**Evaluaciones**

En este capítulo se detallaran las diferentes evaluaciones realizadas sobre los datos para seleccionar la estrategia optima que maximice la correlacion entre la informacion histórica captada kilómetros adentro en el oceano y las observaciones visuales en la costa y que al mismo tiempo minimice los margenes de error. Para realizar estas evaluaciones utilizamos diferentes algoritmos de regresión (redes neuronales, maquinas de soporte vectorial, regresion lineal, procesos gaussianos) buscando el que mejores resultados generaba.

Básicamente, la información que utilizamos para entrenar a los clasificadores consiste en dos grandes grupos de datos. El primero es un histórico de datos del modelo WW3, situados en cuatro coordenadas alrededor de OAHU a unos cuantos kilómetros de la costa, del cual disponemos de lecturas del comportamiento del océano (periodo de ola, altura de ola, dirección de ola, dirección del viento y velocidad del viento) cada 3 horas durante los 365 dias del año desde 1997 hasta la fecha. El segundo grupo es una base de datos de observaciones visuales de la altura de las olas de diferentes playas alrededor de toda la costa de Oahu desde el año 1987 hasta el año 2004 inclusive. Estas observaciones realizadas por un experto en el area nos brinda por cada dia del año la altura de la ola más grande que arribo a la costa ese dia. Además estas observaciones sabemos (ya que la fuente lo especifica) que fueron realizadas durante el periodo de luz solar de dia en esa playa.

A partir de estos dos grupos de datos, comenzamos a implementar diferentes estrategias de filtrado y utilizacion de los mismos hasta encontrar una que optimice los aciertos en la prediccion de las olas en las diferentes costas de la isla.

**Filtrado basico de la informacion**

Dado que las lecturas de WW3 se actualizan cada 3 horas, tanto durante el dia como la noche, y debido a que las observaciones visuales se han realizado en horarios en que se cuenta con luz solar, el primer filtro aplicado a la informacion que disponemos del modelo WW3 fue eliminar todas las lecturas que no correspondian con un horario de luz natural en Oahu.

Por otra parte, las observaciones realizadas nos indican la altura de la ola más grande que arribo a la costa cada dia del año, es decir que disponemos de una lectura diaria, por esto el segundo filtro aplicado a las lecturas de WW3 fue quedarse unicamente con la lectura diaria con la mayor altura de ola. De esta forma ambos conjuntos de datos cuentan con una lectura diaria que indica la ola más grande captada y durante una hora en Oahu en la que hay luz solar.

Los filtros anteriormente descriptos fueron los que finalmente se utilizaron ya que luego veremos en las evaluaciones que con estos obtuvimos los mejores resultados. Otros filtros que tuvimos en cuenta para realizar pruebas (descartados por no mejorar los resultados) fueron entre otros, utilizar las lecturas diarias del WW3 en el que la altura de la ola se acerca más al promedio de altura de olas de cada dia, filtrar las lecturas cuya dirección de ola no estaba dirigida a la costa a evaluar, agrupar las lecturas en dos grupos (dos rangos de meses) que distinguen la temporada de olas grandes de la temporada de olas pequeñas, etc.

**Entrenamiento del clasificador (o no se que nombre lleva esta sección todo este párrafo esta muy verde)**

Para llevar a cabo el entrenamiento del clasificador??? Y testear los diferentes algoritmos de regresión utilizamos el software WEKA. El cual recibe como entrada un archivo con formato arff el cual contiene un conjunto de instancias.

Cada instancia es un conjunto de variables predictoras (atributos) más uno o más atributos de clase que representan la/las variable/s a predecir. Para entrenar el clasificador el sistema recibe instancias con ambos tipos de atributos con valores ya definidos (utilizando información historica). Una vez que el clasificador es entrenado, debe ser capaz de recibir únicamente instancias con las variables predictoras y devolver por cada una el valor del o los atributos de clase a predecir. En el caso especifico de este análisis, las variables predictoras podrían ser las lecturas del modelo WW3, (altura de ola, periodo, dirección, etc) y el atributo de clase o variable a predecir seria la observación visual en la costa, una vez que tenemos entrenado el clasificador, solo bastaría entregarle instancias con solo los datos del modelo WW3 (el cual es capaz de predecir hasta 180h a futuro), y este debería retornarnos por cada una el valor que tendría la observación visual en la costa, es decir la altura de una ola que rompe en la playa.

