Aerolíneas ¿Patrones definidos o pura coincidencia?

Métodos Númericos

Presentación: Joaquín Arribas, Ignacio Lebrero, Jésica Vázquez

Presentación

Los servicios de transporte requieren diferentes tipos de planificaciones a lo largo del tiempo, para poder optimizar su uso. En este trabajo nos focalizaremos en el servicio de transporte aéreo, utilizando datos de los aeropuertos de Estados Unidos entre los años 1998 y 2008.
El presente trabajo pretende dar una noción respecto de la cantidad de ejes distintos a analizar en base a estos datos, y cómo se podrían utilizar para mejorar el servicio de transporte aéreo a futuro.

Posibles estudios

Por ejemplo, algunos estudios que se pueden realizar en base a estos datos son:

- ¿Como varían la cantidad de delays a lo largo de los años?
- Las cancelaciones de vuelos ¿Suelen ser por alguna razón particular? ¿Como varían a lo largo de los años?
- El año en el que fue construido el avión ¿Afecta su *performance*?
- ¿Existe alguna diferencia entre los aeropuertos de distintas zonas?
- ¿Qué granularidad es necesaria para obtener datos representativos?

Análisis hechos

Analizaremos dos ejes en particular:

- ¿Cómo varían a lo largo de los años las cancelaciones por clima? ¿Afecta la zona en la cual este situado el aeropuerto?
- ¿Cómo varían a lo largo de los años la cantidad de vuelos?

Cancelaciones por clima

Para este estudio consideramos inicialmente el aeropuerto Ronald Reagan en Washington [1] dado que es uno de los aeropuertos más al norte de Estados Unidos. Buscamos inicialmente un aeropuerto de estas características ya que nos interesaba ver si las cancelaciones por clima reflejaban las condiciones climáticas que sufren estos estados en invierno. Analizamos en un principio qué granularidad usar:

