Aerolineas ¿Patrones definidos o pura coincidencia?

Joaquín Arribas, Ignacio Lebrero y Jésica Vázquez [1,2,3]

Departamento de Computación
Universidad de Buenos Aires
Buenos Aires, Argentina

Abstract

Todos los años se llevan a cabo miles de vuelos aéreos a nivel mundial, creando un sistema de viaje tanto nacional como internacional a gran escala. Esto genera un volumen de datos considerable con respecto a la información de los vuelos que se llevaron a cabo. En este trabajo nos proponemos encontrar patrones en base a distintos ejes que analizaremos a continuación. Sobre estos aplicamos *Cuadrados Mínimos* para encontrar una función que se ajuste a los patrones y poder ver si estos comportamientos son predecibles en tiempos posteriores o en otros aeropuertos.

Keywords: Cuadrados Mínimos, Análisis de patrones, Aeropuertos, Predicción de datos, Clima

[1] Email: joacoarribas@hotmail.com
[2] Email: ignaciolebrero@gmail.com
[3] Email: jeesivazquez@gmail.com

1 Introducción

La información utilizada en este trabajo son los datos de los vuelos nacionales de distintos aeropuertos de Estados Unidos. Dado que la cantidad de información acerca de los vuelos es extremadamente amplia, son varios los estudios que se pueden realizar. En este trabajo analizamos dos ejes:

- La cantidad de vuelos por semana.
- La cantidad de cancelaciones por semanas.

Para el primer eje nos propusimos observar como varían a través de los años la cantidad de vuelos semanales de varios aeropuertos. Para esto elegimos los aeropuertos con más caudal de vuelos según [1].

Para el segundo eje analizamos el comportamiento de la cantidad de cancelaciones por semana durante un año, a lo largo de varios años. Para esto elegimos dos aeropuertos, uno situado al noreste y otro al noroeste de Estados Unidos, y analizamos su comportamiento tomando como referencia uno para poder realizar estimaciones sobre el otro.

En estos experimentos logramos predecir la mayoría de los patrones[4] e incluso predecir a futuro el comportamiento de los mismos aeropuerto. Desarrollaremos con más profundidad estos temas en las próximas secciones.

2 Cantidad de vuelos por semana

Para el primer análisis decidimos ver la cantidad de vuelos por semana a lo largo de varios años. En el caso de lograr una predicción ajustada, esta información podría ser utilizada para calcular insumos necesarios en los próximos años dependiendo de la cantidad de vuelos que se realizarán.

2.1 Primeras hipótesis y bases del análisis

Nuestra hipótesis respecto de este eje es que la cantidad de vuelos de cada aeropuerto debería aumentar a lo largo de los años dado que el tráfico aéreo aumenta a la par. Es por esta razón que realizamos los experimentos sobre los aeropuertos que más vuelos programados tienen.

Para ver qué conjunto de datos nos convenía utilizar para hacer los experimentos graficamos todos los años entre 1998 y 2008. Consideramos los datos

[4] El método utilizado para las estimaciones fue *Cuadrados Mínimos Lineales*, implementación provista por la librería *scipy* de *Python*.

a partir de 1998 ya que antes de este año no hay muchos datos respecto de los aeropuertos. Por ejemplo, uno de los resultados fue el siguiente:

