

# Análisis de la serie temporal de salidas del Aeropuerto de Menorca

Jordi Vanrell Forteza

24/6/2021

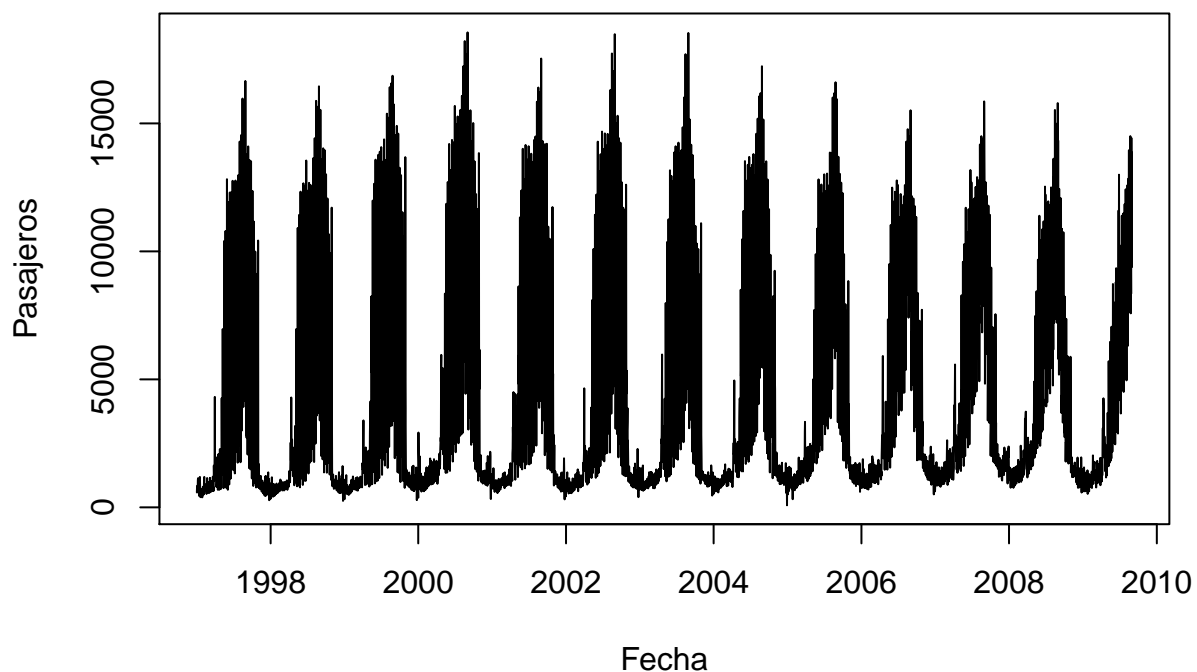
## 1. Datos, lectura y manipulación previa

Se descarga el archivo asignado a este trabajo (en formato `.xlsx`) y se procede, en primer lugar, a leerlo con el paquete `readxl`.

El archivo consiste en una serie temporal en la que se representan las salidas de pasajeros diarias del aeropuerto de Menorca entre el 1 de enero de 1997 y el 31 de agosto de 2009. El formato fecha en que se presentan no es homogéneo. Aunque esto no suponga un impedimento en la acometida de un correcto análisis de series temporales, se procede a normalizar al formato YYYY-MM-DD por facilidad de visualización y consulta durante los pasos del análisis.

Una vez que se dispone de las fechas en formato homogéneo se procede a visualizar la serie por primera vez.

### Salidas del aeropuerto de Menorca



Se aprecia fácilmente una fuerte estacionalidad con picos durante los veranos y valles en invierno, típica de destinos turísticos como Menorca. A parte de la estacionalidad anual, otros patrones habituales de este tipo

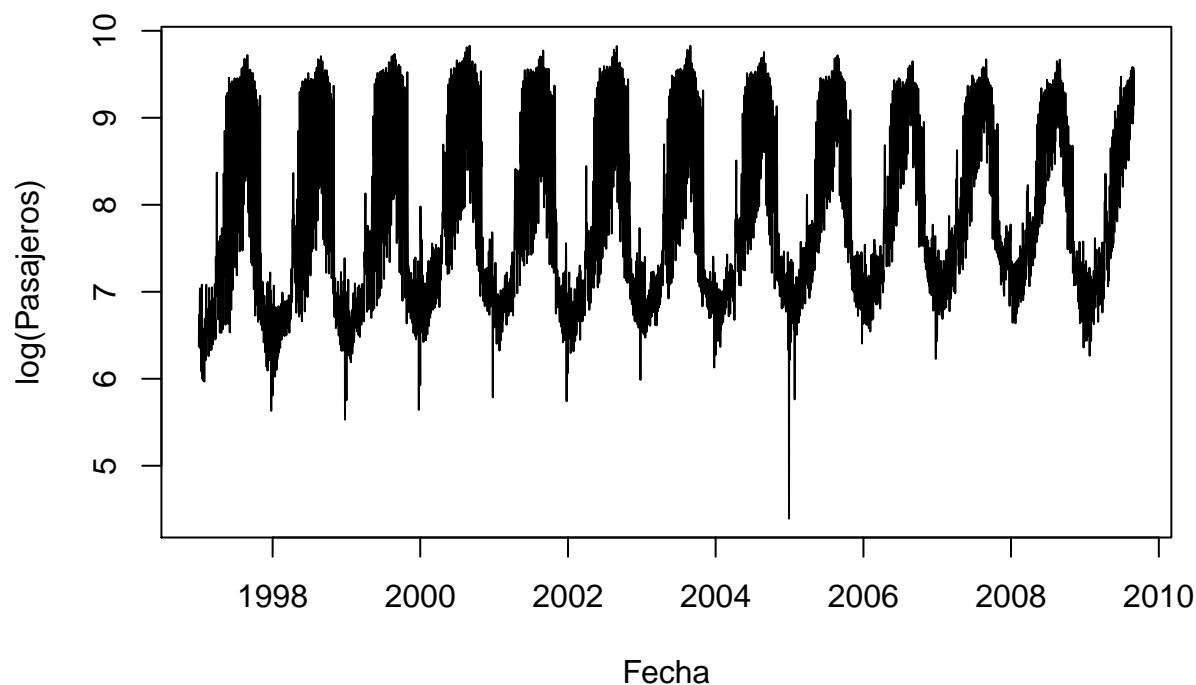
de destinos y susceptibles de afectar a la serie son, por un lado, la estacionalidad semanal y, por otro, el efecto de las fiestas de Pascua. Esta última es, quizás, la más compleja de analizar porque sigue un patrón basado en el ciclo lunar que no casa con el ciclo solar con que normalmente se representan las fechas. Las fiestas de Pascua caen cada año en una fecha distinta, en concreto alrededor del domingo siguiente a la primera luna llena de primavera. Aunque podrían tratarse los datos de Pascua como *outliers*, en realidad no lo son desde el momento en que pueden anticiparse de alguna forma.

Examinando los datos se aprecia que, en el aeropuerto de Menorca, el día de las fiestas de Pascua que más sobresale de los valores previos y posteriores es el Lunes de Pascua. Se sabe que este día marca el último día del periodo vacacional de algunas comunidades autónomas cercanas que suelen mandar viajeros, como Cataluña y, a la vez, se especula que, pasadas las fiestas religiosas, también es el lunes de una semana no lectiva que algunos menorquines aprovecharían para viajar.

Tras una breve búsqueda en calendarios pasados se recopilan las fechas correspondientes a los Lunes de Pascua de la serie, se almacenan en una variable y se incorporan como una variable *dummy* a la tabla original.

La representación de la serie en logaritmos resalta ligeramente la presencia de un pico menor justo antes del crecimiento veraniego de la serie que se correspondería con esta fecha.

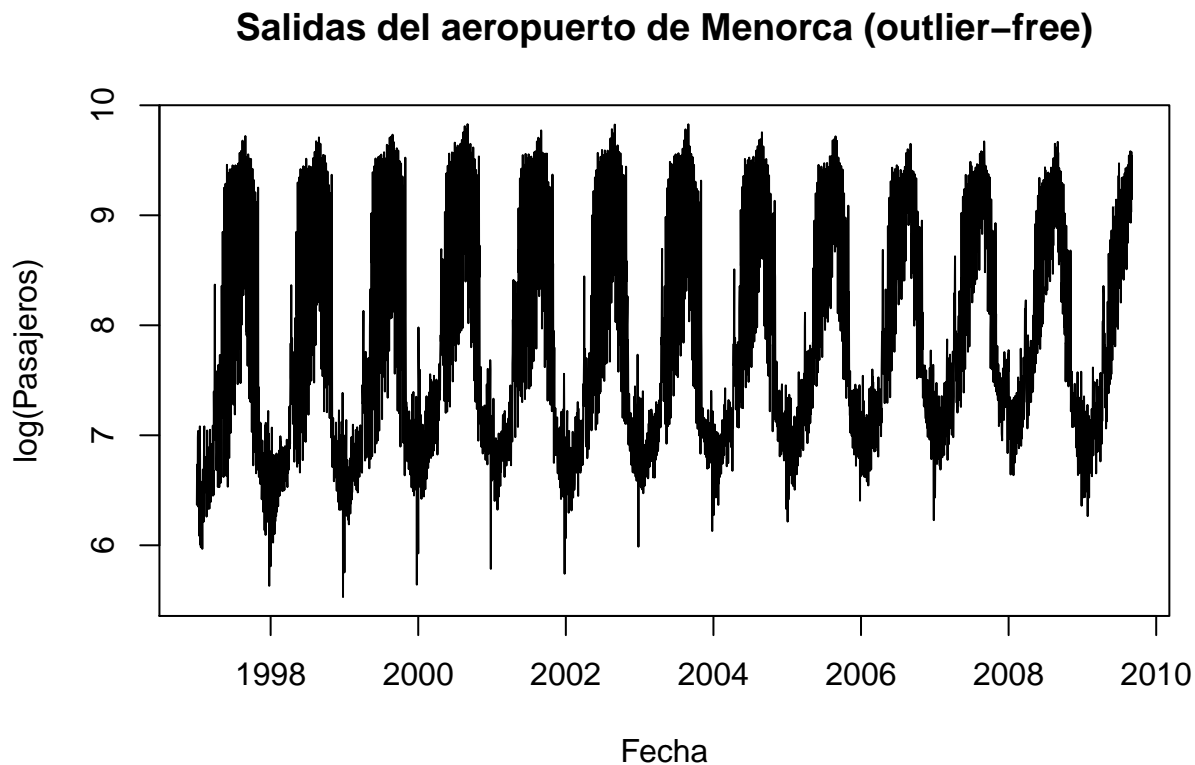
### Salidas del aeropuerto de Menorca



Al mismo tiempo la transformación pone al descubierto algunos valores extremos. Mirando la serie original se identifican en el 29 de diciembre de 2004, por un temporal de viento extremo, y en el 26 de enero de 2005, por causa de una gran nevada. A diferencia de los Lunes de Pascua, estos dos datos concretos sí son estrictamente *outliers*, no obedecen a ningún patrón discernible.