[1] IATA: DCA

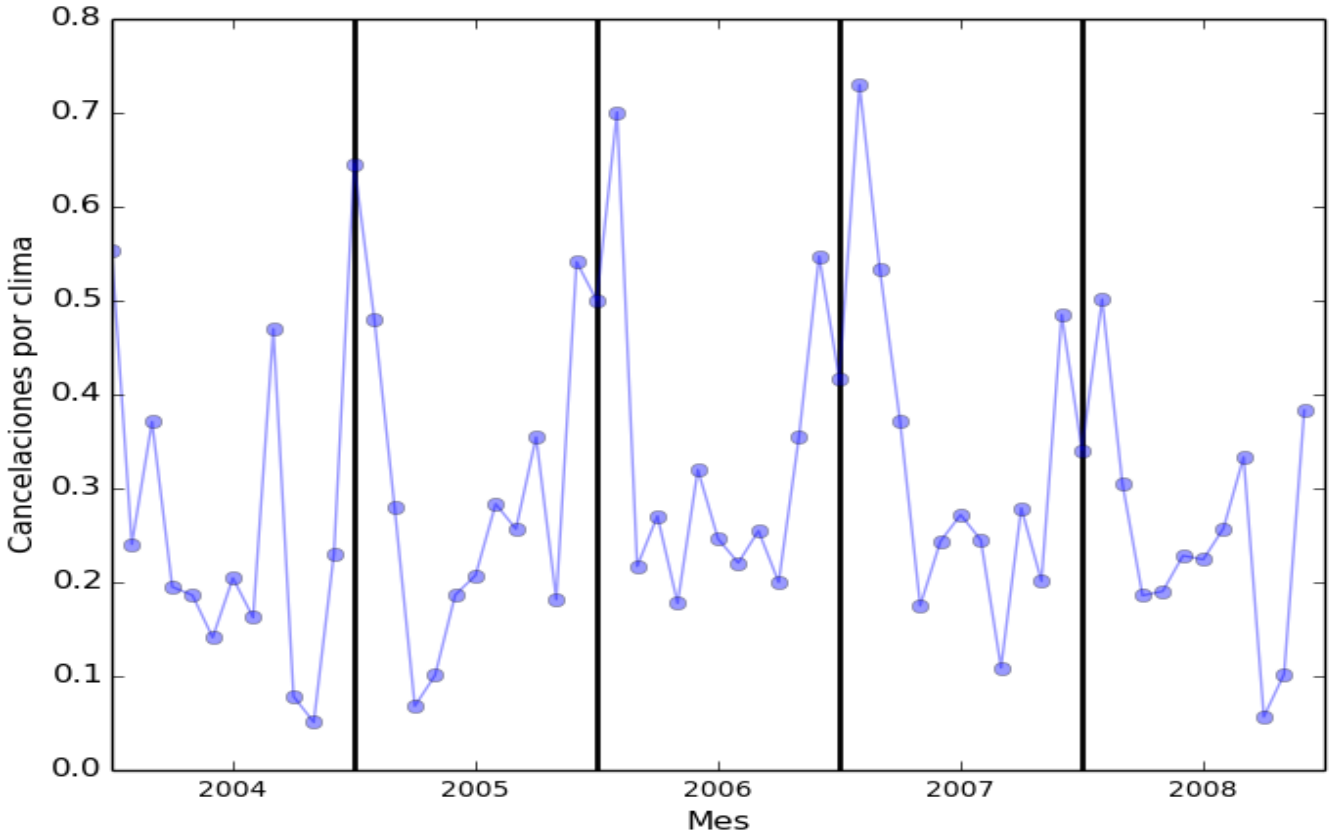


Figura: Cancelaciones de clima por mes DCA

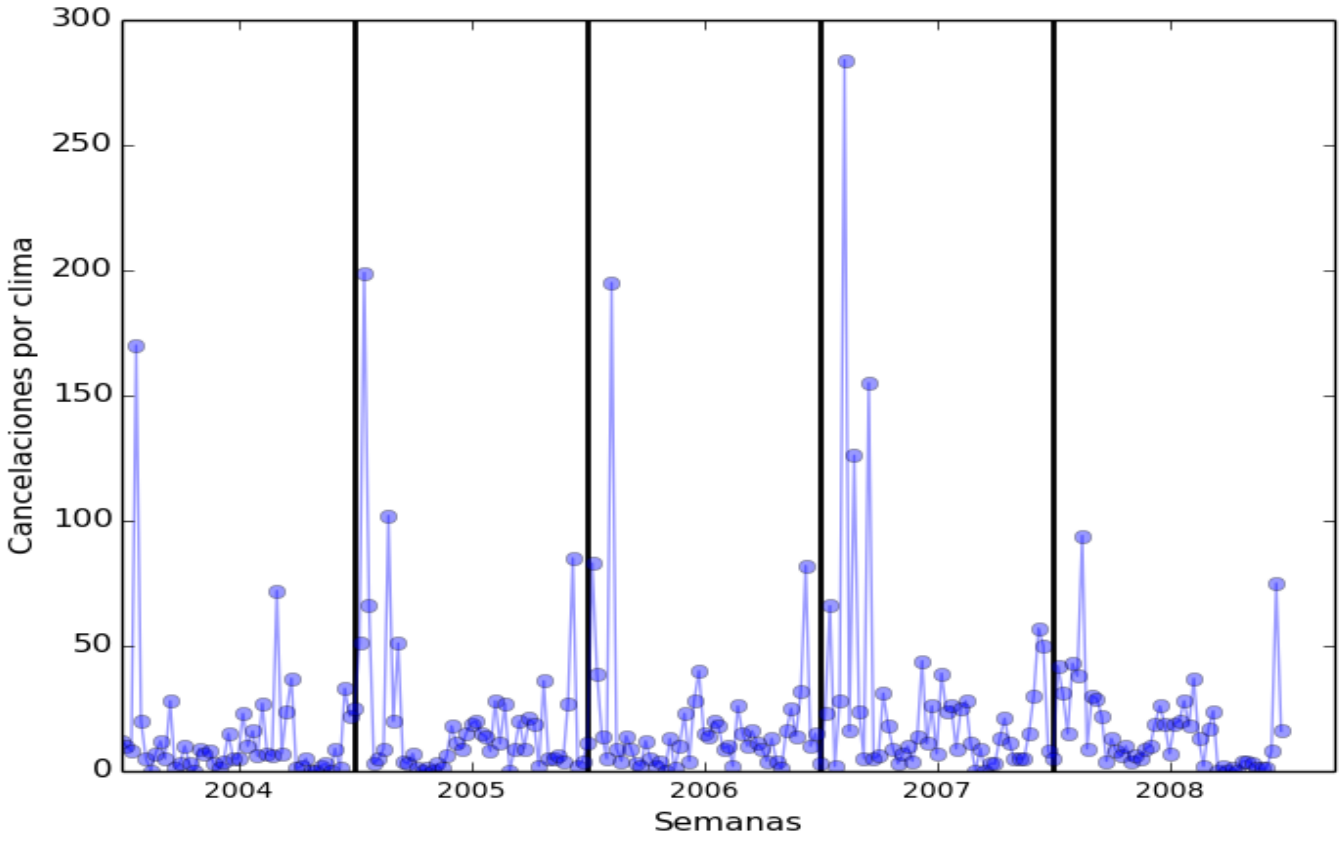
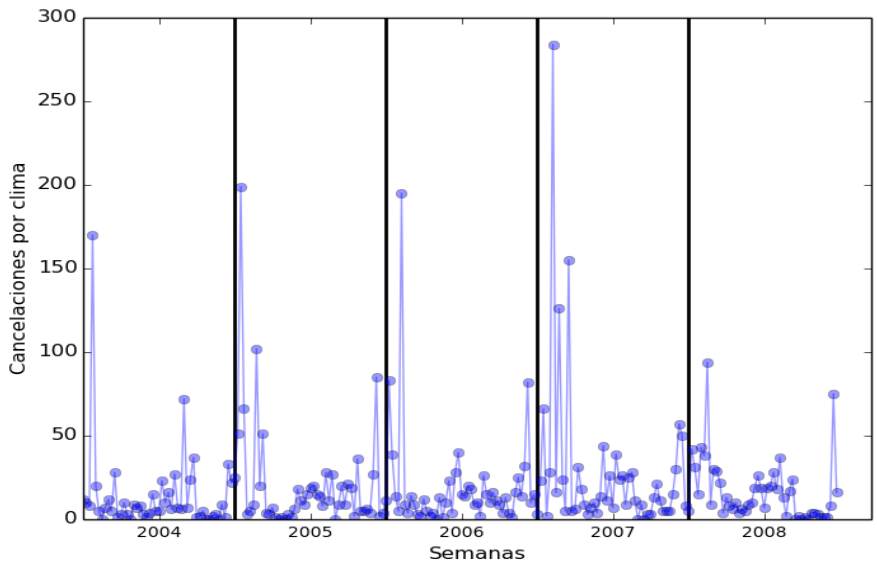


Figura: Cancelaciones de clima por semana DCA

Observaciones de los datos

- En épocas de invierno se ven picos que podrían representar las cancelaciones debidas a las fuertes nevadas y lluvias propias de la etapa.
- En los meses de Junio a Agosto de evidencia un aumento en la cantidad de cancelaciones que creemos que podría estar ligada a las frecuentes tormentas de la época.
- La cantidad de cancelaciones fuera de esas temporadas parece similar.



Estimaciones

Para realizar las estimaciones utilizamos 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.

Error cuadrático medio

Para analizar qué tan robusto es el modelo utilizaremos la métrica de evaluación Error Cuadrático Medio. Sea y_i el elemento del set de datos, n la cantidad de datos y y_i' el elemento de la estimación provista por CML se define este error como:

$$1/n * \sum_{i=1}^{n} (y_i - y_i')^2$$

Esto es, el promedio de las distancias entre los datos originales y la estimación.

Resultados

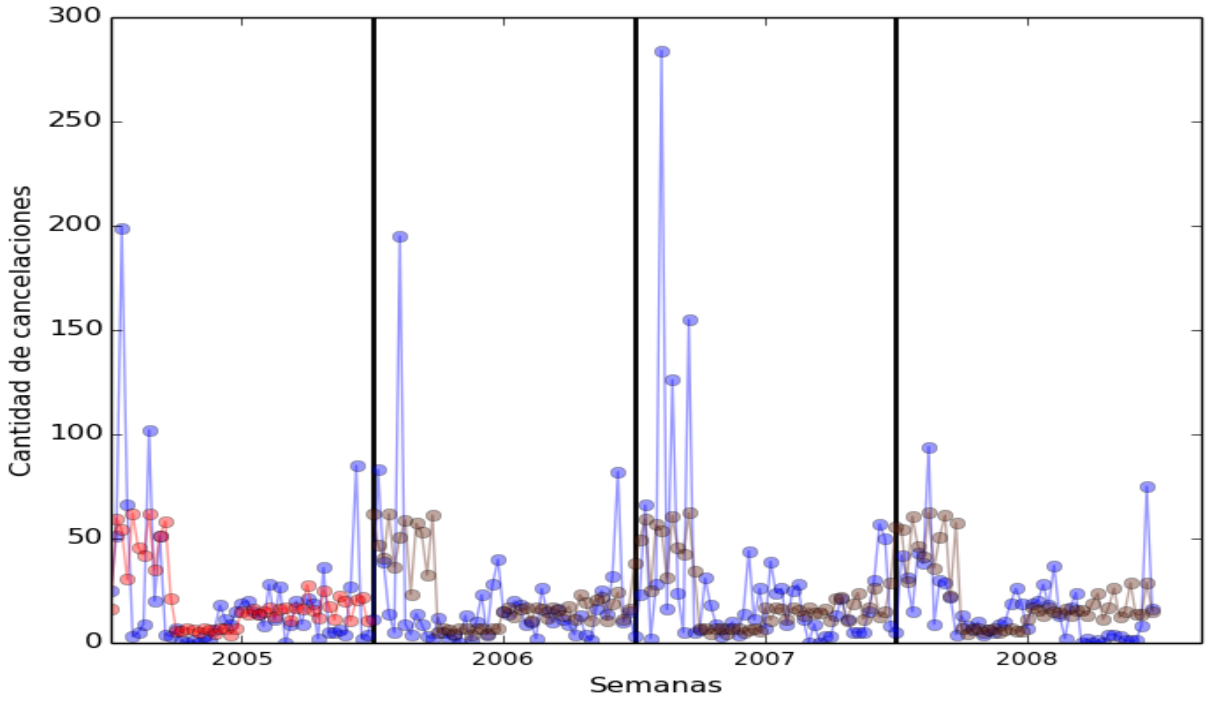


Figura: Cantidad de cancelaciones por clima en DCA entrenando con 2005 y estimando 2006-2008

El error cuadrático medio de las aproximaciones a futuro es de 1007,82.

