Hoy te convertís en (Junior) Data Scientist

Cortés Lucas , Lamela Emanuel , Zimenspitz Ezequiel [1,2,3]

Universidad de Buenos Aires
Buenos Aires, Argentina

Abstract

En este trabajo analizamos la temática de los desvíos en los vuelos en realación a principalmente 2 ejes: la época del año y la distancia del trayecto.

En el caso de la época del año, se puede observar que para distintos Estados de USA la curva de desvíos tiene picos en distintas épocas del año, mientras que para el caso de la distancia del trayecto se observa que al aumentar esta, existe un aumento paulatino de desvíos. Por otro lado, encontramos que existen ciertos trayectos con picos exponenciales de desvíos.

Por último creamos modelos predictivos para cada una de nuestras métricas utilizando cuadrados mínimos lineales y creando familias de funciones adecuadas a cada modelo particular.

Keywords: Modelos predictivos, Data fitting, Cuadrados mínimos lineales, Desvíos aéreos

[1] lucascortes@me.com, LU: 302/13
[2] emanuel93_13@hotmail.com, LU: 021/13
[3] ezeqzim@gmail.com, LU: 155/13

1 Introducción

Imagínemos un fenómeno sobre el cuál se quisiera poder "predecir" su comportamiento en base a aspectos o propiedades del mismo. Como ejemplos pueden ser: las ventas de un(os) producto(s) particular(es), el movimiento de una partícula o hasta la reacción de la gente. *A priori*, parecen problemáticas difíciles de modelar a la perfección, y probablemente lo sean. Por eso se suelen plantear modelos predictivos que no son exactos y se los trata de ajustar lo más posible, en otras palabras achicar el error lo más posible.

Una metodología es la de ***cuadrados mínimos***. Denominamos a los datos como pares (x_i, y_i), $i = 1, ..., m \in \mathbb{N}$, en base a los cuales planteamos **familia de funciones F** tal que $F = \{\alpha_1\phi_1 + ... + \alpha_n\phi_n : \alpha_1, ..., \alpha_n\}$, $n \in \mathbb{N}$, donde los ϕ_i son funciones que aceptan a los x_j tales que $Im(\phi_i) \in \mathbb{R}$. Los ϕ_i están fijos y lo que buscamos son los coeficientes que los acompañan. Por lo tanto, buscamos **una función en la familia F** hallando los coeficientes tales que se alcance $\min_{\alpha_1,...,\alpha_n \in \mathbb{R}} \sum_{i=1}^{m} (\alpha_1\phi_1(x_i) + ... + \alpha_n\phi_n(x_i) - y_i)^2$. Esta última expresión se denomina el **criterio de cuadrados mínimos**.

Tal y como está planteado, encontrar una solución parece rebuscado e inclusive no está claro que exista. Se puede ver que la expresión se puede transformar en $\min_{\vec{\alpha}} ||A\vec{\alpha} - b||$ [1], donde:

- $A \in \mathbb{R}^{m \times n}$, con $fila_i(A)^t = (\alpha_1(x_i) ... \alpha_n(x_i))$.
- $\vec{\alpha} \in \mathbb{R}^n$, con $\vec{\alpha}^t = (\alpha_1 ... \alpha_n)$.
- $b \in \mathbb{R}^m$, con $b^t = (y_1 ... y_m)$.

Este enfoque permite ver que siempre existe un vector de coeficientes tal que se realiza el mínimo del criterio [1].

Retomando lo que corresponde netamente a este trabajo, los fenómenos a analizar son *delays*/cancelaciones/desvíos en vuelos de avión. Trabajamos sobre información de vuelos realizados entre los años 1987 a 2008 desde y hacia los Estados Unidos [2]. Nuestro objetivo va a ser construir **modelos predictivos**, utilizados *cuadrados mínimos*, sobre aspectos de los vuelos y poder extraer información en consecuencia.

2 Desarrollo

El enfoque central sobre el cuál vamos a ahondar va a ser los desvíos en vuelos. En términos más técnicos, utilizaremos la técnica de *cuadrados mínimos* para hallar modelos predictivos, utilizando parámetros que creemos relevantes. Veremos, para cada experimento, la razón que sustenta esa elección de parámetros

y también compararemos distintas funciones predictoras a través del ***Error Cuadrático Medio***.

