

# CML para OTP en aerolíneas

Andrés Nahuel Antola, Martín Morán, Juan Fernández  
Zaragoza

6 Diciembre, 2019

- Análisis de datos presentados
- Evaluación de CML

## 2 Consideraciones Generales

### 3 Experimentación y resultados

- Análisis de Datos presentados 1998 - 2008
- Estudio de CML como predictor

## 4 Conclusión

## Análisis de datos presentados

## 1 Vuelos cancelados como indicador de calidad

## Análisis de datos presentados

- 1 Vuelos cancelados como indicador de calidad
- 2 OTP como indicador de calidad.

## Análisis de datos presentados

- 1 Vuelos cancelados como indicador de calidad
- 2 OTP como indicador de calidad.
- 3 Causas posibles de retrasos.

## Estudio de CML como predictor

- 1 Rendimiento de CML para predecir valores de OTP.
- 2 Análisis en términos de:
  - Regresores utilizados (tipo y cantidad).
  - Granularidad de datos a aproximar.
  - Plazos establecidos.
  - Contexto de aplicación (periodo en que se lo usa).

# Consideraciones Generales

## 1 Cálculo de OTP.

# Consideraciones Generales

- 1 Cálculo de OTP.
- 2 Datos seleccionados.



# Consideraciones Generales

- 1 Cálculo de OTP.
- 2 Datos seleccionados.
- 3 Outliers

# Consideraciones Generales

- 1 Cálculo de OTP.
- 2 Datos seleccionados.
- 3 Outliers
- 4 Familias de regresores consideradas.

# Cálculo de OTP

$OTP = \frac{A_a + A_s}{2 * V_t}$ . Con  $A_a$  y  $A_s$  atrasos de arribos y salidas,  $V_t$  vuelos totales.

# Datos seleccionados

Intervalo temporal 1998 - 2008.

A partir de 1995 el Departamento de Transporte estadounidense empieza a controlar los delays.

# Outliers

El sistema CML es sensible a outliers.

Para calcular outliers asumimos distribución normal:

$x$  es considerado outlier si  $abs(x - mean) > 2 * standard\_deviation$

Filtramos los outliers de los datos salvo en casos específicos por requerimiento de diseño de los experimentos.

# Familias de regresores consideradas

- Intuición primera: Polinomios. Pueden llegar a aproximar muy bien intervalos pequeños.
- Problema: Un polinomio no constante tiende a  $+\infty$  o  $-\infty$  cuando crece  $x$ .
- Funciones periódicas: sinusoides.
- Caso 1: Función constante, función lineal, función seno y función coseno.
- Caso 2: Variar longitud de ondas. Funciones de la forma  $\sin\left(\frac{2\pi x}{k}\right)$  reemplazando al par de funciones seno y coseno.
- Salvo la de seno y coseno, son conjuntos de familias, que determinan una familia cuando se especifica una cantidad de funciones  $n$ .

# Familias de regresores consideradas

- Familia senoide: Fijado un  $n$  (cantidad de funciones) se tiene:
- La función 0 es la constante.
- La función 1 es la lineal.
- La función  $k > 1$  es la función senoide formada por el parámetro  $k$ .
- Por esta razón el  $k$  empezaba en 2. Para valores pequeños de  $k$ , con algunos conjuntos de datos, los coeficientes eran demasiado grandes.
- Modificamos la fórmula a  $\sin(\frac{2\pi x}{k+2})$  (salvo cuando se especifique lo contrario) por ser este el mínimo valor natural del denominador que no dio este error.





## Vuelos cancelados como indicador

- Pregunta: ¿Vuelos cancelados es un buen indicador de calidad?
- Ver si existe variación en %vuelos-cancelados .
- Experimento: Realizamos un gráfico de %Vuelos cancelados por mes.

