Trabajo Práctico III

Jonathan Seijo, Lucas De Bortoli, Roberto Grings y Agustín Penas

Departamento de computación
Universidad de Buenos Aires
Buenos Aires, Argentina

Resumen

En este trabajo utilizaremos el método de cuadrados mínimos lineales para realizar predicciones sobre los vuelos en Estados Unidos.

En primer lugar analizaremos la variación de las cancelaciones por clima a través del tiempo. En segundo lugar, analizaremos algunas aerolíneas por separado e intentaremos averiguar cómo se comporta nuestra familia de funciones de cuadrados mínimos en cada caso.

Keywords: Aeropuertos, Aerolíneas, Cancelaciones, Clima, Cuadrados Minimos Lineales, Predicción, Retrasos, Vuelos

1. Ejes de estudio

Los ejes de estudio en los que nos centraremos son:

- ¿Cómo varían las cancelaciones por clima a través del tiempo? ¿Cómo influye el aeropuerto de origen?
- ¿Cómo se comporta nuestro modelo de cuadrados mínimos con la cantidad de retrasos en diferentes aerolineas? ¿Podemos predecir alguna mejor que otra?

En el primer eje trataremos de encontrar un patrón a las cancelaciones por clima a través del tiempo. ¿Se sigue un patrón regular?, ¿Hay fechas en las cuáles siempre hay cancelaciones? Veremos qué sucede en algunos aeropuertos particulares.

En el segundo eje analizaremos la cantidad de retrasos por aerolíneas. Nos centraremos en algunas más representativas y veremos las regularidades (o irregularidades) que poseen. ¿Hay alguna más dificil de predecir que otras? ¿Cómo se comporta una misma familia de funciones de cuadrados mínimos con distintas aerolineas? ¿Será necesario adaptarlo cada vez?

2. Cancelaciones por clima

2.1. Preliminares

En esta sección veremos cómo varían las cancelaciones por clima a través del tiempo, y veremos si podemos encontrar algún patrón. Nos será muy util contar con los motivos de las cancelaciones para poder diferenciar las que nos interesa, pero sin embargo, los datos previos a 2003 no cuentan con esta información. Es por eso que tomaremos los datos desde 2003 en adelante.

Con respecto a los gráficos que mostraremos, en un principio agrupamos los datos por mes pero con ello perdíamos información: hay valores que varían semana a semana dentro de un mismo mes. Por ello decidimos agrupar nuestros datos por semanas.

2.2. Experimentación

En primer lugar tomaremos las cancelaciones por climas de manera general. Esperamos que haya una regularidad periódica, dónde para un mismo mes en diferentes años se registren cancelaciones comparables, pues tenemos en mente que hay épocas marcadas con tormentas fuertes o huracanes.