Particularmente, atacaremos dos ejes de estudio:

(i) Desvíos de vuelos en relación a la época del año.

(ii) Desvíos de vuelos en relación a la distancia entre origen y destino.

Conocer las tendencias sobre las desviaciones podría permitir a los aeropuertos poder organizar el cronograma de vuelos con mayor eficiencia teniendo en cuenta estos resultados. Más aún, esto no sólo afecta la planificación propia de cada aeropuerto sino de aquellos que interactúen con el mismo puesto que los planes de vuelo deben ser conocidos de antemano. Saber que un vuelo tiene cierta inclinación a sufrir un desvío, facilita saber qué nivel de alerta deben tener los aeropuertos a esta ocurrencia.

En los experimentos planteados analizaremos este fenómeno y su ligadura a los aspectos particulares de cada eje de estudio.

3 Experimentos y Discusión

3.1 Desvíos

En este experimento realizamos un análisis sobre los desvíos de los vuelos en base al tiempo. Dividimos al año en 12 meses y tomamos el período 2000-2008. Para generar un análisis representativo del comportamiento a escala país decidimos tomar 2 estados con características particulares: deben tener gran cantidad de vuelos de llegada y debía ser uno de cada costa. De este modo tomamos a los estados de California y Florida para realizar el análisis.

3.1.1 California

En el siguiente gráfico se puede observar el porcentaje de vuelos desviados en el período mencionado para vuelos con destino a California.

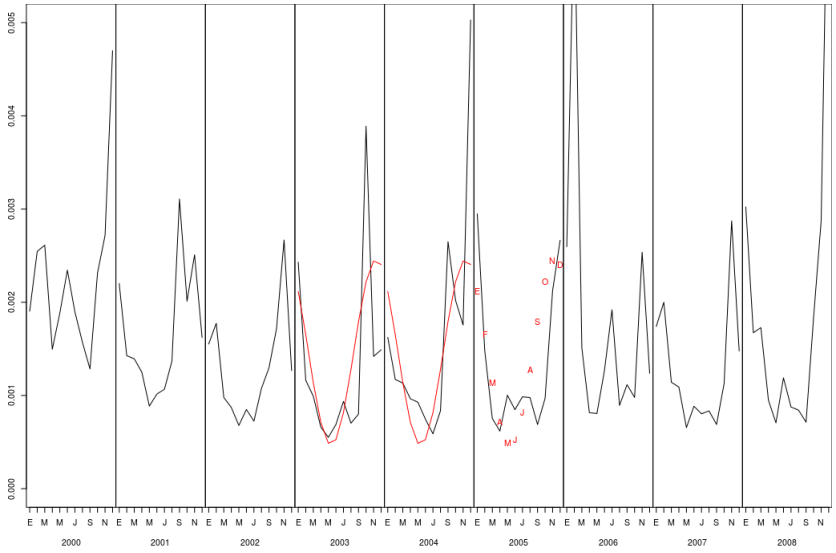

Fig. 1. Diverted arrivals - California

Para aproximar a la función observamos algunas cosas:

- Hay cierta periodicidad en la función. Posee picos en los meses correspondientes a las vacaciones de verano del hemisferio norte y descensos en el resto. El período es de 12 meses.
- Aunque la función tiene picos marcados, su forma se asemeja a la de *sin* y *cos*.

Dados estos indicios nos proponemos encontrar una familia de funciones para aproximar nuestro gráfico usando cuadrados mínimos. Estas funciones

serán una combinación de *sin* y *cos* con período 12. Las siguientes 2 familias de funciones responden a estas características y aproximan relativamente bien a nuestra función, dados α_i correspondientes.

$F_1 = \alpha_1 * abs(sin(\frac{\pi}{12} * x) * cos(\frac{\pi}{6} * x)^2) + \alpha_2$

$F_2 = \alpha_1 * sin(\frac{\pi}{6} * x) + \alpha_2 * cos(\frac{\pi}{6} * x) + \alpha_3$

La primera multiplica al *sin* y *cos* y toma el valor absoluto para eliminar los picos negativos. La segunda realiza la suma de los *sin* y *cos*. No es relevante acá tomar el valor absoluto ya que no hay picos marcados negativos. Luego a ambas funciones le sumamos una constante para que la curva se desplace en dirección vertical. Observamos que sumar una variable lineal no tenía impacto apreciable en la aproximación.

