
SERIES TEMPORALES

ANÁLISIS DE DATOS
PROYECTO

GRADO EN CIENCIA E INGENIERÍA DE DATOS
UNIVERSITAT POLITÈCNICA DE CATALUNYA

ADRIÁN CEREZUELA HERNÁNDEZ - 48222010A

RAMON VENTURA NAVARRO - 21785256R

Abril 2023

Contenidos

1 INTRODUCCIÓN	1
1.1 Serie temporal - ConsumElec	1
1.2 Metodología	2
2 RESULTADOS E INTERPRETACIÓN	3
2.1 Modelo inicial	3
2.1.1 Identificación	3
2.1.2 Estimación	6
2.1.3 Validación	7
2.2 Estabilidad, capacidad de previsión y selección de modelo	10
2.2.1 Previsiones	11
2.3 Efectos de calendario	11
2.3.1 Análisis de Intervención	12
2.3.2 Validación del modelo	14
2.4 Tratamiento de atípicos	16
2.4.1 Identificación y estimación	17
2.4.2 Validación	18
2.4.3 Previsiones	19
3 CONCLUSIONES	20

Lista de gráficos

1 Representación gráfica de la serie X_t	1
2 Análisis gráfico de la varianza de la serie	3
3 Análisis gráfico de la varianza del log de la serie	3
4 Análisis gráfico del patrón estacional del log de la serie	4
5 Representación gráfica de $(1 - B^{12})\log X_t$	4
6 Aplicación de diferenciaciones regulares	5
7 ACF y PACF de W_t	5
8 Análisis de la varianza residual de $mod1$	7
9 Análisis de la varianza residual de $mod2$	7
10 Análisis de la normalidad residual de $mod1$	8
11 Análisis de la normalidad residual de $mod2$	8
12 ACF y PACF residual de $mod1$	8
13 ACF y PACF residual de $mod2$	9

14	Análisis de la independencia residual de <i>mod2</i>	9
15	Predicción de la serie a un año vista a partir de <i>mod1</i>	11
16	Comparación gráfica entre X_t y X_{lin_t}	13
17	Representación gráfica de los residuos de <i>modEC</i>	14
18	ACF y PACF de los residuos de <i>modEC</i>	14
19	Test de Llung-Box para <i>modEC</i>	15
20	Predicción de la serie a un año vista a partir de <i>modEC</i>	15
21	Comparación entre la serie y el efecto sobre ella de los outliers	17
22	ACF y PACF de X_{lin_t}	17
23	Representación gráfica de los residuos de <i>modEClin</i>	18
24	Test de Llung-Box para <i>modEClin</i>	19
25	Predicción de la serie a un año vista a partir de <i>modEClin</i>	19

1 INTRODUCCIÓN

Este es un proyecto evaluable para la asignatura de Análisis de Datos, del Grado en Ciencia e Ingeniería de Datos. El objetivo principal de este es buscar un modelo útil para hacer predicciones basadas en la última observación de una serie temporal dada. El primer paso para ello será buscar posibles modelos que se adecuen a la serie, estimarlos y realizar una validación de estos. Una vez validados, será necesario tratar las observaciones atípicas antes de llevar a cabo las predicciones. En primer lugar, se presentará la serie escogida para ello.

1.1 Serie temporal - ConsumElec

La serie temporal a analizar en este proyecto será la que trata el consumo interior bruto de energía eléctrica mensual en España, la cual ha sido seleccionada del documento "Casos propuestos" proporcionado en la asignatura. La serie ha sido extraída de la página web del Ministerio de Industria, Comercio y Turismo del Gobierno de España [Espa]. Estos datos son recogidos y procesados por la Red Eléctrica de España [Red; Ene], empresa que opera y gestiona la red de transporte de energía en nuestro país, la cual se encarga de, entre otras muchas funciones, asegurarse de hacer una previsión diaria de la energía necesaria, evitando así cortocircuitos por exceso o falta de energía.

El principal interés es ver cómo afecta la componente temporal al consumo energético por parte de los habitantes de nuestro país, tanto a nivel anual como a nivel mensual dentro de un mismo año, así como poder predecir cuál será el consumo en un año concreto a partir de las observaciones.

El conjunto de datos cuenta con 347 observaciones, divididas mensualmente desde el año 1990 hasta 2018. La energía está medida en Gigawatios-hora. Estos aspectos se pueden ver en la siguiente representación gráfica.

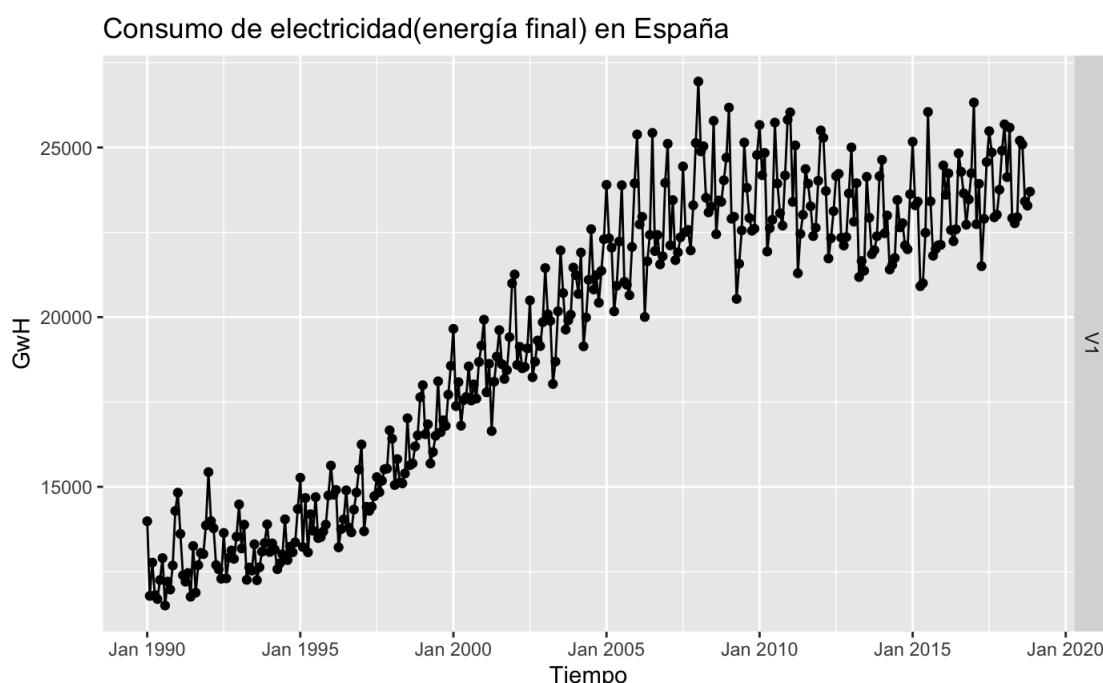


Figura 1: Representación gráfica de la serie X_t

A priori, se puede observar una diferencia de varianzas entre los diferentes años, teniendo picos más altos y picos más bajos dentro de cada año. Más tarde se profundizará sobre las épocas donde ocurren estos desniveles. Debido a la existencia de estos repetidos picos en las mismas épocas año tras año, se puede intuir un patrón estacional.

Sorprende ver cómo en la mayoría de series trabajadas a lo largo del curso ha habido un declive alrededor del año 2008, debido a la crisis financiera en España, pero en este caso no se observa. Según un análisis del mercado eléctrico entre 2008 y 2014 [For], realizado por la empresa AleaSoft, mientras que tras la crisis financiera los precios de la electricidad cayeron y se fueron recuperando posteriormente, el periodo comprendido entre estos años se caracteriza por la baja demanda eléctrica en nuestro país. Esta es la explicación del pequeño valle que se observa en la Figura 1 entre esos mismos años.

