DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Metodos Numericos
Trabajo Practico 3
Cuadrados Minimos Lineales

Alvarez Mon Alicia

aliciaysuerte@gmail.com
LU.: 224/15,
FCEN, UBA,
CABA, Argentina

Ansaldi Nicolas

nansaldi611@gmail.com
LU.: 128/14,
FCEN, UBA,
CABA, Argentina

Castro Luis

castroluis1694@gmail.com,
LU.: 422/14,
FCEN, UBA,
CABA, Argentina

Suarez Romina

romi_de_munro@hotmail.com,
LU.: 182/14,
FCEN, UBA,
CABA, Argentina

Resumen

En este trabajo buscamos extraer información relevante de vuelos en EEUU y usarla para aproximar el comportamiento de los mismos en el futuro, para esto usaremos regresión lineal por cuadrados mínimos. Además buscamos ver las causas que producen las demoras en los vuelos y analizarlas para determinar los factores influyentes en las mismas.

Keywords: CML, RMSE, OTP, Vuelos.

---

1. Introducción

En este trabajo práctico utilizaremos los datos de los vuelos realizados en Estados Unidos, entre 1987 a 2008, y los analizaremos para intentar predecir el comportamiento a futuro de los vuelos. Como ejes centrales tomamos el OTP (on-time performance)[1] que es una métrica que describe el desempeño del servicio, la idea con este eje es ver la puntualidad a lo largo del tiempo para poder tener una visión más clara de las variables que afectan a los vuelos, ya sea por cosas propias a ellos como ajenas. Y, como segundo eje, las cancelaciones de los vuelos a lo largo del tiempo, ya que estas nos permiten ver cuales son las épocas del año más conflictivas para los vuelos, como para ver períodos donde sucesos ajenos a los mismos producen un cambio significativo en su actividad.

2. Desarrollo

Como dijimos la idea de este trabajo es predecir datos de vuelos. Para esto tomaremos 2 ejes principales, el primero es OTP que representa el desempeño de un vuelo medido en delay o retardo y el segundo es sobre como las cancelaciones afectan a los mismos. La forma de ver si un vuelo tiene un buen desempeño es comparar el horario efectivo de partida contra el horario previsto de partida y el horario efectivo de llegada contra el horario previsto de llegada; si alguna de las 2 comparaciones da un retardo de 15 o más minutos se considera que el vuelo tuvo retraso con lo cual no tuvo un buen desempeño. En cuanto a las cancelaciones nos interesa ver como estas varían en el tiempo.

2.1. Nested Cross-validation

Nosotros implementamos el Cross-validation[2] para verificar la exactitud que tiene nuestro programa al aproximar con las muestras de entrenamiento, evitando overfittear los datos. El problema es que no toda muestra de entrenamiento se puede usar para predecir ya que las mismas pueden tener problemas con la dependencia temporal. En esta versión tomamos una partición de tiempo anterior al tiempo que queremos predecir y predecimos la primera partición. En la siguiente iteración, hacemos lo mismo con la siguiente partición pero la nueva muestra de entrenamiento pasa a ser la que teníamos antes mas la partición que predijimos anteriormente. Hacemos esto hasta que ya no queden muestras para predecir.

2.2. Máscara

Para encontrar los valores más relevantes utilizamos los cuartiles de las muestras. Dada la muestra ordenada nos quedamos con el cuartil 0.25 y 0.75 al cual llamamos q1 y q3 respectivamente, calculamos la distancia intercuartil como d = q3-q1 y nos quedamos con los valores

que estén en el intervalo [q1-d*1.5 , q3+d*1.5] ya que consideramos que todo lo que este afuera de ese intervalo es un valor no representativo para nuestra muestra. La idea de usar cuartiles en lugar del desvío estándar es que estos son más estables que el desvío ya que no dependen de los valores de la muestra en sí, sino a la cantidad de los mismo[5].