Resultados

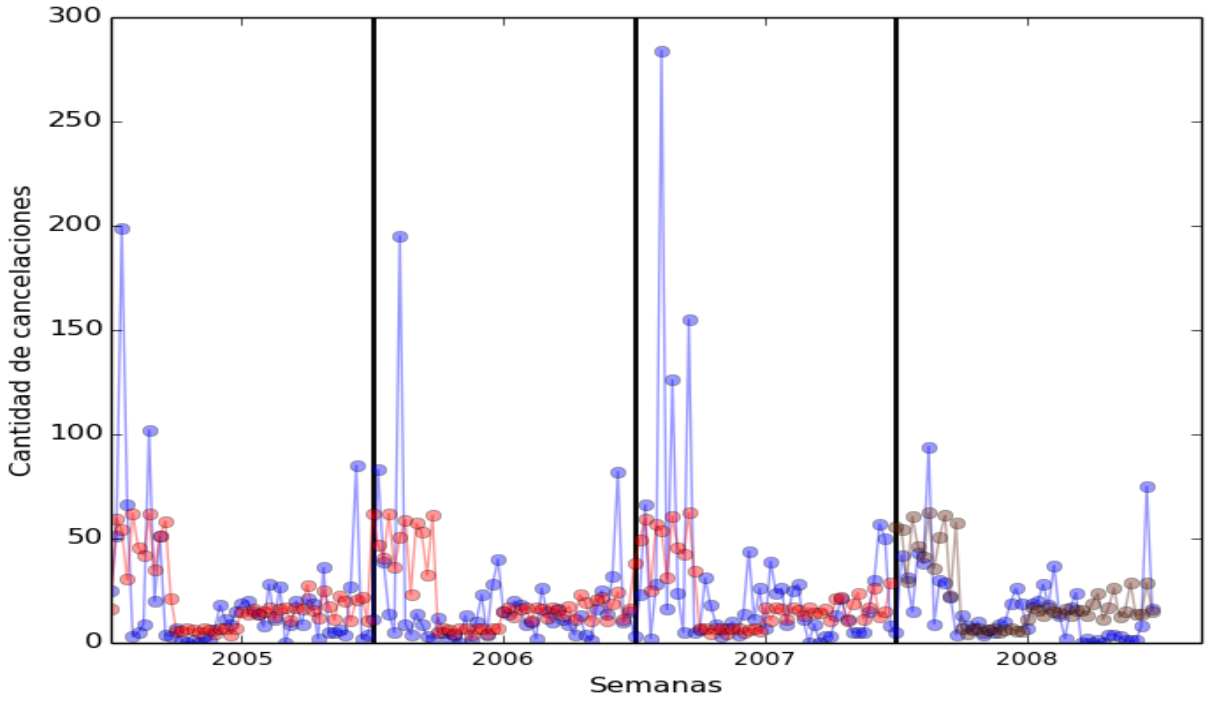


Figura: Cantidad de cancelaciones por clima en DCA entrenando con 2005-2007 y estimando 2008

El error cuadrático medio de las aproximaciones a futuro es de 378,85.

Resultados

Probamos el modelo en un aeropuerto de características similares al anterior (el aeropuerto internacional Boston Logan, situado al noreste del país):

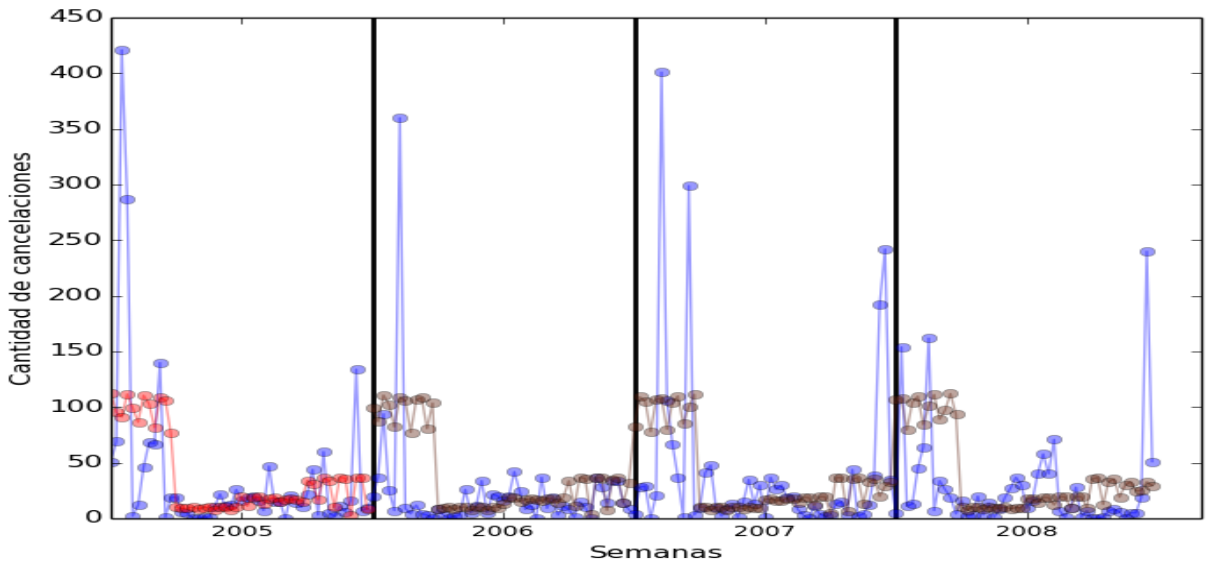


Figura: Cantidad de cancelaciones por clima en BOS entrenando con 2005 y estimando 2006-2008.