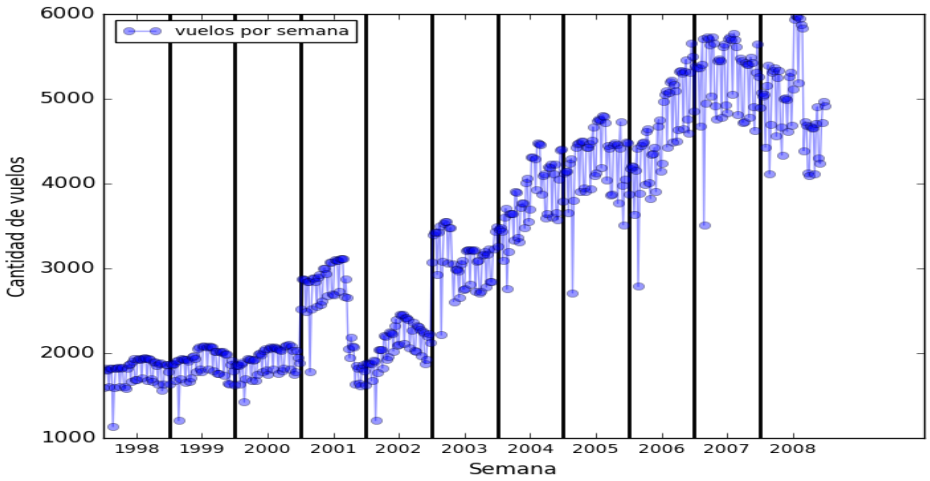

Fig. 1. Cantidad de vuelos por semana en el aeropuerto JFK [5] desde 1998 hasta 2008

Podemos observar en este gráfico cómo a partir del 2001 cambian drásticamente los datos, por lo cual no creemos que tenga sentido realizar experimentos donde se consideren los años inmediátamente previos y posteriores al 2001. Pensamos esto ya que lo sucedido en septiembre de ese año no fue un hecho predecible, y por lo tanto la variación de datos ocurrida no podría ser estimada. Por ende los experimentos realizados en este eje fueron a partir del 2003, año en el cual consideramos que se estabilizan los datos.

Una observación al respecto de este gráfico y todos los siguientes es que en todos los años la última semana de febrero tiene un pico bajo de vuelos. No pudimos encontrar la razón por la cual sucede esto, lo único que pudimos observar es que esa fecha coincide con el intervalo de clases entre dos períodos de vacaciones: el *president's day* y *spring break*.

2.2 Datos concretos y estimaciones

Para realizar los gráficos acumulamos por cada semana de cada año la cantidad de vuelos sucedidos. Por ejemplo, uno de los experimentos realizados fue sobre

[5] John F. Kennedy, aeropuerto de New York

el aeropuerto *John F. Kennedy* (JFK), variando los años de entrenamiento y estimando los siguientes años:

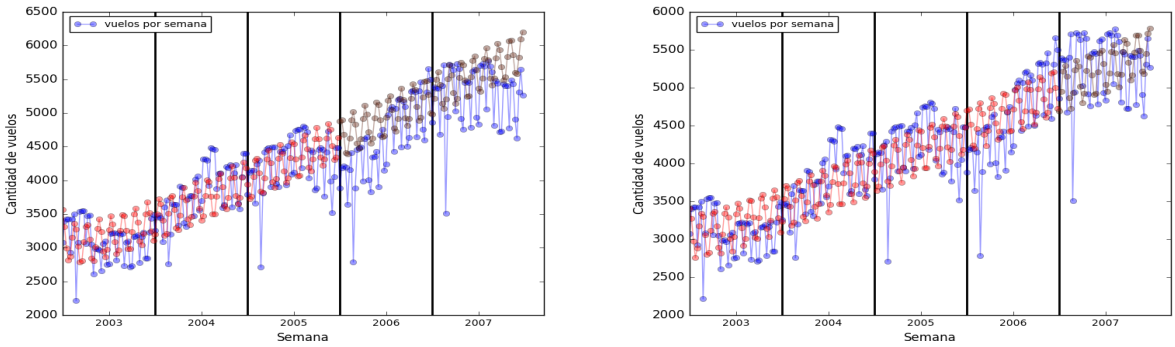

Fig. 2. Estimación 2006-2007 entrenando con 2003-2005 (a la izquierda) y estimación 2007 entrenando con 2003-2006 (a la derecha)

Los errores cuadráticos medios para estas estimaciones fueron 365916.778 (izquierda), y 251920.254 (derecha). La función utilizada con CML para esta estimación fue:

$$f(x) = ax + b|cos(x)| + c|sin(x)| + 300cos(x) + sin(x) + dlog(x+1) + e$$

La intuición que tuvimos para encontrar la función fue:

- Los cosenos y senos fueron utilizados por la frecuencia y amplitud de los datos. Variamos los valores que los multiplicaban para ver como se ajustaba mejor la función.
- La función logaritmo fue utilizada dado que notamos que los últimos años los valores se estabilizaban.
- La funcón lineal fue utilizada dado que la relación que encontramos año a año respecto de la cantidad de vuelos era lineal.

