TP3-Metodos Numericos

Andrés Nahuel Antola [1] Martin Moran [2] Juan Fernández Zaragoza [3]

Departamento de Computación
Universidad de Buenos Aires
Buenos Aires, Argentina

Abstract

En este trabajo se combinan el uso de los datos de aerolíneas de Estados Unidos y el método de cuadrados mínimos lineales (CML) para analizar la puntualidad (OTP).

Keywords: OTP, punctuality predict, cuadrados mínimos lineales, regresiones lineales.

1 Introducción

En este trabajo se analizara la puntualidad en aerolíneas, también conocida como *On-Time Performance* (OTP).

Note 1 *Los datos utilizados en este trabajo fueron obtenidos de la DataExpo 2009* http://stat-computing.c

Para llevar a cabo este análisis, el trabajo se divide en dos partes. Por un lado, un estudio respecto de los datos sobre los cuales se trabaja, y por otro, una evaluación de CML como método para predecir OTPs.

2 Estructura de los Experimentos

Los experimentos que realizaremos estarán divididos en dos secciones. Por un lado se hará análisis de los datos presentados. Y por el otro lado se evaluara la predicción mediante CML. A continuación se encuentra una breve descripción de las preguntas que se intentaran responder en cada sección.

En la sección de análisis de los datos, se pondrá atención a los siguientes interrogantes:

- ¿OTP es un buen indicador para evaluar la calidad de aerolíneas, aeropuertos y pares de ciudades?
- ¿A que se deben la mayoría de los retrasos?
- ¿Hay periodos de tiempo donde hay mayor cantidad de retrasos?
- ¿La cantidad de vuelos cancelados es un buen indicador de performance?

[1] Email: andresnahuel35@gmail.com
[2] Email: martinmoran1994@gmail.com
[3] Email: juanfernandezzaragoza@hotmail.com

En la sección de evaluación de CML, se investiga el rendimiento de CML para predecir valores de OTP y se lo evalúa en función de los regresores que utiliza (en tipo y cantidad), la granularidad temporal respecto de la cual se establece la aproximación, los plazos establecidos (en cantidad de puntos sobre los cuales aproximar y predecir) y los momentos en el tiempo evaluados.

2.1 Análisis de datos

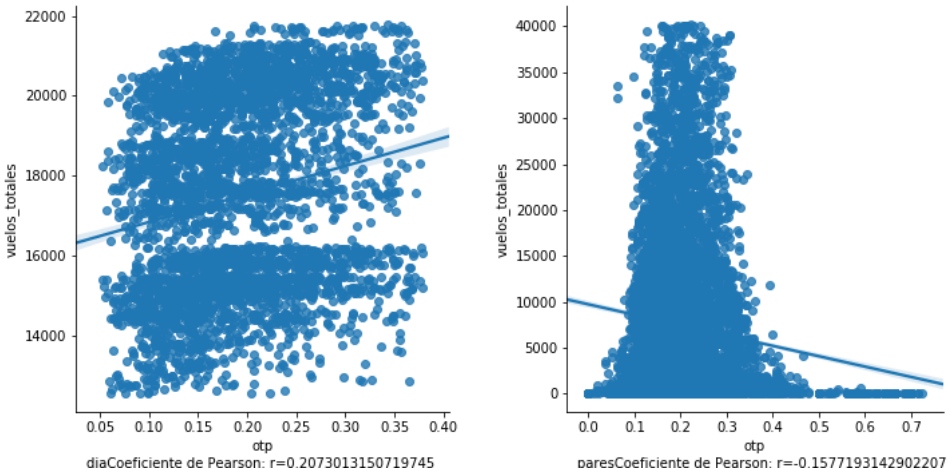
2.1.1 ¿OTP es un buen indicador para evaluar calidad?