2.3. Cuadrados Mínimos Lineales

Para poder predecir el comportamiento de los datos utilizamos distintos familias de funciones, aplicándole ecuaciones normales. Entre ellas elegimos funciones polinomiales, trigonométricas y la unión de ambas a la cual vamos a llamar fusión. También implementamos la función predecir, que junto al cross-validation, se encargan de auto-configurar el programa para que este dé como resultado el menor RMSE[4] posible, eligiendo el mejor grado del polinomio o función trigonométrica. Analizamos el grado de 0 a 20 para polinomios y de 0 a 2 para las funciones trigonométricas (éste último siempre multiplicado por pi).

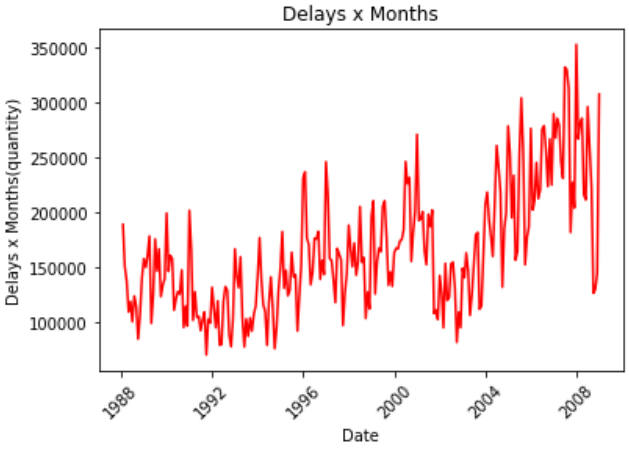
3. Experimentación

En las siguientes secciones descartaremos el año 1987 debido a que solo tiene registros en los últimos 4 meses del año.

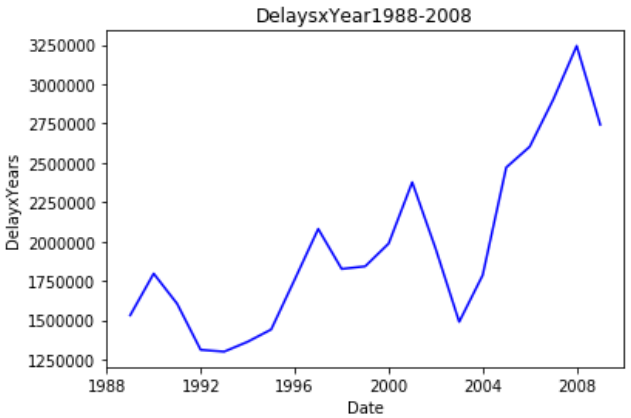
3.1. Eje 1: On-Time Performance

Para calcular la cantidad de delay, o demora, sumamos la cantidad de vuelos que tienen delay en su partida y en su arribo. Si ambas tienen demora, las sumamos.

3.1.1. Función de aproximación



(a) Delays por mes de 1988-2008

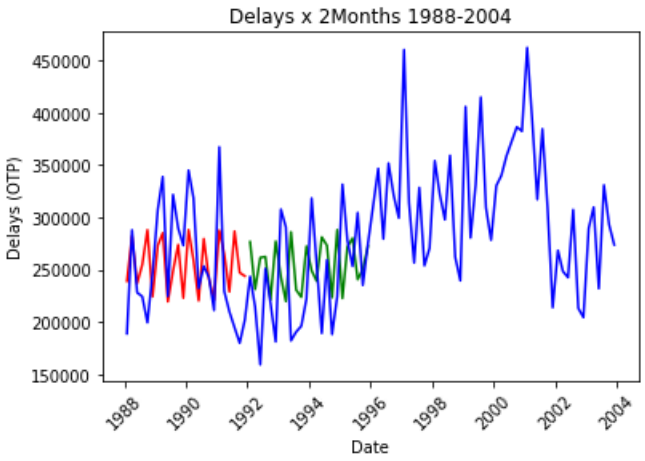


(b) Delays por año de 1988-2008

Esta es nuestra distribución inicial. En esta figura graficamos la cantidad de delays por cada mes, de 1988 a 2008. Vemos que la cantidad de picos es muy grande, por lo que para ver mejor, también graficamos en función de años. Tratamos de fittear la anual a una polinomial y aunque el error de fitteo o RMSE es poco, el error de predicción es muy grande. Lo mismo ocurre con nuestras funciones trigonométricas y de fusión (aunque esta última da mejor).