Experimentamos con varias otras familias de funciones, por ejemplo combinando senos y cosenos con funciones polinómicas de mayor grado que una lineal. En ninguna oportunidad pudimos disminuir el error cuadrático medio, a su vez que en el gráfico también se notaba que la estimación era peor.

Experimentamos disminuyendo la cantidad de años con la cual entrenamos el método, pero como la relación que tienen los primeros dos años no es la misma que el resto (ver años 2003 y 2004), las estimaciones no fueron buenas.

Otro experimento realizado en base a estos datos fue entrenar el método utilizando como base de datos los años 2003 al 2006 de JFK, y tratar de

estimar a futuro que es lo que sucedería en aeropuertos que se comportan de manera parecida en relación a la cantidad de vuelos. El experimento que mejor resultados dió fue sobre el aeropuerto internacional de San Francisco (SFO):

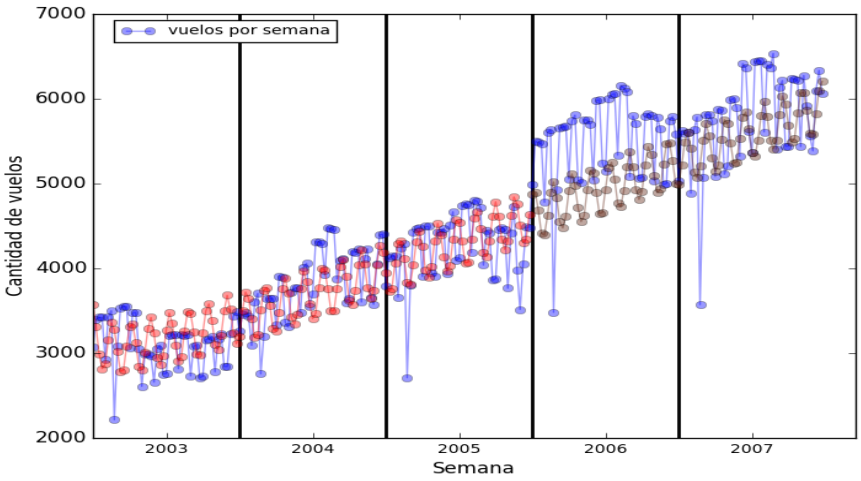

Fig. 3. Estimación 2006-2007 SFO entrenando con JFK 2003-2005

El error cuadrático medio en esta estimación fue 446323.5060

3 Cancelaciones por clima

Para el segundo análisis decidimos ver la cantidad de cancelaciones semanales por año. Estas situaciones se pueden dar por varios factores, entre otras:

- Problemas climáticos.
- Problemas de seguridad.
- Tráfico de aviones.
- Problemas propios del avión.

En particular analizaremos las cancelaciones surgidas por clima, donde observaremos los cambios a lo largo del año, prestando particular atención a la época invernal.

3.1 Primeras hipótesis y bases del análisis

Nuestra hipótesis respecto de este eje es que la cantidad de cancelaciones por condiciones climáticas de los vuelos tendería a cambiar cerca de la época invernal debido al clima inestable de esas fechas.

Para el análisis inicial decidimos tomar como punto de partida el aeropuerto Ronald Reagan [6] en Washington D.C., ubicado al noroeste de los Estados Unidos. Como es un aeropuerto nacional, evitamos el *ruido* que podrían llegar a generar los vuelos internacionales sobre los datos y por otro lado, al ser muy concurrido, podemos tener una *base significativa para los datos*. En particular, en el invierno de los estados del norte se experimentan fuertes nevadas y lluvias heladas.