En las diferentes evaluaciones a continuación se probarán no solo diferentes tamaños de los conjuntos de instancias (una por día) para el entrenamiento sino también diferentes atributos y cantidades de atributos que conforman cada instancia. Al algoritmo encargado de generar los diferentes conjuntos de instancias los denominamos Estrategia (Strategy).

**Evaluaciones**

Para encontrar una estrategia óptima de predicción que maximice la correlación entre las lecturas del WW3 y las observaciones visuales, y a su vez minimicen los márgenes de error entre una futura predicción de altura de la ola y la altura real que esta alcanzara en la costa, se aplicaron diferentes algoritmos y pre-procesamientos sobre los datos que vamos a ir presentando a continuación, en orden desde la peor estrategia hasta la estrategia escogida como la optima para entrenar al clasificador del sistema final. Siempre teniendo en cuenta que todas las estrategias a continuación utilizan como conjunto de entrada datos ya filtrados como se explicó en la sección anterior (Filtrado básico de la información).

**North shore**

WW3 One Day Strategy

Esta estrategia utiliza los datos especificados anteriormente ya filtrados para poder entrenar el clasificador. Cada instancia del conjunto de datos de entrada esta compuesto por los siguientes atributos:

WW3 Model:

* Altura de ola (metros)
* Periodo de ola (segundos)
* Dirección de ola (grados)

Observación visual:

* Altura de ola (metros)

Cada instancia corresponde a un determinado día, y no tiene en cuenta lo que haya sucedido días atrás.

Las lecturas del modelo WW3 se tomaron de una ubicación en frente a la costa norte de Oahu, aproximadamente a 87 kilómetros de la misma. La Fig. 4-1 muestra la ubicación donde se tomaron las lecturas del modelo mar adentro y al mismo tiempo la ubicación en la costa del observador.