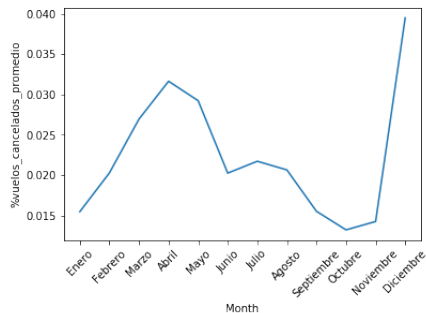
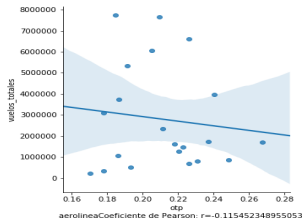
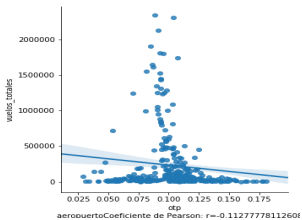
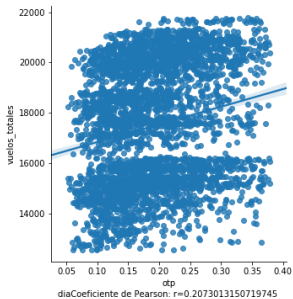
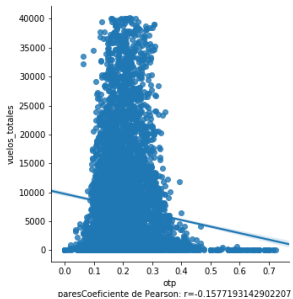
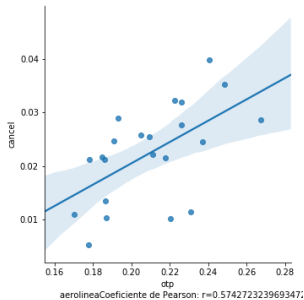
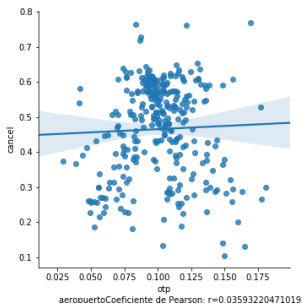
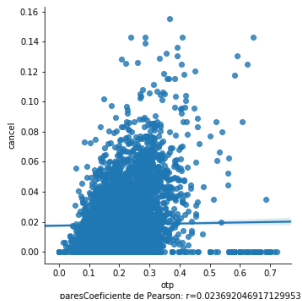
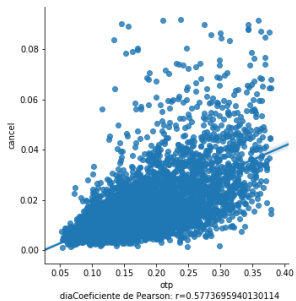


Figure: Porcentaje de vuelos cancelados por mes

# OTP como indicador de calidad

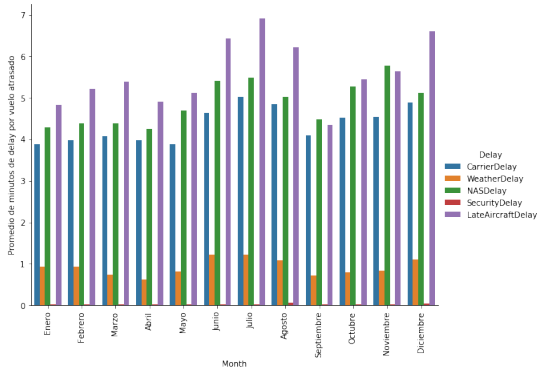
- Pregunta: ¿El indicador OTP resulta eficiente para evaluar la calidad de las aerolíneas?
- Queremos probar:
  - OTP es informativo: No constante.
  - Incluir OTP como indicador: No redundancia.
  - Elegir OTP como único indicador: Informativo respecto del resto.
- Experimento: Realizamos gráficos de correlación. Y coeficiente de Pearson.