Para estos experimentos utilizamos los datos de los aeropuertos entre los años 2004-2008. En todos los aeropuertos que experimentamos no encontramos datos previos al 2004 respecto de las cancelaciones por clima. Para experimentar seleccionamos aeropuertos al norte de Estados Unidos dado que pensamos que la mayoría de las cancelaciones por clima deberían suceder al norte del país (por las condiciones inestables mencionadas previamente). Por ejemplo, observamos en los siguientes gráficos la cantidad de cancelaciones en dos aeropuertos distintos del país, uno situado al noroeste (DCA) y el otro al sureste (MIA):

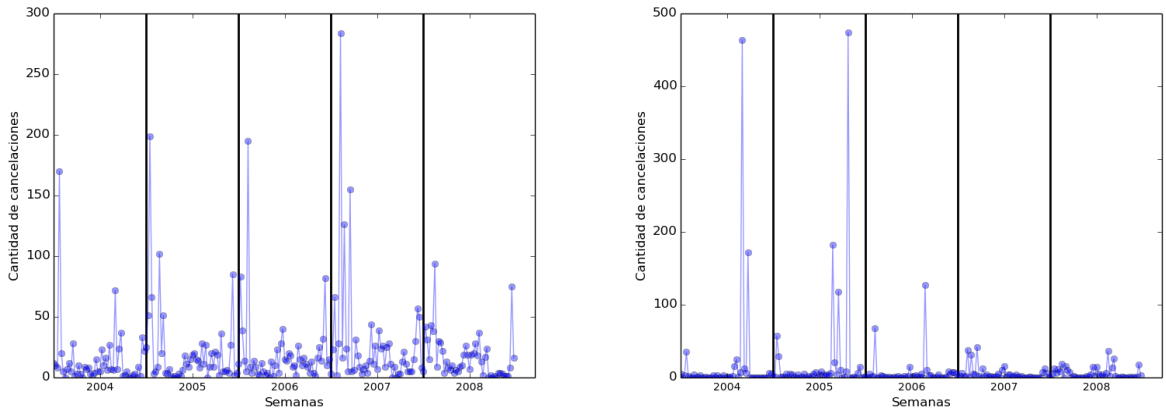

Fig. 4. Cantidad de cancelaciones por clima en DCA (a la izquierda) y en MIA (a la derecha) [7] desde 2004 hasta 2008

[6] Codigo IATA: DCA
[7] Aeropuerto internacional de Miami

Podemos ver como la cantidad de cancelaciones por clima difieren bastante en estos aeropuertos. En DCA vemos como la cantidad de cancelaciones se mantiene estable a lo largo del año (entre 0 y 50 cancelaciones por semana) evidenciando picos alrededor del invierno. Por otro lado, el aumento de cancelaciones que se puede observar en la época de verano podría estar dada por las frecuentes tormentas que además ocasionalmente producen tornados [2].

Por el otro lado, en MIA se evidencian poca cantidad de cancelaciones (entre 0 y 20) que creemos que se debe al clima más estable del sur, a pesar de las frecuentes lluvias de la zona. Se puede ver un aumento de cancelaciones en los primeros meses del año, que creemos que puede ser causado por las tomentas eléctricas y la brisa marina que hay en esa época. Para los picos aislados, no encontramos una razón particular, pero Miami sufre de abundantes lluvias entre mitad de Mayo e inicios de Octubre, con los cual los picos podrían coincidir con ellas [3].

3.2 Datos concretos y estimaciones

Los siguientes gráficos fueron hechos en base a la cantidad de cancelaciones por semana en el aeropuerto DCA, variando la cantidad de años con la cual se entrena al método:

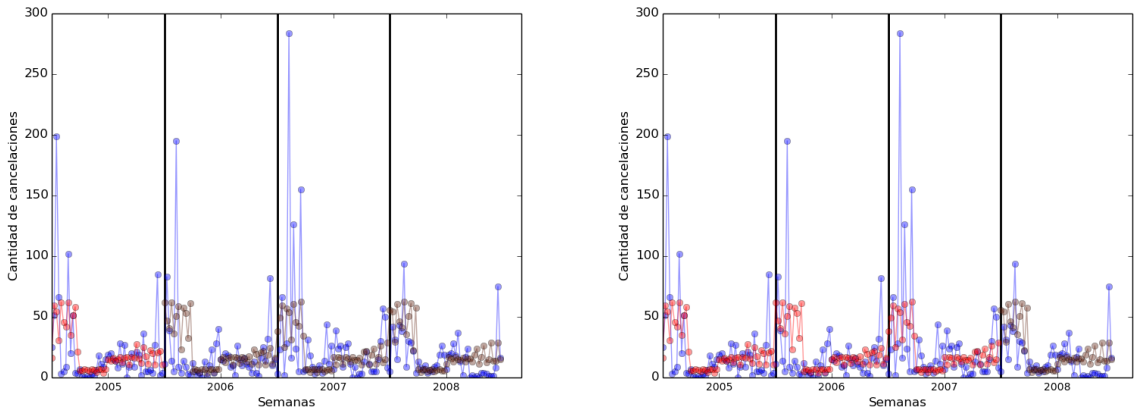

Fig. 5. Cantidad de cancelaciones por clima en DCA entrenando con 2005 y estimando 2006-2008 (a la izquierda), y entrenando con 2005-2007 y estimando 2008(a la derecha)

Esta estimación fue hecha partiendo el año en cuatro porciones (una por cada estación del año) y utilizando el método CML sobre dichas particiones. Consideramos necesario separar los datos en intervalos dado que estos se com-

portan de manera distinta dependiendo de la época del año. Luego, para estimar los siguientes años, se utilizaron para cada estación los parámetros que el método encontró para la siguiente función:

$$f(x) = a|cos(x)| + b|sen(x)| + c$$

La intuición que tuvimos para encontrar la función fue:

- El coseno y el seno fueron utilizados por la frecuencia y amplitud de los datos. Notamos que los datos cumplen cierta periodicidad dentro de cada estación.
- Aplicamos valor absoluto ya que los valores a estimar son mayores o iguales que cero.

Inicialmente los experimentos fueron pensados utilizando los datos desde el año 2004, pero luego de aplicar cross validation sobre los experimentos, decidimos tomar los datos del 2005 para entrenamiento, ya que en el 2004 hubo un comportamiento anormal a fin de año que no pudimos explicar y los del 2005 fueron los que mejor sirvieron para aproximar a futuro. El error cuadrático medio de las aproximaciones a futuro en Fig.5 fueron 1007,82 (izquierda) y 378,85 (derecha).

Experimentamos con otros aeropuertos que cumplan con las mismas características (situarse al norte del país), utilizando la misma función, y el efecto fue el mismo. El segundo aeropuerto utilizado fue el aeropuerto internacional Boston Logan ubicado en Boston al noreste del país. Los resultados fueron los siguientes:

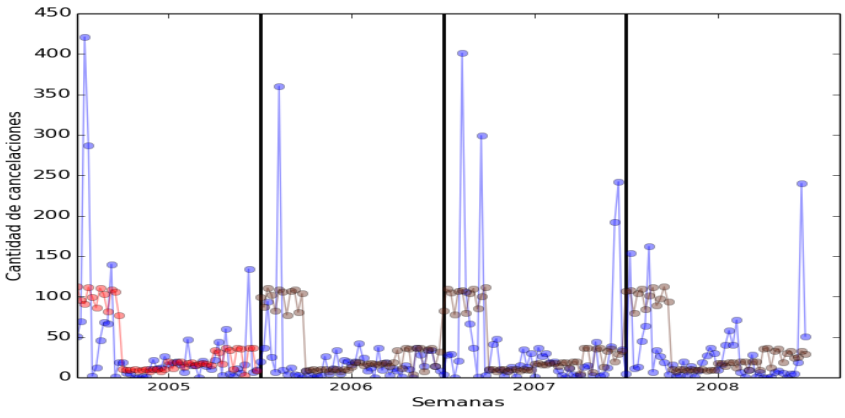

Fig. 6. Cantidad de cancelaciones por clima en BOS entrenando con 2005 y estimando 2006-2008.