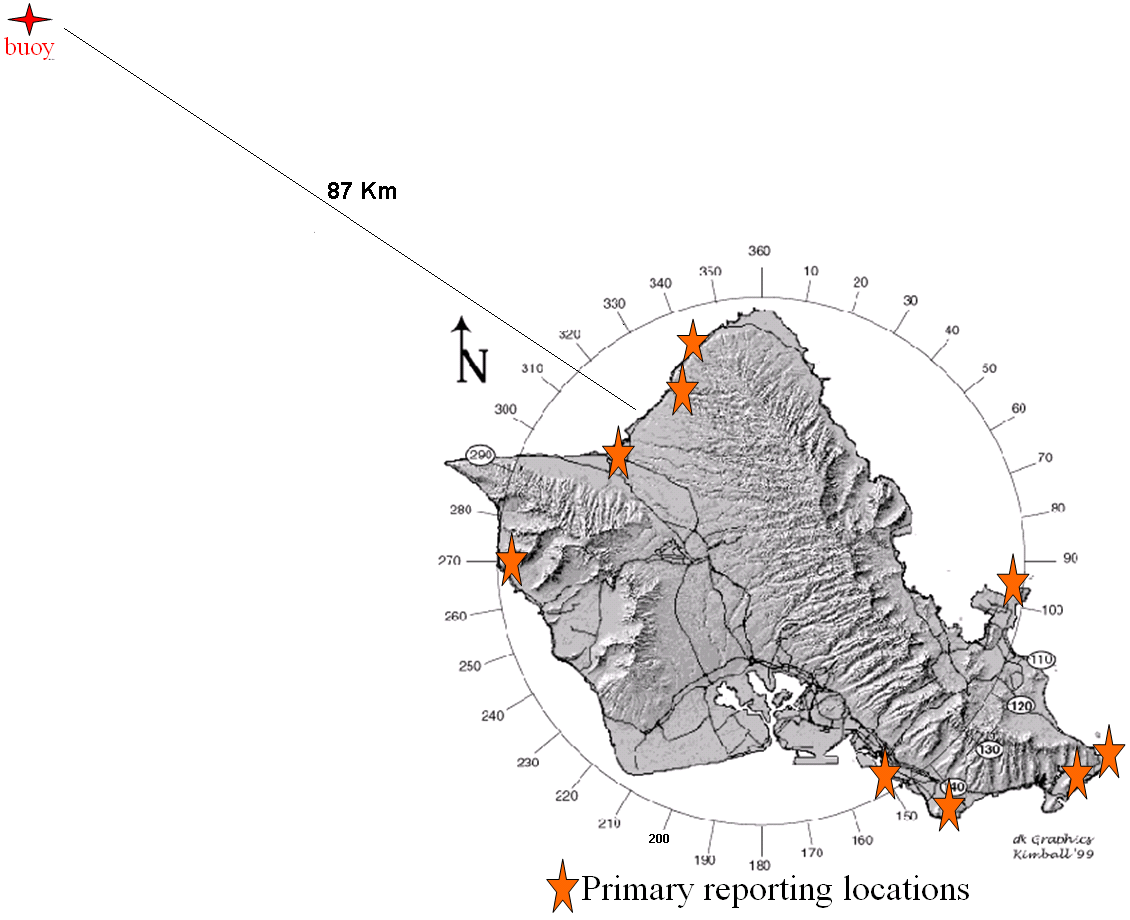


Fig. 4-1

**Prueba 1**

Para esta prueba se utilizaron los datos desde el 01-01-2003 hasta el 30-09-2004. Con los que se generaron 638 instancias.

**Sin aplicar ningún algoritmo de regresión se han obtenido los siguientes resultados:**

* Coeficiente de correlación 0.8217
* Error absoluto medio 0.88

La Fig. 4-2 a continuación muestra gráficamente el valor que reporto el modelo WW3 para la altura de la ola más grande, y la altura de la mayor ola captada en la costa el mismo día para una muestra de 63 instancias al azar del conjunto de datos original.

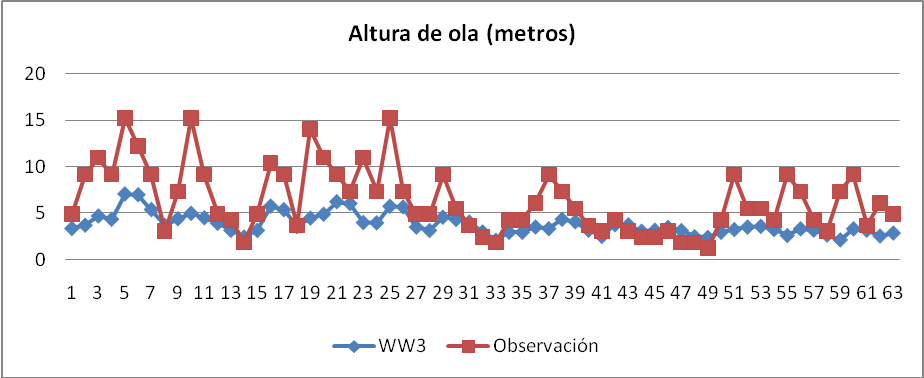


Fig. 4-2

La figura anterior nos demuestra que si bien la correlacion de los datos es bastante buena, los errores son importantes y no asegura un modelo confiable de prediccion del oleaje.

Por otra parte, en la siguiente imagen, Fig. 4-3, vemos una tabla de dispersión que relaciona los reportes del modelo con las observaciones costeras, brindando mejor visibilidad del margen de error entre las observaciones y el reporte WW3.