Podemos intuir que la serie no es estacionaria, es decir, que no se comporta igual independientemente del tiempo, y puede tener cambios en la varianza o en la media.

1.2 Metodología

Antes del análisis, se describirán de forma sintética la metodología empleada para analizar el conjunto de datos. Sin embargo, paralelamente al análisis y resultados obtenidos posteriormente, se desarrollarán más extensamente los conceptos que aparecen en esta sección.

A lo largo de este proyecto, se empleará la metodología de Box-Jenkins tratada en la asignatura de Análisis de Datos [Sán23]. Ésta es una herramienta útil, sistemática y rigurosa para la predicción, análisis de series temporales y construcción de modelos ARIMA que consiste esencialmente en tres etapas iterativas: Identificación y selección del modelo, Estimación de los parámetros y Comprobación del modelo.

La identificación del modelo adecuado para la serie temporal implica una revisión cuidadosa de la propia serie y, principalmente, la determinación de si es estacionaria o no estacionaria. Es decir, si varía o no en el tiempo. Se requiere asegurar que la serie en cuestión es estacionaria, en caso contrario se aplicarán transformaciones para estabilizar la varianza y la media, a la vez que eliminar los posibles patrones estacionales. Una vez se cuente con la serie estacionaria, mediante la confección y análisis de los gráficos de autocorrelación (ACF) y autocorrelación parcial (PACF), se definirán los parámetros de los posibles modelos a estimar, tal como se explicará en la sección correspondiente.

Una vez identificados potenciales modelos, se deben ajustar a los datos. Este segundo paso consiste en la estimación de los parámetros del modelo mediante algoritmos de cálculo con la finalidad de obtener los coeficientes que mejor ajusten el modelo ARIMA seleccionado. En este caso se realizará la estimación mediante las funciones integradas en R.

El tercer y último paso consiste en realizar todo un seguido de pruebas de diagnóstico, basadas en un análisis de los residuos del modelo. Principalmente, se comprobarán las hipótesis de homocedasticidad, normalidad e independencia residual. En caso de no cumplirse estas hipótesis, se debe volver a iniciar el proceso desde la primera etapa.

Finalmente, en el caso de tener un modelo ajustado y validado, ya se podrán realizar pronósticos para valores futuros de la serie temporal. Los pronósticos se basan en los parámetros estimados del propio modelo y los valores pasados o históricos de la serie. Se evaluará la precisión de las predicciones mediante la utilización de medidas como el error absoluto medio porcentual (MAPE) y el error de raíz cuadrada media (RMSE).

2 RESULTADOS E INTERPRETACIÓN

2.1 Modelo inicial

2.1.1 Identificación

Para poder predecir será necesario un modelo adecuado para la serie. De manera que ese será el primer paso, plantear como mínimo dos modelos candidatos a partir de los cuales poder trabajar. En caso de que ninguno de los propuestos en un inicio fueran eficaces, sería necesario encontrar otro que se ajuste a los datos.

La identificación de estos se realizará a través del ACF(Auto-Correlation Function) y el PACF(Partial Auto-Correlation Function) de la serie. Para que esta identificación pueda llevarse a cabo, la serie deberá de ser estacionaria. Así tendrá la media y varianza constantes, y la estructura de correlación no depende del origen. De este modo, se pueden estimar los parámetros del modelo con la serie temporal. Si no lo fuera, el número de parámetros sería excesivo y no se podría realizar una estimación eficiente. Tal como se ha visto al introducir la serie, ésta no lo es, por lo que serán necesarias transformaciones. En primer lugar, será necesario realizar un análisis gráfico de la varianza.

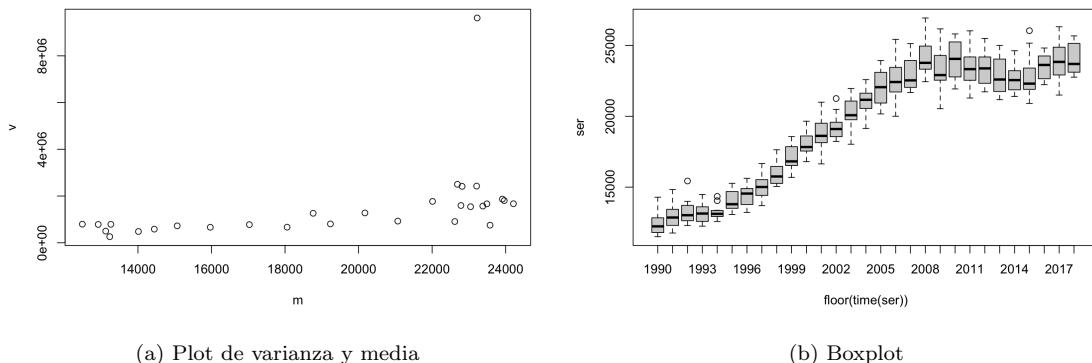


Figura 2: Análisis gráfico de la varianza de la serie

El ancho de las cajas del boxplot es una medida de dispersión robusta, ya que dentro se encuentran los 6 meses centrales. Se puede observar como la varianza es muy variable a lo largo de los años, por lo que será necesario aplicar logaritmos a la serie.

Una vez aplicado el logaritmo, vemos como ancho de las cajas [Figura 2b] presenta una estabilidad mayor que el de la serie original, por lo que se puede considerar que la varianza se ha homogeneizado, y es constante. En el boxplot se puede observar también la presencia de algunas observaciones atípicas, las cuales serán tratadas adecuadamente más tarde.

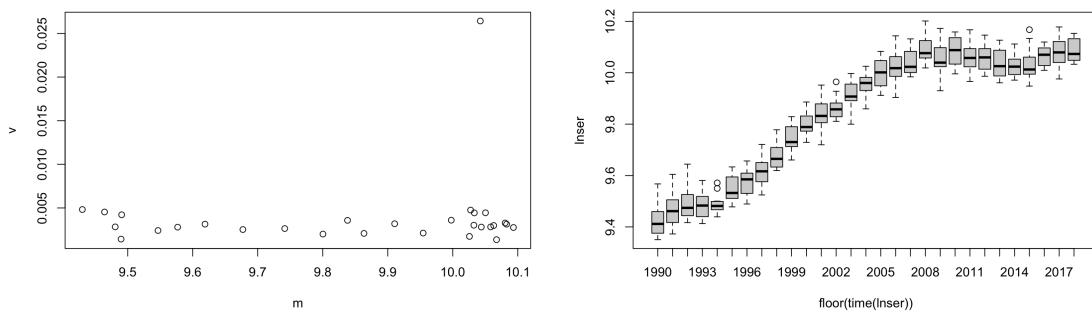


Figura 3: Análisis gráfico de la varianza del log de la serie

Una vez tratada la varianza, se ha de comprobar si la serie presenta un patrón estacional, es decir, si presenta subidas y bajadas periódicas que se presentan de forma regular en la serie temporal. Para ello, se utilizan las representaciones gráficas por meses. Las mayores subidas de consumo eléctrico se sitúan en Enero y Julio, los meses centrales de la invierno y verano respectivamente, por lo que tiene sentido que estos sean los meses de mayor consumo, con el objetivo de combatir las altas o bajas temperaturas, dependiendo de la estación.