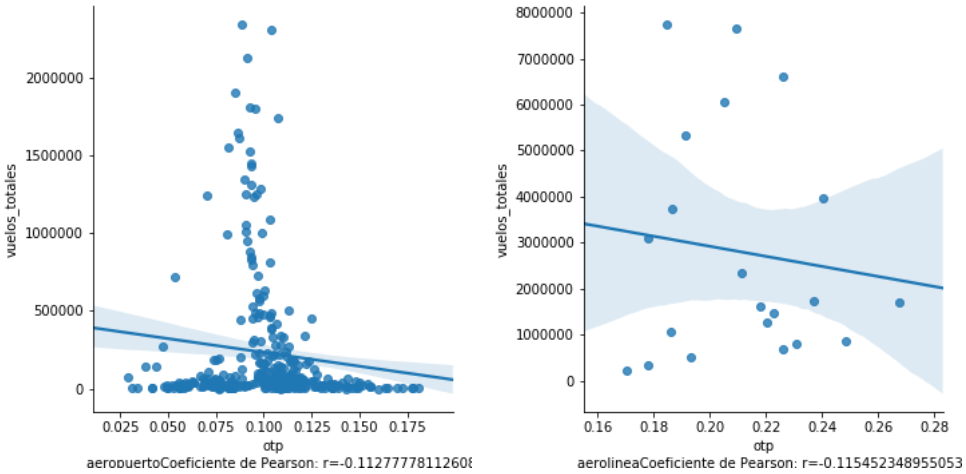
Como esta mencionado en la sección anterior el primer interrogante que nos interesa evaluar es si OTP es un buen indicador para distintos casos. Creemos que podría ser un buen indicador, ya que una persona que viaja en una aerolínea se ve afectado directamente por este factor en su opinión sobre una aerolínea. Pero, para confirmar nuestra hipótesis primero queremos descartar que tenga correlación, tanto positiva como negativa, respecto a la cantidad de vuelos, ya que la cantidad de vuelos no necesariamente es un buen indicador de calidad. Además nos gustaría ver si tiene correlación respecto a la cantidad de vuelos cancelados. Ya que ese si es un indicador de un buen servicio aéreo.

Los gráficos de correlación que realizamos utilizan la información desde el año 1998 a 2008 consiguiendo OTP del promedio.Además se filtran los elementos utilizando la siguiente formula: "$absoluto(elemento - mean <= 2 * standar\_deviation$" Por ultimo calculamos el coeficiente de Pearson de cada gráfico para tener un mejor indicador.

Ejemplo: Para el gráfico de correlación cantidad vuelos/aerolíneas. Calculamos para cada aerolínea el promedio de OTP y la cantidad de vuelos totales promedio. Y usamos el par de cada aerolínea como un punto del gráfico de correlación.

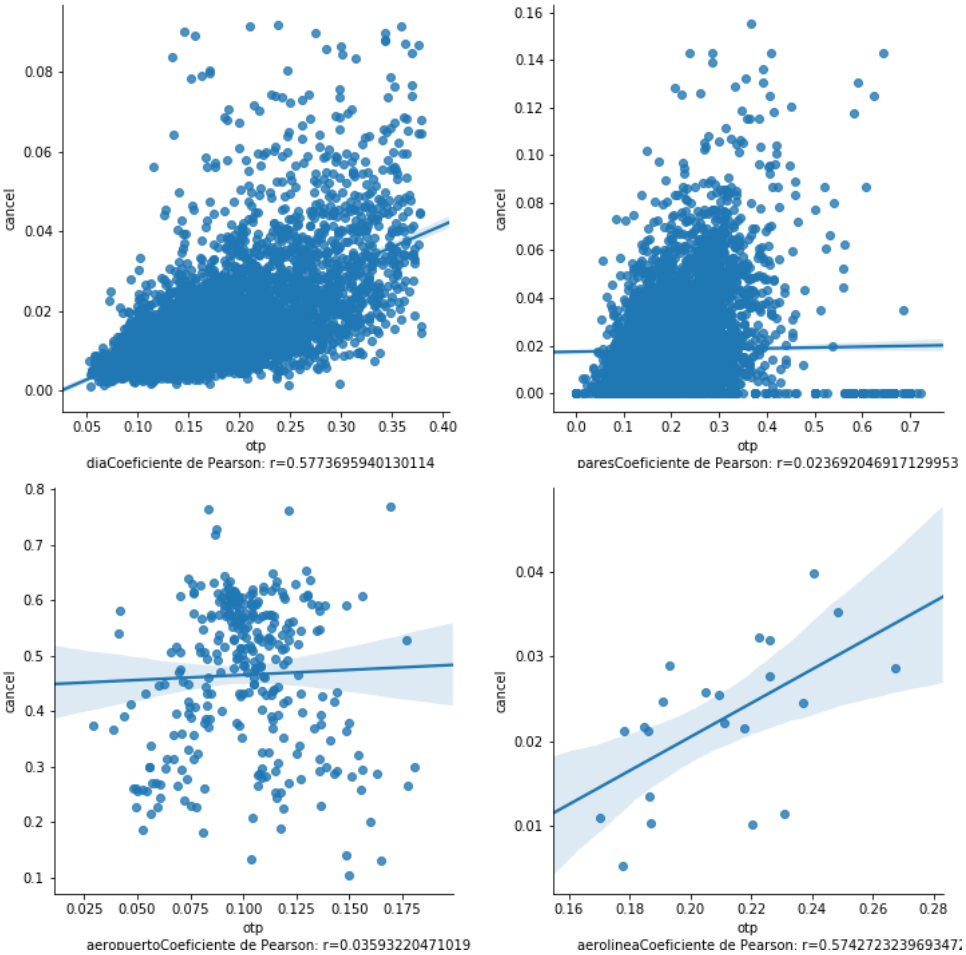
A continuación están los gráficos de correlación de cantidad de vuelos/OTP.