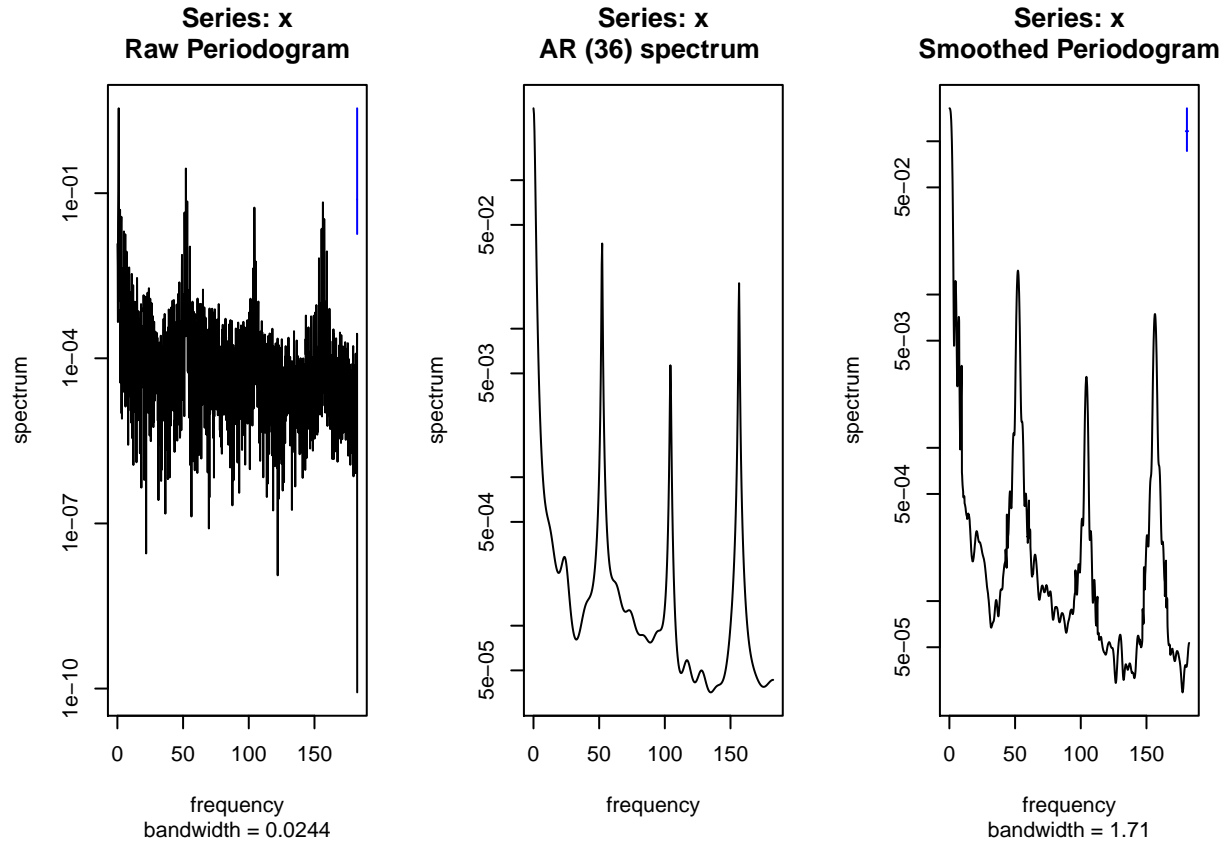
## 2. Transformación en serie temporal, tratamiento de outliers y estacionalidad.

Antes de empezar a aplicar modelos es necesario disponer de los datos en forma de serie temporal. Sin embargo, antes incluso de esto, cabe plantear la aproximación con que se tratarán los *outliers*. De los métodos disponibles, el que brinda `tsoutliers`, siguiendo el procedimiento de Chen & Liu (1993), es quizás el más completo: detecta y reemplaza los *outliers* de toda la serie. Pero tiene un gran coste computacional y, a pesar de que puede generar muy buenas estimaciones de las salidas que se hubieran producido de no haber habido temporales, probablemente también detectaría los Lunes de Pascua como *outliers* y se perdería esa información. Por otra parte, la función `na.interp` del paquete `forecast` permite interpolar los valores faltantes de una serie temporal de manera sencilla, sin apenas coste computacional, y sin que los demás valores de la serie se vean afectados. Por eso se ha considerado más conveniente considerar primero los dos *outliers* como datos faltantes y luego interpolar sus valores con las prestaciones del paquete `forecast`. Probablemente no sea el método que permita obtener los mejores resultados, pero sí resultados mejores que sin tratar los dos *outliers*.



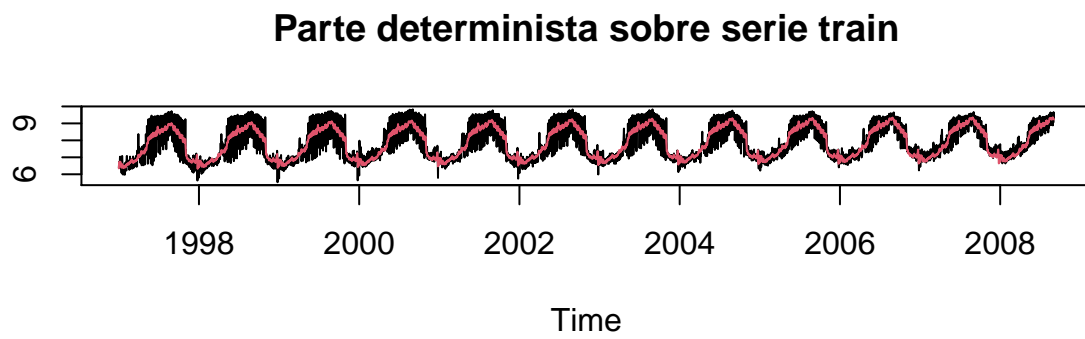
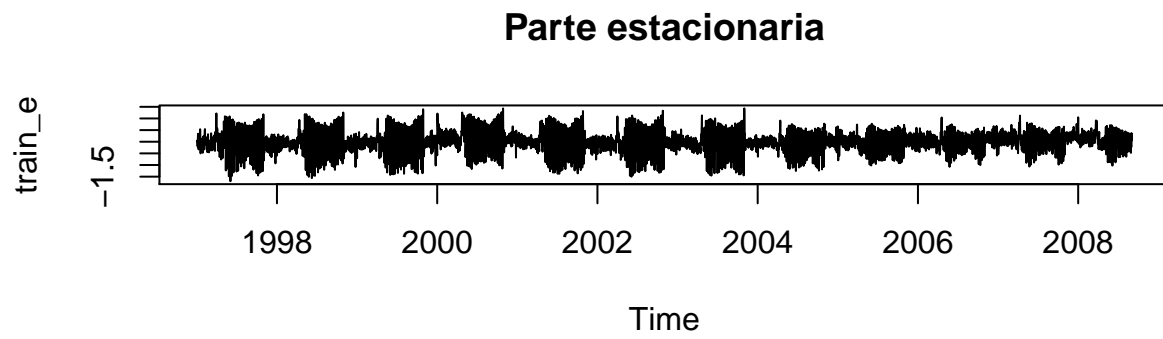
A continuación se extraen de la serie dos sub-series, una de entrenamiento (**train**) y una de validación (**test**). Concretamente, **train** contiene los datos de la serie hasta la observación 4261, correspondiente al 31 de agosto de 2008 (día 244 de 2008); **test** se solapa parcialmente con **train** y contiene los datos entre la observación 3896, correspondiente al 1 de septiembre de 2007 (día 244 de 2007), hasta el final: un total de 2 años. La muestra **test** tiene una longitud de 731 y los datos de solapamiento entre ambas muestras son 366 porque 2008 fue año bisiesto.

Se representan los espectros de la serie **train**. La forma del espectro confirma la existencia de estacionalidad semanal, a parte del ciclo anual.

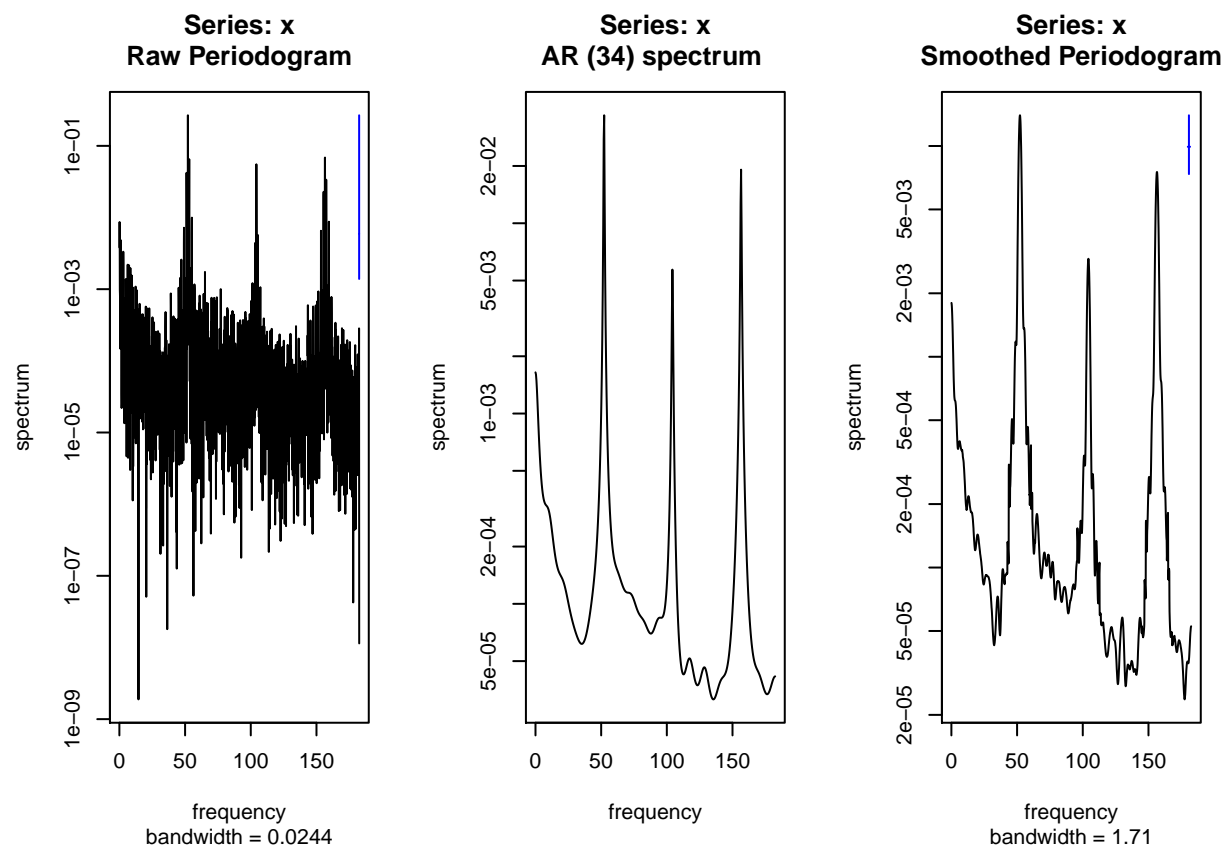


Se desea tener la serie **train** desestacionalizada para poder modelizar los errores. Con este fin se genera una variable que recogerá la parte determinista de la serie, que sigue la forma oscilatoria de una función trigonométrica. Luego se estima por mínimos cuadrados ordinarios una regresión de la parte determinista y el tiempo en sus dimensiones lineal y cuadrática sobre la serie. Los errores de esta regresión conforman la serie estacionaria a modelizar, la parte que explica la regresión corresponde a la parte estacional.

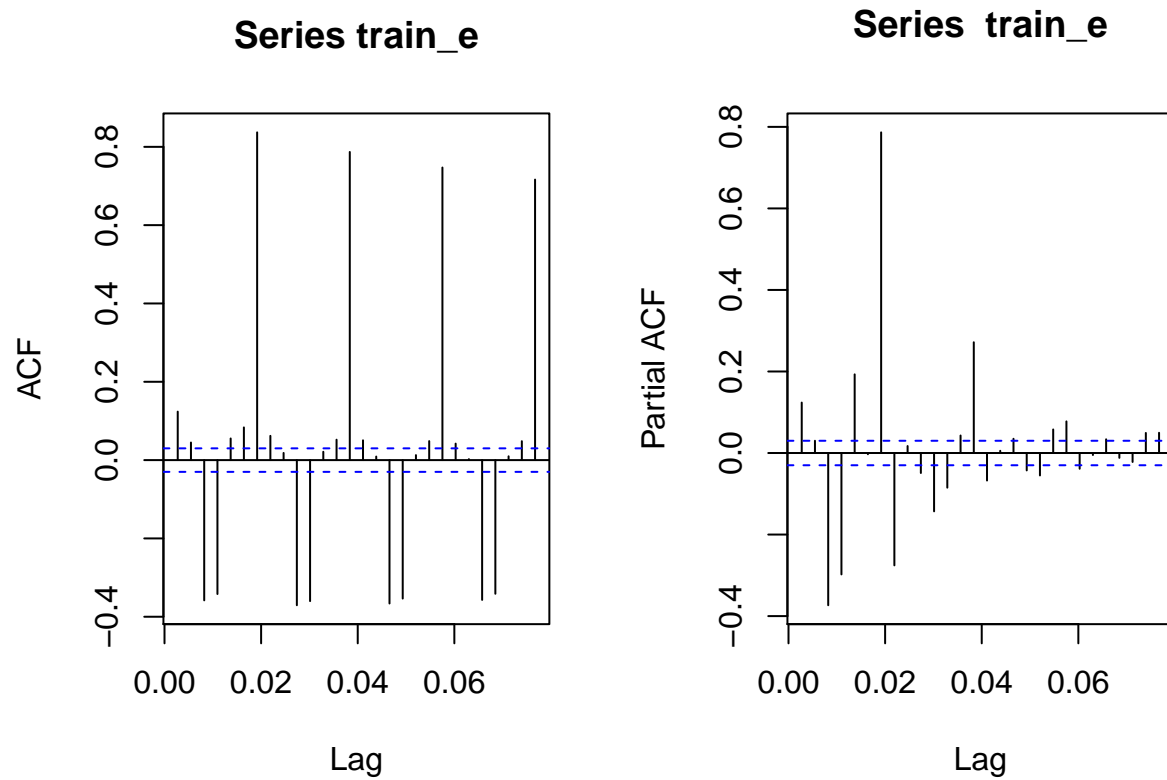
Representadas gráficamente:



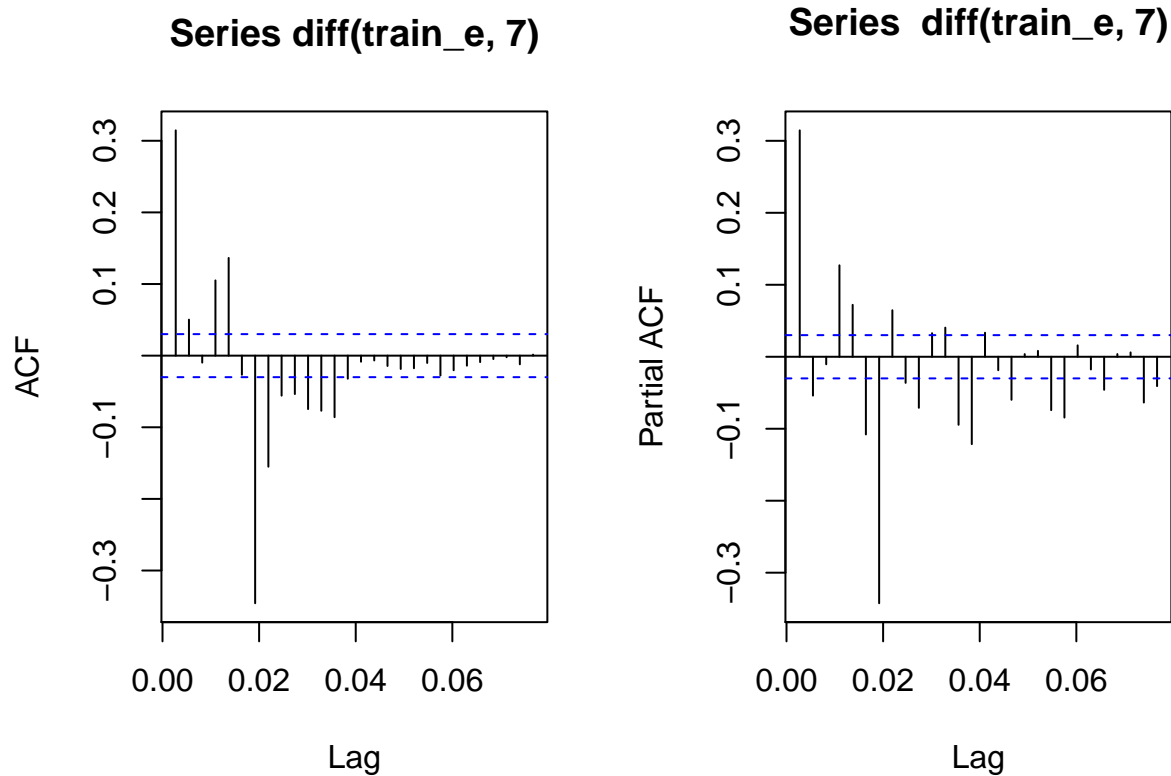
Ahora, analizando los espectros de la parte estacionaria se observa que se conserva la estacionalidad semanal de la serie original.



Esta percepción concuerda con lo que se observa en sus correlogramas:



Mirando la ACF puede decirse que existen correlaciones significativas en todos los retardos hasta el octavo y, además, la permanencia de coeficientes muy significativos con retardos muy elevados denotan muy bien la existencia de esa estacionalidad semanal. Para acabar de confirmarlo, véase la representación de los correlogramas cuando se toman primeras diferencias en el séptimo retardo.



### 3. Modelos ARIMA con parte SARIMA

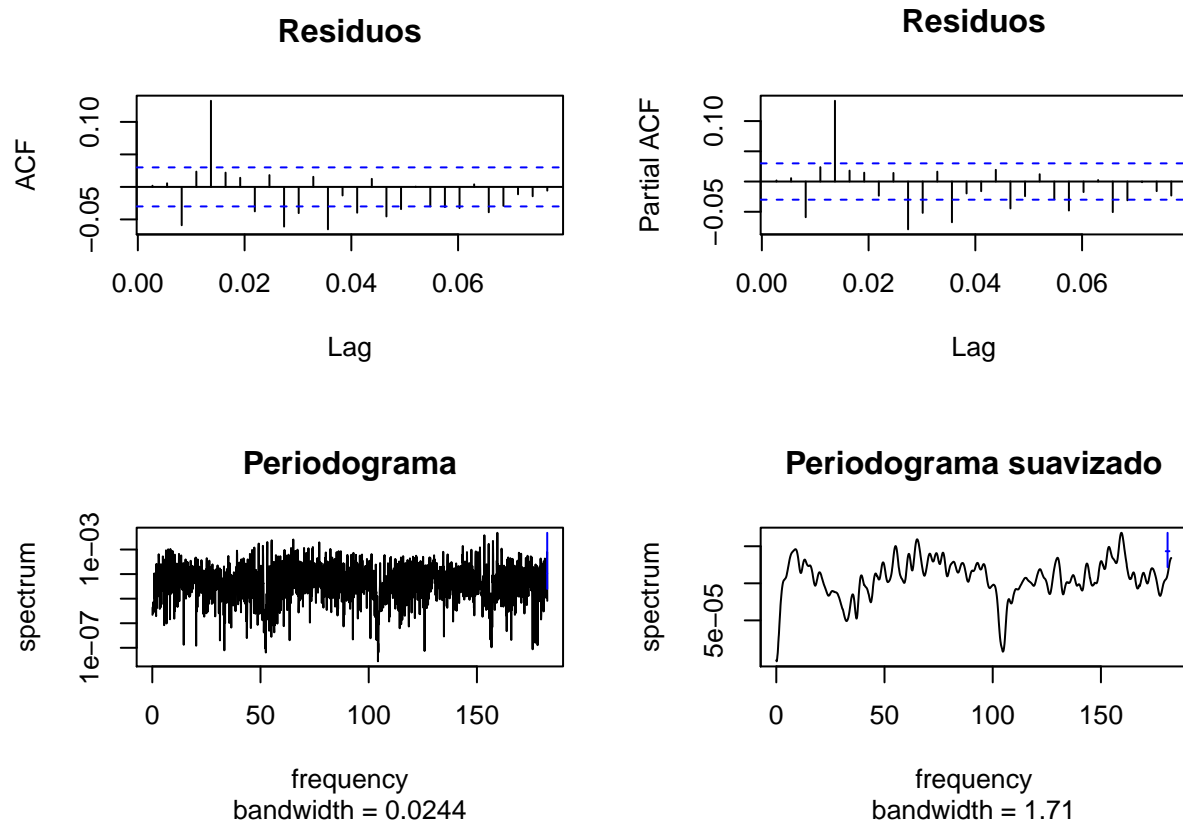
Los modelos ARIMA describen el comportamiento de una serie como cierta combinación lineal de sus retardos, un operador de diferencias y procesos de ruido blanco. El componente estacional puede incluir otra combinación análoga pero siguiendo una frecuencia predeterminada. El modelo más adecuado en cada caso puede inferirse examinando los correlogramas (los valores de las autocorrelaciones y las autocorrelaciones parciales) de la serie.

Dados los últimos disponibles en el punto 2, parece que la serie podría tener un componente autorregresivo porque los primeros coeficientes parecen descender rápidamente. Además, la significatividad del séptimo retardo en la ACF y la PACF parecen sugerir que podría existir una media móvil estacional, además de las diferencias estacionales introducidas.

#### 3.1. Modelo 1: ARIMA(1,0,0) con SARIMA(0,1,1) + *dummies* para los Lunes de Pascua

Con esto se construye el primer modelo como un ARIMA(1,0,0) con SARIMA(0,1,1), añadiéndole la variable *dummy* correspondiente a los Lunes de Pascua. El modelo arroja unos coeficientes que son todos significativos.

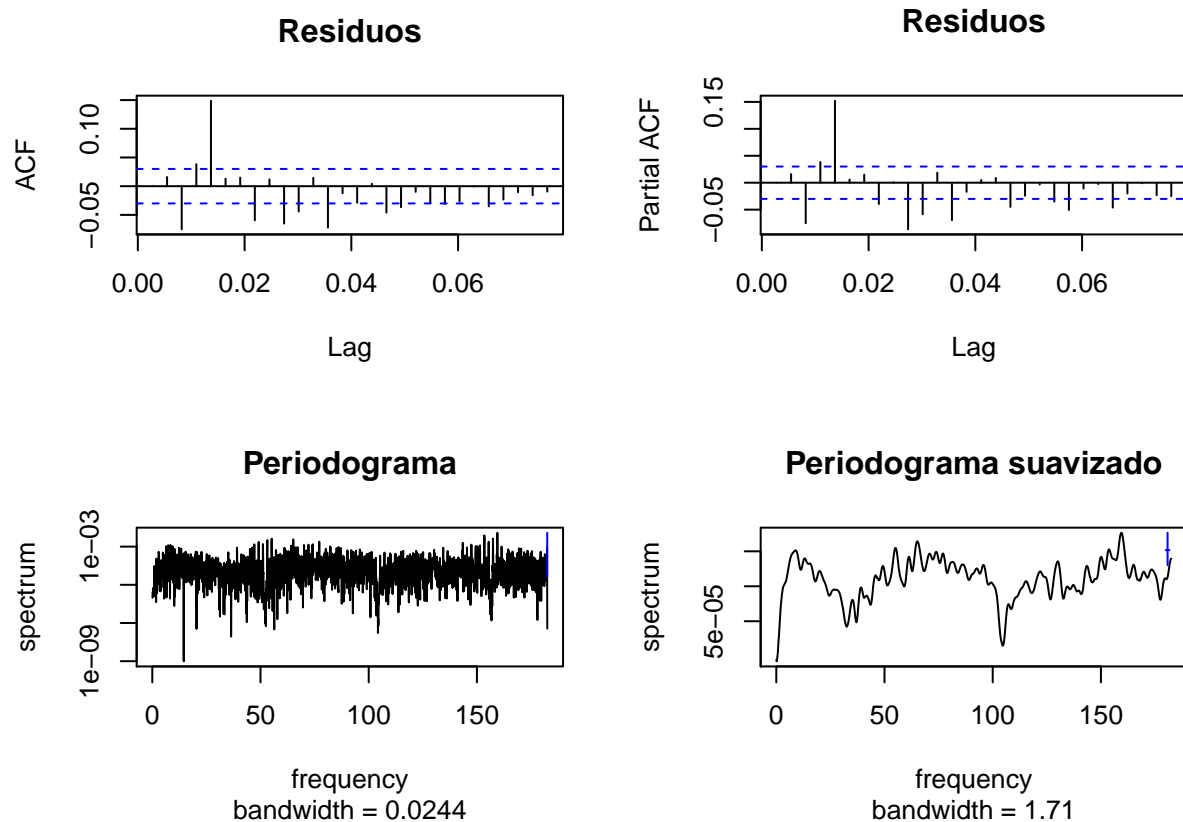




Aunque una parte de los residuos quede sin modelizar, los efectos residuales son pequeños y los periodogramas se mueven en valores muy pequeños, próximos a cero.