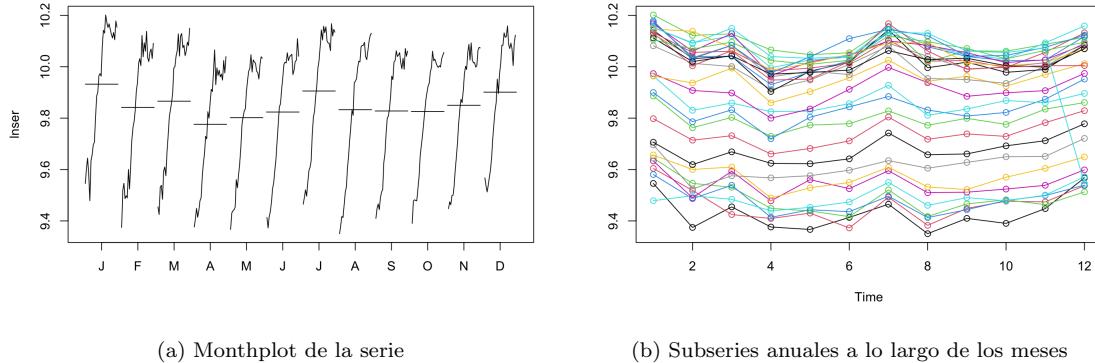


Figura 4: Análisis gráfico del patrón estacional del log de la serie

Por tanto, la serie presenta un patrón estacional, y será necesario aplicar una diferenciación estacional $(1 - B^{12})$, y trabajar con un incremento de la estación o mes respecto del año anterior.

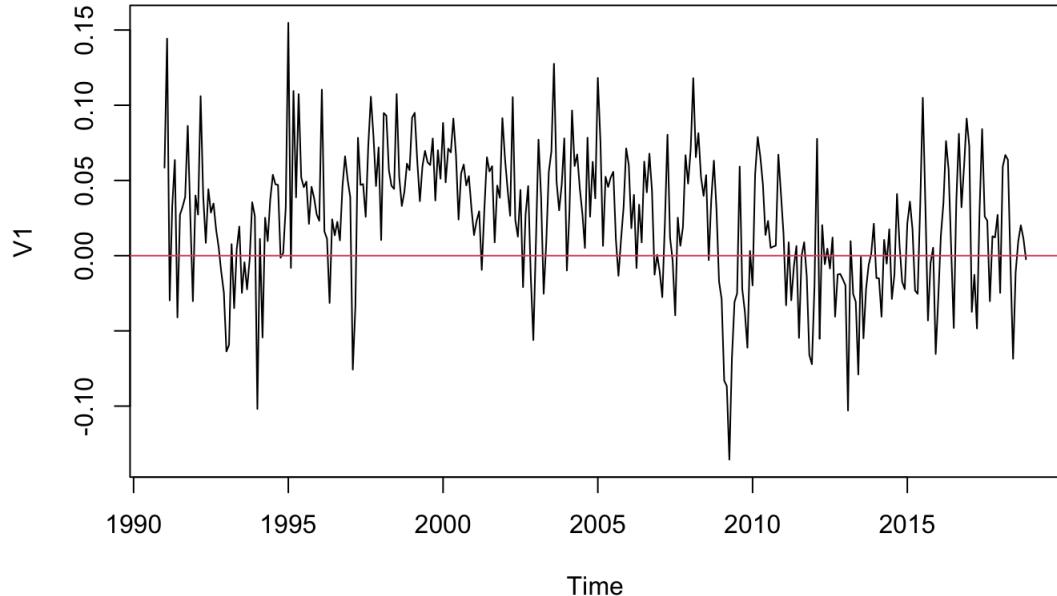


Figura 5: Representación gráfica de $(1 - B^{12})\log X_t$

De esta manera queda eliminado el patrón estacional, y se puede pasar a comprobar si la media es constante. Claramente, a través del gráfico anterior, puede verse que no lo es. Por tanto, será necesario aplicar tantas diferenciaciones regulares $(1 - B)$ como sea necesario con tal de corregirla. Una vez se consiga esto, deberemos ver qué diferenciación es la que da una mejor relación entre mantener la media constante y no aumentar la varianza en exceso.

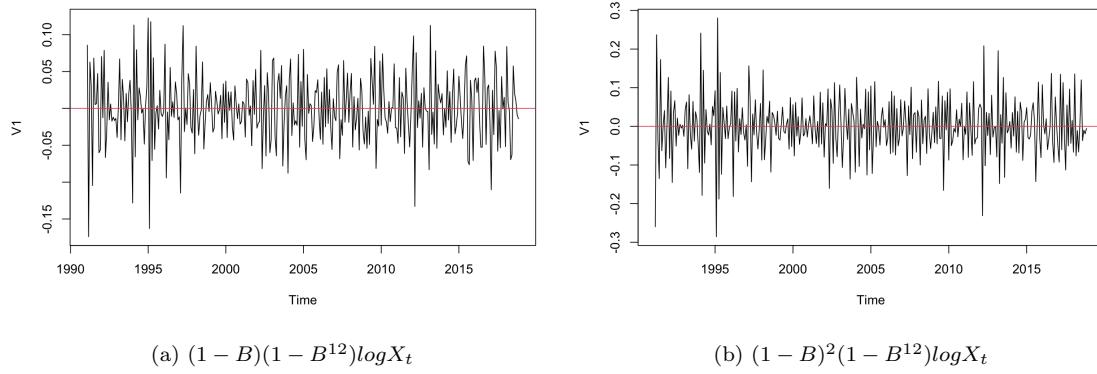


Figura 6: Aplicación de diferenciaciones regulares

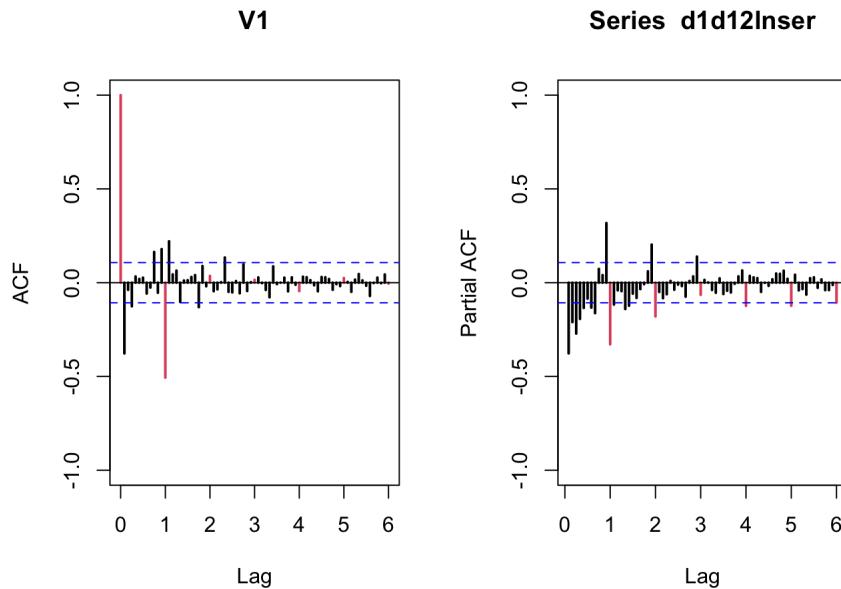
W_t	Varianza
$\log X_t$	0.05765
$(1 - B^{12})\log X_t$	0.00200
$(1 - B)(1 - B^{12})\log X_t$	0.00220
$(1 - B)^2(1 - B^{12})\log X_t$	0.00604

La varianza de la serie tras la última diferenciación regular ha aumentado considerablemente. Tras aplicar una sola diferenciación regular también aumenta la varianza, pero es un aumento mínimo, por lo que se puede tomar a costa de hacer constante la media. Por tanto, la última diferenciación regular no se tendrá en cuenta y, tras realizar las transformaciones, se trabajará con la serie estacionaria W_t , que es de la forma

$$W_t = (1 - B)(1 - B^{12})\log(X_t)$$

A continuación, se presentará el ACF y el PACF de la serie transformada, con el fin de decidir cuáles serán los parámetros de los modelos a presentar.