En estos gráficos se puede ver que fácilmente que no hay correlación en ninguno de los gráficos ya que la mayoría de los puntos están muy lejos de la recta. Además el coeficiente de Pearson es cercano a 0 lo cual confirma que no hay una correlación lineal entre estos dos factores en ninguno de los casos. Esto reafirma nuestra hipótesis de que es un buen indicador.

A continuación se encuentran los gráficos de correlación de %vuelos cancelados/OTP.



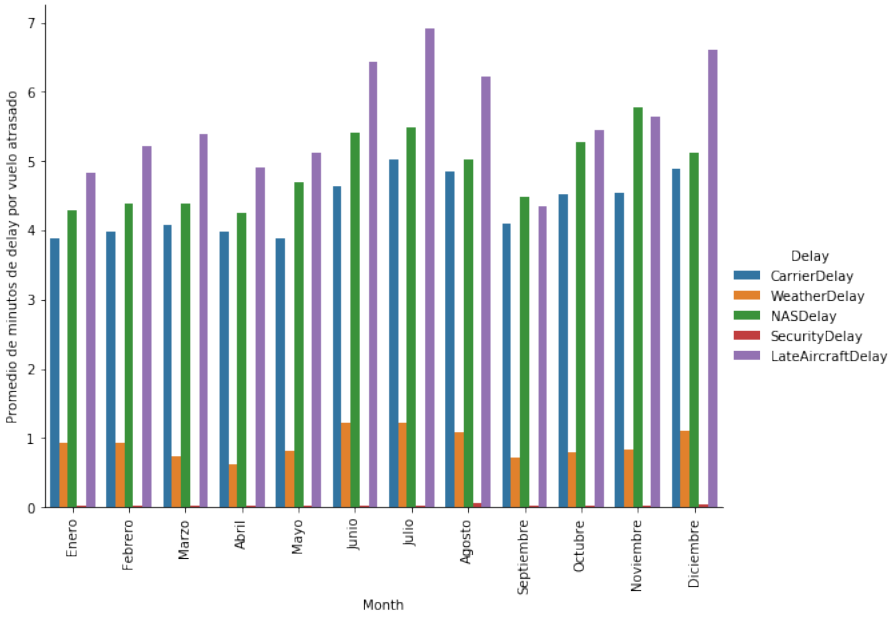
En los gráficos por día y por aerolínea de arriba se puede ver que OTP tiene correlación positiva moderada con respecto a la cantidad de vuelos cancelados. Esto sumado a los resultados del anterior experimento muestran que OTP es un muy buen indicador para medir aerolíneas y la eficiencia general por día. Pero no es tan buen indicador para pares de ciudades y aeropuertos.

2.1.2 Razones del delay

En este sección analizaremos las razones de los retrasos(delay). Y en que épocas son mayores. Para esto realizamos el siguiente gráfico de barras por mes por las distintas razones. Este gráfico fue producido al hacer un promedio de los datos de delay desde el año 1998 al 2008.

En ese gráfico se puede ver claramente que la mayor razón de delay en general es por la llegada tardía del avión. Como segunda causa es el delay producido por National Airspace System(NAS). Como tercera la aerolínea. Como cuarta el clima y por ultimo por seguridad.

Otra cosa interesante que se puede observar es que el delay es mayor en Diciembre julio, agosto y junio,Esto puede tener causas en los periodos de vacaciones. Además el delay por clima aumenta en los meses de junio a agosto. Que es el verano en EEUU, por lo tanto parece ser que hay mas tormentas en esa época.