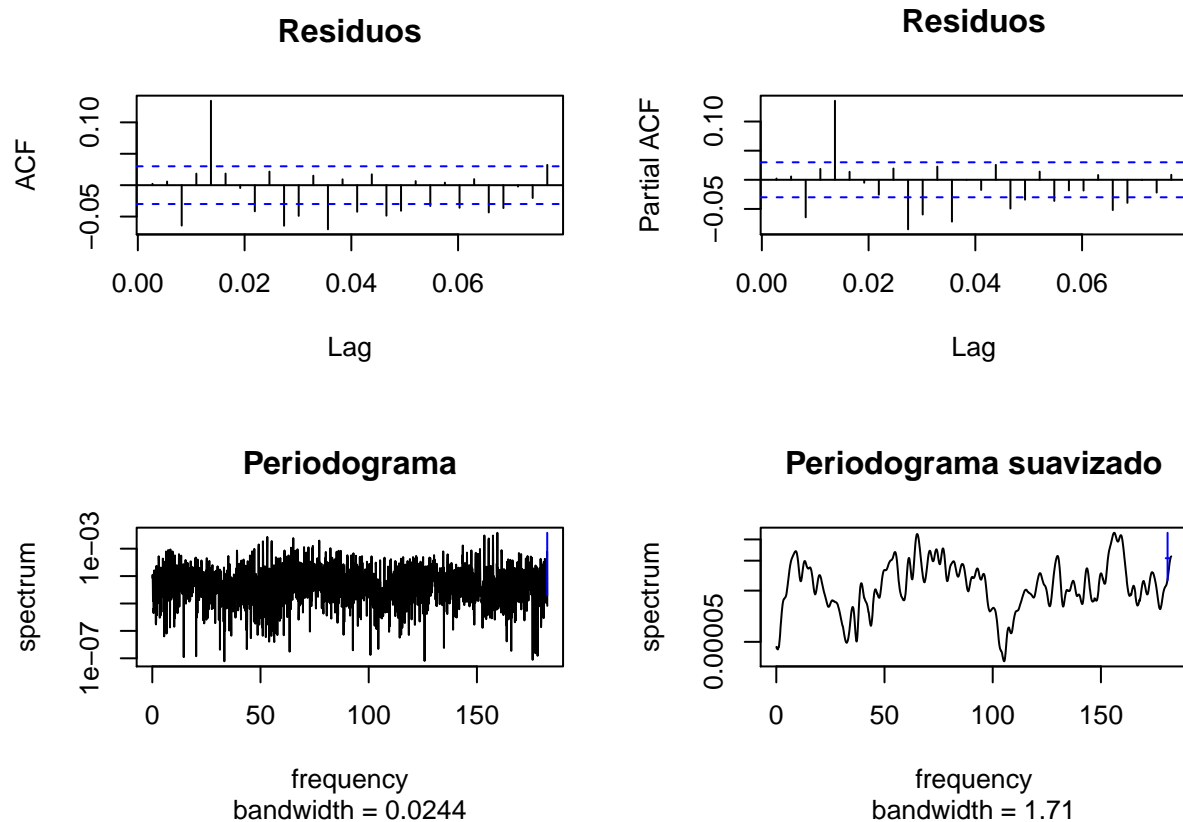
### 3.2. Modelo 2: ARIMA(1,0,0) con SARIMA(0,1,1) (sin *dummies*)

Se repite el modelo anterior pero sin las *dummies*. En este caso el modelo no es tan bueno (el criterio de Akaike devuelve un valor más grande).



### 3.3. Modelo 3: $ARIMA(1,0,0)$ con $SARIMA(1,1,2)$ + *dummies* para los Lunes de Pascua

Ya se ha apuntado a la existencia de un componente autorregresivo y uno de media móvil estacional, pero podría ser que existiera también un autorregresivo estacional, junto con las diferencias estacionales, y que la media móvil estacional pudiera ser de orden superior. Además, según los criterios de información, parece que las *dummies* mejoran el modelo. En este caso todos los coeficientes siguen siendo significativos y el criterio de Akaike mejora con respecto al modelo del punto 3.1.



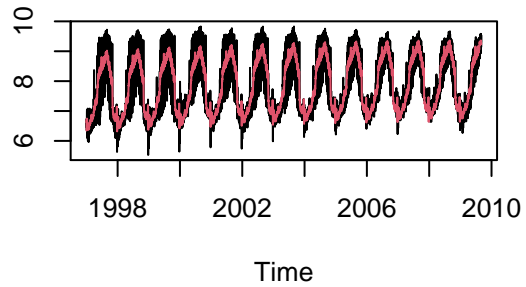
### 3.4. Valoración de los modelos ARIMA

A continuación se define una función que almacenará, por un lado, las series predichas para el periodo de validación a partir de la serie `train` y, por otra, las medidas de error (Error Cuadrático Medio, ECM, y Error Absoluto Medio, EAM) en las muestras de entrenamiento y validación.

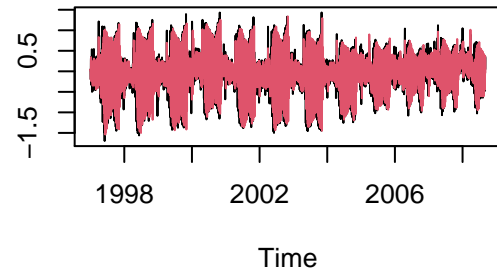
Una vez definida se aplica sobre cada modelo.

### 3.4.1. Resultados del Modelo 1

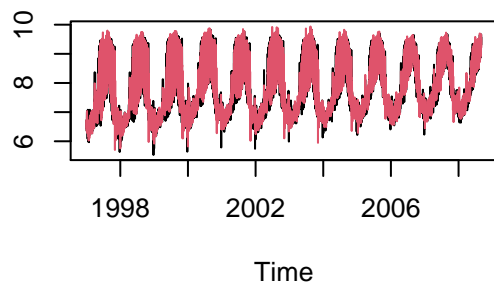
**Ajuste parte determinista (full)**



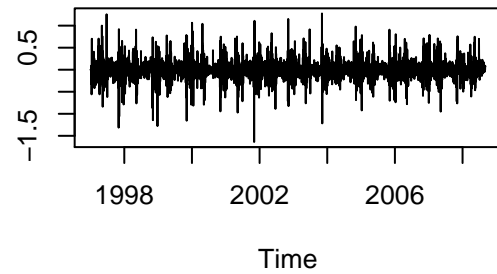
**Ajuste parte estacionaria (train)**



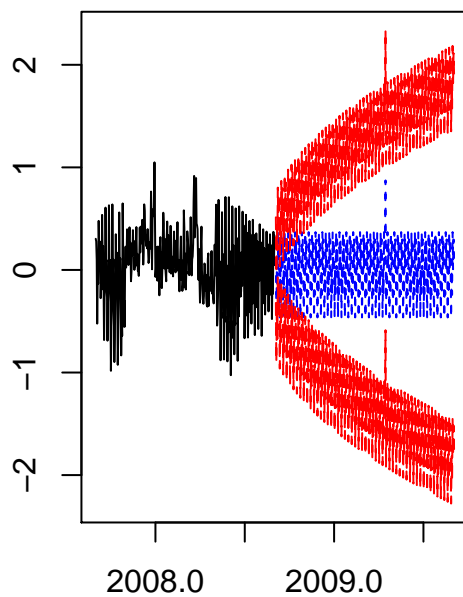
**Ajuste agregado (train)**



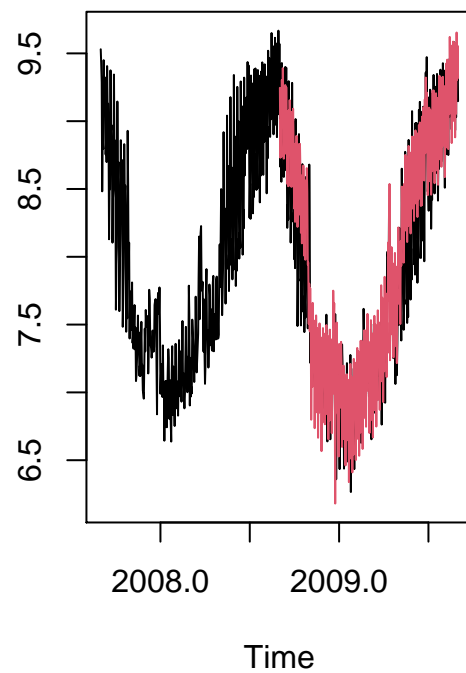
**Errores (train)**



**Predicción (parte estacionaria)**



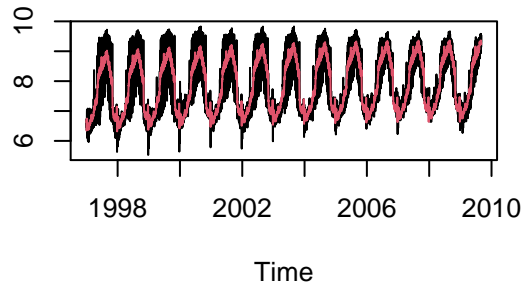
**Ajuste (test)**



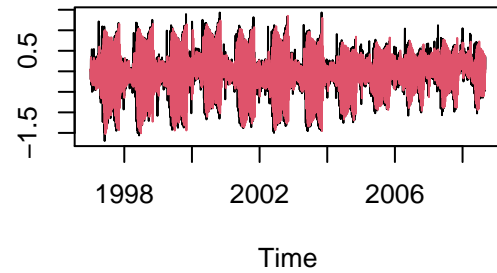
Error	Muestra <b>train</b>	Muestra <b>test</b>
ECM	0.0434	0.1123
EAM	0.1363	0.2666

### 3.4.2. Resultados del Modelo 2

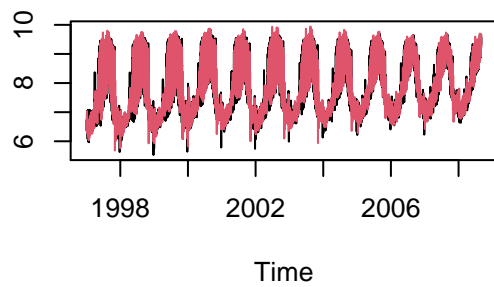
**Ajuste parte determinista (full)**



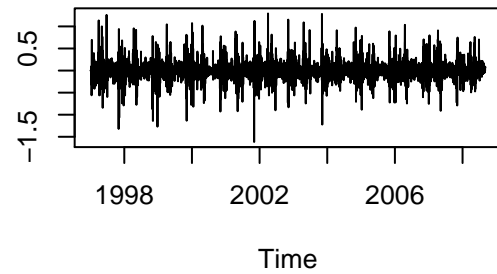
**Ajuste parte estacionaria (train)**



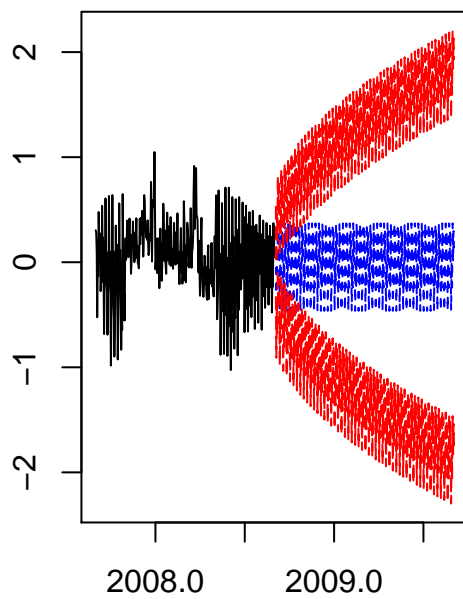
**Ajuste agregado (train)**



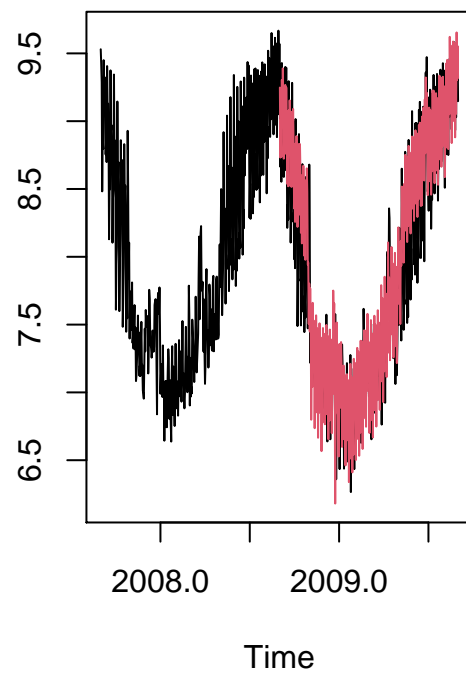
**Errores (train)**



**Predicción (parte estacionaria)**



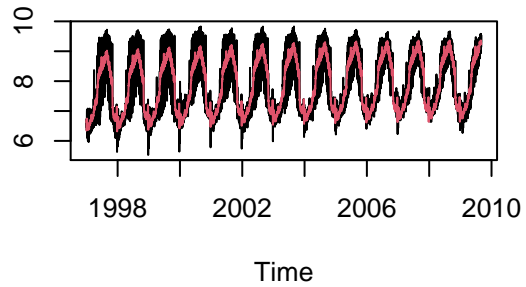
**Ajuste (test)**



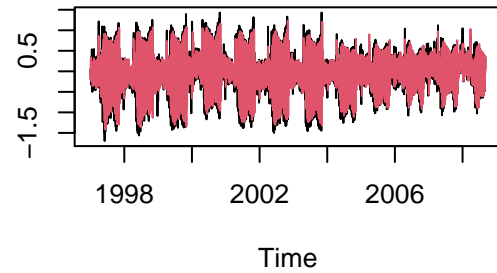
Error	Muestra <b>train</b>	Muestra <b>test</b>
ECM	0.0459	0.1128
EAM	0.1387	0.2672