El ACF representa como de relacionado está el presente con el pasado de hace un período, incluyendo bandas horizontales de confianza. Con pocos retardos significativos, que enseguida convergen a las bandas, se puede afirmar que la serie es estacionaria, tal como ocurre en Figura 7, mostrada a continuación.

Figura 7: ACF y PACF de W_t

Los modelos presentados serán de la forma $ARIMA(p, d, q)(P, D, Q)_{12}$, donde (p, d, q) corresponden a los parámetros de la parte regular del modelo, mientras que (P, D, Q) a los de la parte estacional. Las p 's y q 's son los parámetros de los correspondientes modelos AR o MA , mientras que las d 's son el número de diferenciaciones que se llevan a cabo en las transformaciones.

Para la parte estacional se tendrán en cuenta los retardos significativos múltiples marcados en rojo, mientras que para la parte regular se tendrán en cuenta los 6,7,8 máximos primeros retardos. Los órdenes p y q equivaldrán a la posición del último retardo no nulo.

En cada caso, el parámetro p del modelo $AR(p)$ será la posición del último retardo no nulo en el PACF de la serie, mientras que el parámetro q del modelo $MA(q)$ será la posición del último retardo no nulo en el ACF.

De este modo, tenemos que:

- Parte estacional: En el ACF encontramos un retardo significativo. Por tanto, obtendremos un modelo $MA(1)$ estacional. Para el PACF se observan dos retardos significativos. Después del tercero tenemos dos significativos, pero como sobresalen por poco de las bandas de confianza, se considerarán valores aleatorios fruto del azar, y el modelo que tomaremos será un $AR(2)$ estacional.
- Parte regular: En el ACF encontramos tres retardos significativos. Los posteriores, como antes, se consideraran fruto del azar, considerando también que están muy lejos del inicio. Se considerará un $MA(3)$ regular en este caso. Para el PACF se observan cinco u ocho retardos significativos, según si consideramos los últimos dos valores aleatorios. En un principio se propondrá un $AR(8)$ regular, con la opción de poder reducir los parámetros más tarde en la estimación.

2.1.2 Estimación

Teniendo en cuenta que previamente se ha aplicado una diferenciación regular y una estacional, y considerando los modelos con un $MA(1)$ estacional, los modelos iniciales propuestos serían los siguientes:

$$mod1 \rightarrow ARIMA(0, 1, 3)(0, 1, 1)_{12} \quad (1)$$

$$mod2 \rightarrow ARIMA(8, 1, 0)(0, 1, 1)_{12} \quad (2)$$

Primer modelo: Estimaremos el modelo 1. Obtenemos los siguientes coeficientes para cada parámetro:

Coefficients:			
ma1	ma2	ma3	sma1
-0.5672	-0.1152	-0.0758	-0.7942
s.e.	0.0559	0.0632	0.0577
			0.0383

Para determinar si un coeficiente es significativo en el modelo o no, se realiza el test de t-ratios para cada uno, calculando el estadístico $|t| = |\frac{\hat{\phi}_1 - 0}{S_{\hat{\phi}_1}}|$ y rechazando la hipótesis nula ($H_0 : \phi_1 = 0$) si este es mayor que 2. Si el valor está cerca, no obstante, puede ser necesario considerarlo en el modelo igualmente. De esta manera, fijará a 0 el coeficiente $ma3$. Por tanto, el modelo a considerar será un $ARIMA(0, 1, 2)(0, 1, 1)_{12}$, cuyos coeficientes serán los siguientes:

Coefficients:			
ma1	ma2	ma3	sma1
-0.5861	-0.1636	0	-0.7924
s.e.	0.0522	0.0489	0
			0.0383

Segundo modelo: Procedemos igual para el modelo 2, obteniendo los siguientes coeficientes:

Coefficients:	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	sma1
	-0.5787	-0.4554	-0.4506	-0.3889	-0.2732	-0.1709	-0.1676	-0.1116	-0.7913
s.e.	0.0554	0.0634	0.0668	0.0693	0.0698	0.0670	0.0635	0.0554	0.0394

Realizando el test de significancia, en este caso comprobamos que ninguno de los parámetros debe ser eliminado del modelo, todos son significativos.

2.1.3 Validación

Una vez decididos los modelos sobre los que proceder, el siguiente paso es validarlos. Será necesario analizar tanto sus residuos, como su causalidad e invertibilidad.

A continuación se realizará un análisis completo de los residuos de ambos modelos, centrado en el estudio de la varianza, normalidad e independencia de estos.