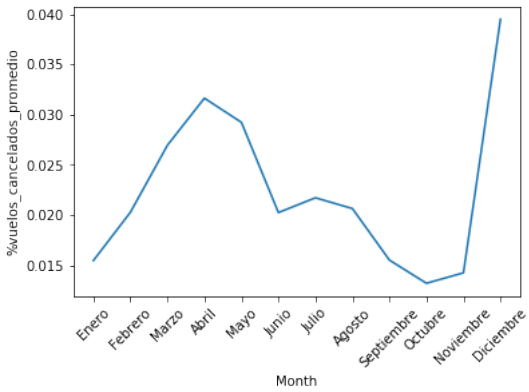


Este experimento nos demuestra que la estación, la época vacacional y otros factores tienen un gran impacto sobre el delay, y por lo tanto sobre OTP.

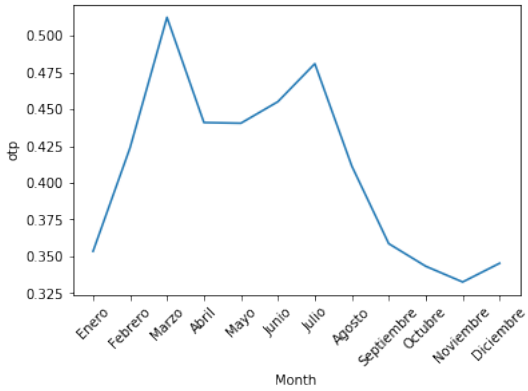
2.1.3 Vuelos cancelados

Otro punto de interés en el análisis de datos son los vuelos cancelados. ¿Son una buena medición al igual que OTP?

Por los gráficos de correlación anteriormente vistos. Parece ser una buena medición, pero primero nos gustaría hacer lo mismo que con OTP, compararlo con la cantidad de vuelos, esperando que haya poco porcentaje de vuelos cancelados en todo el año. Para esto realizamos el siguiente gráfico de linea obtenido del promedio del periodo 1998 al 2008.



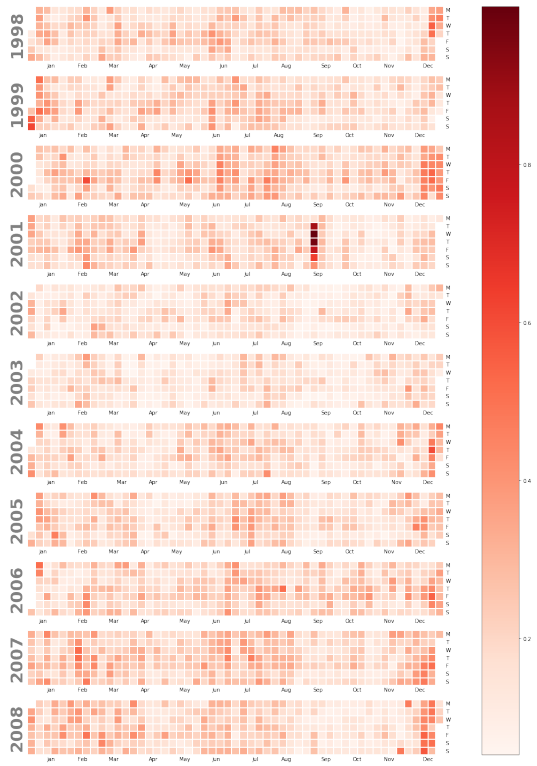
En este gráfico se puede ver que el porcentaje de vuelos cancelados es menor a 0.05 en todo el año. Que es relativamente pequeño. En especial al compararlo con el siguiente gráfico en base a OTP el cual muestra un máximo de 50 por ciento de delay en un mes. Al tomar esto en cuanto mas la correlación entre los mimos. Se puede descartar una correlación entre vuelos cancelados y vuelos totales. Por lo tanto puede es una buena medida de comparación.



Otro dato interesante que sale del gráfico de cancelaciones por mes que el porcentaje de cancelaciones es mucho mayor en invierno. Siendo hasta 3 veces mayor que en otros meses. Esto se lo puede atribuir a el clima, las tormentas, etc. Pero también al periodo vacacional en esa época. Que es lo mismo que pasaba con los delays, solo que en menor medida.

2.1.4 OTP a lo largo de los años

En esta sección queremos mostrar el OTP por día a lo largo de distintos años. Para identificar patrones en el mismo y analizar posibles razones a adjudicar a estos patrones.