Viendo la tendencia(comportamiento a largo plazo), que se puede observar mejor en los Delays por año, vemos que la gráfica tiende a tener períodos de tiempo donde oscila con cierto comportamiento parecido a un función seno (por ejemplo, entre 1988 y 1996, o entre 2001 y 2008).

Por lo tanto, tratamos de predecir usando el modelo trigonométrico y entre los años 1988-1992 y los años 1993-1996. Como por años tenemos pocos datos y por meses hay muchos picos, vemos que ocurre si tomamos los datos usando una frecuencia de 2 meses.



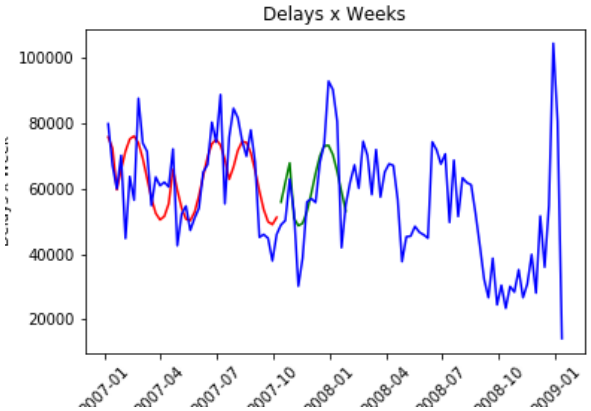
(a) Delays por bimestre 1988-2004

En este gráfico obtuvimos un k = 0.25, RMSE = 347514.04 y un Error de predicción = 3431951334.73: Como podemos ver, los picos nos molestan un poco la predicción, pero el comportamiento es muy parecido. Tratar de predecir más allá de 1996 con esta función nos va a generar mucho más error, lo cuál se condice con la hipótesis de que el comportamiento cíclico cambia a largo plazo.

3.1.2. Función de aproximación en Semanas

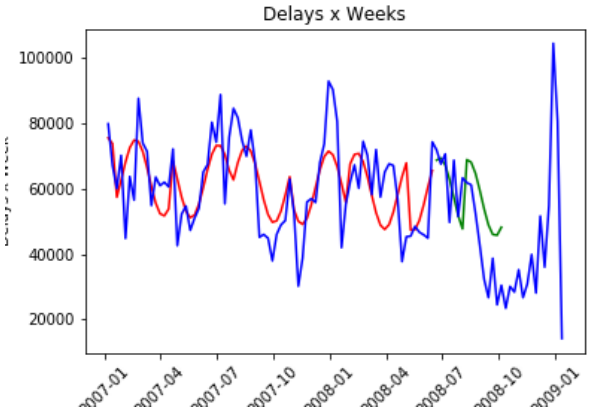
Como vimos anteriormente, ver los gráficos en el mismo período nos genera un gráfico con aún más picos y ruido, por lo que nos interesa usar semanas para ver comportamientos de períodos de tiempo más cortos.

También, habiendo visto que el modelo polinomial es peor para modelar este modelo, nos centraremos en el modelo trigonométrico y el de fusión de ambas. Analizaremos los periodos de 2007-2008 y entrenaremos con los próximos 4 meses el siguiente, con valores promediados por cross-validation:



(a) Delays por semana 2007-2008

(b) Error de predicción: 129204739.98



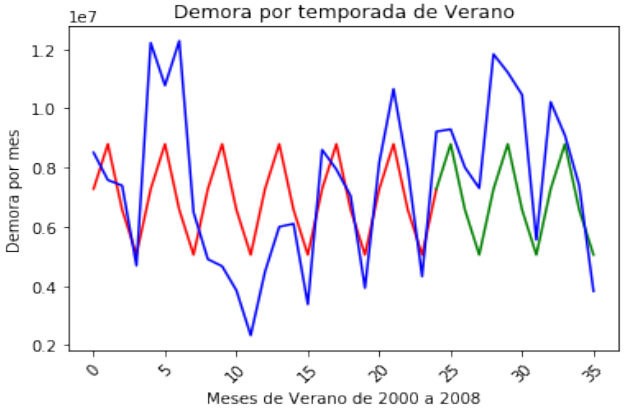
(c) Delays por semana 2007-2008

(d) Error de predicción:179328091

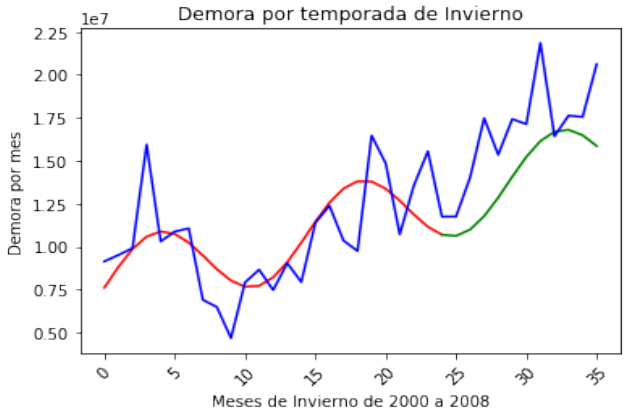
Con estos parámetros, obtenemos un k = 1.93 y un RMSE = 83601.97 (Observamos que hay mejores curvas de fitteo, además, el mejor fitteo se da con un fusion e grado 1 (por la caída que hay que culmina en una caída en el 2008). Hay funciones de fitteo que dan un rsme maś bajo, pero a costa de errores de predicción de fiteo más altos Podemos notar que hay una caída profunda en los meses finales del 2008, tal vez debido a la crisis económica en ese momento.[3] En particular, dado el brusco cambio vemos que nuestro error aumenta si usamos más datos para tratar de predecirlo, como, por ejemplo, cuando usamos 10 meses para predecir los meses futuros.

3.1.3. Demora según temporada climática

Para este experimento quisimos ver que ocurría con la demora de los vuelos según si los mismos se producían en temporada de Verano o de Invierno, siendo nuestra hipótesis que habría más demora en invierno que en verano, debido a las tormentas de nieve y sus complicaciones. Para ambos gráficos utilizamos una función fusión entre trigonométrica y polinómica, entrenando desde 2000 al 2006 y prediciendo desde 2007 al 2008, siendo el eje x de los gráficos, los meses de la temporada a través de los años (4 meses * 9 años = 36 entradas) y el eje de y representa la suma de la demora(salida o llegada mayor a 15 min) de estos meses en un año:



La cantidad de vuelos de Verano analizados fueron 5,535,787, comprendidos en los meses de junio a septiembre y el Error de predicción del gráfico de Verano fue de 5,339,564,495,296.95.



Luego para invierno, la cantidad de Vuelos fue de 5,720,531, comprendidos entre diciembre y marzo y el error de predicción del gráfico fue de 10,201,836,200,868.53.

Suponiendo que la diferencia de cantidad de vuelos a analizar no afecta al análisis, debido a vuelos cancelados, podemos ver entre los dos gráficos que la hipótesis se cumple, dado que la demora de los vuelos en invierno es mayoritariamente más elevada que la de verano. Lo que podemos observar al mismo tiempo, es que la función en verano aproxima mejor que en invierno, debido a que la temporada de verano parece ser más estable y cíclica que invierno.