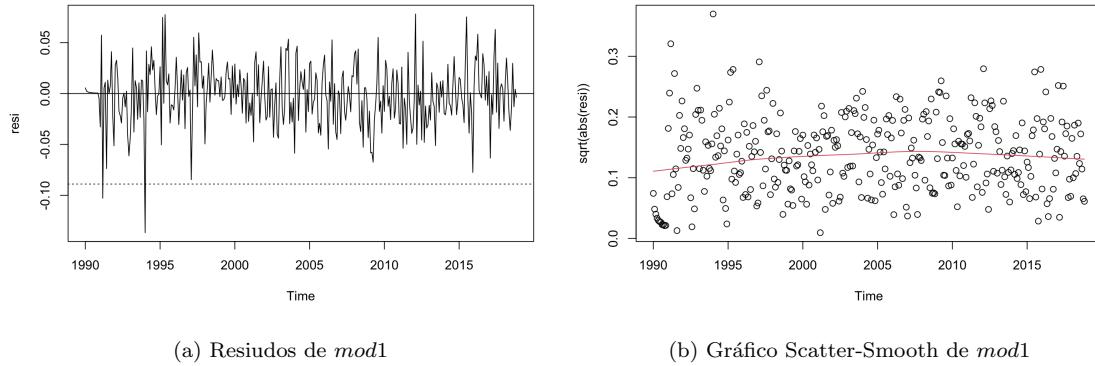


Figura 8: Análisis de la varianza residual de *mod1*

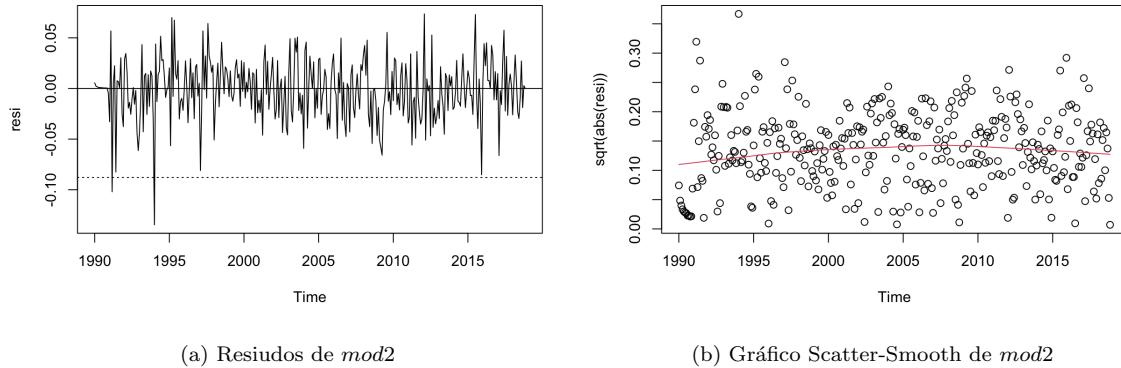
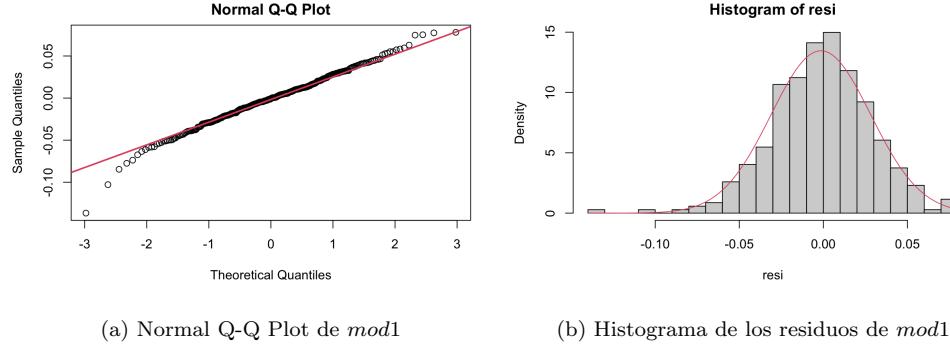
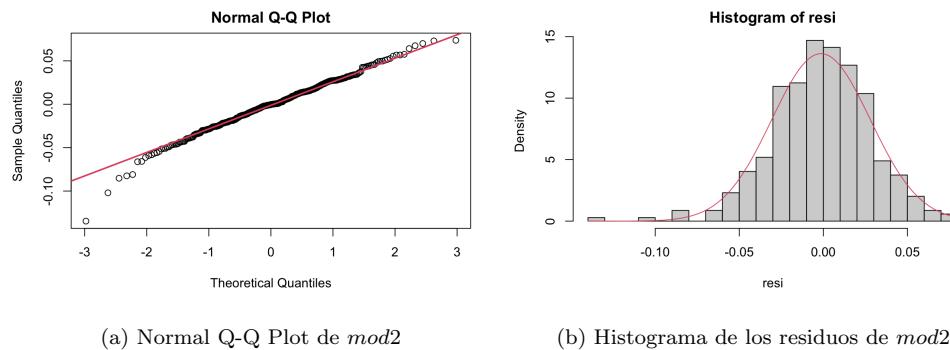


Figura 9: Análisis de la varianza residual de *mod2*

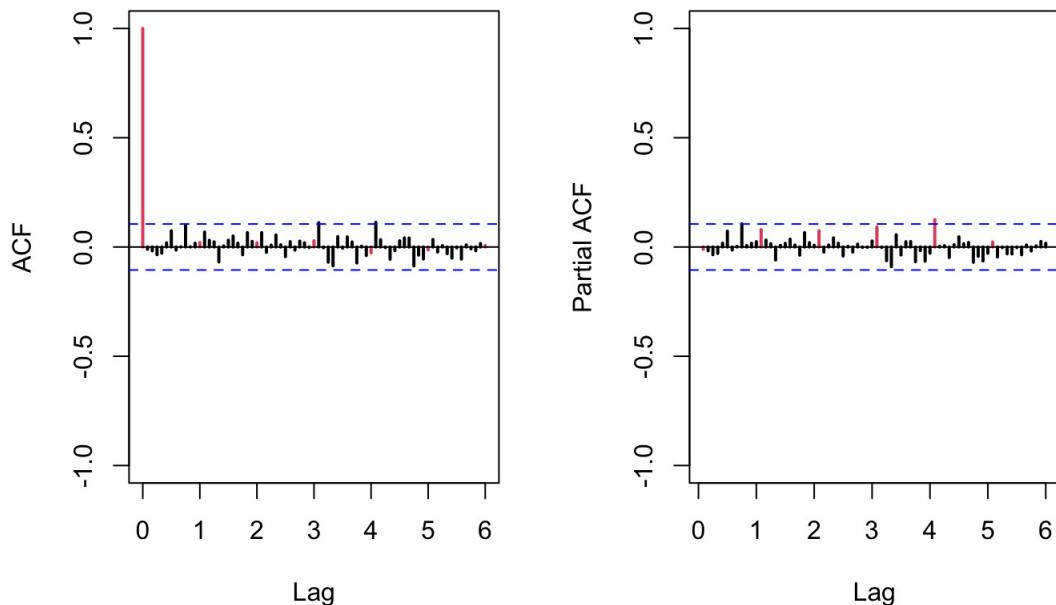
Al realizar una primera exploración de los residuos de los modelos, se observa que los gráficos de sus residuos son muy similares. Sin tener en cuenta el claro pico pronunciado, y un segundo pico que se sale de las bandas de confianza, que aparecen entre los años 1990 y 1995, la variabilidad es similar para toda la serie. Por lo tanto, se puede concluir que este no es un caso de varianza no constante producida por una alta volatilidad en los datos, sino de la presencia de observaciones atípicas, tanto en el modelo 1, como en el modelo 2.

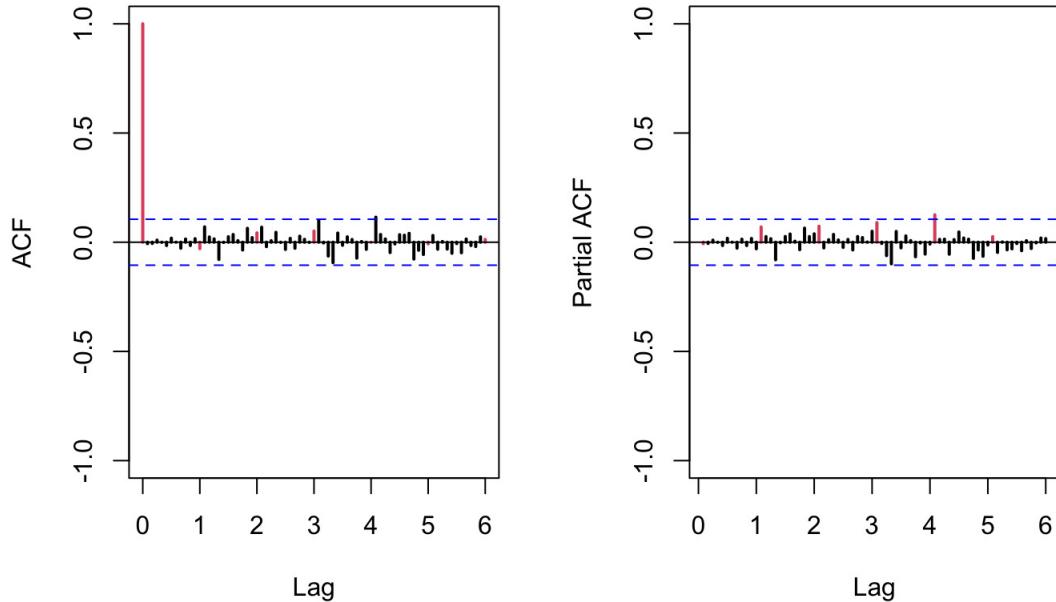
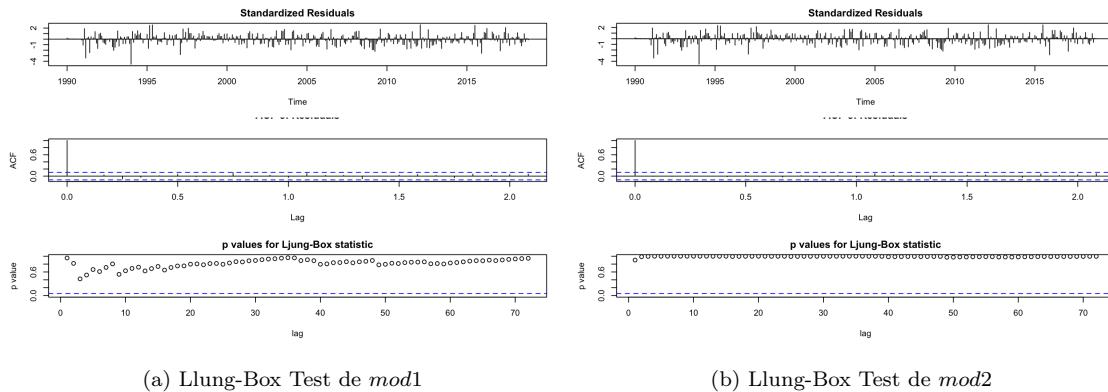
Se observa también como, globalmente la línea roja es horizontal en el Scatter-Smooth y los residuos se distribuyen de manera homogénea alrededor de ella, a excepción de algunas observaciones atípicas. Por lo tanto, la conclusión es la misma, la varianza de los residuos es constante en los dos casos.