Lo mas llamativo del gráfico se encuentra en septiembre de 2001. En esa semana se puede ver claramente que es la que tiene mayor cantidad de atrasos por día. Esto se lo puede adjudicar al atentado del 9/11. Por posibles medidas extra de control para evitar nuevas tragedias similares en los siguientes días.

Otra cosa que llama la atención es que hay una mayor cantidad de delay en el periodo de diciembre año a año. Probablemente debido al invierno.

Por ultimo se puede ver que a lo largo de los años el OTP se fue reduciendo, implicando una mejora general. Esto se debe a las mejoras tecnológicas de año tras año.

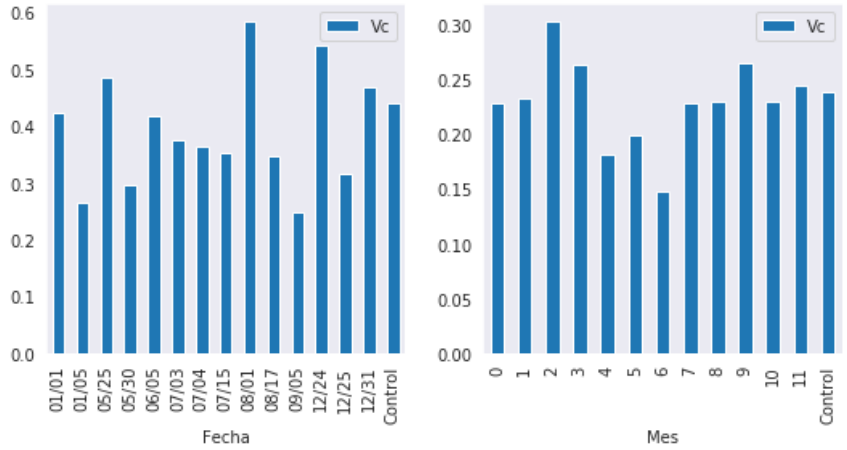
2.2 Análisis de CML

Para esta sección, los datos utilizados son los del rango de años entre 1998 y 2008 porque son los más robustos respecto de la información que contienen. En años anteriores hay casos de información faltante. Salvo que se especifique, los experimentos toman ese rango.

2.2.1 Día/Mes como predictor

En primer lugar, queríamos evaluar, antes de usar CML para hacer predicciones, si acaso era factible predecir el valor de OTP de un día del año o un mes dado conociendo solamente los valores de OTP para los mismos días del año/meses correspondientes a años anteriores.

Como primer enfoque, decidimos evaluar los VC (desvío estándar sobre rango de valores) respecto de los OTPs para los mismos días del año en años distintos (seleccionados arbitrariamente), así como los VC para los mismos meses en años distintos. Se evaluó además la varianza para un grupo de control, construido tomando al azar alguno de los meses/días evaluados para cada año con el fin de comparar si la varianza entre los elementos del mismo conjunto es menor que la del grupo de control (buen aproximador) o no (mal aproximador). A la izquierda, tenemos el gráfico por días, a la derecha, por meses.



Podemos apreciar que en ambos casos, los VC relativos a cada grupo no distan tanto de los VCS del grupo de control. En términos generales son similares. Esto parece indicar que el mes o el día del año no son buenos predictores del OTP. Sin embargo, existen algunos casos en que la varianza es más pequeña, para los cuales haría falta una base de datos más extensa para evaluar si acaso son buenos predictores.

2.2.2 Selección de Familias de funciones

Una observación interesante es que al aproximar usando CML, si la longitud de onda de las funciones senoides utilizadas en la aproximación sinusoidal era de igual o menor tamaño que la espaciación entre los puntos, los coeficientes de dichas funciones tomaban valores del orden de magnitud de 10^{20}. Como las predicciones resultantes de estos casos no eran muy útiles (su error también crecía en magnitud similar) no fueron utilizadas en nuestros experimentos.

Nuestra hipótesis fue que al ser funciones periódicas y sinusoidales, y teniendo en cuenta que la proyección sobre el eje x de los puntos utilizados estaba equiespaciada para los puntos, si tendieran a coincidir las regiones de corte en el eje x de una función sinusoidal y la proyección x de los puntos que aproximar, al maximizar el coeficiente se minimizaría el error (a mayor pendiente, mejor puede aproximarse dos puntos que estén a distintas alturas, cuando la pendiente tiende a infinito, en este caso, tendería a cortar todos los puntos equiespaciados sin importar su valor de y).