El error cuadrático medio de este experimento fue 3725.6399

Resultados

Por último, entrenamos con un aeropuerto y predijimos sobre el otro:

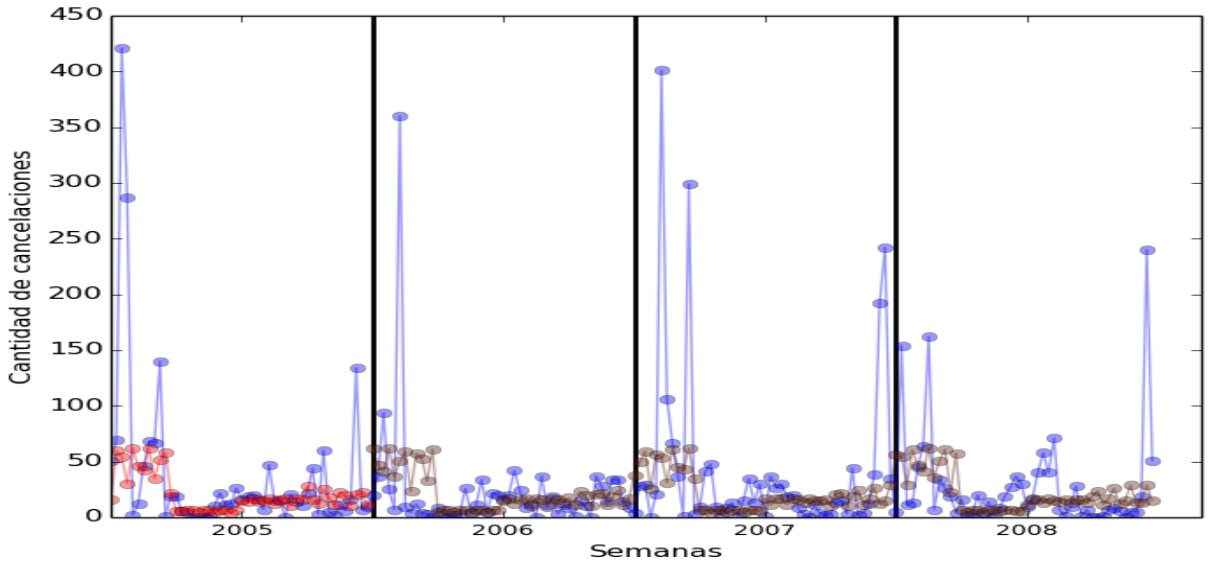


Figura: Cantidad de cancelaciones por clima en BOS entrenando con DCA en el año 2005 y estimando en BOS en el período 2005-2008.

El error cuadrático medio en este caso fue 3400.4013.

Cantidad de vuelos

Para este estudio consideramos los aeropuertos con más caudal de vuelos. Creemos que en base a este análisis se podría estimar como crecerá en años futuros. En base a esto, y combinándolo con otros, se podrían optimizar gastos y servicios de las aerolíneas.