Figura 10: Análisis de la normalidad residual de *mod1*Figura 11: Análisis de la normalidad residual de *mod2*

En cuanto al análisis de la normalidad de los residuos, se observa como en los extremos de las Figuras 10a y 11a aparecen colas que se alejan de la recta, pero no lo suficiente como para estar hablando de volatilidad. Son debidas a la ya mencionada presencia de atípicos. Se puede suponer, de momento, que se cumple la hipótesis de normalidad en ambos modelos. Todo y ser los histogramas únicamente gráficos ilustrativos, se puede ver que a priori se debería de cumplir la hipótesis de normalidad.

Al realizar el Shapiro-Wilks test para ambos modelos, nos encontramos un p-valor de 0.0006275 para *mod1* y de 0.002389 para *mod2*, inferior a 0.05. Por tanto, rechaza la hipótesis nula, concluyendo así que no se puede garantizar que se cumpla la hipótesis de normalidad para los residuos de los modelos.

Figura 12: ACF y PACF residual de *mod1*

Figura 13: ACF y PACF residual de *mod2*Figura 14: Análisis de la independencia residual de *mod2*

Se observa en las Figuras 12 y 13 como en ambos modelos todas las observaciones, a excepción de una en el PACF y un par en el ACF, caen dentro de las bandas de confianza. Los mencionados retardos que se salen de estas, lo hacen por muy poco y podrían considerarse valores aleatorios fruto del azar.

En cuanto al análisis de la independencia, para que se pueda hablar de independencia y ruido blanco, todas las observaciones en el último gráfico del test de Llung-Box deberían quedar por encima de la línea discontinua de color azul, que representa el p-valor de 0.05. Se puede ver como es el caso para ambos modelos, por tanto no se rechaza la hipótesis y concluimos que tanto el modelo 1 como el modelo 2 explican suficientemente bien nuestros datos.

En cuanto al estudio de la causalidad e invertibilidad de los modelos, debemos tener en cuenta que un modelo será causal si y solo si el módulo de todas las raíces de su polinomio característico respecto a la parte autoregresiva son mayores que 1. De igual manera, un modelo será invertible si y solo si el módulo de todas las raíces de su polinomio característico respecto a la parte de *moving average* son mayores que 1. Es el caso para ambos, tanto el modelo 1 como el modelo 2 son causales e invertibles.

Las medidas de adecuación que tendremos en cuenta serán el AIC y BIC:

	<i>mod1</i>	<i>mod2</i>
AIC	-1370.72	-1366.37
BIC	-1355.47	-1328.26

2.2 Estabilidad, capacidad de previsión y selección de modelo

Eliminando las últimas 12 observaciones de la serie y ajustando ambos modelos a esta, obtenemos los modelos cuyos coeficientes son los siguientes:

Coefficients:

	ma1	ma2	ma3	sma1
	-0.5868	-0.1588	0	-0.7951
s.e.	0.0532	0.0498	0	0.0400

mod1.1

Coefficients:

	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	sma1
	-0.5778	-0.4448	-0.4381	-0.3852	-0.2687	-0.1628	-0.1522	-0.1016	-0.7963
s.e.	0.0566	0.0648	0.0679	0.0703	0.0709	0.0683	0.0647	0.0567	0.0411

mod2.1

Realizando la comparación de estos coeficientes con los obtenidos en los anteriores modelos estimados, se observa como tanto que el signo es igual como que el valor es muy similar. Por tanto, podemos concluir que ambos modelos son estables.

	ma1	ma2	ma3	sma1
mod1	-0.5861043	-0.1635852	0	-0.7923814
mod1.1	-0.5867856	-0.1587501	0	-0.7950725

	ar1	ar2	ar3	ar4	ar5	ar6	ar7	ar8	sma1
mod2	-0.5786551	-0.4554162	-0.4505765	-0.3889024	-0.2732165	-0.1709281	-0.16755656	-0.1116351	-0.7913315
mod2.1	-0.5671875	-0.4309319	-0.4173346	-0.3511284	-0.2245491	-0.1202464	-0.09612503	0.0000000	-0.8011776

A continuación, a través de la función *accuracy()* de la librería *forecast* se evaluará la capacidad de previsión de ambos modelos, generando una predicción para los próximos 12 períodos a partir de los modelos ajustados con la serie recortada, y evaluando su precisión respecto a las observaciones originales. La salida es una tabla que contiene diferentes errores de predicción, tales como el MAPE o el MAE. Cuanto menor sean, mayor capacidad de previsión del modelo.

	ME	RMSE	MAE	MPE	MAPE	MASE
mod1.1	-0.004921900	0.0211000	0.01778297	-0.04891289	0.1761067	0.4274488
mod2.1	-0.008017934	0.02232939	0.01959548	-0.07959307	0.1940617	0.4710160

A través de las tablas se puede observar como todas las medidas de error tienen valores similares para los dos modelos, siendo en algunas muy ligeramente superior el primer modelo, así como a la inversa en otras. De todos modos, las medidas de error tienen valores bajos para ambos modelos, así que se puede afirmar que ambos tendrán buena capacidad de previsión. Son valores muy similares, la bondad de precisión no variará apenas de uno a otro.

El último paso en la validación de modelos es escoger con cuál de los dos ajustados se realizarán las predicciones. En primer lugar, respecto al análisis de residuos, se ha podido observar cómo ambos tienen las mismas características: varianza constante, independencia y ruido blanco y, en ninguno de ellos se cumple la hipótesis de normalidad de residuos. Respecto al estudio de la invertibilidad, causalidad y estabilidad de los modelos, también se ha podido concluir que comparten estas características. Ambos son invertibles, estables y causales. Además de todo el parecido en estas características, se ha podido ver en el último análisis como su capacidad de predicción es muy similar. Por tanto, ninguno de estos rasgos será completamente determinante a la hora de decidir entre dos modelos. En lo que sí hay diferencia entre ambos es en las medidas de adecuación. No es una diferencia abismal, pero teniendo en cuenta la importancia de estas, y que en todas las demás facetas son prácticamente iguales, es diferencial a la hora de escoger el primer modelo:

$$mod1 \rightarrow ARIMA(0, 1, 2)(0, 1, 1)_{12}$$

$$(1 - B)(1 - B^{12})X_t = (1 - 0.5861\theta_1 B - 0.1636\theta_2 B^2)(1 - 0.7924\Theta_1 B^{12})Z_t$$

2.2.1 Previsiones

Para el modelo ajustado sin las últimas 12 observaciones, obtenemos las predicciones puntuales, y el correspondiente intervalo de confianza al 95% para el último año, dado por $x_{pred} \pm 1.96se$. Se puede observar a continuación, como la predicción, graficada en rojo, se encuentra dentro del intervalo de confianza, graficado en azul, y se ajusta de una manera precisa a las observaciones originales, graficadas en negro.