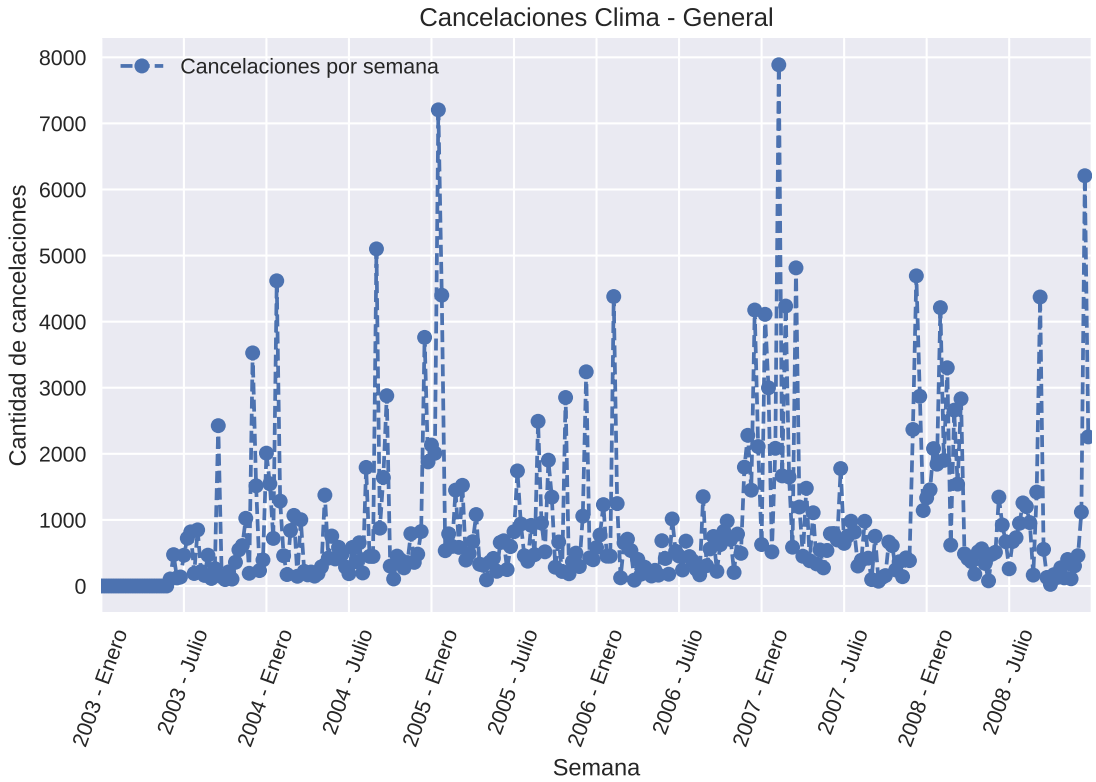

Figura 1: Cancelaciones por clima con respecto al tiempo.

Analicemos un poco los datos. El comienzo de 2003 se ve completamente en cero, esto es por la falta de datos sobre cancelaciones climáticas en ese período. En las estimaciones de cuadrados mínimos no tendremos en cuenta este período sin datos, por lo que no nos será de ningún problema.

En enero y diciembre (invierno de USA) se observan mayor cantidad de cancelaciones. Consideramos que esto puede deberse a dos motivos: Tormentas de nieve y mal clima en general, y el aumento del caudal de gente que viaja en épocas festivas. Con respecto a esto último, creemos que un aumento en la cantidad total de vuelos implica una mayor cantidad de cancelaciones.

En general, en agosto y septiembre se registran gran cantidad de cancelaciones, pero no es así todos los años. Por ejemplo, en 2007 no hay tanta cantidad como en 2004. Creemos que tiene que ver con las temporadas de huracanes, que no todos los años son catastróficas. Con respecto a nuestros datos, en agosto de 2004 el pico puede deberse al huracán Charley [1], mientras que agosto de 2005 se corresponde con el huracán Katrina [2].

Como curiosidad, en agosto de 2003 se encuentra un pico aislado de los demás. Coincide con un apagón en grandes ciudades (Por ej, Nueva York) que duró 24hs en el cuál se reportaron cancelaciones de vuelos. Si bien uno creería que no tiene por qué estar relacionado al clima, el apagón se produjo por una caída del servicio central por las grandes demandas debido a las temperaturas inusuales de hasta 40 grados. [3]

Para la predicción con cuadrados mínimos, la mejor familia de funciones que encontramos fue la siguiente:

$$f(t) = \alpha_1 + \alpha_2 * cos(\frac{\pi}{48}t)^8 + \alpha_3 * sen(\frac{\pi}{24}t) + \alpha_4 * sen(\frac{\pi}{12}t) \qquad (1)$$