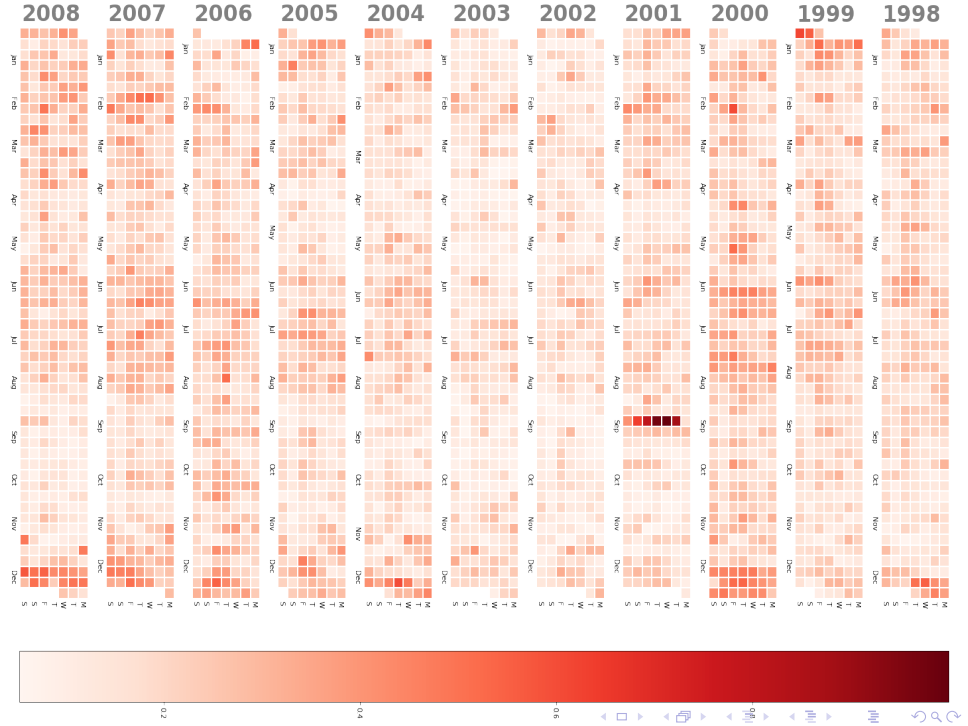






# Razones del delay











## Datos día

- Conjunto de fechas elegidos arbitrariamente.
- Años: 1998 - 2008



## Datos Mes

- 12 meses.
- Años: 1998 - 2008



# Día/Mes como predictor

- Respuesta: En general, el VC de los conjuntos es algo menor al del control.
- Sucede menos con el mes.
- Varía según las fechas y los meses.
- No podemos concluir que sean de por sí buenos predictores, es útil considerar otras formas de predicción (CML).

# Regresores a considerar

- Pregunta: ¿Qué familias de regresores nos es conveniente tomar?
- ¿En las familias de tamaño variable, qué valor de  $n$  nos conviene?
- Observación: Esta sección no estudia la variación respecto del denominador en sinusoides mencionada en la introducción, ese paso fue previo y se decidió por  $k+2$ .

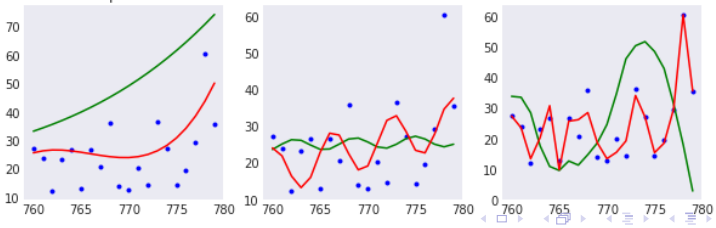
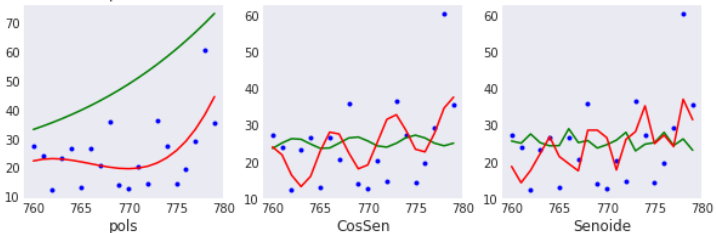
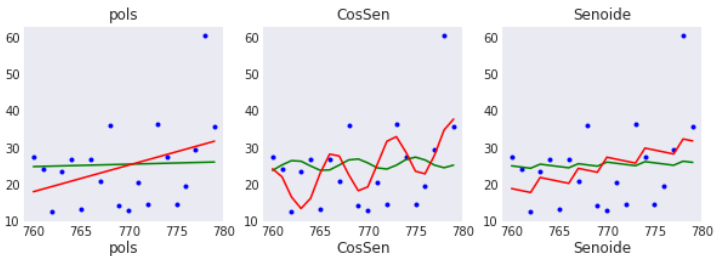


# Regresores a considerar

Idea: Ver cómo se relaciona, en un conjunto de datos, la función que los predice con la función fitteada ad hoc para esos datos.

Si difieren mucho, estamos en un caso de overfitting (mala selección de parámetro  $n$ ) o de mala selección de familia de funciones.