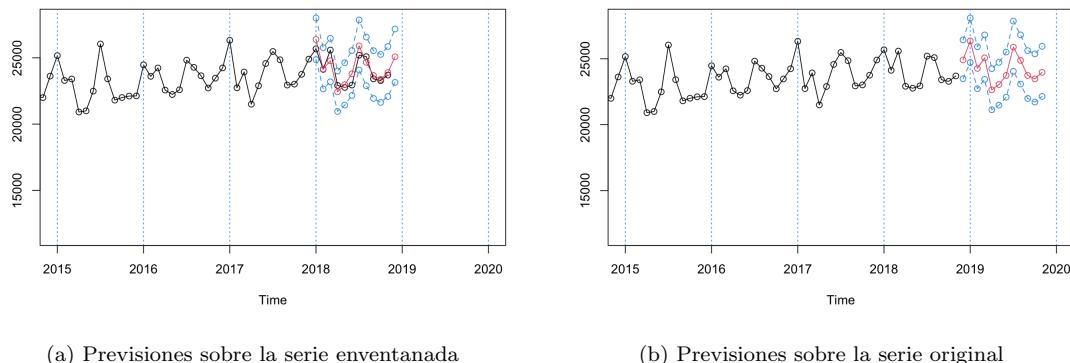


Figura 15: Predicción de la serie a un año vista a partir de *mod1*

	RMSE	MAE	RMSPE	MAPE	MEAN(ull-ll)
<i>mod1</i>	517.8877	434.0531	0.02119971	0.0178454	3441.743

2.3 Efectos de calendario

En las series temporales hay ciertas configuraciones en el mes que pueden afectar a las predicciones, son conocidas como efectos de calendario. Se considerarán dos casos principales: Semana Santa [Ide] y la configuración de los días de la semana [365].

El primero porque dependiendo del año, la Semana Santa puede caer en el mes de Marzo, en el mes de Abril, o en ambos, y eso podría no verse reflejado en las predicciones y que estas fallaran. Con tal de lidiar con este problema, se intentará cambiar la serie con tal de que la Semana Santa sea considerada la mitad de días para cada mes.

El segundo será también considerado porque según la proporción de días laborables y fines de semana

a lo largo de los meses de un año puede afectar también a las predicciones hechas sobre la serie. En este caso, se lidiará con el problema estableciendo la proporción de Trading Days/Weekends a 5/2 en todos los meses.

A continuación, se ajustará un modelo para cada efecto de calendario, además de otro con ambos efectos juntos. De esta manera, se podrá estudiar su efecto respecto al modelo principal ajustado anteriormente.

Modelo con efecto de calendario *Easter*: *modEa*

Coefficients:				
	ma1	ma2	sma1	vEa
	-0.5659	-0.1794	-0.7863	-0.0222
s.e.	0.0515	0.0482	0.0386	0.0063

Modelo con efecto de calendario *Trading Days*: *modTD*

Coefficients:				
	ma1	ma2	sma1	vTD
	-0.5718	-0.1727	-0.7877	0.0015
s.e.	0.0518	0.0485	0.0390	0.0004

Modelo con efecto de calendario conjunto: *modEC*

Coefficients:					
	ma1	ma2	sma1	vEa	vTD
	-0.5531	-0.1877	-0.7830	-0.0201	0.0014
s.e.	0.0513	0.0480	0.0391	0.0062	0.0004

	<i>modEa</i>	<i>modTD</i>	<i>modEC</i>
AIC	-1380.54	-1380.6	-1388.72
$\hat{\sigma}^2$	0.0008777	0.0008774	0.0008518

Después de ajustar los tres modelos con efectos de calendario, se puede observar como a cada cual disminuye tanto el AIC como la varianza estimada. En este caso, el efecto causado por los Trading Days es prácticamente idéntico al causado por la Semana Santa, pero el efecto provocado por los dos juntos es aun mayor que los dos anteriores por separado. En este último modelo el estadístico test de significancia de los dos parámetros correspondientes a los efectos de calendario tiene un valor de -3.23 para la Semana Santa y de 3.20 para los Días de Traspaso. Es decir, ambos son mayores que 2, en valor absoluto y, por tanto, significativos.

2.3.1 Análisis de Intervención

Una vez estudiados los efectos de calendario sobre el modelo, es conveniente estudiar también el efecto de alguna circunstancia concreta a lo largo de los años que comprende la serie que pudiera haber causado cambios en esta.

Entre 1990 y 2018 destacan dos factores que pudieran tener efecto sobre la misma:

- La liberalización del mercado eléctrico de 1998 [New], en búsqueda de una mayor competencia.
- La introducción del Plan de Acción de Energías Renovables en 2005 [Espb], el cual se introdujo con objetivo de aumentar la proporción de energías renovables en el consumo eléctrico del país.

Se ajustará un modelo añadiendo el efecto de estos, donde $vLib$ será el parámetro que contenga el efecto de la liberalización del mercado y $vPlan$ el parámetro que contenga el efecto causado por la aprobación del Plan de Energías Renovables. De esta manera podrá verse si el efecto causado en la serie es suficientemente significativo como para ser considerado en el modelo.

```
Coefficients:
ma1      ma2      sma1     vEa      vTD      vLib      vPlan
-0.5462  -0.1777  -0.7849  -0.0201  0.0014  -0.0152  -0.0243
s.e.      0.0523   0.0490   0.0387   0.0062  0.0004   0.0215   0.0209
sigma^2 estimated as 0.0008472: log likelihood = 701.26, aic = -1386.52
```

Después del ajuste realizado, se puede ver como, por un lado, el AIC de este último modelo es mayor que el del modelo que contiene solo efectos de calendario. Mientras que, por otro lado, los estadísticos del test de significancia para los parámetros añadidos tienen un valor de -0.71 para $vLib$ y de -1.16 para $vPlan$. Estos son menores que 2, en valor absoluto, por lo tanto no serán parámetros influyentes.

Pese a que después de un análisis exhaustivo de circunstancias que pudieran afectar al consumo eléctrico en España se hubieran encontrado los mencionados como factores principales, se concluye que estos no son significativamente distintos de 0 y, por tanto, no deben ser incluidos en el modelo. Su efecto asociado a la serie temporal no es diferencial a la hora de realizar buenas predicciones.