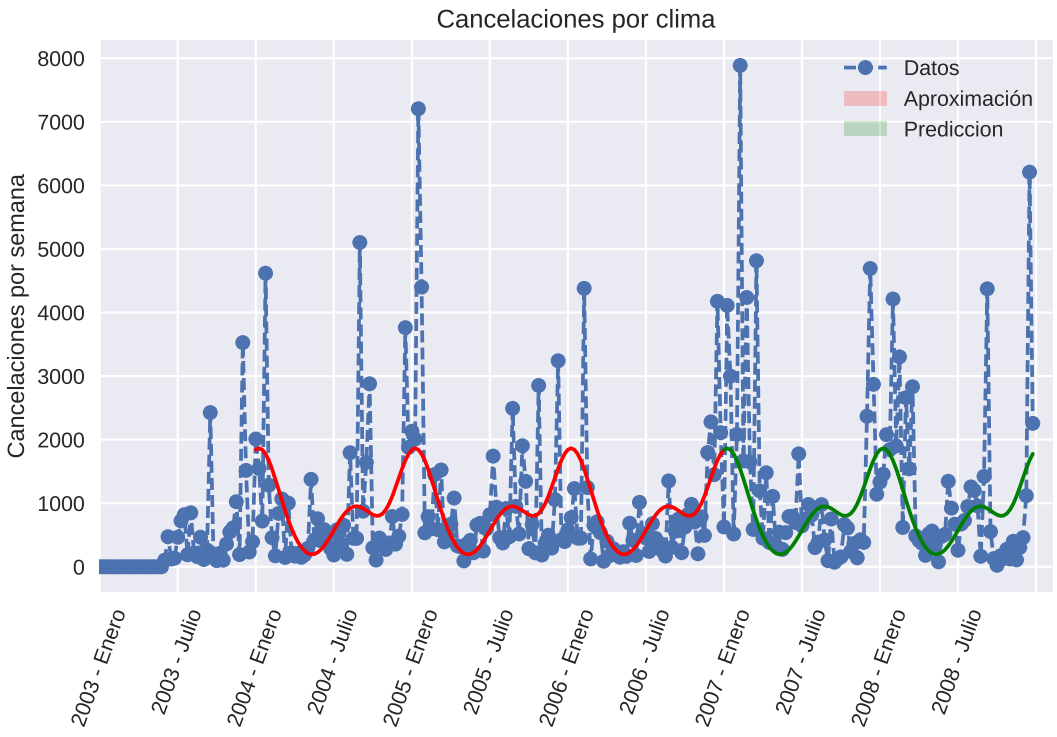

Figura 2: Cancelaciones por clima, 3 años de entrenamiento y 2 de predicción.

Los períodos elegidos se corresponden con la cantidad de semanas. Por ejemplo, como sabemos que hay picos pronunciados en enero, tomamos $cos(\frac{\pi}{48}t)$ para que alcance un máximo cada 48 semanas. (En nuestros datos, cada mes está dividido en 4 semanas). La potencia a la cuál esta elevada fue elegida

para que estos picos sean pronunciados en esas fechas. Análogamente con los períodos de los senos, queremos que se tengan en cuenta los ciclos de seis meses y tres meses respectivamente.

Con esta familia de funciones obtuvimos (promediando con Cross Validation) un ECM de 1264403. Las cancelaciones por clima en general no fueron sencillas de predecir de manera exacta, y no obtuvimos resultados tan precisos. Diferentes aeropuertos podrían estar sumando cancelaciones en períodos diferentes, es por esto que en el siguiente experimento mostramos dos aeropuertos particulares: los aeropuertos de Miami y Los Ángeles. La razón de la elección es porque ambos se encuentran en costas opuestas del país, y porque Miami suele sufrir cancelaciones por mal clima sobre todo en época de huracanes (Agosto).

Una aclaración importante: en los datos de Miami quitamos dos valores outliers que representaban mas de 450 cancelaciones esa semana pues distorsionaban el gráfico completamente. Creemos que correspondían a los huracanes Charley y Katrina. De todos modos, sus respectivos picos en esas fechas se siguieron manteniendo.

Para ambos aeropuertos consideraremos la misma familia de funciones que en el experimento anterior, y realizamos Cross Validation para medir los errores. Para Miami obtuvimos un ECM de 227. Con Los Ángeles, el ECM fue de 180.