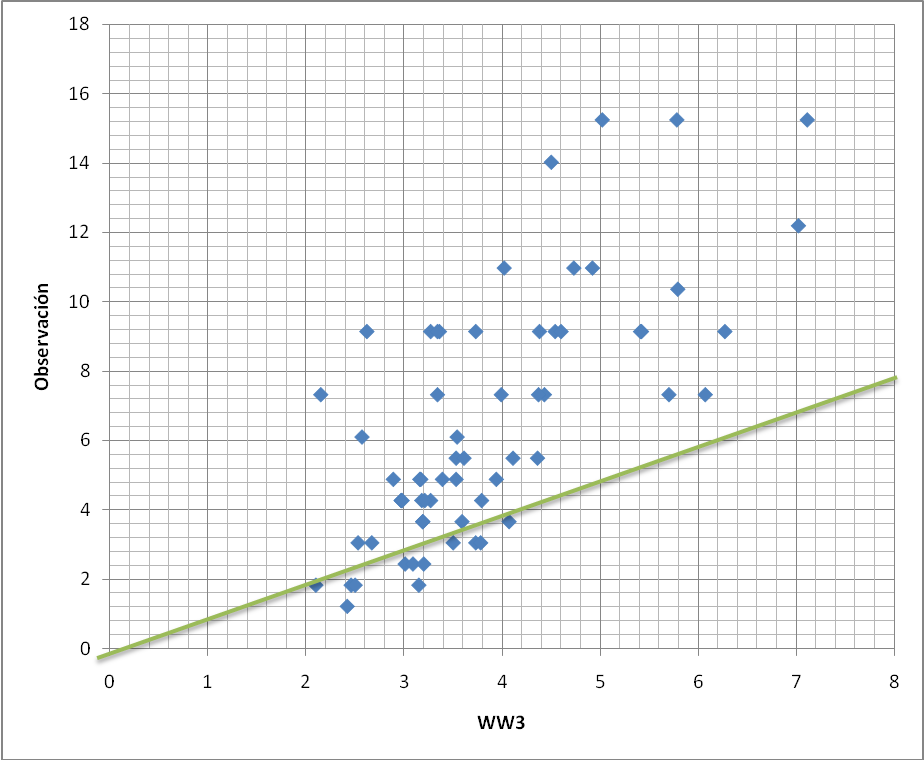


Fig. 4-3

Teniendo en cuenta que la linea verde representa los valores ideales, en que los reportes del modelo son identicos a los sucedido en la costa, vemos que los errores son frecuentes y con un margen grande por lo que no sería bueno despreciarlo.

Con el fin de mejorar los resultados anteriores se entreno el clasificador con los siguientes algoritmos: Regresión Lineal, Redes Neuronales y Máquinas de Soporte Vectorial obteniendo los siguientes resultados:

**Regresión Lineal:**

* Coeficiente de correlación 0.8586
* Error absoluto medio 0.9799
* Error absoluto cuadrático medio 1.3716

En base a los datos anteriores vemos como la correlación de los datos es del 85%, lo cual es muy buena, y el error absoluto medio es de 98 cm, lo cual no es abultado en olas de alrededor de 7 u 8 metros en adelante, pero si es malo en olas de menor tamaño.

La imagen a continuación, Fig. 4-4 muestra nuevamente en un gráfico de líneas la relación entre la altura reportada por el modelo y la observación visual en la costa, bajo el algoritmo de regresión:

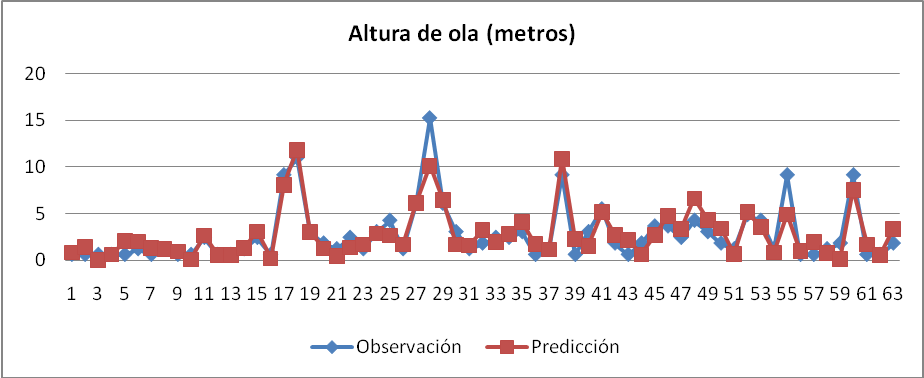


Fig. 4-4

Claramente se puede apreciar ahora como han disminuido los márgenes de error, obteniendo dos líneas que son muy similares, por lo que en este caso podemos ver que la regresión lineal a mejorado mucho a la utilización de los datos en crudo. En la imagen siguiente, Fig. 4-5, usaremos nuevamente una gráfica de dispersión para reforzar esta mejoría con el algoritmo.