### 3.4.3. Resultados del Modelo 3

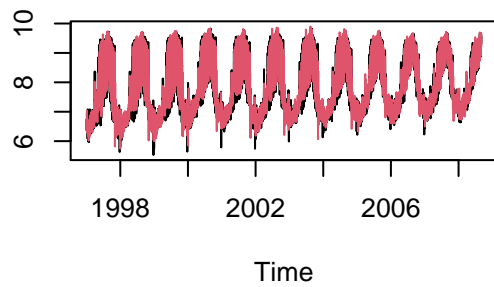
**Ajuste parte determinista (full)**



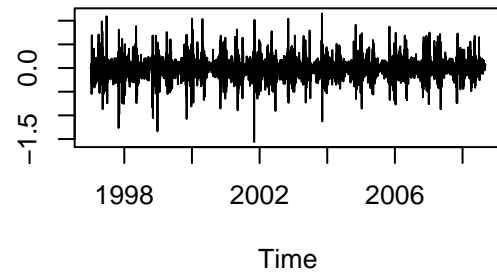
**Ajuste parte estacionaria (train)**



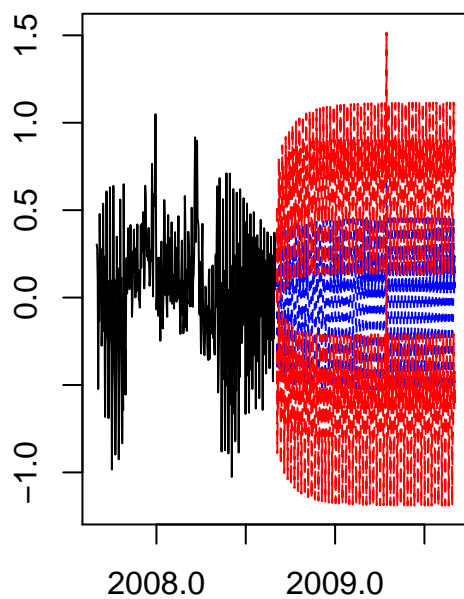
**Ajuste agregado (train)**



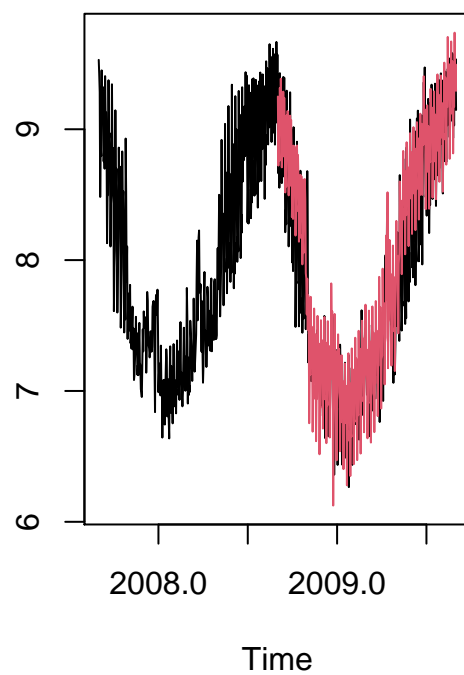
**Errores (train)**



**Predicción (parte estacionaria)**



**Ajuste (test)**

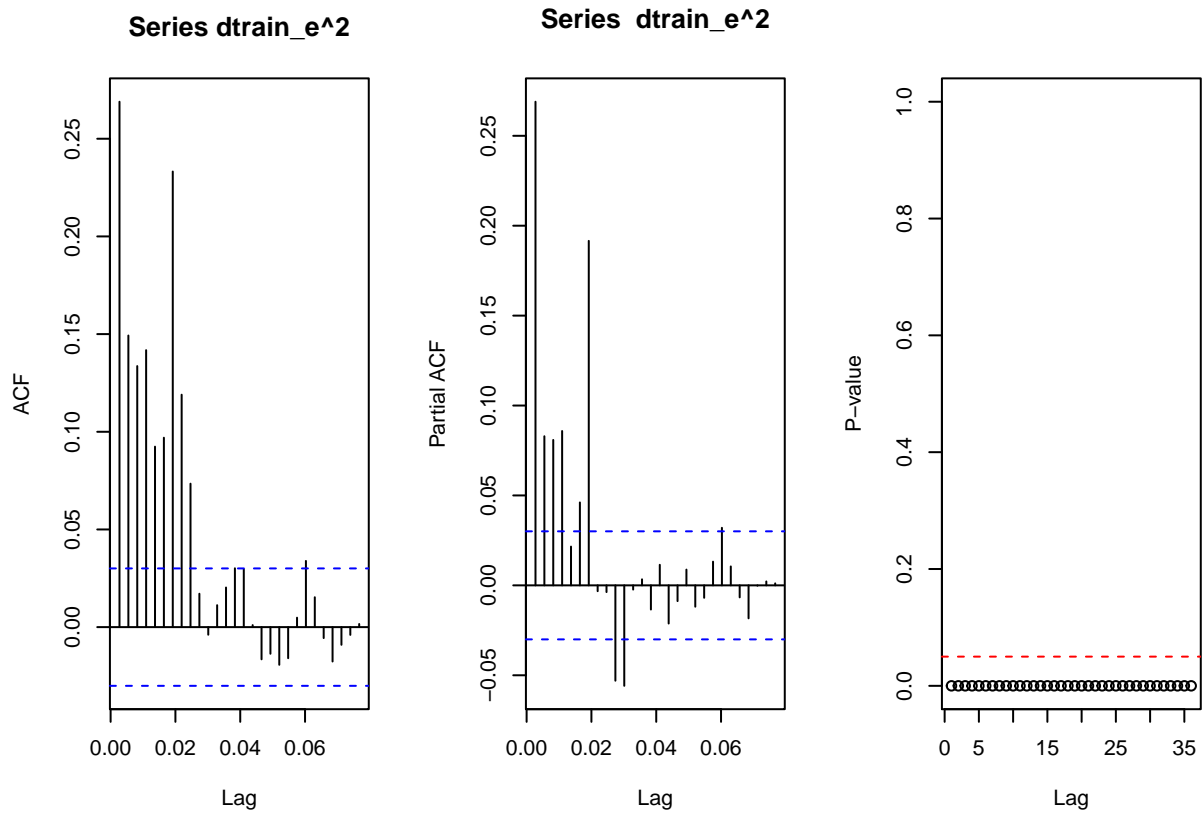




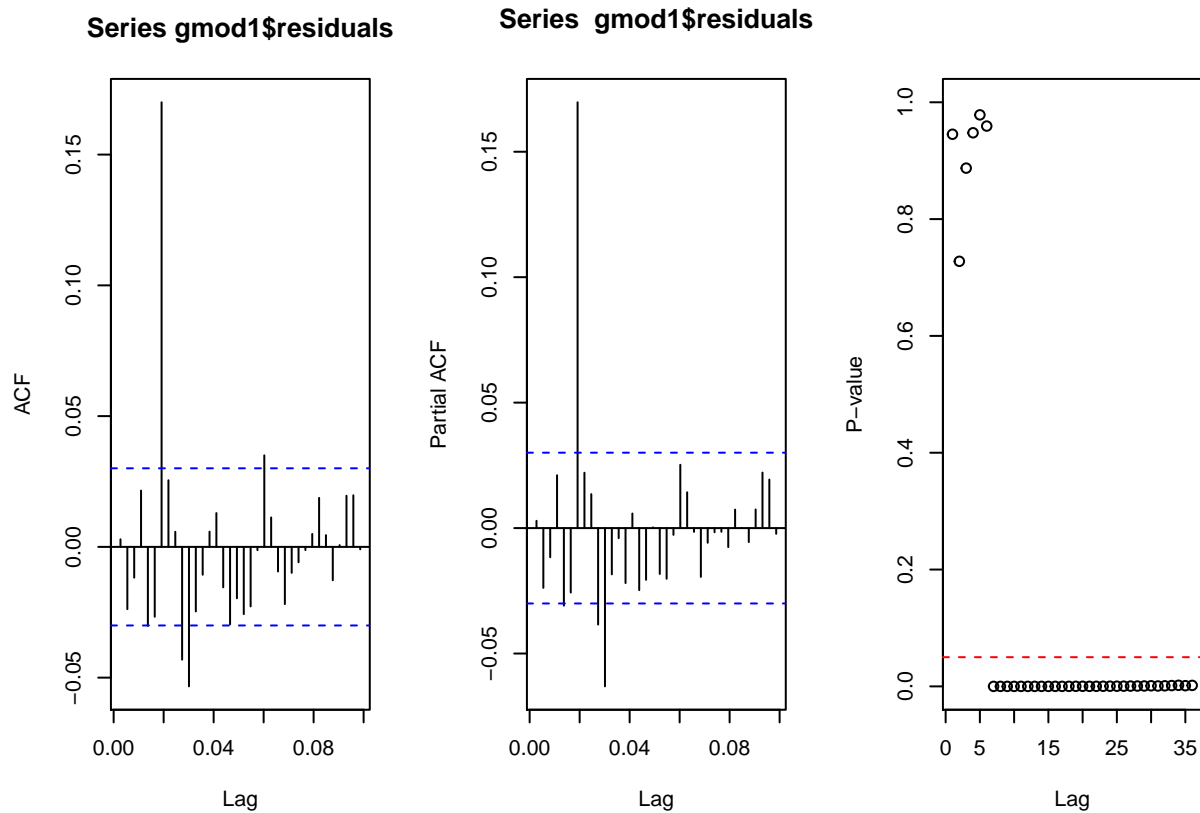
Error	Muestra <b>train</b>	Muestra <b>test</b>
ECM	0.0413	0.1088
EAM	0.1346	0.2615

## 4. Análisis de la volatilidad

Analizar la volatilidad de la serie tiene poco sentido para el análisis que se efectúa aquí. Dicho esto, la volatilidad se mueve entre unos máximos y mínimos más o menos constantes a lo largo de los años. Sin embargo, puede ser interesante analizar la volatilidad de la serie desestacionalizada. La desestacionalización pasa, como se ha mencionado al final del punto 2, por sacar diferencias en el séptimo retardo. Se visualiza el análisis de correlación del cuadrado de las séptimas diferencias de los errores.



La prueba de McLeod & Li establece como hipótesis alternativa ( $H_0$ ) que no hay efectos GARCH antes de determinado retardo. En este caso se rechaza  $H_0$ ; hay volatilidad. De acuerdo con lo observado se intenta modelizarla como un ARMA (o, en otras palabras, se modeliza un GARCH) con diferentes valores en el componente AR y en el MA. Por ejemplo, se modeliza como un GARCH(2,1).