Luego resolvimos cuadrados mínimos para ambas funciones y calculamos el error cuadrático medio de cada una. Como training tomamos a los años 2003 y 2004 e intentamos predecir 2005.

Los errores cuadráticos medios son: $ECM(F_1) = 0.0003236866$ y $ECM(F_2) = 0.0002451076$, siendo la segunda función una mejor aproximación que la primera. Los valores son pequeños ya que las mediciones son sobre un porcentaje pequeño.

Por lo tanto se puede ver en el gráfico anterior la aproximación que F_2 realiza en la muestra, prediciendo cómo será 2005. Se ve que respeta el comportamiento general y aproxima el pico que hay a principio y fin de año.

3.1.2 Florida

Realizamos el mismo análisis para los vuelos dirigidos a Florida. La función sigue teniendo un comportamiento periódico año a año, pero, como se puede apreciar en 2, vemos que por algún motivo hay un desplazamiento horizontal de la curva: los picos positivos se encuentran a mitad de año.

Por otro lado, el porcentaje de vuelos desviados está en el mismo rango que en California. Dado este conjunto de similitudes y diferencias con el caso anterior, nos interesó ver cómo se comportaban nuestras familias de funciones anteriores para aproximar a nuestra nueva curva.

Para esto realizamos cuadrados mínimos con las siguientes dos familias de funciones:

$F_1 = \alpha_1 * abs(sin(\frac{\pi}{12} * x) * cos(\frac{\pi}{6} * x)^2) + \alpha_2$

$F_2 = \alpha_1 * sin(\frac{\pi}{6} * x) + \alpha_2 * cos(\frac{\pi}{6} * x) + \alpha_3$

y vimos que los errores cuadráticos medios en este caso fueron de $ECM(F_1) = 0.0003098236$ y $ECM(F_2) = 0.0003590168$, o sea, bastante similares al anterior.

Se puede observar la curva en el siguiente gráfico:

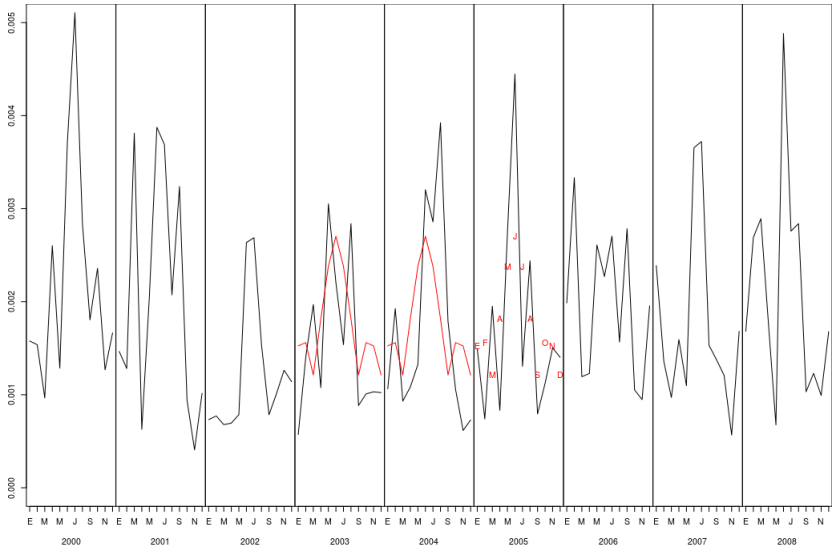

Fig. 2. Diverted arrivals - Florida

Una razón que puede ser atribuída a este fenómeno es el hecho de que la temporada de huracanes coincide con los meses donde se presentan los picos (el verano del hemisferio norte). Con lo cuál, tiene sentido que se presente un mayor porcentaje de desvíos para sortear estas condiciones climáticas adversas. Quisimos crear una familia de funciones parecida a las que teníamos pero que considerase esa particularidad. Probando un poco nos dimos cuenta que las potencias de *cos* tienen un comportamiento de este estilo. Por ejemplo $5 * cos(x)^{500} + 1$ tiene el siguiente gráfico:

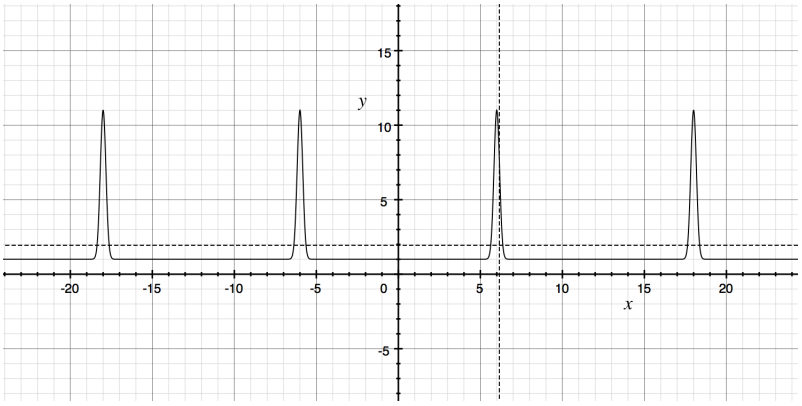

Fig. 3. $5 * cos(x)^{500} + 1$

Por lo tanto llegamos a la siguiente familia de funciones que usa esta nueva técnica.

$F_3 = \alpha_1 * abs(sin(\frac{\pi}{12} * x) * cos(\frac{\pi}{6} * x)^2) + \alpha_2 * cos(\frac{\pi}{12} * x - \frac{\pi}{2})^{500} + \alpha_3$

El error cuadrático medio de esta función es $ECM(F_3) = 0.00030524$.

Podemos ver a F_3 representada en el gráfico de Florida en rojo. En ese caso se entrenó cuadrados mínimos con 2003 y 2004 y se predice 2005.

3.2 Desvíos por distancia

Otro eje de análisis que nos pareció interesante fue investigar qué ocurría con la cantidad de vuelos que se desvían en relación a la distancia del viaje. Este eje surge de la idea de que a mayor distancia del viaje, mayor probabilidad de que surjan problemas en el trayecto, ya sean desperfectos técnicos, mal clima o situaciones impredecibles.

Para esto realizamos el siguiente gráfico que muestra el porcentaje de vuelos desviados según la distancia del vuelo entre 2005 y 2008. Como los vuelos tienen hasta 5 mil millas de distancia, partimos a nuestro conjunto en 10 subconjuntos (0 a 500 millas, 501 a 1000 millas, ..., 4501 a 5000 millas).

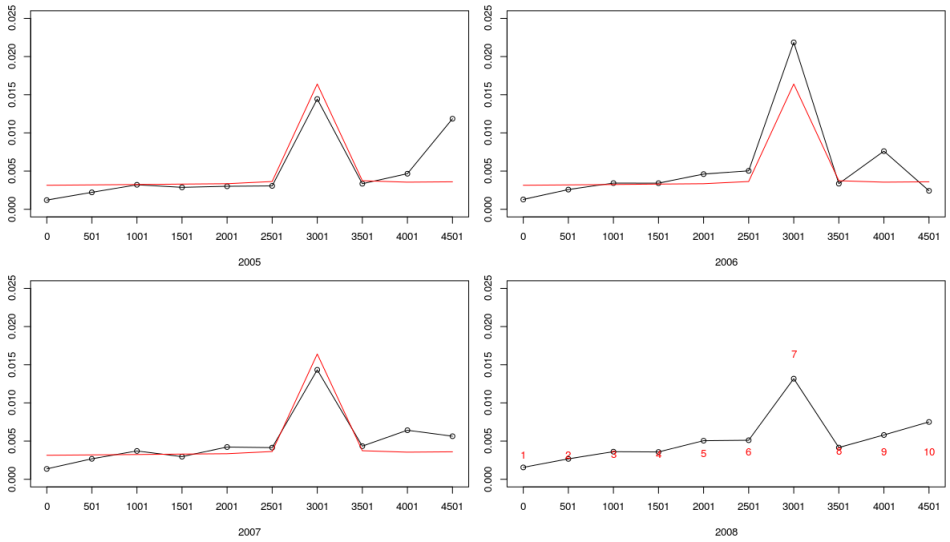

Fig. 4. Diverted by distance

Hay algunas cosas que se pueden ver rápidamente en el gráfico. Por un lado vemos que año a año la curva respeta cierto patrón lineal y de pendiente positiva, lo cual acompaña la hipótesis del aumento de desvíos según la distancia del vuelo, a pesar de que la pendiente sea bastante pequeña. Por otro

lado, la más llamativa de las características del gráfico es que en el segmento de 3001 a 3500 millas hay un pico muy llamativo, donde la cantidad de vuelos desviados llega a triplicar su valor respecto a los segmentos aledaños. Lo más extraño es que aunque pareciera ser un outlier, este evento se da año a año.