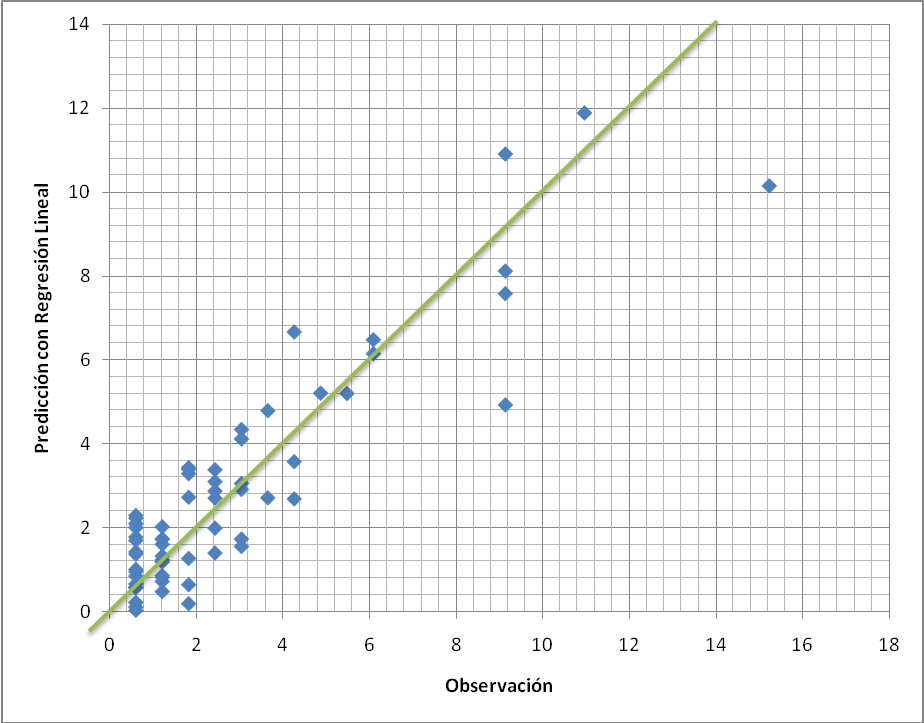


Fig. 4-5

Nuevamente podemos ver como al usar regresión lineal para entrenar al clasificador, las predicciones en relación con las observaciones se acercan bastante al ideal (linea verde), mucho más confiables generando márgenes de error mucho menores.

**Red Neuronal (parámetros por defecto):**

* Coeficiente de correlación 0.8187
* Error absoluto medio 1.1522
* Error absoluto cuadrático medio 1.5608

En este caso la correlación de los datos es de aproximadamente el 82%, lo cual es muy buena, pero ha disminuido un poco en comparación con el algoritmo de regresión lineal, el error absoluto medio es de 115 cm, lo cual también ha empeorado.

Gráficamente:

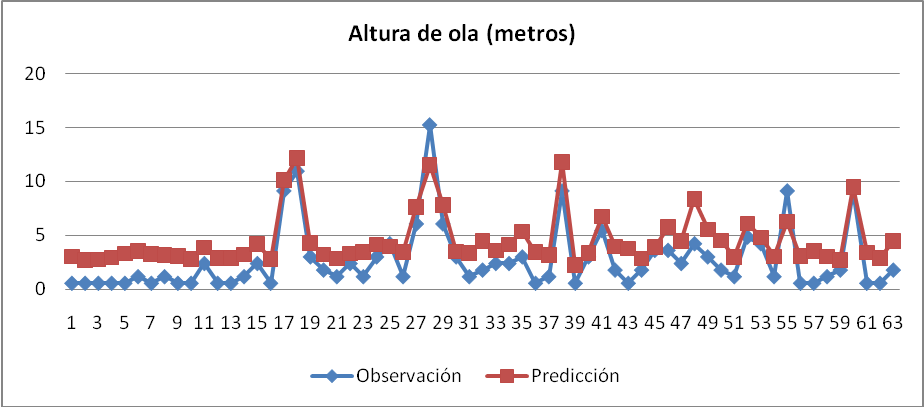


Fig. 4-6

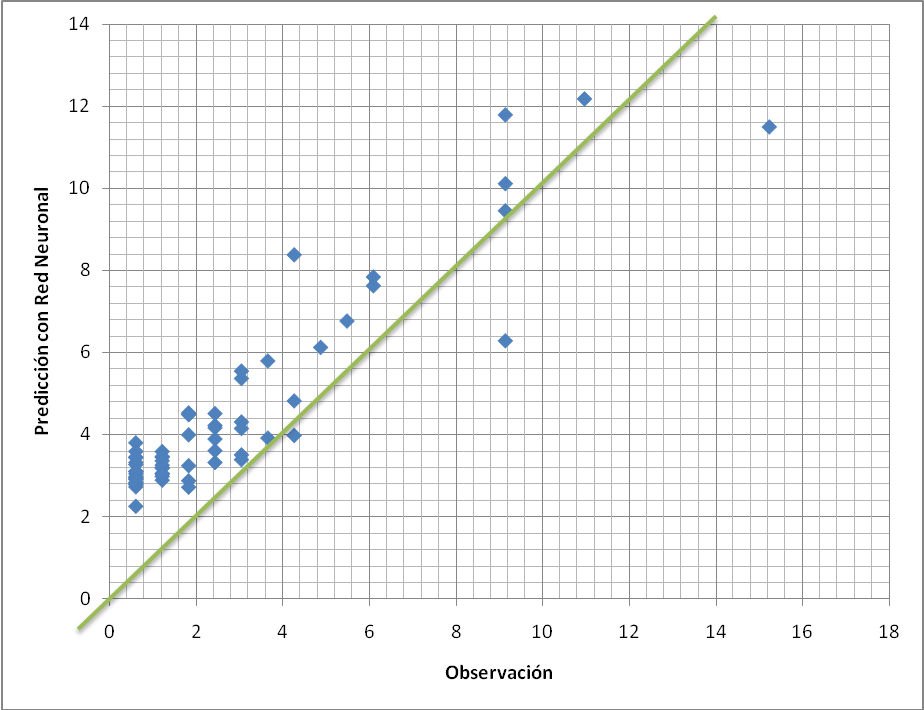


Fig. 4-7

Si comparamos estos resultados gráficos con el algoritmo anterior de regresión lineal se nota claramente que la prediccion ha empeorado un poco pero al mismo tiempo ha mejorado en ampliamente con respecto a intentar predecir sin utilizar algoritmos de regresion. En la segunda gráfica vemos como las prediciones se acercan bastante al ideal, pero al mismo tiempo se muestran un poco mas desfazadas que con la utilización de regresión lineal.

**Red Neuronal (parámetros óptimos):**

* Coeficiente de correlación 0.8838
* Error absoluto medio 0.8662
* Error absoluto cuadrático medio 1.2519

En este caso la correlación de los datos es de aproximadamente el 88,4%, también muy buena, mejorando todos las evaluaciones previas, el error absoluto medio es de 86,6 cm, lo cual ha disminuido casi 12 cm en comparación con la regresión lineal, este algoritmo ha dado resultados muy confiables en la evaluación de esta costa de la isla.

Gráficamente:

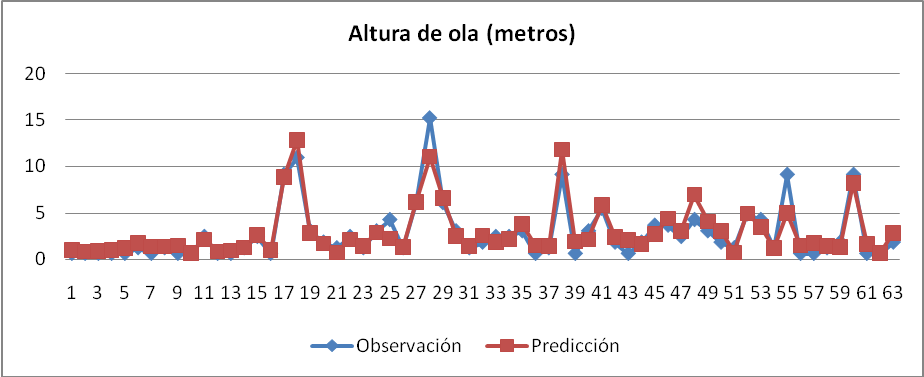


Fig. 4-8

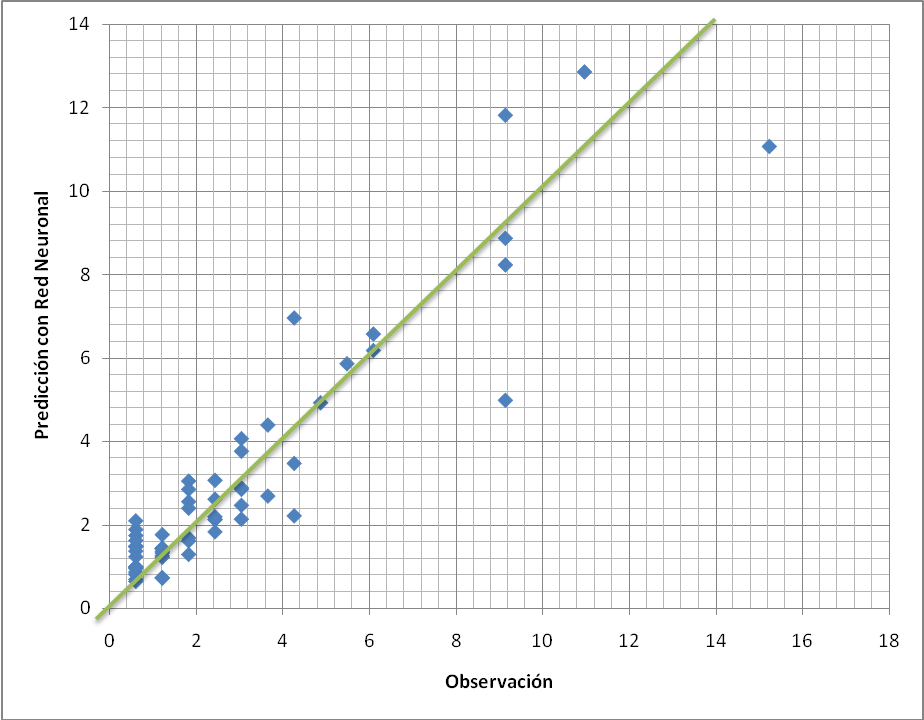


Fig. 4-9

Son claras las mejoras con respecto al algoritmo de regresión lineal en ambas graficas, en la primera la linea roja que representa a las predicciones se superpone practicamente sobre la linea azul de observaciones, y vemos como en la tabla de dispersión los datos se acercan todavía un poco más al ideal, asi mismo, la mejoria con respecto a la red neuronal sin parametrizar es notoria.

**Máquina de soporte vectorial (parámetros por defecto):**

* Coeficiente de correlación 0.8588
* Error absoluto medio 0.9634
* Error absoluto cuadrático medio 1.3969

Los resultados en este caso son muy similares al algoritmo de regresión lineal, un poco por debajo del algoritmo anterior.

Gráficamente:

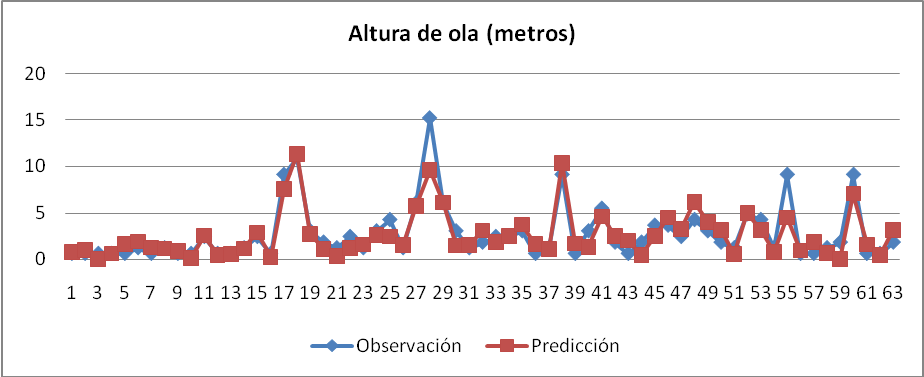


Fig. 4-10

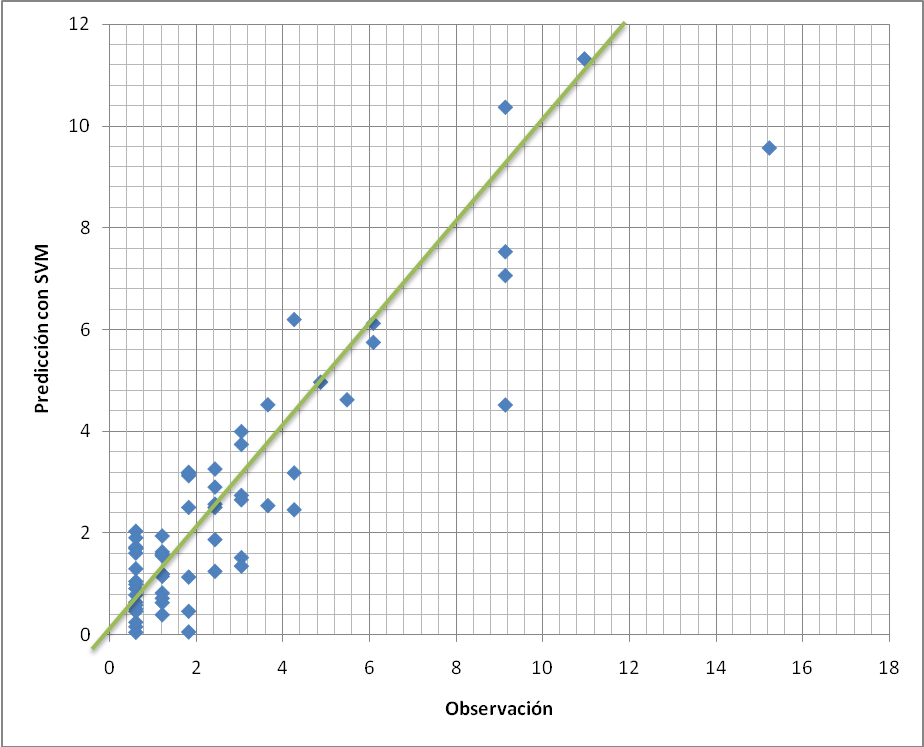


Fig. 4-11

Presenta gráficas muy similares a las del algoritmo de regresión lineal, al igual que los resultados numéricos, los cuales son muy buenos.