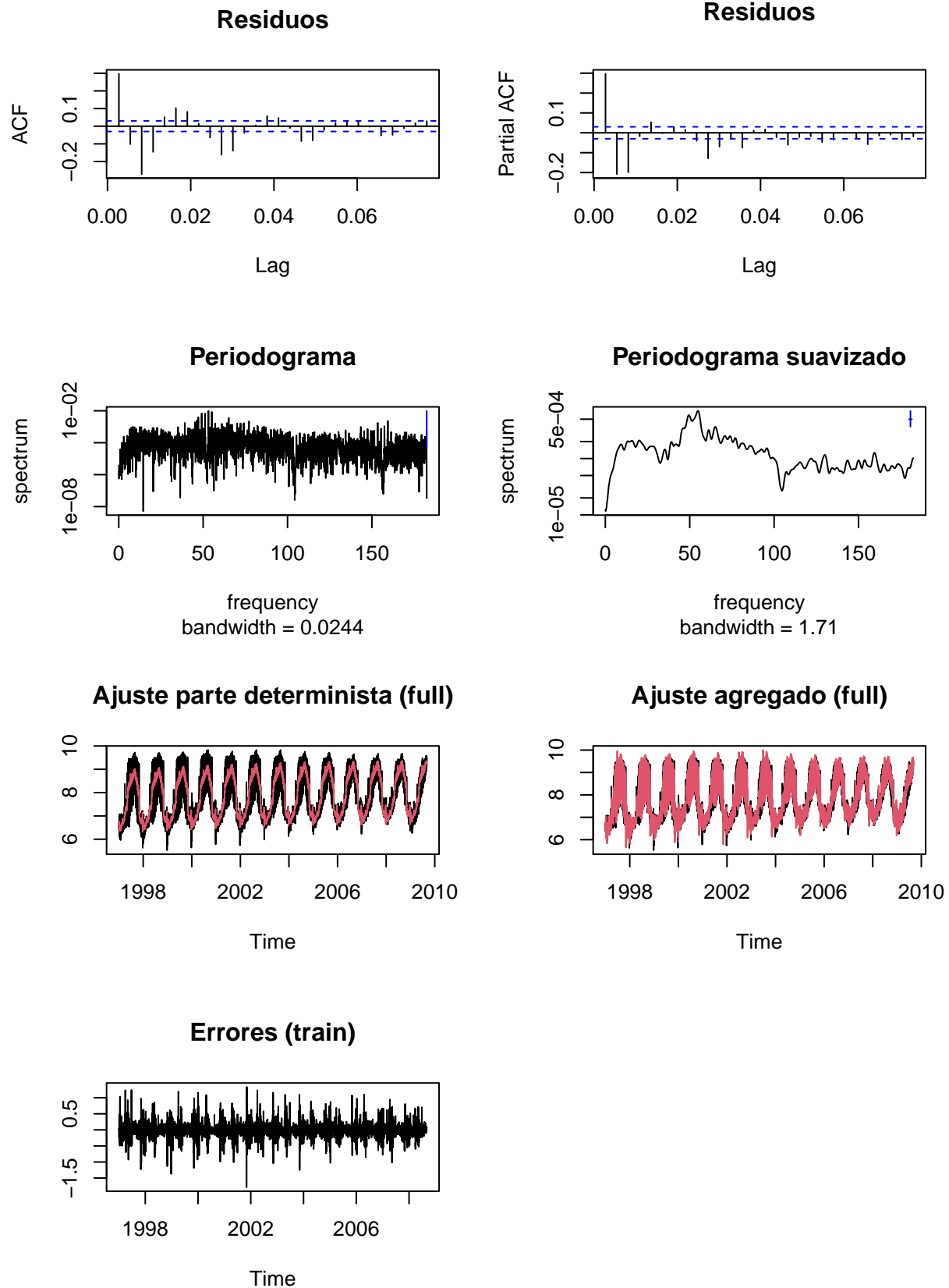
De acuerdo con el resultado de la prueba de McLeod & Li no puede rechazarse que no haya efectos GARCH por debajo del séptimo retardo, pero sí más allá. O, en otras palabras, el cambio en las oscilaciones es despreciable de una semana a otra, pero a partir de la segunda semana se rechaza que no pueda ser diferente.

## 5. Modelos estructurales de series temporales (componentes no observables)

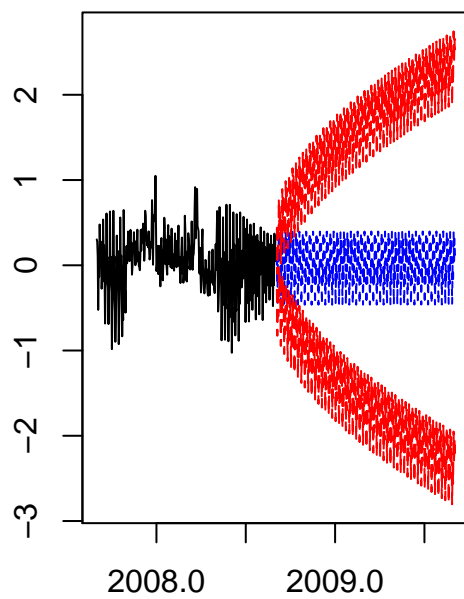
En los modelos estructurales de series temporales busca determinarse la forma que siguen los componentes no observables de la serie temporal (la tendencia, el ciclo y la estacionalidad). Existen dos maneras de especificar el modelo, según la forma en que se escribe el espacio de los estados: por variables instrumentales o con la especificación trigonométrica. Como en este trabajo ya se ha destendializado y desestacionalizado la serie en su componente anual en el punto 2 (y el ciclo no parece relevante), el modelo se aplica sobre la parte de la serie que queda sin explicar.

Tal como se ha hecho en el punto 3.4., se define una función para almacenar las series de predicción y las medidas de error. Luego se aplica sobre los distintos modelos.

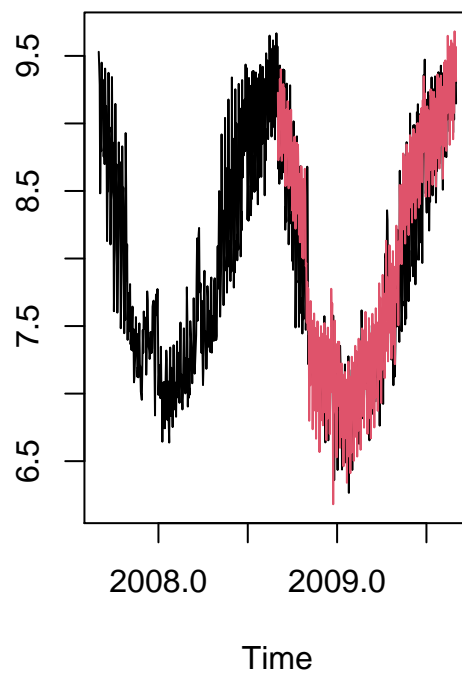
## 5.1. Resultados del Modelo 4: Modelo estructural por variables instrumentales



**Predicción (parte estacionaria)**

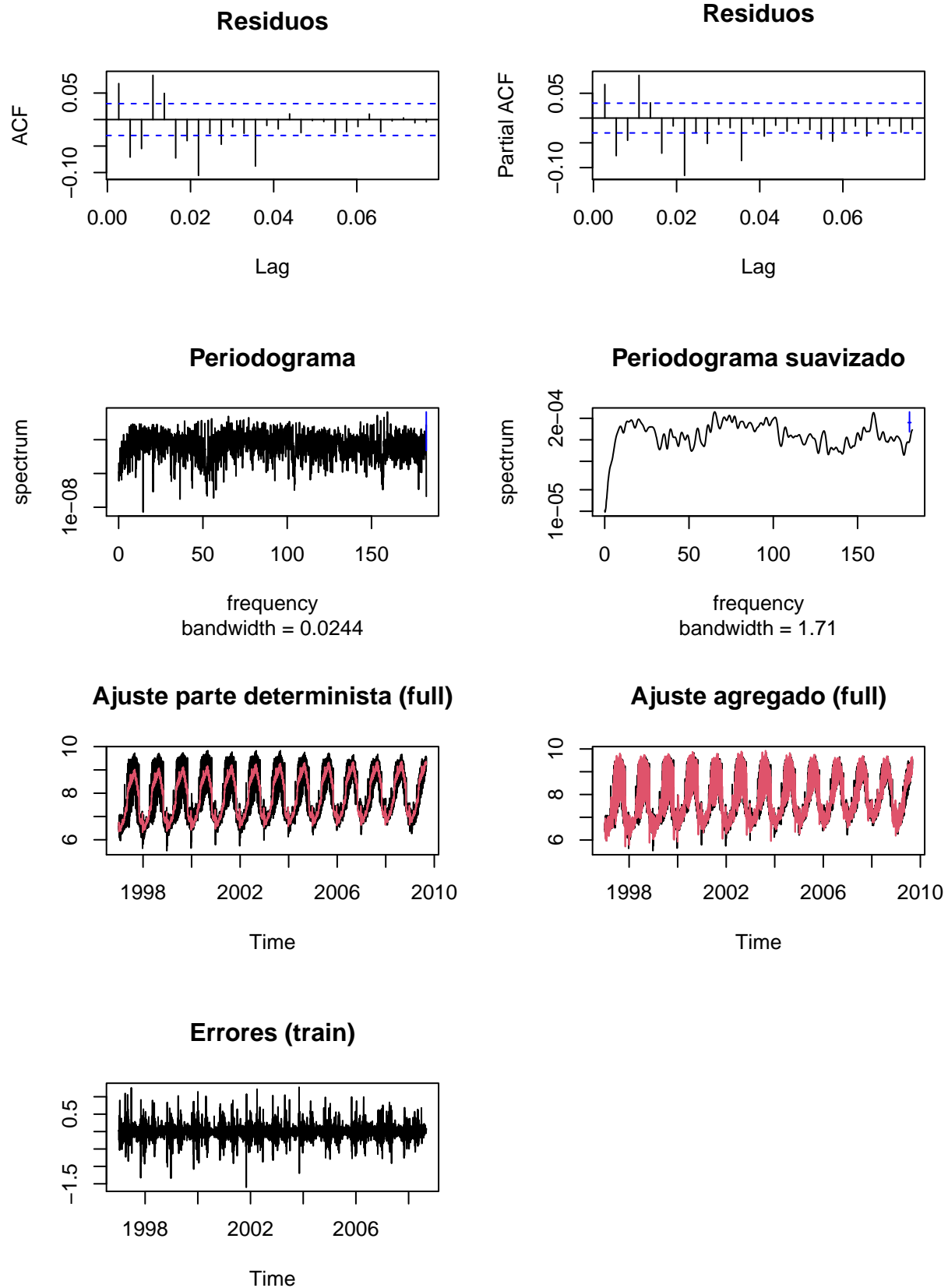


**Ajuste (test)**

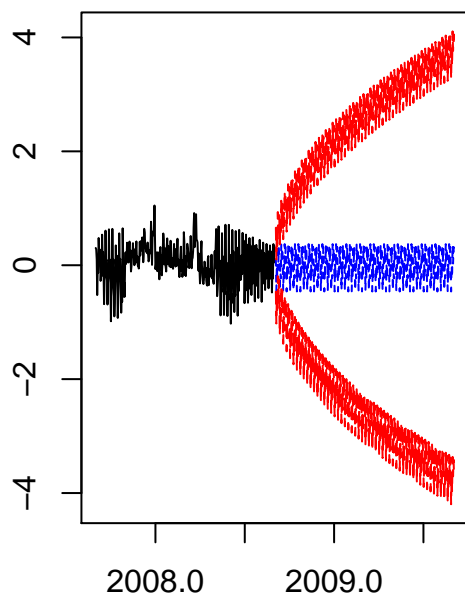


Error	Muestra train	Muestra test
ECM	0.0548	0.115
EAM	0.1558	0.2708

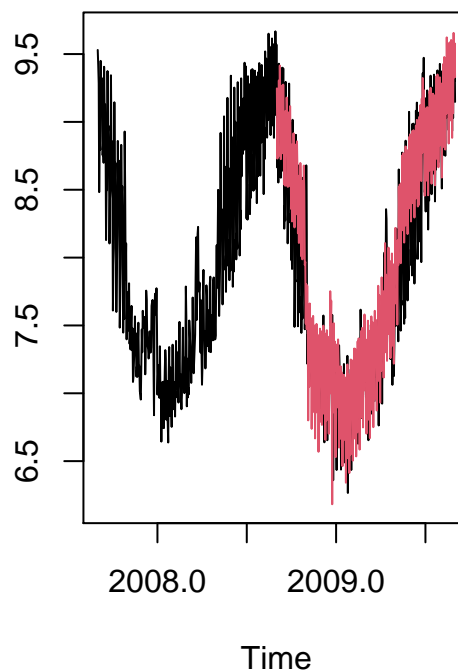
## 5.2. Resultados del Modelo 5: Especificación trigonométrica del modelo estructural



