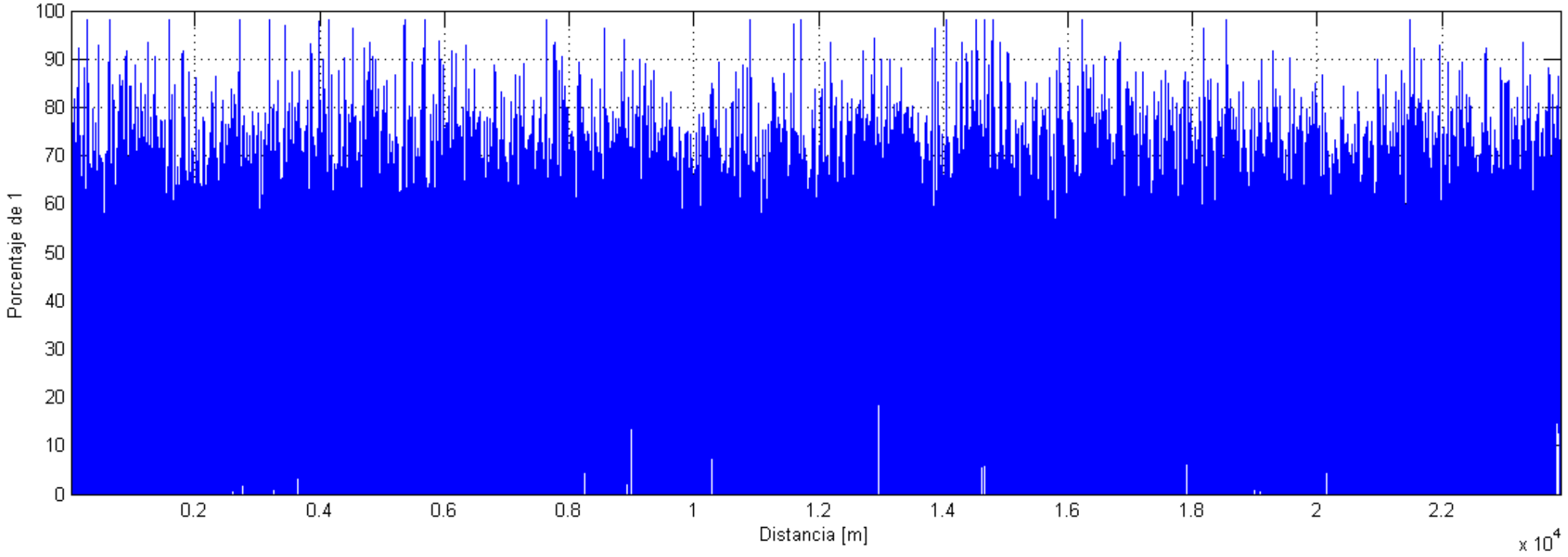


Figura 32. Promedio del ventaneo con los 51 T para el resultado del clasificador ADL, señal 11 de FFM

CORRECCIÓN DE FALSOS POSITIVOS

- Selección ancho de ventana (σ)