**Máquina de soporte vectorial (parámetros óptimos):**

* Coeficiente de correlación 0.8873
* Error absoluto medio 0.8259
* Error absoluto cuadrático medio 1.2527

En este caso, con una máquina óptimamente configurada obtuvimos los mejores resultados con esta estrategia para los años 2003 y 2004 en el North Shore (costa norte de Oahu). Por muy poco mejora a la red neuronal configurada.

Gráficamente:

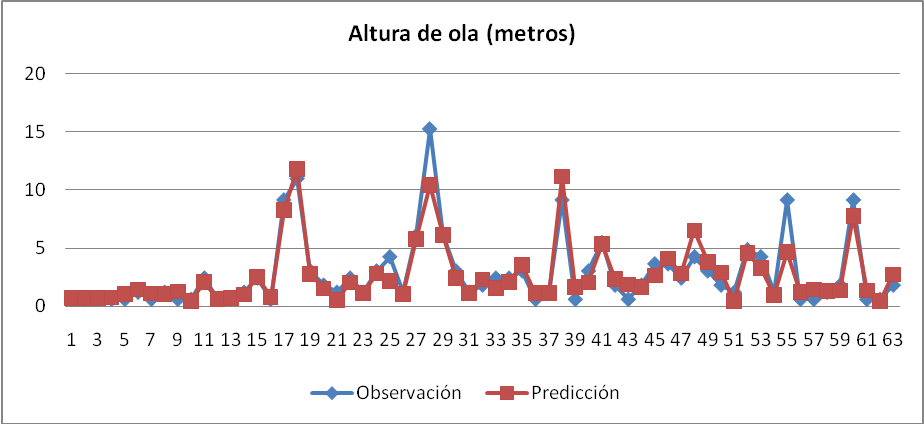


Fig. 4-12

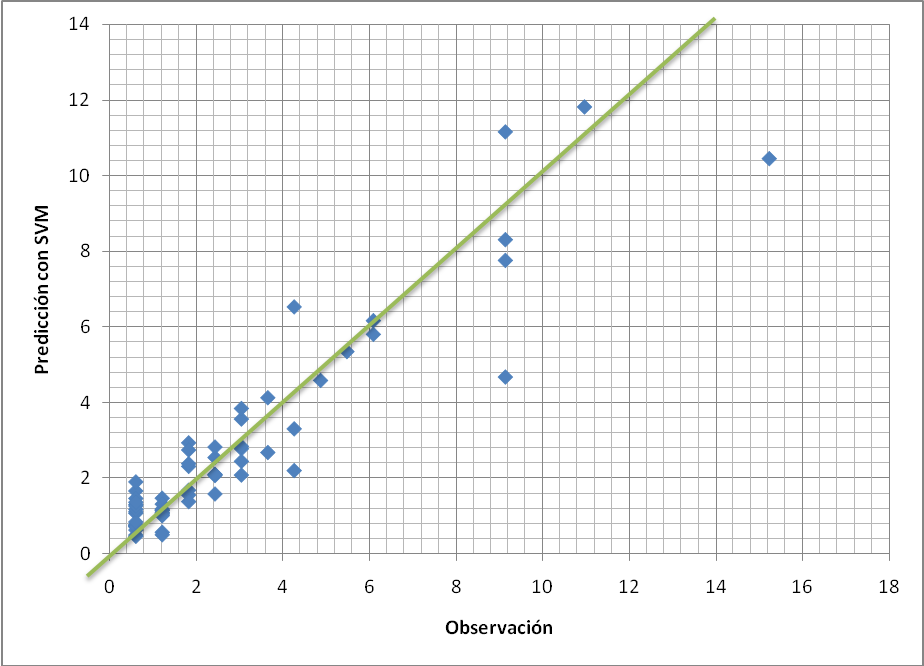


Fig. 4-13

Estas mejoras también se hacen presentes en el gráfico de dispersión donde vemos que las predicciónes se acercan bastante al ideal, salvo en algunos casos aislados que fallaron de forma similar en la mayoría de los algoritmos y configuraciones. Estos pocos casos parecieran ser los que definen el grado de error medio y cuadrático de los datos.