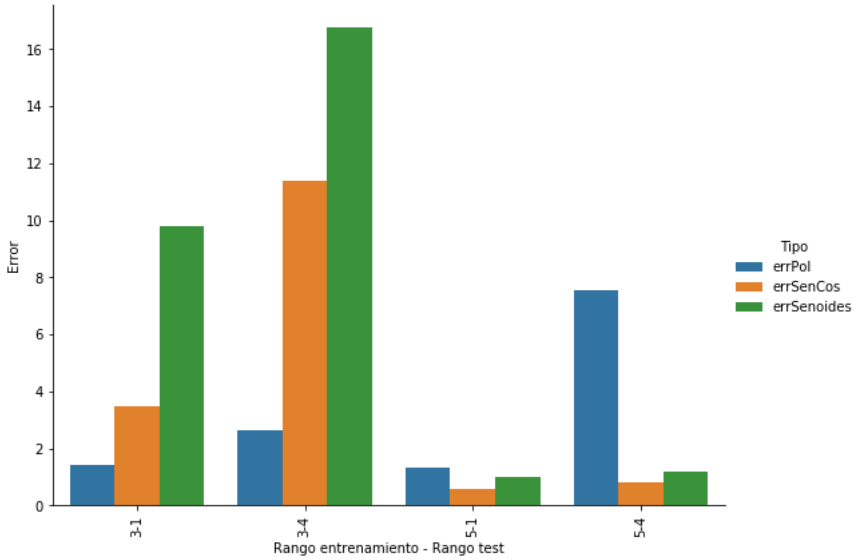
Por esta razón, al determinar la familia de funciones para la aproximación con senoides, pusimos una cota mínima (4) para la longitud de onda, determinada por el valor k que divide a x en cada una de las funciones senoides. Elegimos esta cota porque es el menor de los valores enteros para los que no obtuvimos errores de ordenes de magnitud tan grandes.

Una vez fijado este valor inicial, decidimos comparar las distintas familias de funciones, con distintas cantidades de coeficientes, para distintas granularidades, con el fin de evaluar qué regresores, y de qué manera, se comportan mejor para cada caso.

2.2.3 Granularidad Años

A continuación, vemos los errores correspondientes a predicciones cuya granularidad es del nivel del año. Evaluamos, para los años 1998 y 2000, el promedio de error de aproximaciones para cada par entre rangos de entrenamiento 3,5 y rangos de test 1,4. Fijamos la aproximación en grado 3, y la senoide en grado 5.

Observamos los errores de aproximación para cada una de las 3 familias de funciones, es decir, polinomiales, funciones de seno-coseno, y suma de senoides de distintas frecuencias.

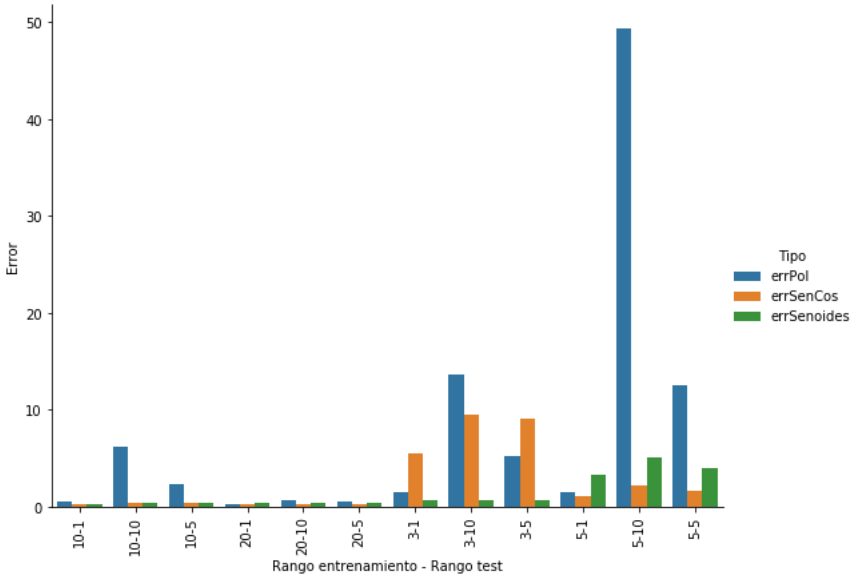


Puede observarse que la variación en el error es muy grande entre los pares de rangos. El error de aproximación, con un conjunto tan pequeño de puntos, es grande porque las bases de entrenamiento son pequeñas. Para los casos con las bases de entrenamiento más chicas la mejor aproximación es la polinomial. En cambio, cuando estas son relativamente más grandes su rendimiento baja en relación al resto de las familias.