Podemos ver que a través de los años los patrones de invierno y fines de otoño se repiten y el modelo logra predecir el comportamiento de creciemiento y decrecimiento de los mismos. El error cuadratico medio de este experimento fue 3725.6399

Otro experimento consistió en reutilizar los parámetros obtenidos por CML en el primer aeropuerto (DCA) y aplicar el método en el segundo (BOS) en el rango de años 2006 hasta 2008. Es decir, entrenamos con los datos del aeropuerto DCA en el año 2005 y predecimos las cancelaciones de aeropuerto de Boston del 2006 al 2008. Queríamos ver si con las estimaciones realizadas en un aeropuerto con las mismas características que otro se podía estimar los años futuros.

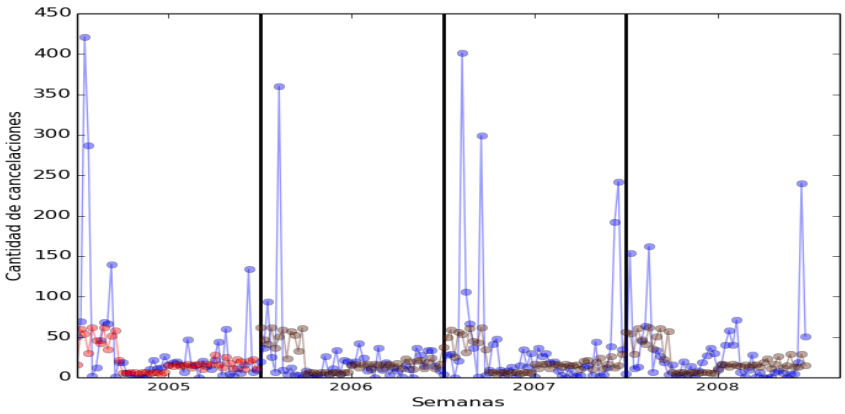

Fig. 7. Entrenamiento con el aeropuerto DCA con el 2005 y estimación del aeropuerto BOS entre 2006 y 2008

Como se puede observar la predicción se comporta bastante parecida a los datos reales. Si bien no se acopla exactamente con los datos, los patrones de crecimiento y decrecimiento son muy similares. Por otro lado, aunque los valores que predicen los picos no se acercan mucho a los valores reales, el modelo se asemeja bastante en lo que respecta a en qué momento es más probable que ocurra un aumento de cancelaciones. El error cuadratico medio en este caso fue 3400.4013, hecho curioso dado que entrenando y estimando sobre el mismo aeropuerto, el error cuadrático medio fue mayor.

4 Conclusión

En este trabajo mostramos dos posibles ejes a analizar que evidencian características comunes sobre algunos aeropuertos de Estados Unidos. Creemos que logramos predecir con cierta efectividad el comportamiento de estos ejes a futuro.

En cuanto a cantidad de vuelos por año pudimos observar un aumento a lo largo del tiempo con predicciones bastante acertadas. Suponemos que esto se mantendrá por un tiempo y eventualmente cambiará por motivos externos (saturación del espacio aéreo, capacidad del aeropuerto, etc).

Por otro lado, la cantidad de cancelaciones mostraron un incremento en las épocas invernales, lo cual se pudo predecir con cierta eficacia por la naturaleza del eje. Esto suponemos que se mantendrá similar a lo largo de los años variando la escala de las cancelaciones debido al aumento progresivo de vuelos visto en el otro eje.

Sin embargo es posible que aún no tengamos las herramientas necesarias para hacer un análisis profundo de los datos. Como mencionamos en la introducción son varios los estudios que se pueden hacer sobre los datos y utilizando herramientas más avanzadas podríamos encontrar patrones más complejos.

En cuanto a las predicciones con cuadrados mínimos lineales, al utilizar un modelo simple se logran predicciones aproximadas por el tipo de familias de funciones que pueden ser aplicadas. Esto brinda una herramienta de rápido uso a la hora de encontrar patrones simples en datos no muy complejos.

References

[1] Aeropuertos con más vuelos:
https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States

[2] Clima en Washington:
https://en.wikipedia.org/wiki/Washington,_D.C.#Climate

[3] Clima en Miami:
https://en.wikipedia.org/wiki/Climate_of_Miami