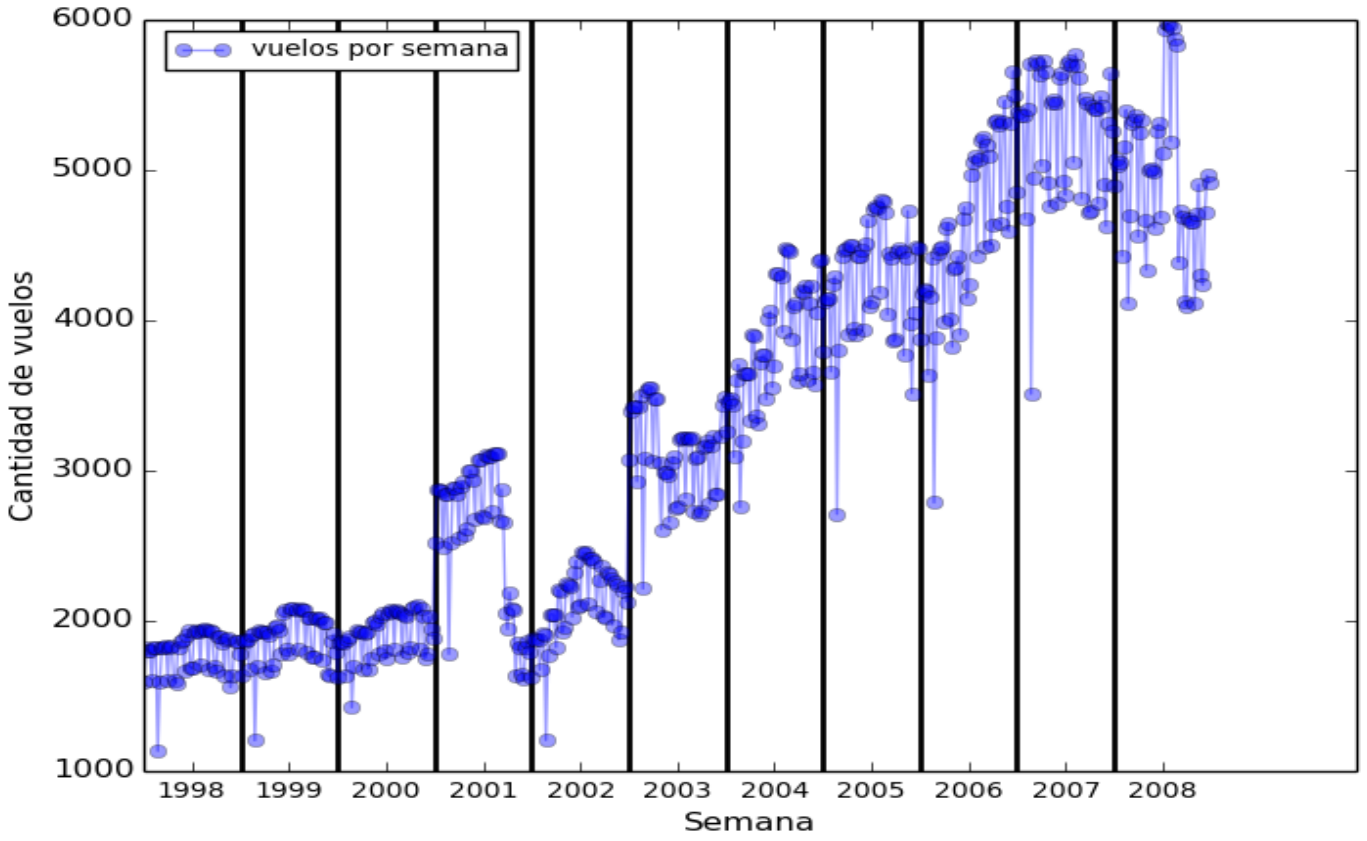


Figura: Cancelaciones de clima por semana JFK

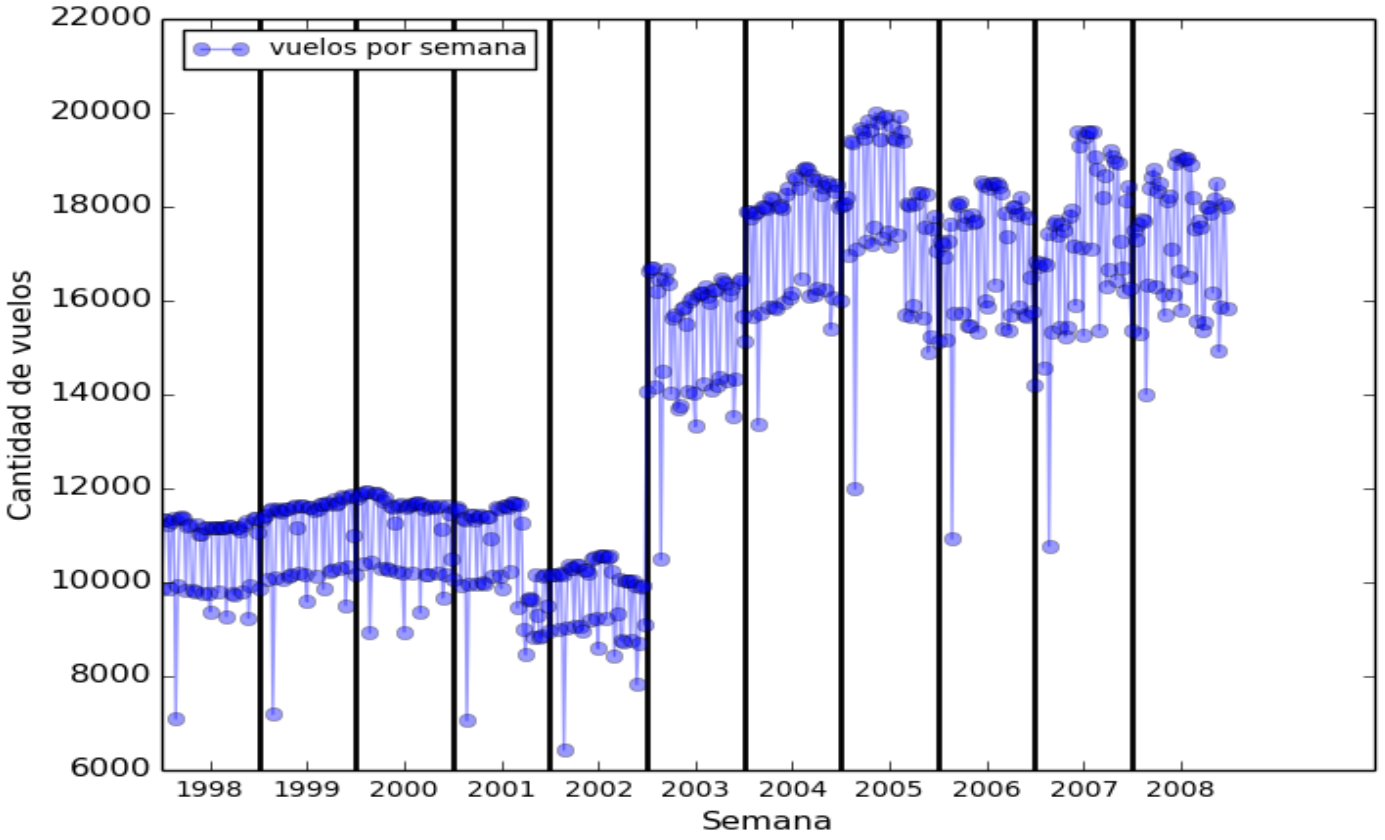


Figura: Cancelaciones de clima por semana ATL

Estimaciones

Para realizar las estimaciones utilizamos la siguiente función:

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

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 función lineal fue utilizada dado que la relación que encontramos año a año respecto de la cantidad de vuelos era lineal.

Resultados

Estos experimentos fueron realizados sobre el aeropuerto John F. Kennedy(JFK) variando los años de entrenamiento y estimando los siguientes años.

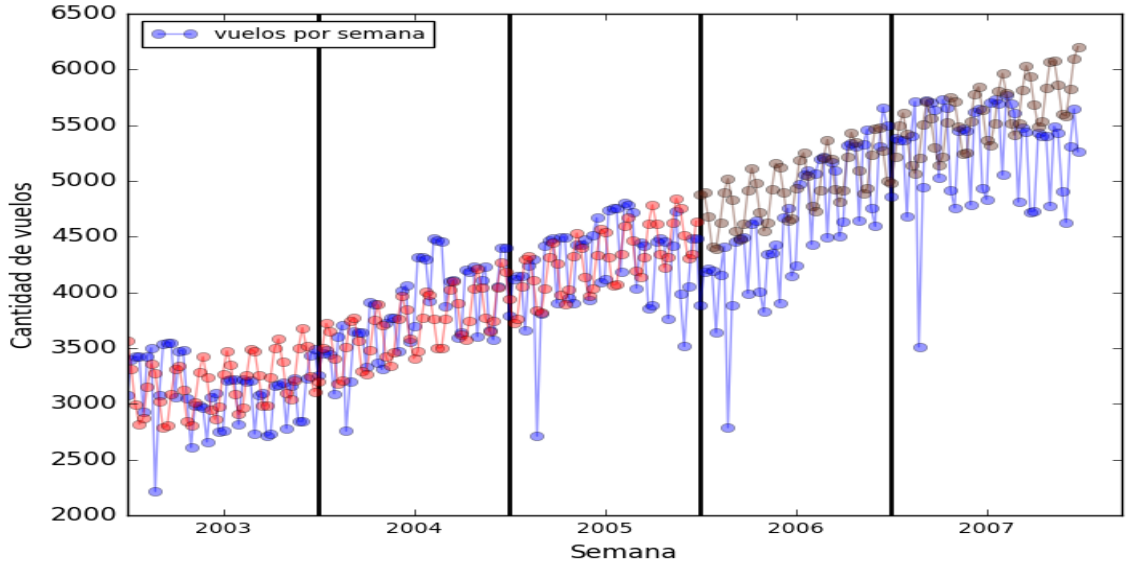


Figura: Estimación 2006-2007 entrenando con 2003-2005.

El error cuadrático medio para esta estimación fue 365916.778.

Resultados

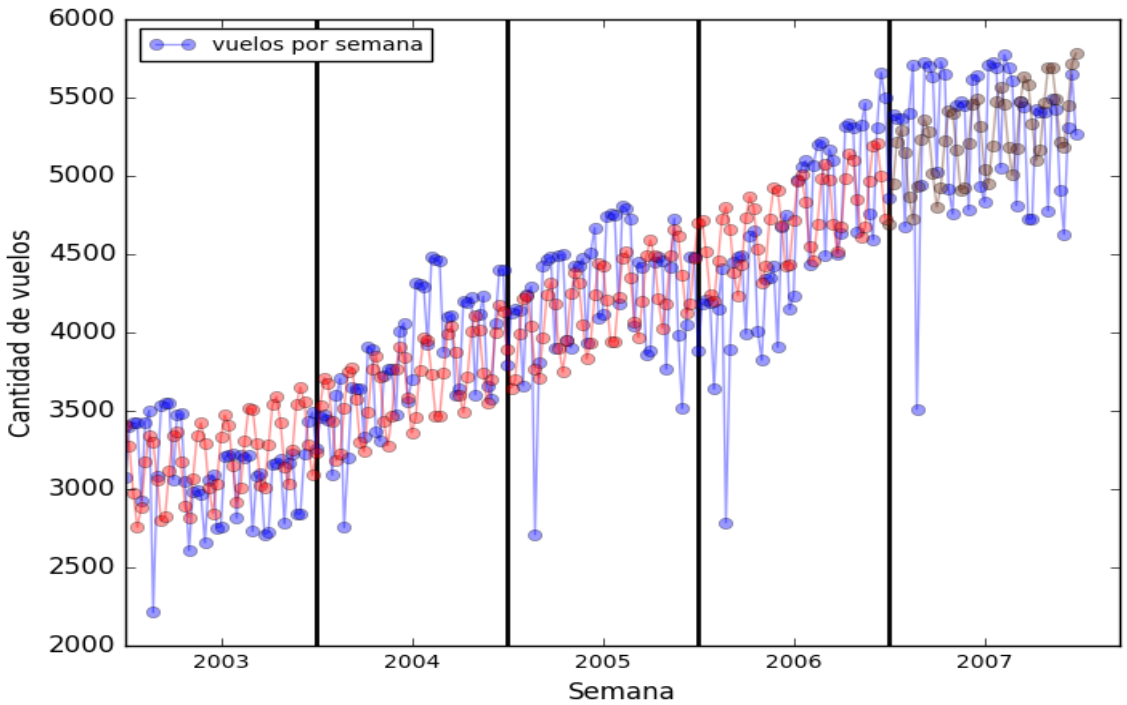


Figura: Estimación 2007 entrenando con 2003-2006.