**Predicción (parte estacionaria)**



**Ajuste (test)**



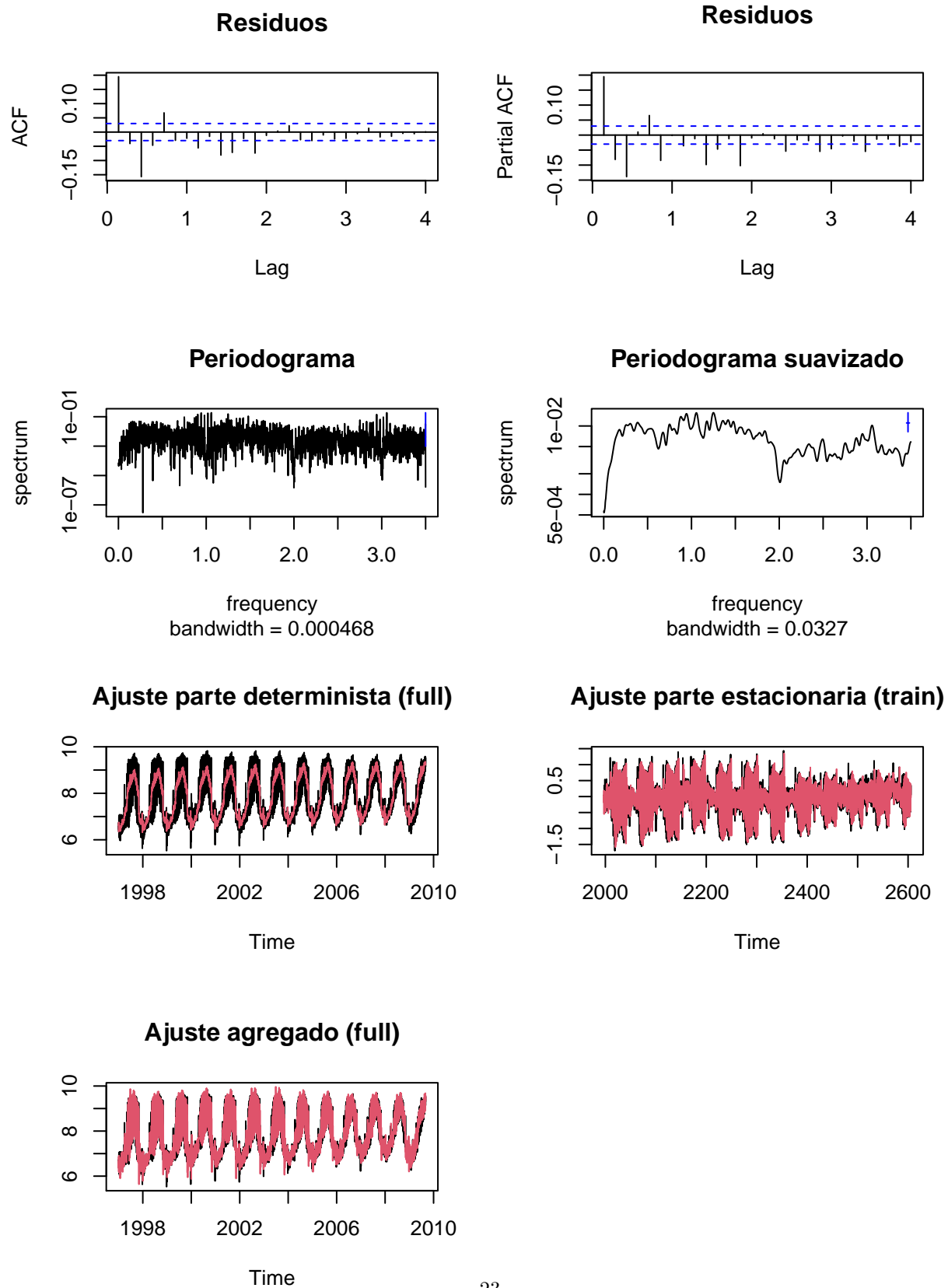
Error	Muestra <code>train</code>	Muestra <code>test</code>
ECM	0.0464	0.1167
EAM	0.1403	0.2736

## 6. Alisado exponencial

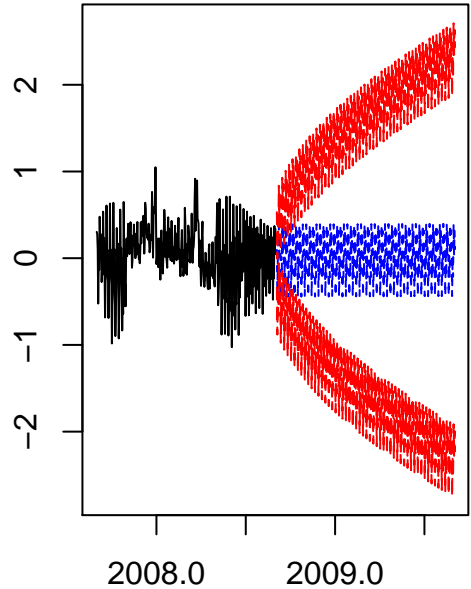
El alisado exponencial es un método de predicción basado en obtener sucesivas predicciones a partir de una media de los valores pasados de la serie ponderada por una constante de alisamiento obtenida por Mínimos Cuadrados No Lineales. Existen diversas variantes del método según incorporen al análisis de series y cómo traten la tendencia, la estacionalidad y la corrección de errores. En R esto último se traduce en especificar el enfoque (aditivo, multiplicativo o automático) que debe tomar en cada caso.

Tal como se ha hecho en los puntos 3 y 5, se define una función que almacenará los errores y las series para la representación de los resultados de los análisis. Luego se aplica sobre los modelos.

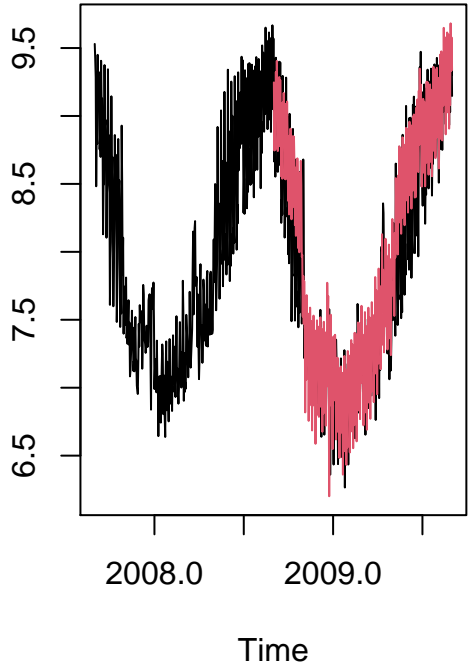
### 6.1. Resultados del Modelo 6: Alisado exponencial por State Space (coeficientes aditivos)



**Predicción (parte estacionaria)**



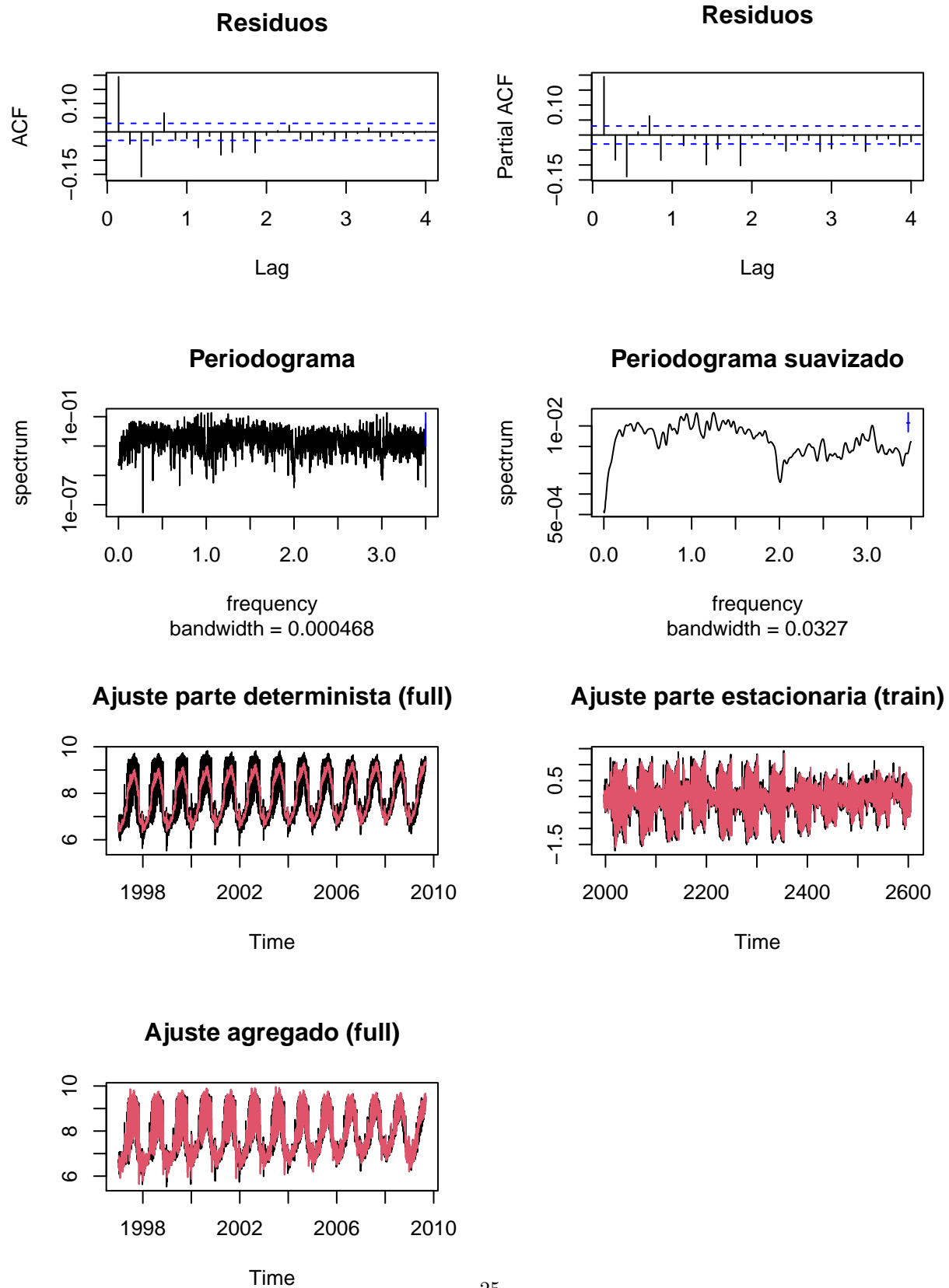
**Ajuste (test)**



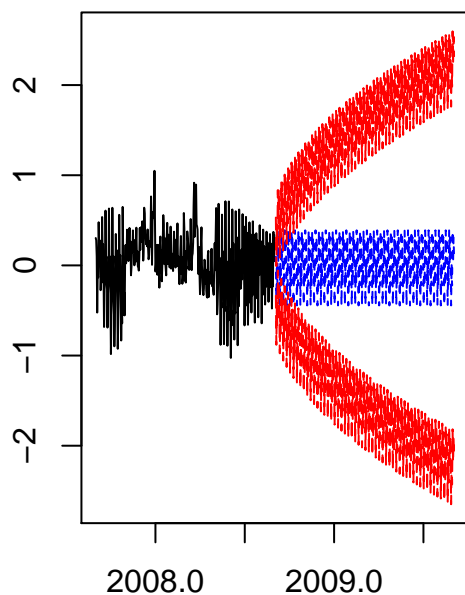
Error	Muestra <b>train</b>	Muestra <b>test</b>
ECM	0.0489	0.1177
EAM	0.1432	0.2744



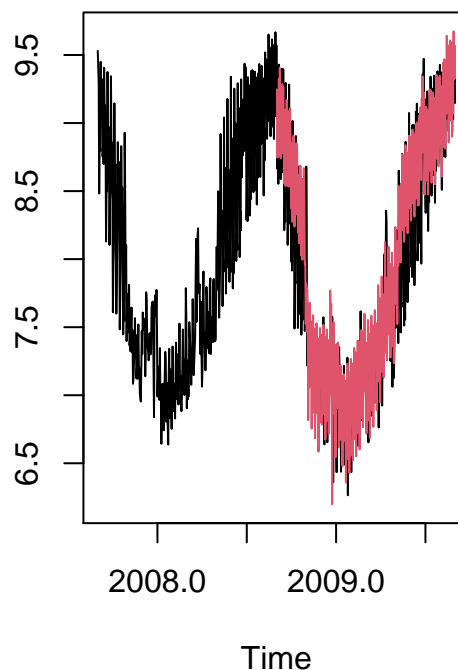
## 6.2. Resultados del Modelo 7: Alisado exponencial por State Space (coeficientes automáticos)