3.2. Eje 2: Cancelaciones

En esta sección vamos a analizar las cancelaciones en el tiempo para determinar el desempeño de los vuelos en general, teniendo como hipótesis que el incidente del 09/11 provocará muchas cancelaciones. Primero, graficamos por meses las cancelaciones:

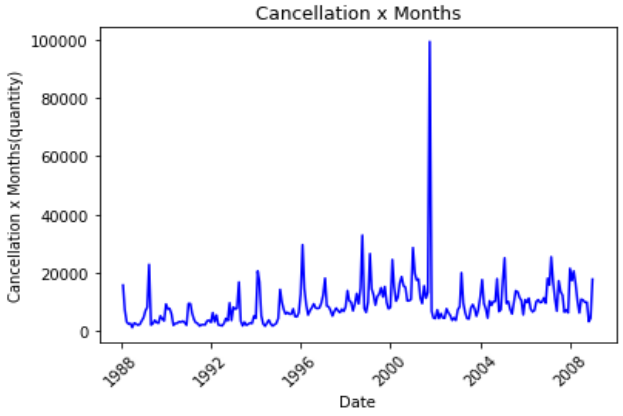
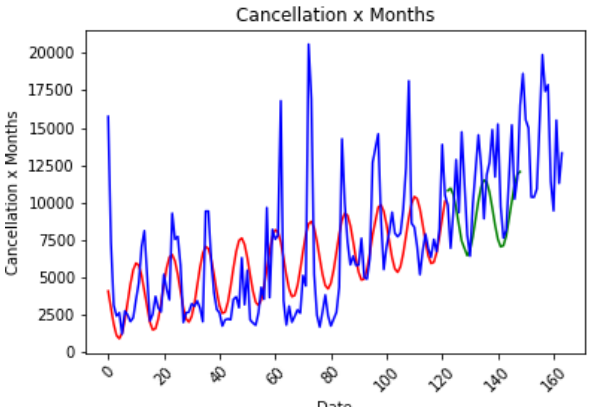


Figura 4. Vista general de cancelaciones entre 1988-2008

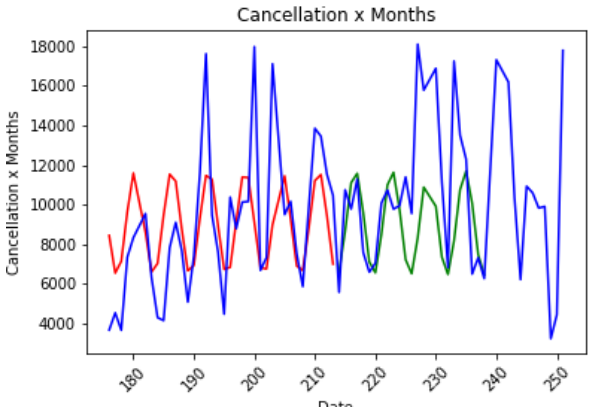
Como podemos ver, hay un pico en el 2001 dado que aumentan las cancelaciones por el incidente. Viendo el gráfico podemos ver que nos conviene usar funciones periódicas para aproximar su comportamiento. Además, cualquier predicción que tome el punto del 2001 nos va a causar mucho error, por lo que conviene ver antes y después del 2001.

Nos damos cuenta de que parece que las cancelaciones siguen un período cíclico determinado con una tendencia definida, por lo que decidimos que tal vez un modelo de función compartida entre trigonométrica y polinómica nos daría mejor resultado, donde f = a.sen(x*k) + b.cos(x*k)

+ d + poli, donde poli es una función polinómica. Notamos además que el polinomio no debería superar el grado 3, porque representa la tendencia del comportamiento cíclico.(aumenta linealmente en el tiempo). Luego intentamos ver el grado de mejor ajuste de esta función fusionada:



(a) Función fusión antes del 9/11



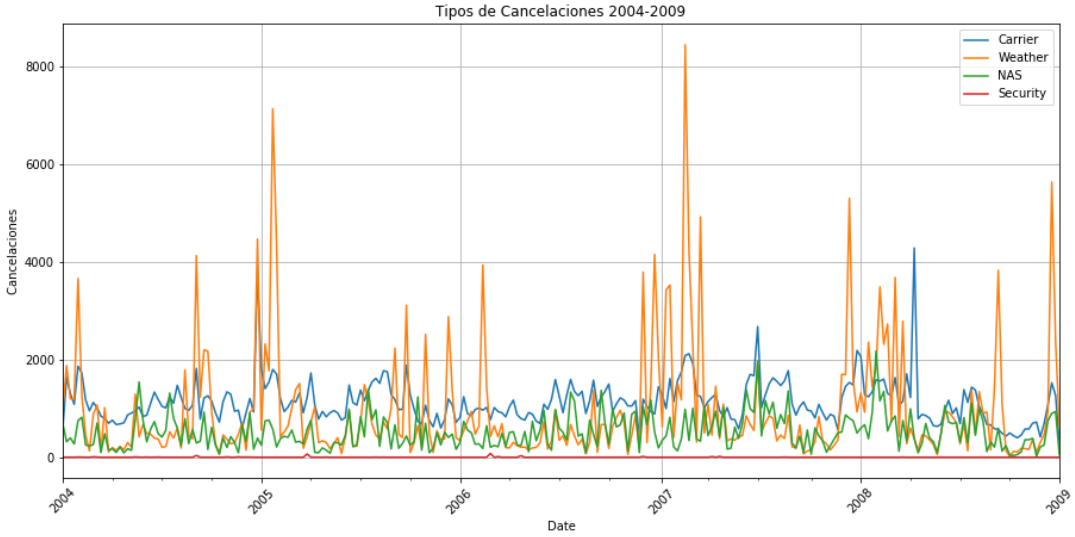
(b) Función fusión después de noviembre de 2002

En el período 1988-2001 (antes del 9/11), si entrenamos por 5 años para predecir los siguientes 2, desde los últimos días del 2001, obteniendo que el mejor grado polinómico era una recta, el mejor k = 0.58, el RMSE = 21864.62 y el error de predicción = 11,395,325.53. parece que la mejor predicción es usando una función fusión con una recta no decreciente, pero luego del 2001, el incremento cíclico se estanca, y los ciclos en sí mismos tienen más amplitud que en el período del 2001.

Decidimos tomar desde el año siguiente (nov 2002) y la mejor predicción se da en cambio con una función fusión de grado 0 (trigonométrica pura), donde k = 0.10, el RMSE = 15110.32 y el Error de predicción: 13,724,212.79.

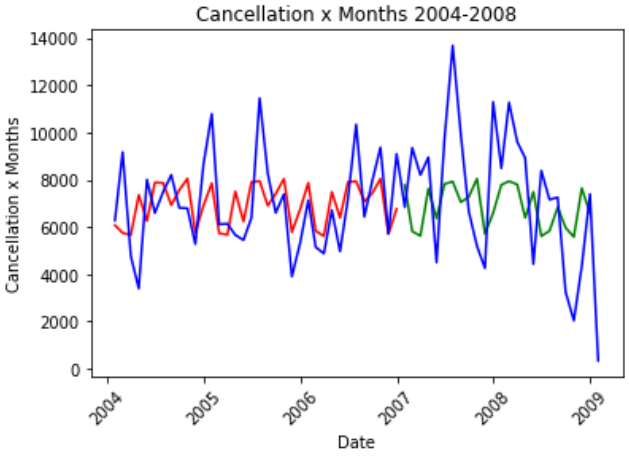
3.2.1. Causas de Cancelaciones

Desde 2004-2008, comienza a haber registros de las causas de las cancelaciones en los datos.Por lo tanto, hicimos un gráfico modelando esos años (por semana) y cómo se distribuyen las mismas, obteniéndose:



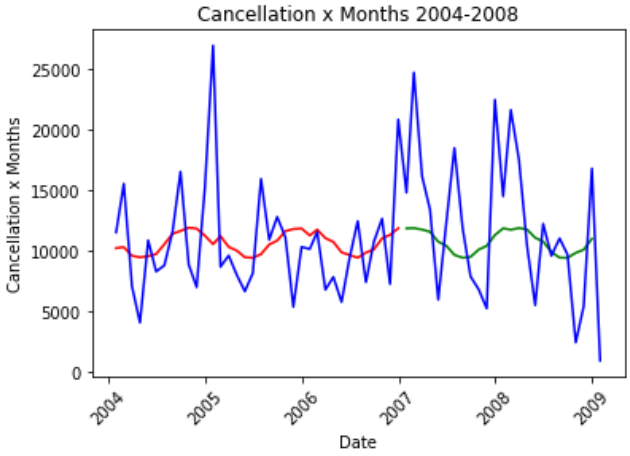
Como podemos observar, los picos de cancelaciones más elevados ocurren debido al clima. En particular hubo uno muy elevado en 2007, posiblemente debido a la Temporada de huracanes en el Atlántico en ese mismo año [6]. Si usamos las cancelaciones sin contar las que se deben al clima que es el más afectado por el efecto estacional, el resto de las cancelaciones debería poder ser fiteado mejor por una trigonométrica.