Parece ser que las funciones que utilizan senoides funcionan mejor en casos de bases de entrenamientos mayores con este nivel de granularidad, pero el experimento no parece ser conclusivo. Por el tamaño de la base de datos, a este nivel no se cuenta con un conjunto grande de puntos como para poder establecer conclusiones más robustas. Experimentos subsiguientes podrían utilizar datos de años anteriores a 1998, asumiendo que las erratas y elipsis no afecten significativamente los resultados dado este nivel de granularidad.

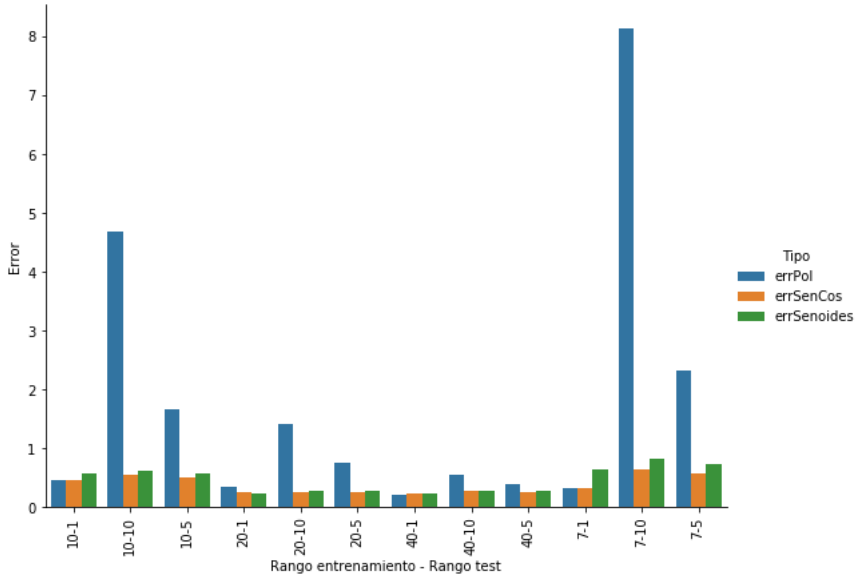
2.2.4 Granularidad Meses

De la misma manera construimos el gráfico para evaluar la granularidad de acuerdo a los meses, pero al contar con más puntos usamos más combinaciones entre rango de entrenamiento y rango de test (3, 5, 10, 20x1, 5, 10 respectivamente). Los meses iniciales fueron seleccionados arbitrariamente distribuidos entre la totalidad de los meses.



2.2.5 Granularidad Días

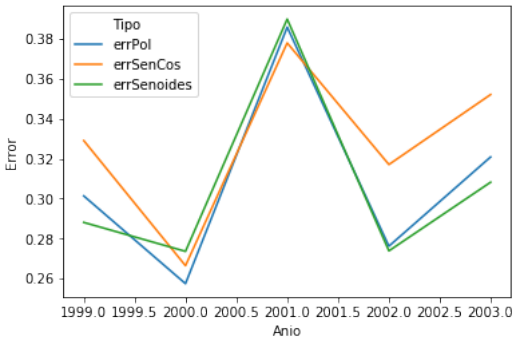
Por último, repetimos el mismo proceso para los días, usando los rangos de training 7, 10, 20, 40 y rangos de testing 1, 5, 10. Los días fueron arbitrariamente seleccionados, distribuidos por el conjunto de los días en los datos.



En ambos casos puede observarse que la aproximación mediante polinomios puede tener un error muy grande cuando las bases de test crecen, sobre todo si las bases de entrenamiento son relativamente pequeñas. Sim embargo, en casos en que las bases de entrenamiento son más grandes que las de testing, tienen relativo buen funcionamiento. En términos generales, para granularidades de días y meses, las aproximaciones que mejor funcionan son las de suma de senoides y una función lineal.