Probamos para los 3 géneros de familias considerados, variando el parámetro de  $n$  para polinomios y sinusoides.



# Regresores: Respuesta

- Vemos que en el caso de polinomios difieren mucho más cuando aumenta  $n$ .
- En este caso, la función que predice sobre los puntos graficados funciona mejor en el caso lineal.
- Los polinomios, al no ser periódicos, pueden fittear bien con  $n$  alto pero son malos predictores.
- La familia seno coseno puede ajustarse poco a los datos (pocos grados de libertad).
- La familia de senoides es buena prediciendo, pero sensible a overfitting cuando crece  $n$  a valores cercanos a 20.
- Esto acota los valores de  $n$  a considerar en los siguientes experimentos.

# Granularidad

- Pregunta: ¿Qué nivel de granularidad en función del tiempo es conveniente tomar?
- Experimento: Evaluar el error de aproximación normalizado para distintas granularidades, donde  $ECM = \sqrt{\text{mean} \sum_1^n e^2}$ , con  $e = y - y_{pred}$  y  $NECM = \frac{ECM}{y_{max} - y_{min}}$
- Probarlo para distintas familias de regresores.

# Granularidad

- Pregunta: ¿Qué nivel de granularidad en función del tiempo es conveniente tomar?
- Experimento: Evaluar el error de aproximación normalizado para distintas granularidades, donde  $ECM = \sqrt{\text{mean} \sum_1^n e^2}$ , con  $e = y - y_{pred}$  y  $NECM = \frac{ECM}{y_{max} - y_{min}}$
- Probarlo para distintas familias de regresores.
- Probarlo para distintos tamaños de conjunto de training/testing.

# Granularidad

El método para elegir los conjuntos de entrenamiento y de training de distintos tamaños es el siguiente.

Se cuenta con una fecha inicial del conteo.

Se cuenta con un conjunto de índices de fechas iniciales de los conjuntos de training.

Se cuenta con un conjunto de rangos (tamaños) de bases de training y de testing.

para cada combinación, elegimos  $Tr$  tal que comience en [fecha inicial del conteo + índice  $i$ ] y termine [rango training  $i$ ] días/meses/años después. La base de testing son los [rango testing  $i$ ] días/meses/años posteriores al conjunto de training.

# Granularidad de datos a aproximar (Día)

Fecha inicial: 01/01/1998

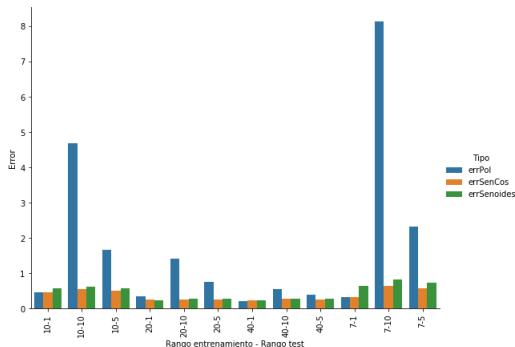
Inicios de conjuntos de entrenamiento: A partir de la fecha inicial, los días número 0, 150, 300, 450, 600, 1200, 3000

Rangos de training: 7, 10, 20,40.

Rangos testing: 1, 5, 10 días.

Valores de n fijos en 3 para el polinomio y 5 para el senoide.

# Granularidad de datos a aproximar (Día)

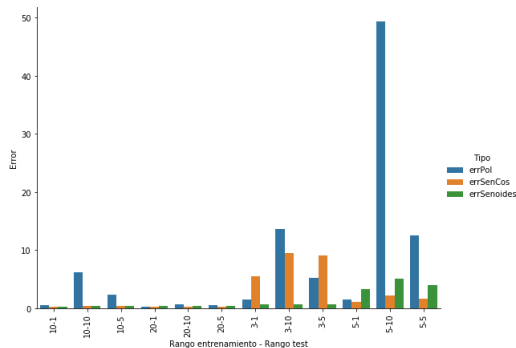




# Granularidad de datos a aproximar (Mes)

- Fecha inicial: 01/01/1998 (del conteo de meses)
- Inicios de conjuntos de entrenamiento: A partir de la fecha inicial, los meses número 3, 20, 40, 60.
- Rangos de training: 3, 5, 10, 20.
- Rangos testing: 1, 5, 10 meses.
- Valores de n fijos en 3 para el polinomio y 5 para el senoide.