Debido a que por semana muestra muchos picos, utilizamos meses de medida y fiteamos viendo con cuadrados mínimos y validando con cross-validation una función de fusión, y comparamos con hacer lo mismo agregando las cancelaciones de clima. Donde la de menor error nos queda de grado 0 (trigonométrica) y prediciendo con 3 años, los próximos 2. Queda:



(a) Cancelaciones sin clima 2004-2008

(b) k = 1.47, RMSE = 10082.21
Error de predicción = 7,276,537.64



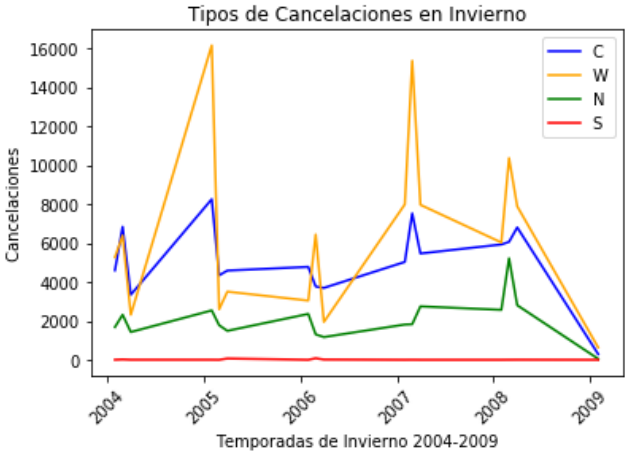
(c) Cancelaciones con clima 2004-2008

(d) k = 1.59, RMSE = 21476.94
Error de predicción = 31,378,081.95

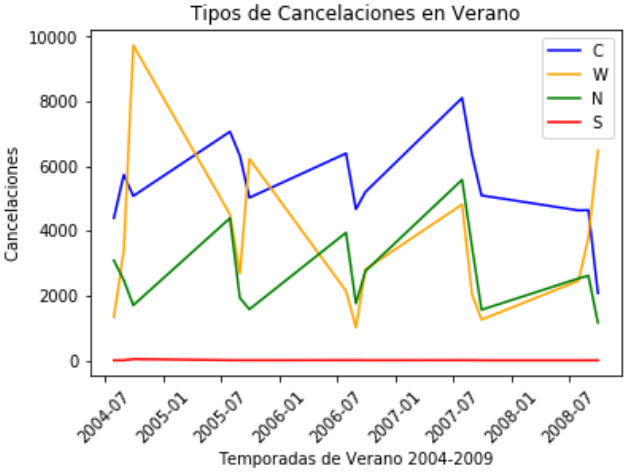
Podemos ver que la primera predicción es mucho mejor que la predicción de cancelaciones en el mismo período con el clima agregado, que tiene error de predicción = 31,378,081.95, lo que nos dice que muchos picos de la función que la desvían del comportamiento trigonométrico son efectivamente producidas por el clima.