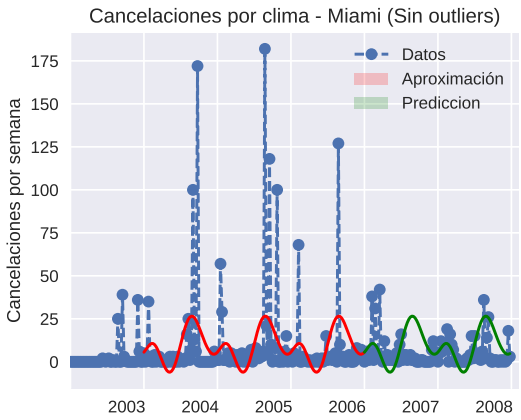

Figura 3. Clima - Miami

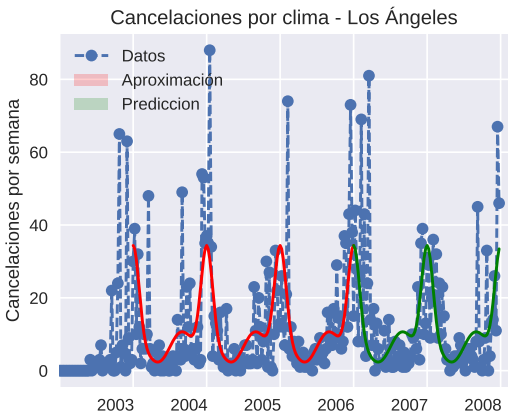

Figura 4. Clima - Los Ángeles

Diferentes aeropuertos tienen diferentes cantidades de cancelaciones, sin embargo en los aeropuertos en los que experimentamos encontramos resultados similares, incluso en los que no mostramos. En los meses de diciembre y enero podemos encontrar siempre los mayores picos de cancelaciones por clima. Además, en los meses de julio y agosto también suelen registrarse picos

(aunque de menor altura).

Sin embargo hay que destacar la diferencia entre Miami y Los Ángeles. Podemos apreciar cómo la curva en el grafico de Los Ángeles acompaña a los datos de mejor manera que en el grafico de Miami. Creemos que esto se da por que Miami tiene un clima mas impredecible y por ende tiene picos mas altos en sus cancelaciones, lo que hace al método de cuadrados mínimos mas inexacto.

En conclusión, gracias a la periodicidad anual y semestral de las cancelaciones por clima, la familia de funciones que fijamos al comienzo de la sección sirve como un buen predictor para distintos aeropuertos, aunque no de forma exacta.

3. Aerolíneas

3.1. Preliminares

Para nuestro segundo eje analizaremos la variación de la cantidad de retrasos con respecto a diferentes aerolíneas. Intentaremos ver si los mismos patrones se repiten, y si será necesario utilizar una familia de funciones diferente para cada aerolínea en particular.

Las aerolíneas tienen vuelos en diferentes lugares del país, que a su vez tienen distinta cantidad de vuelos, con condiciones climáticas diferentes. Tomando únicamente las aerolíneas como la variables fijas, hay gran cantidad de ruido en los datos que hace que sea muy difícil la realización de predicciones. Es por este motivo que en los experimentos a continuación, el aeropuerto de origen se encuentra fijo.

Si bien realizamos experimentos en varios aeropuertos diferentes, decidimos mostrar únicamente dos aeropuertos distintos, pues en general los resultados son similares. Elegimos en primer lugar el aeropuerto de Los Ángeles, pues hasta 2009 manejaba la mayor cantidad de pasajeros de origen y destino en todo el mundo [4]. El segundo aeropuerto que tomamos fue el de Atlanta, que es también uno de los más grandes del país.

Debido al tamaño de los aeropuertos, creemos que las aerolineas cuentan con datos los suficientemente representativos como para realizar predicciones. Esperamos que una misma aerolínea tenga los mismos patrones de retrasos, anuales o semestrales, independientemente del aeropuerto, aunque es claro que las cantidades totales serán diferentes.

A diferencia de la sección anterior, sobre retrasos en general existen datos previos al 2003. Sin embargo, los datos mas antiguos parecen seguir un patrón diferente a los de los ultimos años. Consideramos que para realizar una predicción para el año siguiente, tomar los últimos 5 años es representativo, por lo que los datos que mostramos seguirán siendo a partir de 2003.

Sobre los retrasos, consideraremos que un vuelo tiene retraso cuando su tiempo de demora es superior a los 15 minutos. Esto se corresponde con la métrica de *On Time Performance* (OTP) propuesta en el enunciado del trabajo.

3.2. Experimentos

La primer aerolínea que consideramos será United Airlines. En todos los aeropuertos encontramos que hay mucha variación en los retrasos semana a semana, lo que hace que sea dificil encontrar una familia de funciones que aproxime bien a todos los datos. Sin embargo, aunque las predicciones no hayan sido del todo buenas, queremos destacar que fue la misma familia de funciones la que logró minimizar el ECM en los distintos aeropuertos. Esta familia fue distinta (aunque similar) a la del primer eje.

$$f(t) = \alpha_1 + \alpha_2 * t + \alpha_3 * cos(\frac{\pi}{48}t) + \alpha_4 * cos(\frac{\pi}{24}t) + \alpha_5 * cos(\frac{\pi}{12}t) \qquad (2)$$

Agregamos un término lineal pues a lo largo del tiempo las cancelaciones no se encuentran sobre el mismo nivel. Los períodos de los cosenos se corresponden a un año, seis meses y tres meses respectivamente, medidos en cantidad de semanas. Creemos que con estos períodos se logra representar la información sobre cancelaciones. Períodos mas pequeños y mas grandes producen peores resultados en cuanto al error cuadrático medio.