# Granularidad de datos a aproximar (Mes)

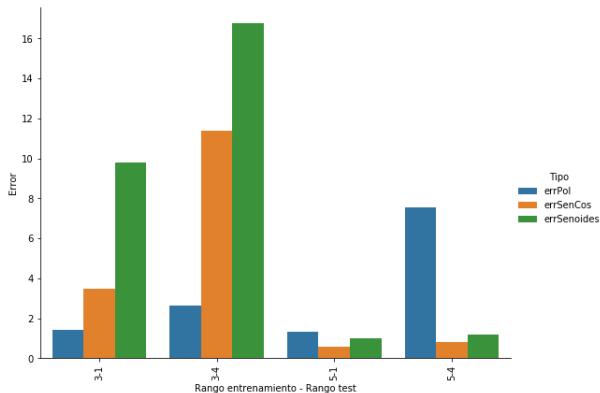


# Granularidad de datos a aproximar (Año)

Intervalos tomados(Training, Testing).

- Año inicial: 1998 (del conteo de años)
- Inicios de conjuntos de entrenamiento: A partir del año inicial, los años número 0, 2.
- Rangos de training: 3, 5.
- Rangos testing: 1, 4.
- Valores de n fijos en 3 para el polinomio y 9 para el senoide.

# Granularidad de datos a aproximar (Año)

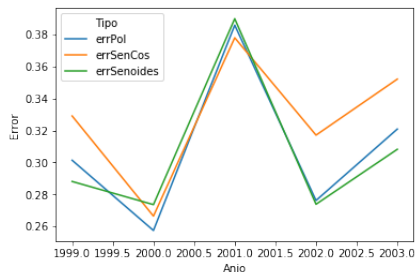




# Evento Catastrófico

- Pregunta: ¿Cómo afectó el atentado al World Trade Center en el 2001 a los modelos predictivos?
- Experimento: Para un conjunto de años alrededor del 2001, ver con qué nivel de error desde los 90 días anteriores al 11 de septiembre se predicen los 30 siguientes.
- Hacer esto para 3 familias de funciones diferentes, una por cada regresor (valores de  $n$  8 y 4 para polinomios y senoides respectivamente).

# Evento Catastrófico



# Evento Catastrófico

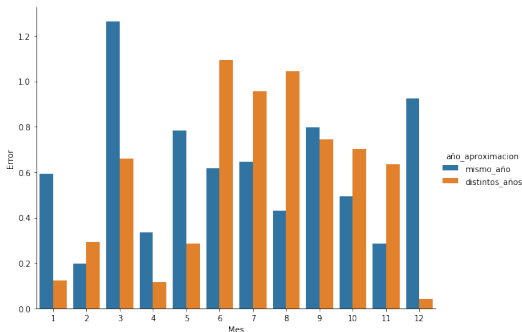
- Respuesta: Hay un aumento del error en la capacidad predictiva en ese año.
- Eventualmente cuando el modelo toma bases de training posteriores se estabiliza el error.
- Problema grave cuando el evento que modifica las tendencias se encuentra justo entre la base de training y la de testing.



# Predicción: ¿Tres anteriores o módulo 12?

- La pregunta va en la línea de si nos sirve la predicción usando el mismo mes del año pasado.
- En este caso, queremos saber, si usando 3 meses para predecir un cuarto mes, es mejor usar los inmediatamente anteriores o los mismos pero correspondientes a años pasados.
- Se eligió el año 2007 para evitar la coincidencia de meses anteriores con el 2001.
- Para polinomio, aproximación lineal, para senoide,  $n = 5$ .

# ¿Tres anteriores o módulo 12?



## ¿Tres anteriores o módulo 12?

- Varía respecto del mes en particular.
- Conclusión parcial: depende del caso.
- Experimento en un solo año y con pocos puntos, pero algunas observaciones pueden hacerse.
- Observación interesante: En el caso de 2007, los datos no coinciden tanto con los esperados por los valores de VC estudiados para meses (sin embargo, el experimento tuvo pocos datos).

# Conclusión

- OTP es un buen indicador de calidad.
- CML para predecir OTP funciona bien en condiciones normales.
- CML es, sin embargo, muy sensible a anomalías.
- Es mejor usar funciones periódicas para predecir sobre conjuntos medianamente grandes.