**Predicción (parte estacionaria)**



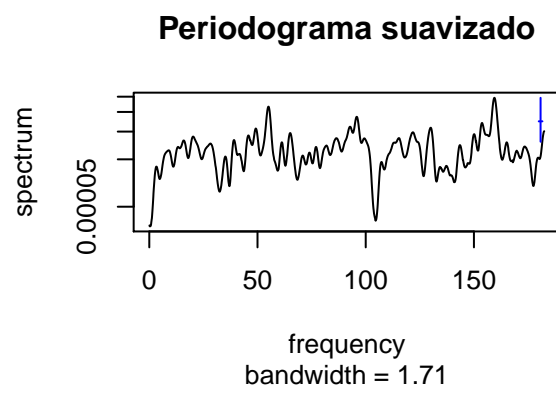
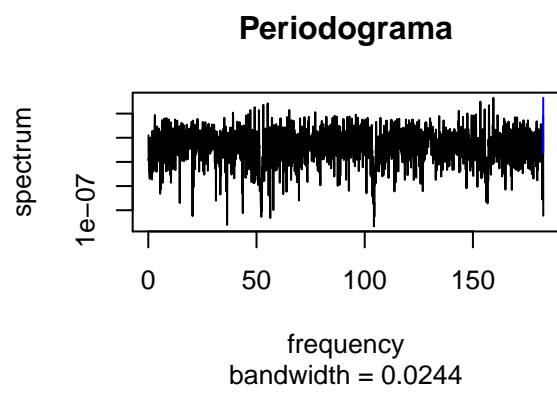
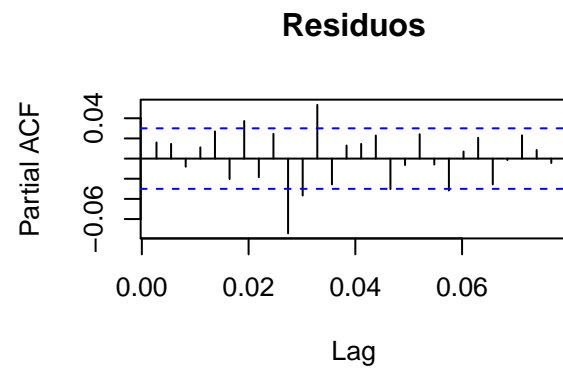
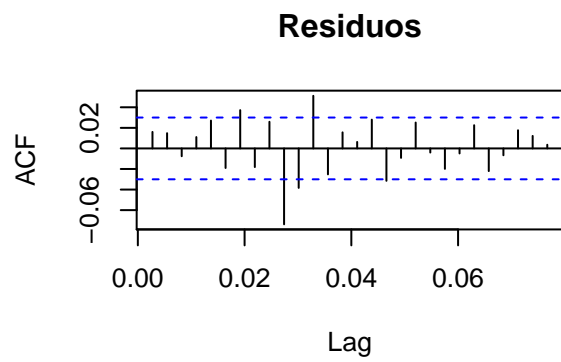
**Ajuste (test)**



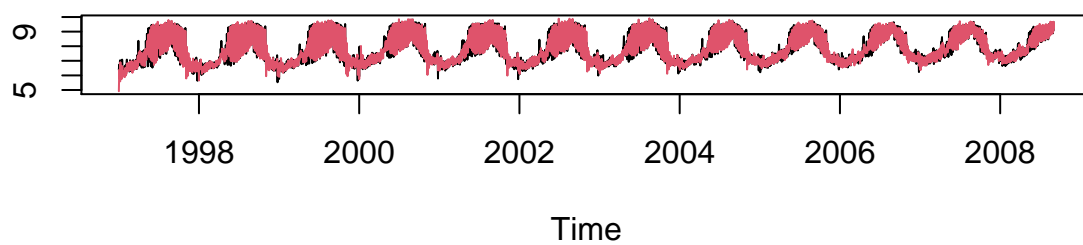
Error	Muestra train	Muestra test
ECM	0.0488	0.1167
EAM	0.1431	0.2729

### 6.3. Resultados del Modelo 8: Doble alisado exponencial

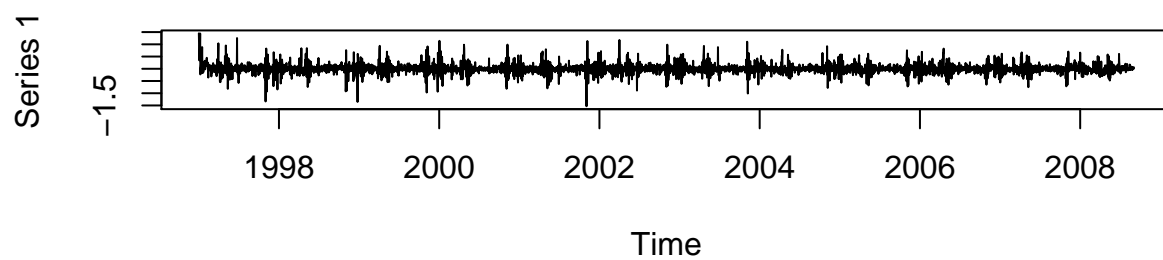
El doble alisado exponencial es una generalización del alisado exponencial que permite la existencia de dos periodos estacionales distintos, con lo cual es especialmente interesante en el caso de este análisis, en el que coexisten estacionalidades semanal y anual. La permisividad del método hace que pueda aplicarse a la serie original en logaritmos (convertida a un formato de serie temporal definido con dos frecuencias), sin necesidad de desestacionalizarla.



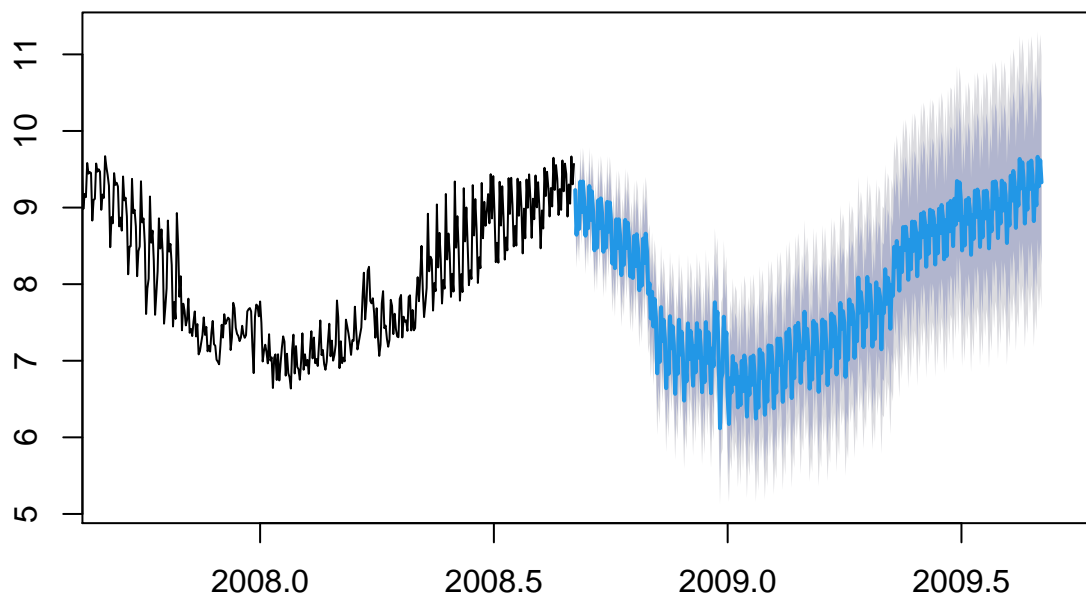
**Ajuste agregado (entrenamiento)**



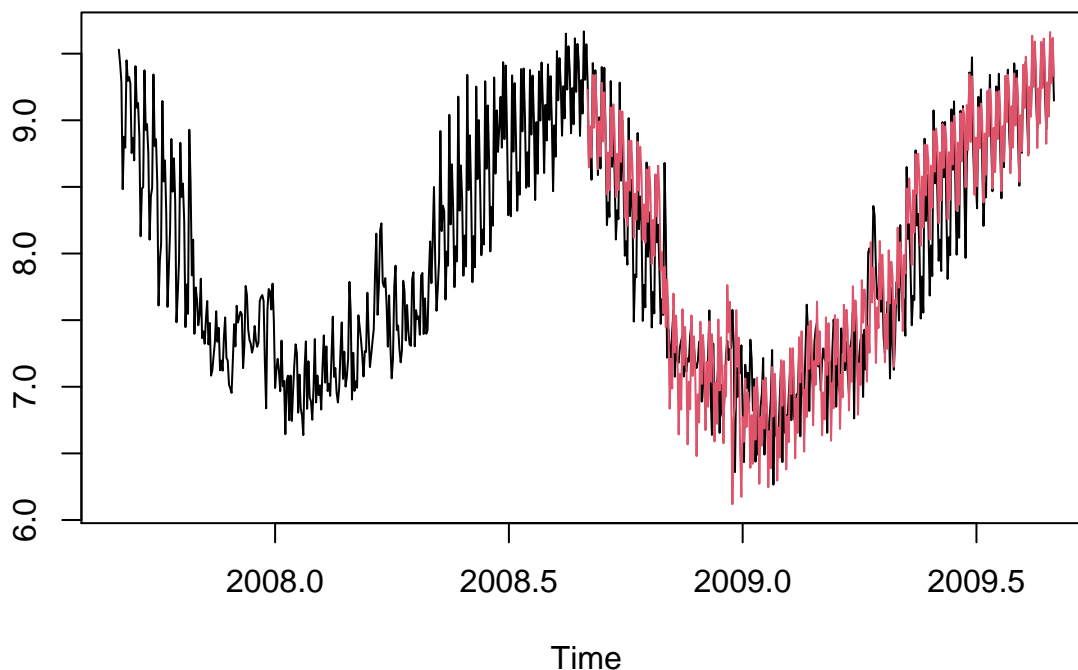
**Errores (entrenamiento)**



## Serie y predicción



## Ajuste sobre la serie test



Error	Muestra train	Muestra test
ECM	0.0432	0.1099
EAM	0.1355	0.2596

## 7. Elección del modelo y conclusiones

A modo de síntesis, en el cuadro siguiente se exponen los modelos con sus correspondientes errores, tanto en entrenamiento como en validación:

Modelo	ECM train	EAM train	ECM test	EAM test
1: ARIMA(1,0,0), SARIMA(0,1,1) + dummies	0.0434	0.1363	0.1123	0.2666
2: ARIMA(1,0,0), SARIMA(0,1,1)	0.0459	0.1387	0.1128	0.2672
3: ARIMA(1,0,0), SARIMA(1,1,2) + dummies	0.0413	0.1346	0.1088	0.2615
4: Estructural, dummies	0.0548	0.1558	0.115	0.2708
5: Estructural, trigonométrico	0.0464	0.1403	0.1167	0.2736
6: Alisado exponencial, aditivo	0.0489	0.1432	0.1177	0.2744
7: Alisado exponencial, automático	0.0488	0.1431	0.1167	0.2729
8: Doble alisado exponencial	0.0432	0.1355	0.1099	0.2596

Los mejores modelos parecen ser el modelo 3, que tiene el mejor desempeño en la muestra de entrenamiento y también el Error Cuadrático Medio menor en la muestra de validación. Sin embargo, tampoco puede desdeñarse el modelo 8 que, aunque justo por detrás del modelo 3 en esos indicadores, tiene mejores cifras

que este en lo que respecta al Error Absoluto Medio en la muestra test.

Aunque no es sorprendente el buen desempeño del Doble Alisado Exponencial, sí sorprende que un modelo relativamente sencillo como el 3 se postule como uno de los mejores, por delante de los modelos estructurales. Considerado junto con el modelo 1, que incluye el mismo regresor externo, se pone de manifiesto la importancia de las *dummies* para los Lunes de Pascua en el acierto general de las predicciones. Donde no ha sido posible incluir estas *dummies*, los modelos se resienten hasta el punto de verse superados por aquellos más sencillos. Por otra parte, la carencia de un modelo dominante en todas las medidas de error parece un síntoma más de esto: el modelo 3 puede estar prediciendo en entrenamiento mucho mejor algunas fechas, como las Pascuas, donde el modelo 8 comete grandes errores, pero el modelo 8 puede estar prediciendo el resto con más precisión que el modelo 3.

Una forma de mejorar las cifras podría pasar por usar varios modelos para predecir, en función de patrones en el desempeño de estos en la muestra de entrenamiento. A falta de profundizar más al respecto, esta consideración queda como una posibilidad a explorar en el futuro.