Nos propusimos analizar esta situación para esclarecer sus causas. Para eso tomamos como referencia el año 2006.

Primero calculamos la cantidad de vuelos que hay para cada segmento y vimos que decrece exponencialmente conforme la distancia del vuelo aumenta: mientras que para trayectos de menos de 500 millas hubo más de 3 millones de vuelos, a partir de 3 mil millas los segmentos tienen menos de 5 mil vuelos. Por lo tanto al tener una densidad baja de trayectos, cualquier outlier cobra mucho más peso.

Luego quisimos averiguar qué aeropuertos eran los que más desvíos tuvieron en el segmento del pico y los resultados son llamativos: de 47 vuelos desviados, 43 pertenecen a 2 trayectos en particular, los que vuelan entre DEN y HNL, y los que vuelan entre IAH y ANC. Por lo tanto este pico se debe exclusivamente a características de estos 2 vuelos. Puede ser un tema climático de la ruta que utilizan, o de la administración interna de los aeropuertos con esos vuelos. Sea cual sea el motivo, es responsabilidad de muy pocos y se manifiesta en el gráfico de ese modo debido a la poca cantidad de vuelos de tanta distancia.

Luego nos proponemos a aproximar funciones que representen la curva de desvíos utilizando cuadrados mínimos. Para esto nos propusimos un set de ecuaciones que fueron mejorando paulatinamente el error cuadrático medio. Primero utilizamos una ecuación lineal, de la forma $F_1 = \alpha_1 * x + \alpha_2$ y obtuvimos que $ECM(F_1) = 0.001351388$.

Luego intentamos aproximar mejor a la función utilizando una ecuación cuadrática. De este modo $F_2 = \alpha_1 * x^2 + \alpha_2 * x + \alpha_3$ y obtuvimos que $ECM(F_2) = 0.001290791$.

Finalmente quisimos incluir en nuestras ecuaciones el pico que hay en el segmento de 3001 a 3500 millas. Para esto se nos ocurrió utilizar la función gaussiana con la media y varianza adecuada para que se sitúe en el valor de x correcto. Y para representar el resto del gráfico utilizamos la función lineal. De este modo llegamos a la función $F_3 = \alpha_1 * e^{-\frac{(i-7)^2}{(2*0.25)^2}} + \alpha_2 * x + \alpha_3$ y obtuvimos que $ECM(F_3) = 0.001070443$, la más baja de las tres.

Por lo tanto consideramos que F_3 es la familia de funciones que mejor representa a nuestro dataset. Toma ligeramente su pendiente positiva y el pico del segmento. Por lo tanto es la que elegimos para representar en rojo

en el gráfico anterior. Tomamos 2005, 2006 y 2007 como entrenamiento y testeamos en 2008. Se puede observar que la aproximación es muy buena: es capaz de predecir el pico y los valores generales de la función

4 Conclusiones

Pudimos apreciar el poder que provee la herramienta Cuadrados Mínimos Lineales: un artilugio matemático *simple* de utilizar pero *no fácil*. **Su efectividad está atada directamente a la familia de funciones** provista. Por ende, dedicar recursos a hallar una familia acorde es de vital importancia, ya que una vez hecho eso el CML otorga función con errores pequeños.

En lo relevante a los ejes de estudio, fue claro que la ocurrencia de desvíos en viajes aéreos **efectivamente presenta una relación con la distancia entre origen y destino**, y temporal con **la época del año**. Más aún, pudimos plantear modelos que, como poco, ***muestran tendencias*** en el porcentaje de desvíos en función de los aspectos mencionados. Los aeropuertos, provistos de esta información, pueden realizar un análisis con mayor profundidad en los detalles técnicos, administrativos y/o relevantes a la organización de los vuelos para poder generar cronogramas que tengan en cuenta estas tendencias para reducir su caudal.

References

[1] G. Dahlquist and A. Bjorck, *Numerical methods*, Prentice-Hall. (1974), 167–171.

[2] http://stat-computing.org/dataexpo/2009/the-data.html