2.2.6 Evento Catastrófico

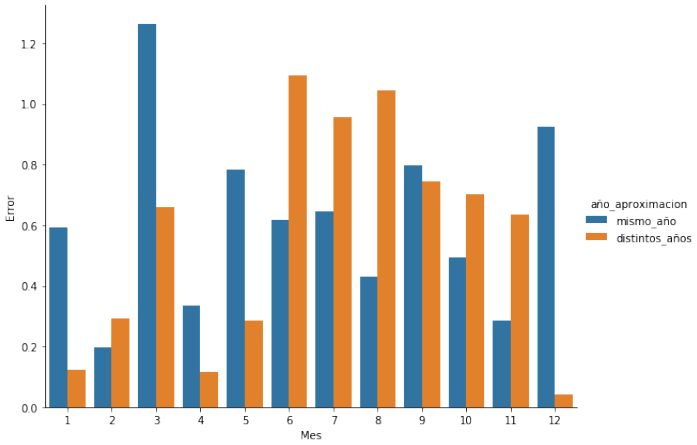
Para evaluar si acaso el atentado al World Trade Center en 2001 afectó a los predictores CML negativamente, decidimos tomar el error de predicción para distintos años (1999, 2000, 2001, 2002, 2003), con base de entrenamiento en los 30 días anteriores al 9/11 de dicho año, y con base de testing en los 10 días posteriores.



Podemos ver en dicho gráfico que el error creció a partir del atentado, y luego se estabilizó (las predicciones se adaptan a períodos sin cambios significativos en sus tendencias, pero cuando el cambio se da entre la base de entrenamiento y la base de testing el error crece).

2.2.7 Predicción Módulo 12

Continuando con el estudio que iniciaba la sección, comparamos si acaso contando con sólo valores de OTP para 3 meses como base de entrenamiento para predecir el OTP de otro mes, era mejor tener los tres meses inmediatamente anteriores al mes evaluado o los mismos meses correspondientes a años anteriores. Evaluamos esto para 2007, para contar con datos cercanos en el tiempo pero que en los 3 años anteriores no estuviera el evento de 2001, y evaluamos el error dada una aproximación lineal. Los resultados obtenidos fueron:



En dicho gráfico puede verse que esto depende significativamente de los meses. Por ejemplo, para diciembre, es mucho mejor predecir a partir de los diciembre anteriores que de los tres meses inmediatamente anteriores. En cambio, para meses como junio o julio, es mejor aproximar usando los inmediatamente últimos meses.

3 Conclusiones

Por el lado de análisis de datos se pudo verificar que OTP y y vuelos cancelados son buenos evaluadores de calidad para pares de ciudades y aeropuertos. Pero son aun mejores para aerolíneas y fechas. Además se pudo ver que los retrasos son mayormente causados por las llegadas tarde de otros vuelos. Otro dato interesante fue que hay mas retrasos en el pe diodo de verano respecto al clima, aun cuando se esperaba que eso ocurriera en invierno. Pero esto puede deberse a que hay mas cancelaciones en invierno y por lo tanto menos vuelos que pueden retrasarse. Otro dato interesante es que se pudo verificar que la cantidad de vuelos cancelados es un buen indicador de performance. Y por ultimo se pudo ver el impacto de eventos externos al OTP. Como por ejemplo: El atentado del 11 de septiembre de 2001 o el avance tecnológico a lo largo de los años.

Respecto de la sección de análisis de CML, puede observarse que los errores obtenidos para la granularidad de los días son las menores, pero son las que más puntos utiliza. Puede observarse también que se puede predecir, con bases de entrenamiento y testing razonables (20 y 5, por ejemplo), meses siguientes con buena confiabilidad a partir de meses anteriores, con lo cual es razonable usar diferentes granularidades en distintos contextos. Además, podemos ver que existen eventos externos que afectan el rendimiento de CML y que en el caso de algunos meses (como diciembre) el mes en el que alguien se encuentra aporta mucha información para predecir el OTP (probablemente por regularidades respecto de cantidades y tendencias de viajes en algunas épocas del año).

Un criterio de análisis sobre los datos con los que se trabaja resulta fundamental a la formular una familia de funciones. Las mismas forman el núcleo de la efectividad del método de CML. A lo largo del presente trabajo, vimos como familias de funciones periódicas van de la mano de datos de tipo temporal. Si bien no siempre es necesariamente así, podemos inclinarnos a favorecer las mismas a la

hora de realizar este tipo de estudios.

References

[1]