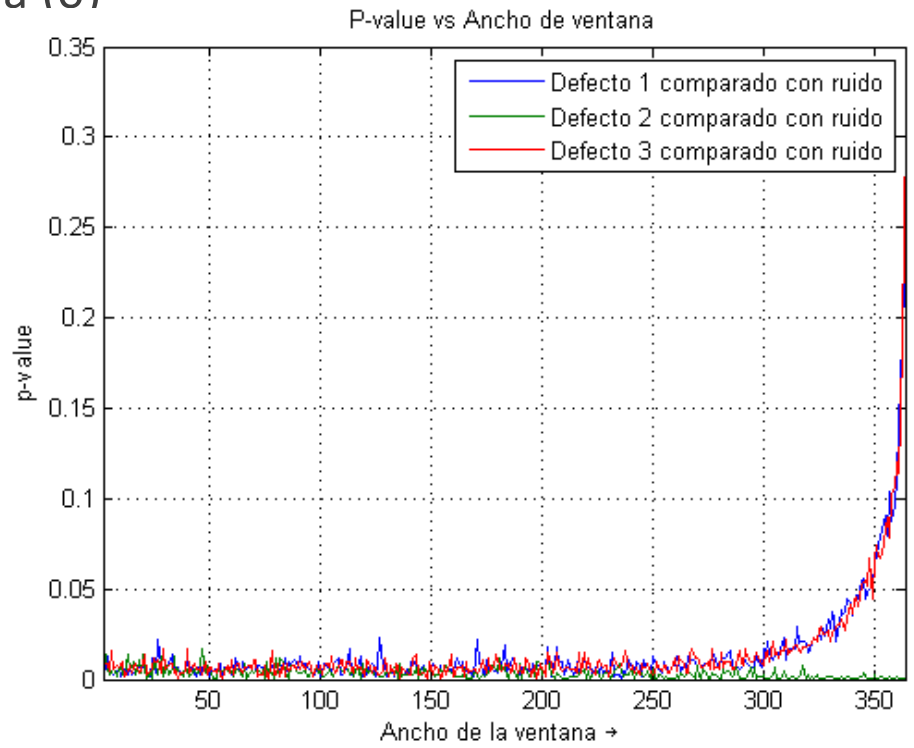


Figura 33. P-Value vs anchos de ventana para los 3 defectos

CORRECCIÓN DE FALSOS POSITIVOS

- **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.

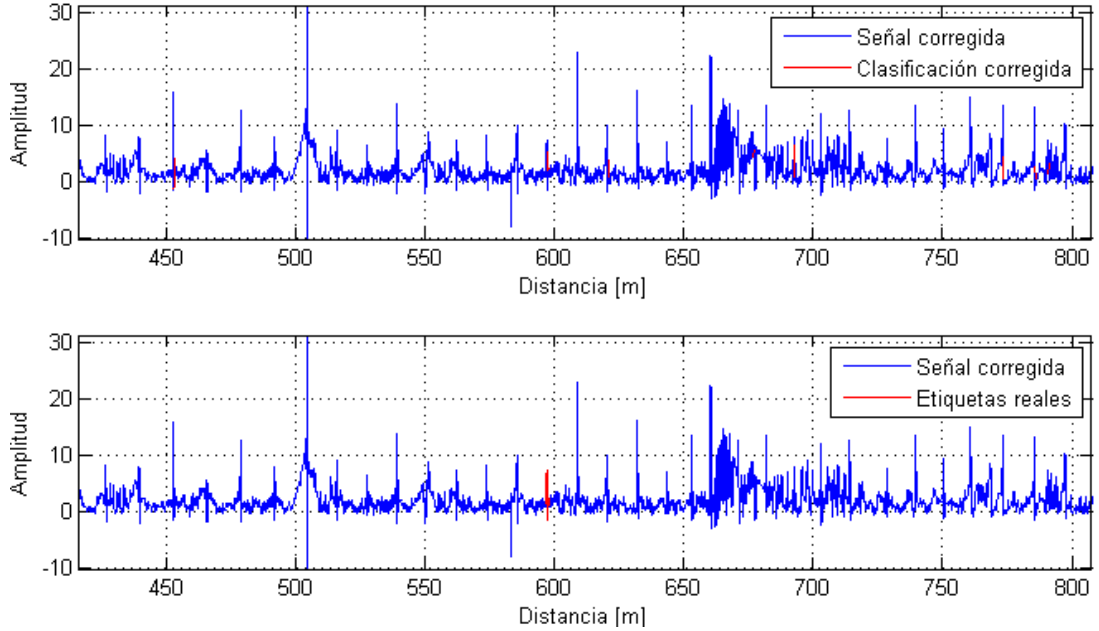


Figura 34. Clasificación corregida, ADL con datos desbalanceados Señal 11 FFM (Defecto 1).