El error cuadrático medio para esta estimación fue 251920.254.

Resultados

En estos ejemplos podemos ver como afecta la cantidad de años utilizados para entrenar:

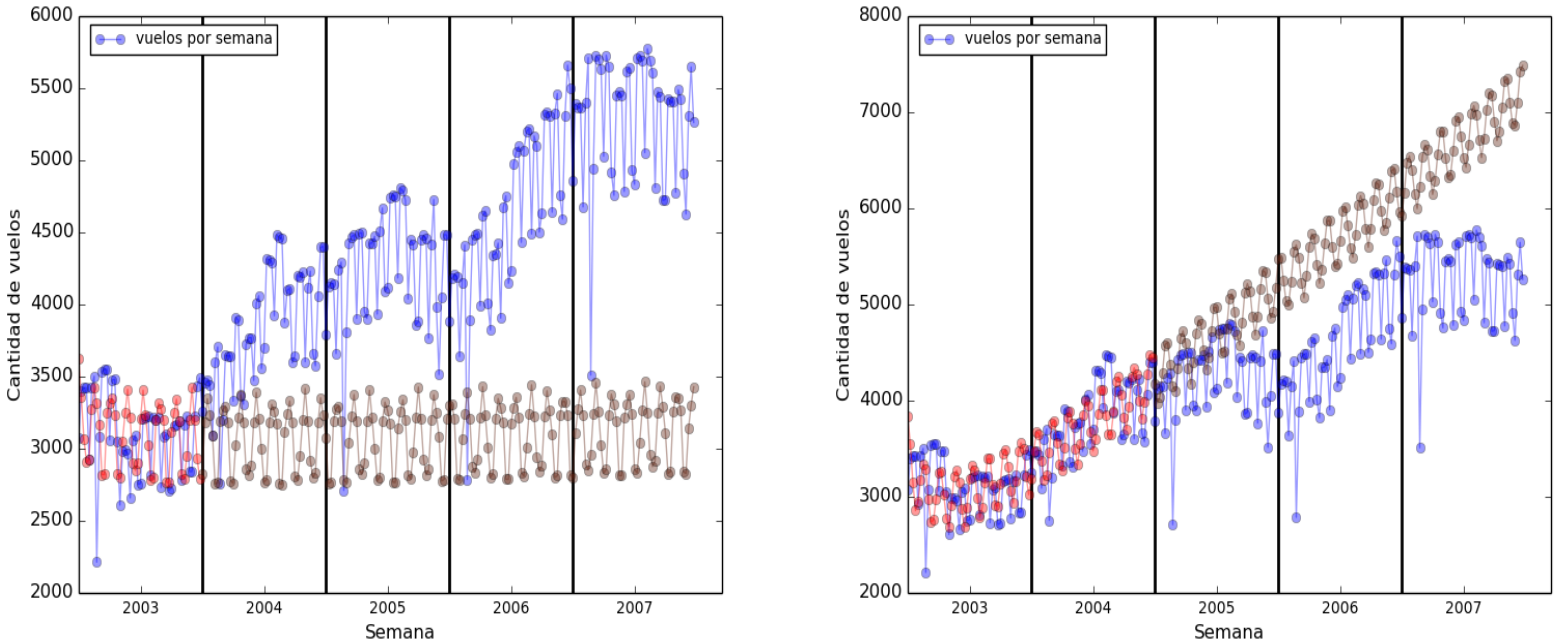


Figura: Estimación 2004-2007 entrenando con el 2003 (a la izquierda) y estimación 2005-2007 entrenando con 2003-2004 (a la derecha).

Resultados

Otro experimento fue entrenar con datos del aeropuerto JFK y predecir datos de un aeropuerto de comportamiento similar (con respecto a la cantidad de vuelos) como lo es el aeropuerto internacional de San Francisco (SFO).

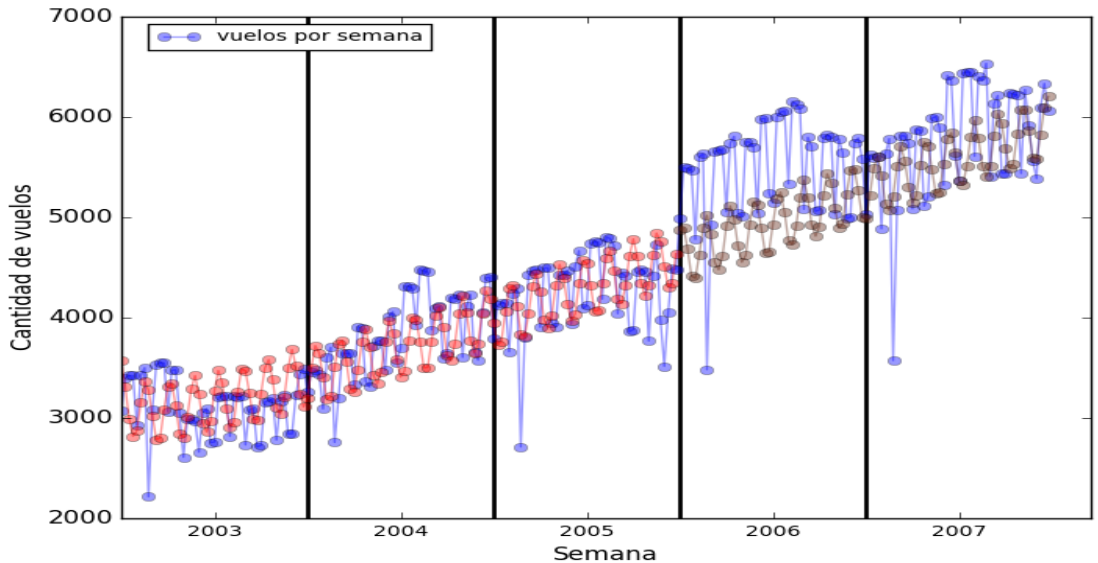


Figura: Estimación 2006-2007 SFO entrenando con JFK 2003-2005

El error cuadrático medio en esta estimación fue 446323.5060

Conclusiones generales (1)

Con respecto a las herramientas utilizadas podemos decir que:

- CML presenta una herramienta útil para estimación simples.
- El error cuadrático medio da una noción cuantitativa sobre la predicción. Sin embargo, esto depende de la semántica del experimento.