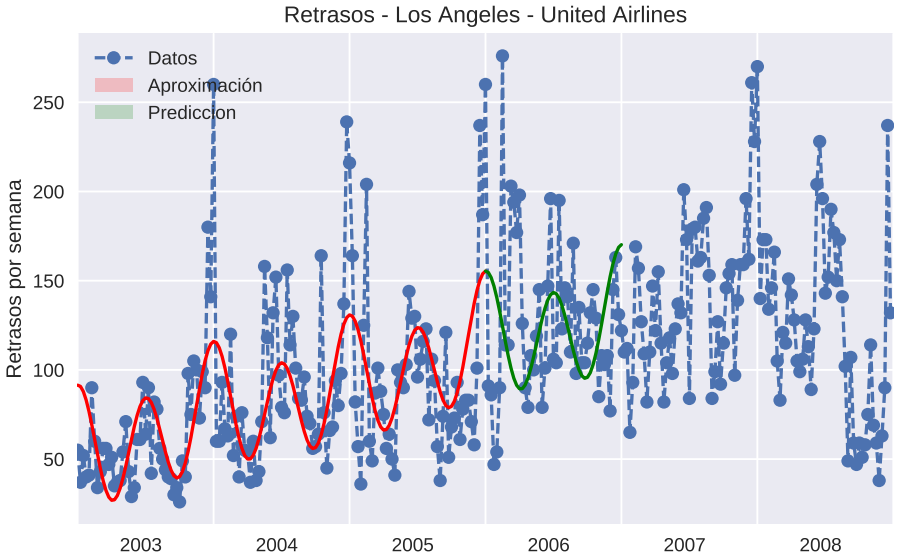

Figura 5: Retrasos - United Airlines - Los Ángeles

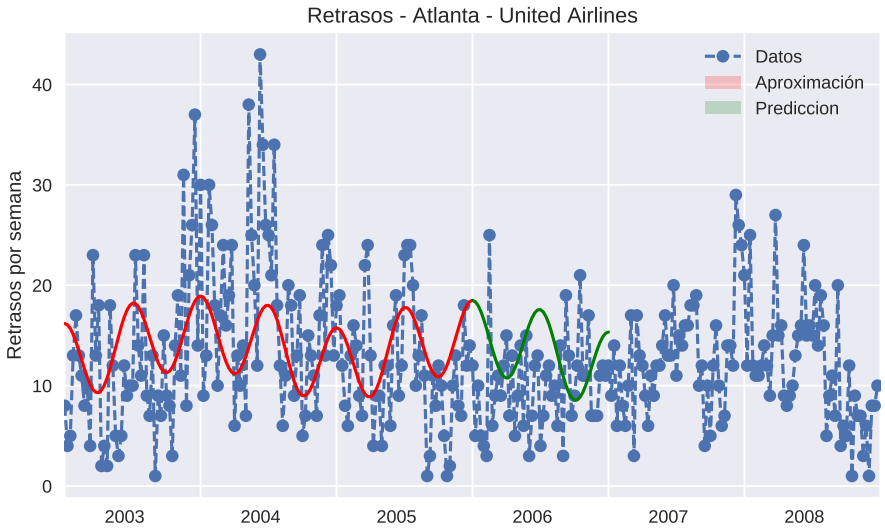

Figura 6: Retrasos - United Airlines - Atlanta

Tanto en Los Ángeles como en Atlanta nos encontramos con que los datos no siguen un patron bien marcado como en el primer eje de estudio. Los ECM, aplicando Cross Validation, fueron de 2619 para Los Ángeles y 56 para Atlanta. Como habíamos mencionado, la misma familia fue la que minimizó los ECM de cada aeropuerto. Esto nos parece importante, porque quiere decir que existen patrones en los retrasos de una aerolínea que no tienen que ver con el lugar dónde se encuentran.

Si bien es ambos casos los picos coinciden con enero y julio, es notorio que en el caso de Los Ángeles los picos son más altos en enero, y eso está ajustado con nuestra función. Tiene sentido entonces que tomar períodos anuales y semestrales con los cosenos dé buenos resultados.