3.2.2. Cancelaciones según temporada climática

Ahora, vamos a fijarnos más atentamente del comportamiento de las cancelaciones en las distintas estaciones del año. Primero, graficamos los comportamientos de los tipos de cancelaciones según la temporada, concentrándonos en invierno y verano:



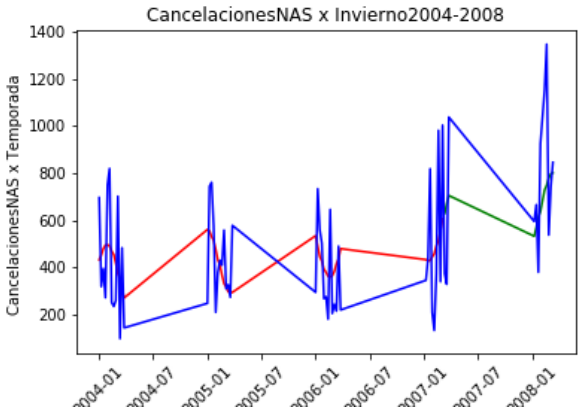
(a) Cancelaciones con clima 2004-2008



(b) Cancelaciones con clima 2004-2008

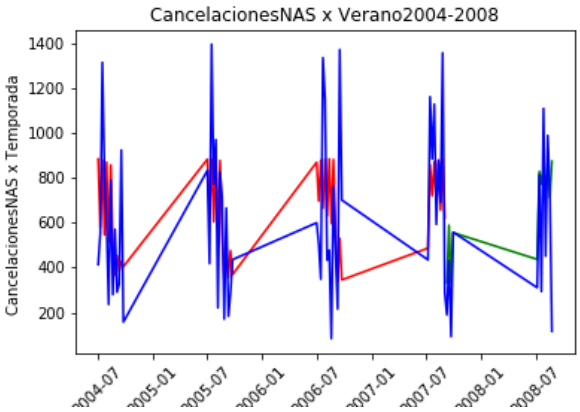
Hay muchas más cancelaciones en invierno, y la mayor parte son resultado del clima. En el verano, las cancelaciones de carrier(aerolínea) parecen ser las que generan una mayor cantidad de cancelaciones, pude deberse al incremento de tráfico de personas durante esas fechas lo que aumente las cancelaciones de aerolínea.

Ahora, parece ser que las cancelaciones de NAS mantienen el mismo comportamiento en los distintos veranos, por lo que podríamos a partir de veranos anteriores aproximar el próximo.



(a) Cancelaciones NAS en Invierno 2004-2008

(b) mejor k = 0.93, RMSE = 1469.13
Error de Predicción = 90,953.01



(c) Cancelaciones NAS en Verano 2004-2008

(d) mejor grado = 1, mejor k = 1.52, RMSE = 1282.34
Error de predicción = 87,445.70

Vemos que podemos aproximar mejor el verano que el invierno, y mirando el gráfico vemos que efectivamente en el verano el NAS también tenía una conducta bastante regular, pero en el invierno el mejor fitting/pred se da con una función trigonométrica, (fusión grado = 0), mientras que en el verano la mejor función tiene una recta (que indica que las cancelaciones

por NAS han ido subiendo cada invierno).

4. Conclusión

Para concluir este trabajo practico, podemos resumir en base a los experimentos observados, que es difícil aproximar una función a datos que dependen de muchos factores, como puede ser un caso tan particular como el 9/11 o casos más generales. Sin embargo, pudimos observar patrones, que fueron reflejados en el uso de regresión lineal por cuadrados mínimos para predecir comportamiento a futuro, que se corresponden a temporadas del año o factores climáticos, como por ejemplo que viajar en verano es mejor que en invierno, si estas en el hemisferio norte.

Referencias

[1] $On - Time\ Performance,\ https://en.wikipedia.org/wiki/On-time\_performance.$

[2] $Cross-validation,\ https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9$

[3] $Crisis\ del\ 2008,\ https://es.wikipedia.org/wiki/Crisis\_financiera\_de\_2008$

[4] $RMSE\ https://en.wikipedia.org/wiki/Root-mean-square\_deviation$

[5] $CuartilesPg146/155http://cms.dm.uba.ar/academico/materias/1ercuat2015/probabilidades\_y\_estadistica\_C/PyEC.pdf$

[6] Temporada de huracanes en el Atlántico de 2007, $https://es.wikipedia.org/wiki/Temporada\_de\_huracanes\_en\_el\_Atl\%C3\%A1ntico\_de\_2007$