Conclusiones generales (2)

Con respecto a los ejes investigados:

- La cantidad de vuelos mostró un crecimiento predecible al menos en los años estimados. Un posible análisis futuro es ver si esto continua o si eventualmente se satura el espacio aéreo del aeropuerto.
- La cantidad de cancelaciones por clima mostró ciertos patrones en épocas invernales. A futuro suponemos que estos patrones se mantendrán cambiando solo la escala debido a lo mostrado en el otro eje.

Otros posibles ejes

Nos dedicaremos en esta sección a mostrar otros ejes que intentamos analizar, pero no encontramos evidencia suficiente como para suponer algo, o no tenemos las herramientas necesarias como para hacer el análisis adecuado.

Delay por hora

Uno de los ejes analizados fue cómo varía el delay de los vuelos en cada hora el día.

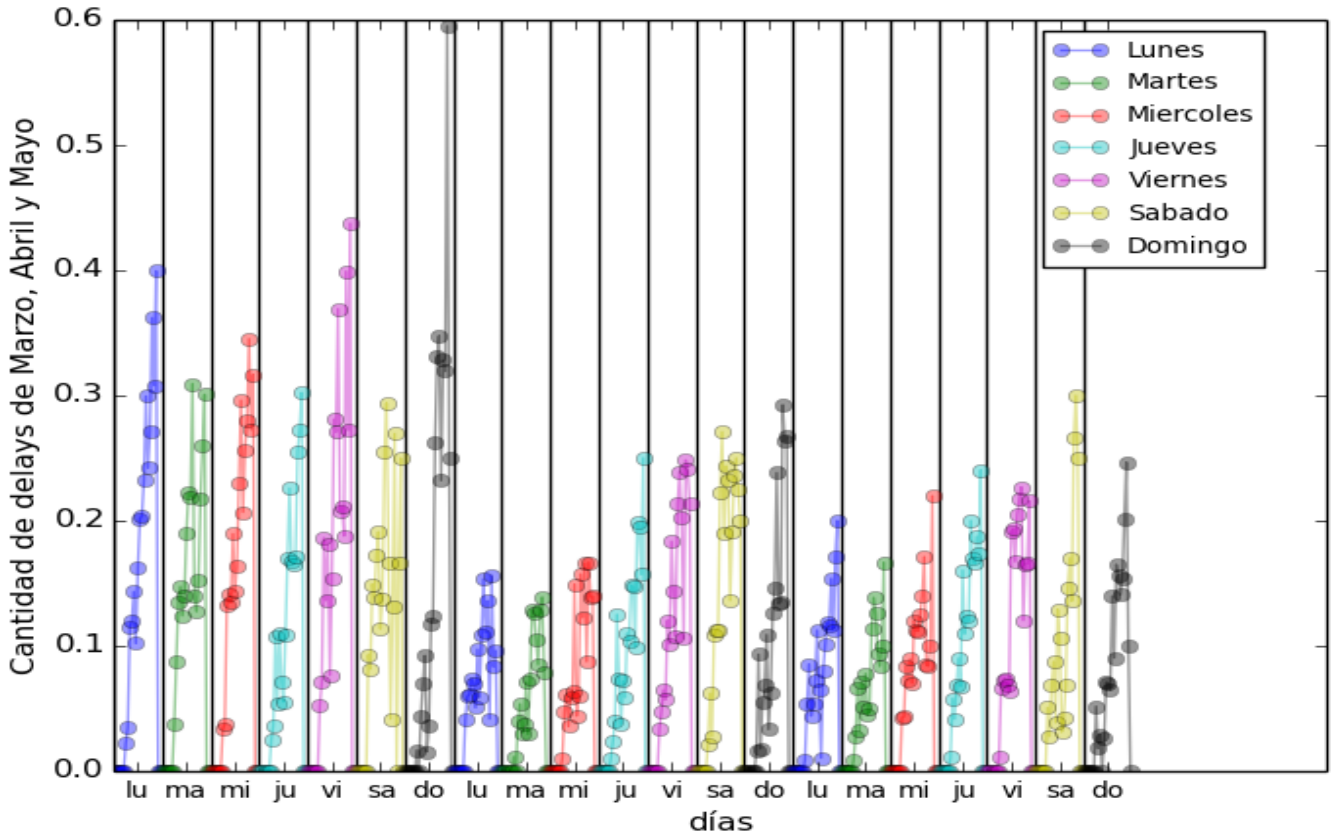


Figura: Delays por hora DCA

Delay por año

En base al anterior eje, notamos que se podría haber estudiado también lo siguiente:

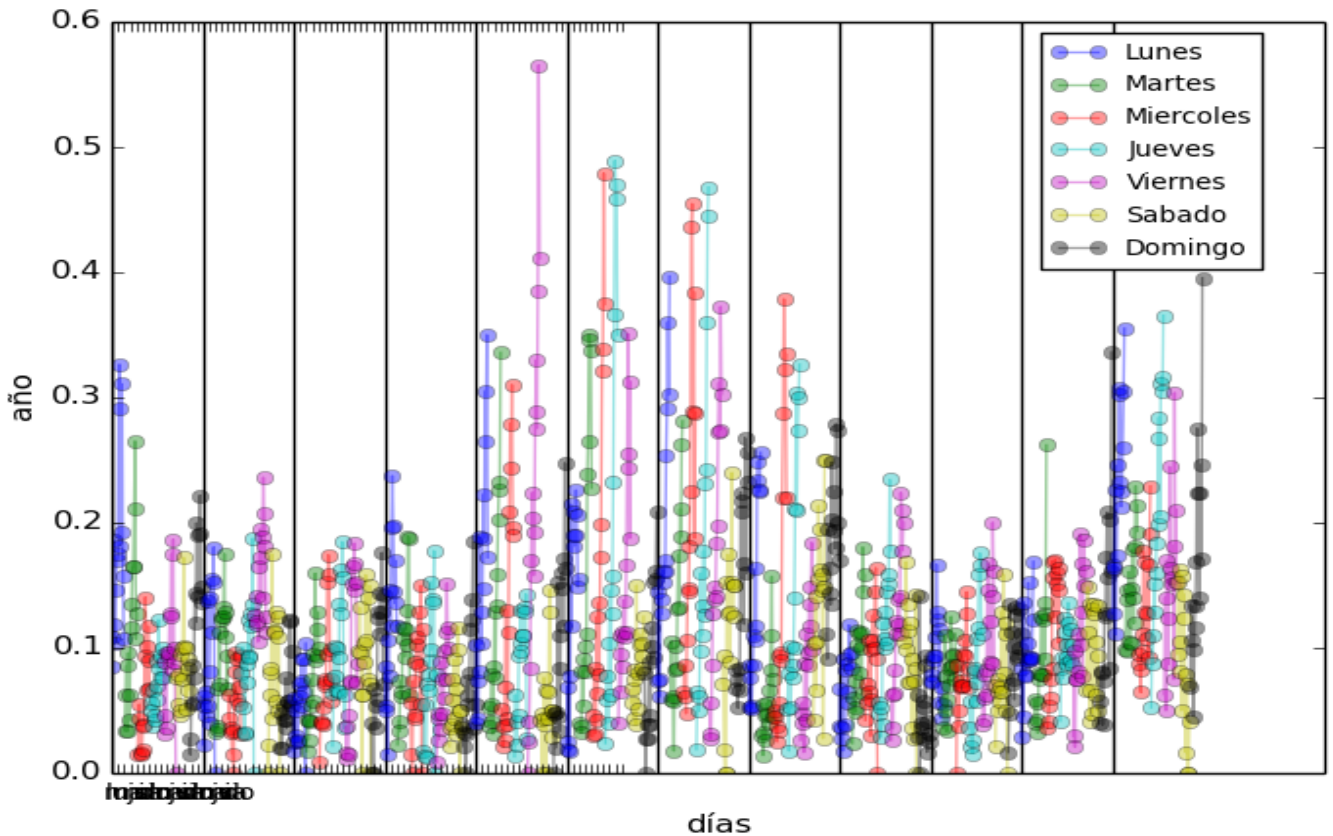


Figura: Delays por hora DCA

Delays por seguridad

Intentamos analizar como varían la cantidad de delays por seguridad, por mes:

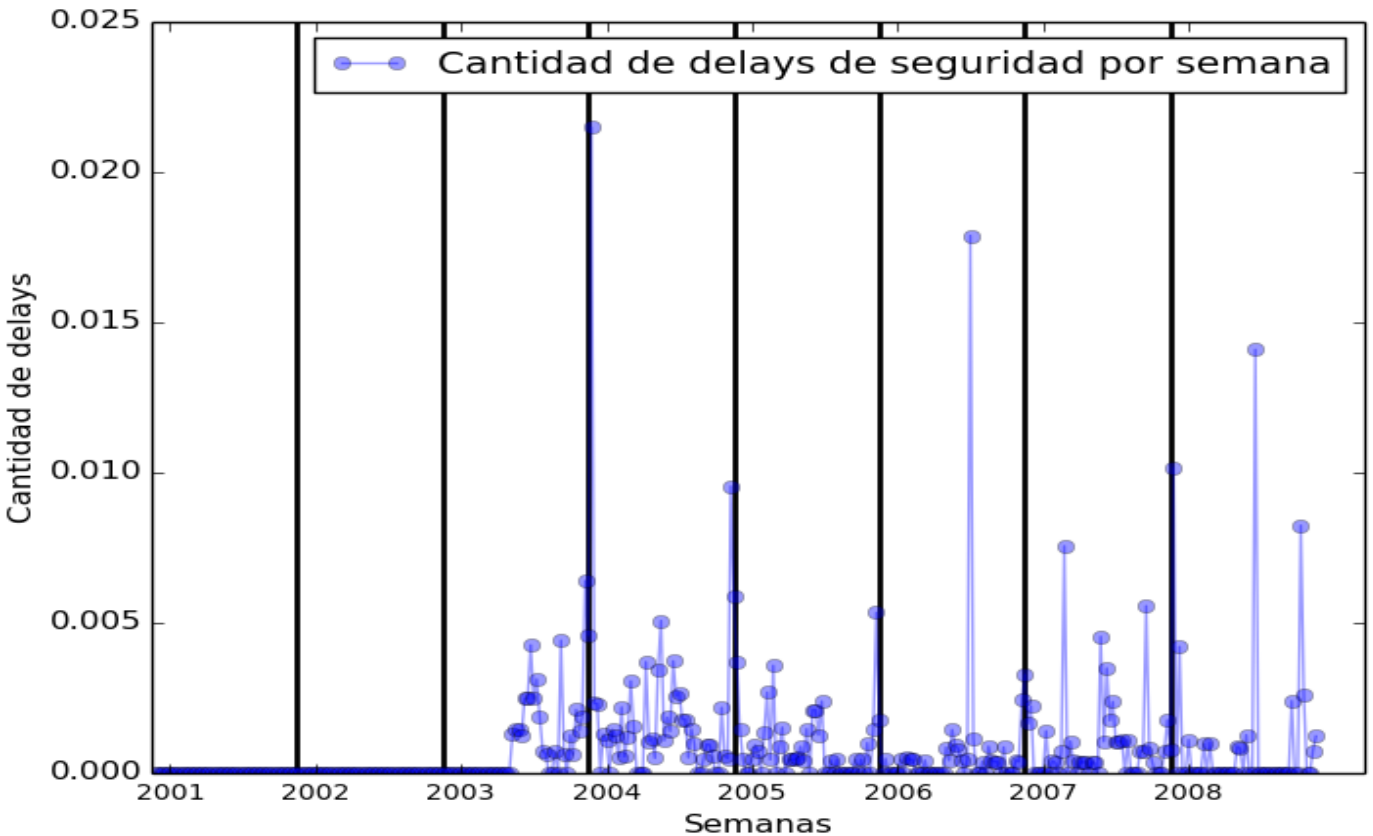


Figura: Delays por hora DCA

Delays por aeropuerto

Intentamos analizar si había alguna relación entre los delays, el lugar en el que se situa un aeropuerto, y el día de semana. Por ejemplo, si había diferencia entre los lugares considerados de *ocio* respecto de los lugares considerado de *negocios*:

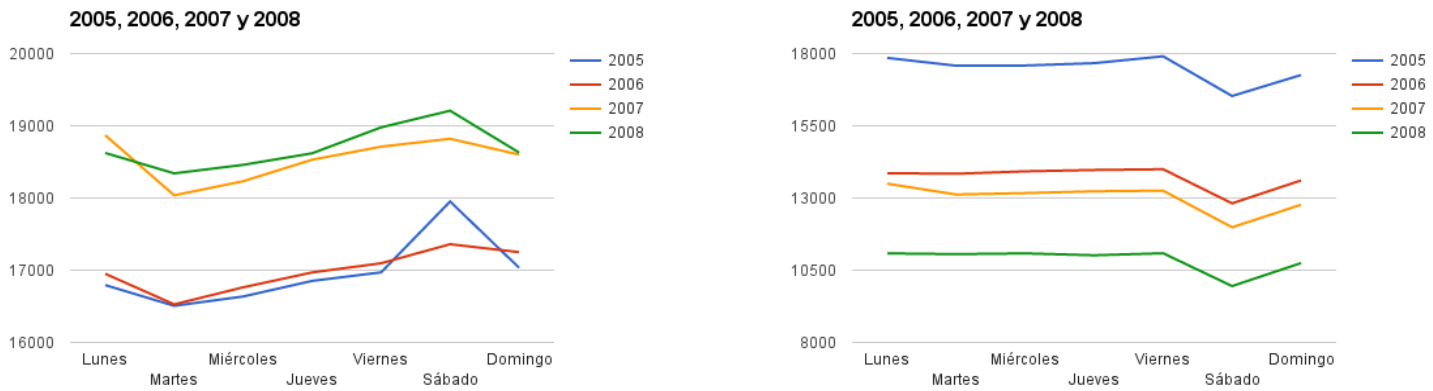


Figura: Cantidad de delays por día (de llegada) a la izquierda Orlando y a la derecha Washington.