Veamos que sucede si tomamos otra aerolínea, como es el caso de American Airlines. Sospechábamos que la mejor familia iba a ser una diferente, pues distintas aerolíneas podrían tener distintos períodos de cancelaciones. Nos encontramos con algo muy interesante: la misma familia de funciones que usamos con United Airlines minimiza también los ECM de American Airlines.

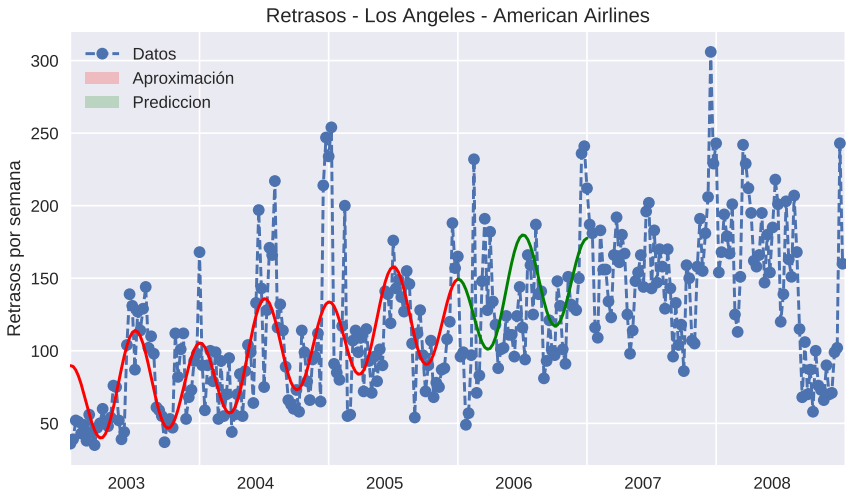

Figura 7: Retrasos - American Airlines - Los Ángeles

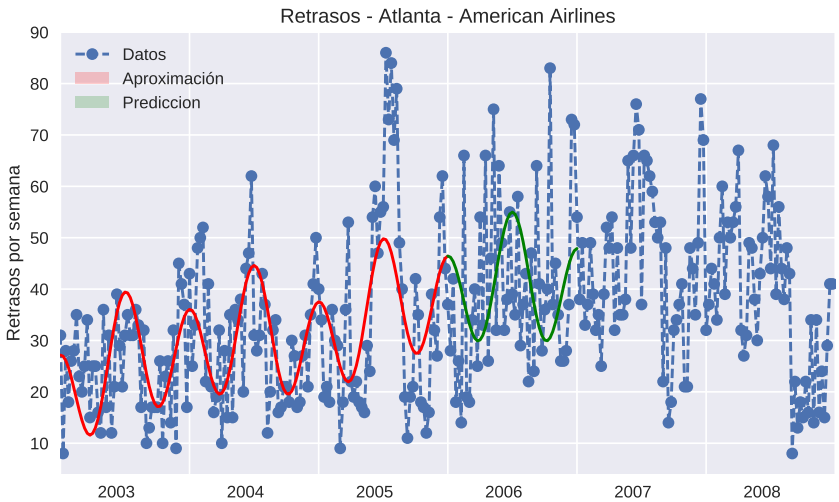

Figura 8: Retrasos - American Airlines - Atlanta

El ECM obtenido para Los Ángeles fue de 2607, mientras que para Atlanta obtuvimos 281.

En general observamos que no hay una periodicidad bien marcada en los retrasos por aerolíneas. Realizamos los mismos experimentos pero considerando cancelaciones, y también cancelaciones únicamente por climas, pero los resultados fueron los mismos: los datos varían muchísimo semana a semana, y los patrones no son claros.

No pudimos realizar aproximaciones *precisas*, sin embargo, podemos identificar patrones anuales y semestrales que se repiten en todas las aerolíneas, en diversos aeropuertos: encontramos una misma familia de funciones que mejor aproximan a distintas aerolíneas. Consideramos que las aproximaciones conseguidas pueden ser útiles si se quiere observar un patrón general.

Referencias

[1] https://es.wikipedia.org/wiki/Hurac%C3%A1n_Charley_(2004)

[2] https://es.wikipedia.org/wiki/Hurac%C3%A1n_Katrina

[3] www.elmundo.es/elmundo/2003/08/14/internacional/1060893592.html

[4] https://web.archive.org/web/20090120121530/http://www.lawa.org/welcome_lax.aspx?id=40