CORRECCIÓN DE FALSOS POSITIVOS

Criterio de corrección

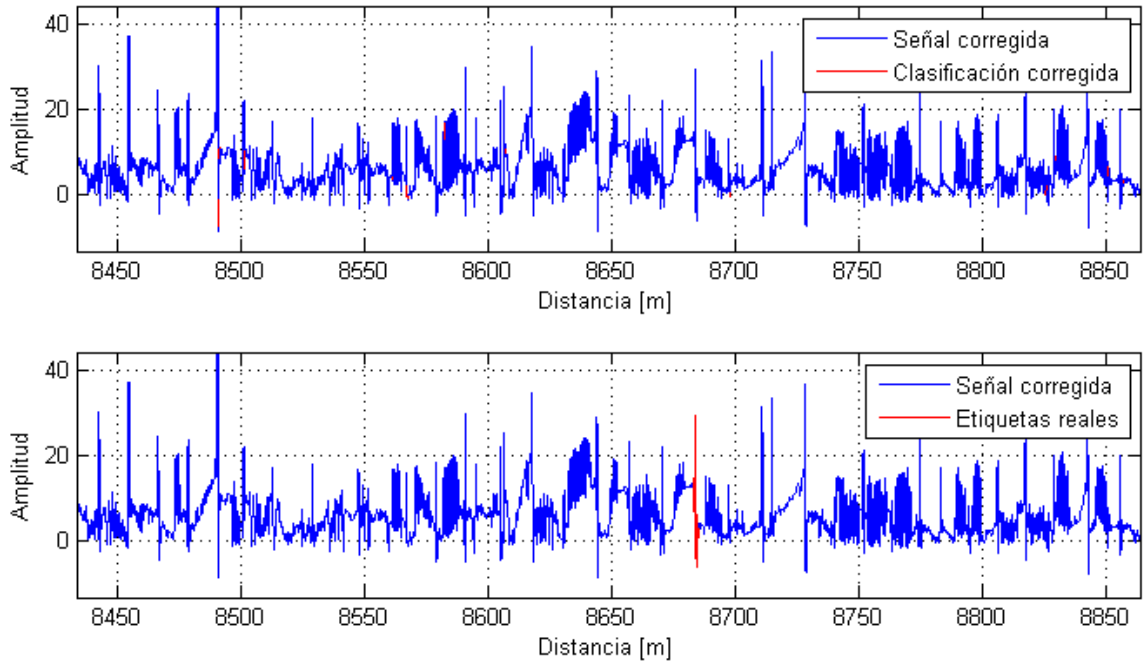


Figura 35. Clasificación corregida, ADL con datos desbalanceados Señal 11 FFM (Defecto 2).

CORRECCIÓN DE FALSOS POSITIVOS

Criterio de corrección

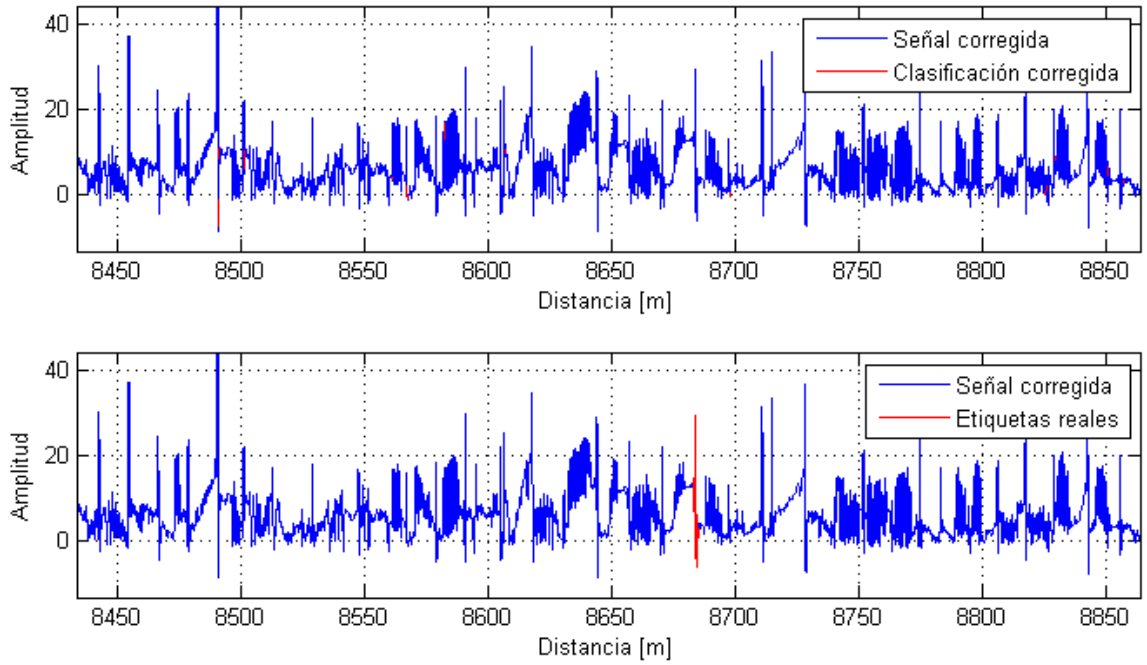


Figura 36. Clasificación corregida, ADL con datos desbalanceados Señal 11 FFM (Defecto 3).

CONCLUSIONES

- 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.