Una vez queda seleccionado qué modelo se ajusta mejor a todos estos fenómenos, podemos ver una comparativa de los valores de la serie original con los valores de la serie que contiene estos efectos [Figura 16], a partir de linealizar la serie, extrayéndole el efecto causado por estas variables.

$$X_{lin_t} = X_t - \omega_{TD} TD_t - \omega_{Ea} Ea_t$$

Comparativa del logaritmo de la serie con y sin efectos de calendario aplicados

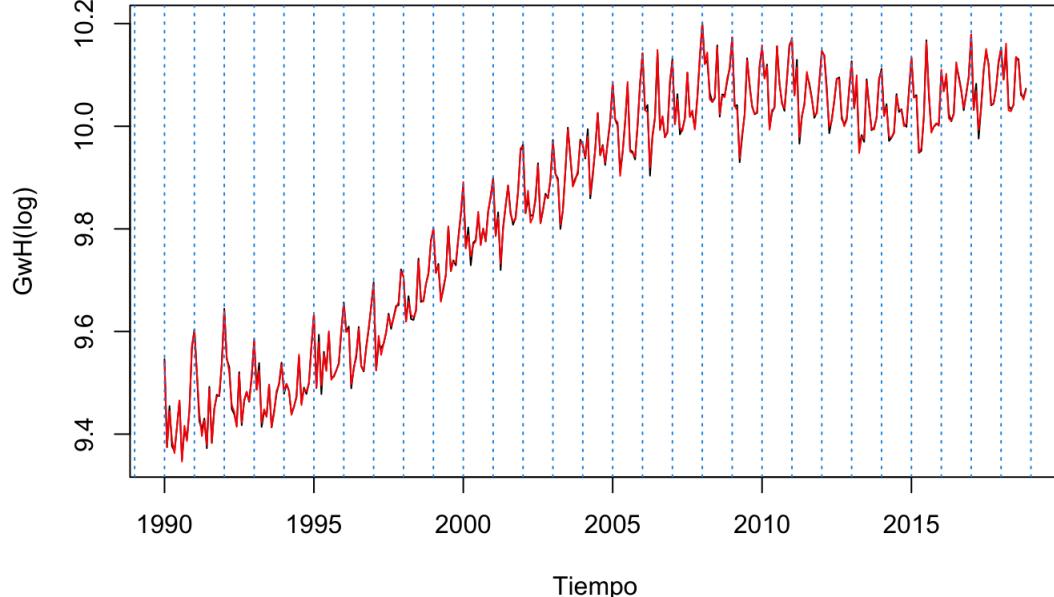


Figura 16: Comparación gráfica entre X_t y X_{lin_t}

Una vez visto que los efectos de calendario no cambian la serie, y teniendo el modelo estimado con estos efectos, debemos comprobar que los residuos siguen cumpliendo las hipótesis, y que sigue siendo causal e invertible. Es decir, es necesario validar este nuevo modelo ajustado.

2.3.2 Validación del modelo

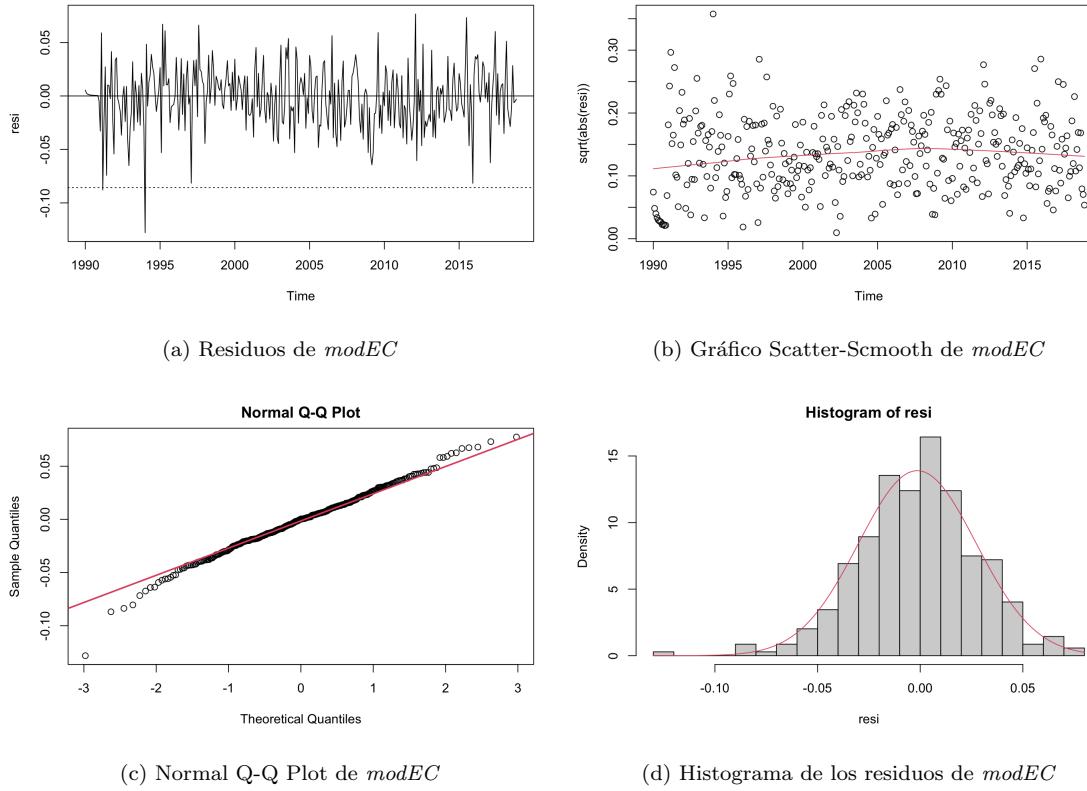


Figura 17: Representación gráfica de los residuos de *modEC*

El análisis gráfico de los residuos es idéntico al realizado en secciones anteriores para el modelo previo. Las Figuras 17a y 17b muestran una varianza constante, mientras que las Figuras 17c y 17d indican que debería cumplirse la hipótesis de normalidad. No obstante, si realizamos el Shapiro-Wilks Test, el p-valor es de 0.02065, menor a 0.05. Por tanto, en este caso tampoco se puede asegurar normalidad en los residuos del modelo.

Tal como muestran las Figuras a continuación, también podemos hablar de independencia residual y ruido blanco en este caso.

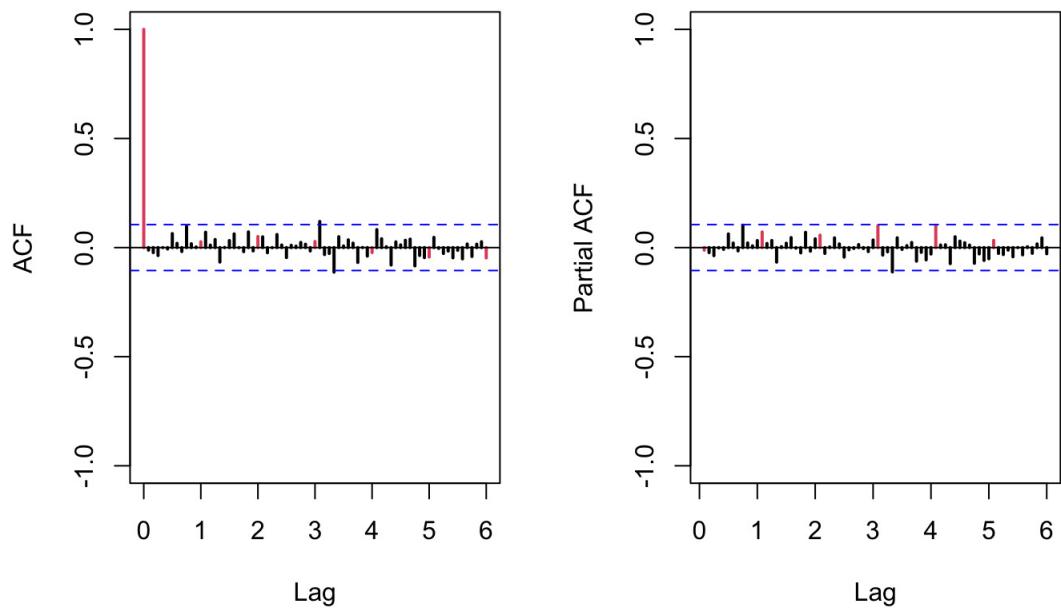
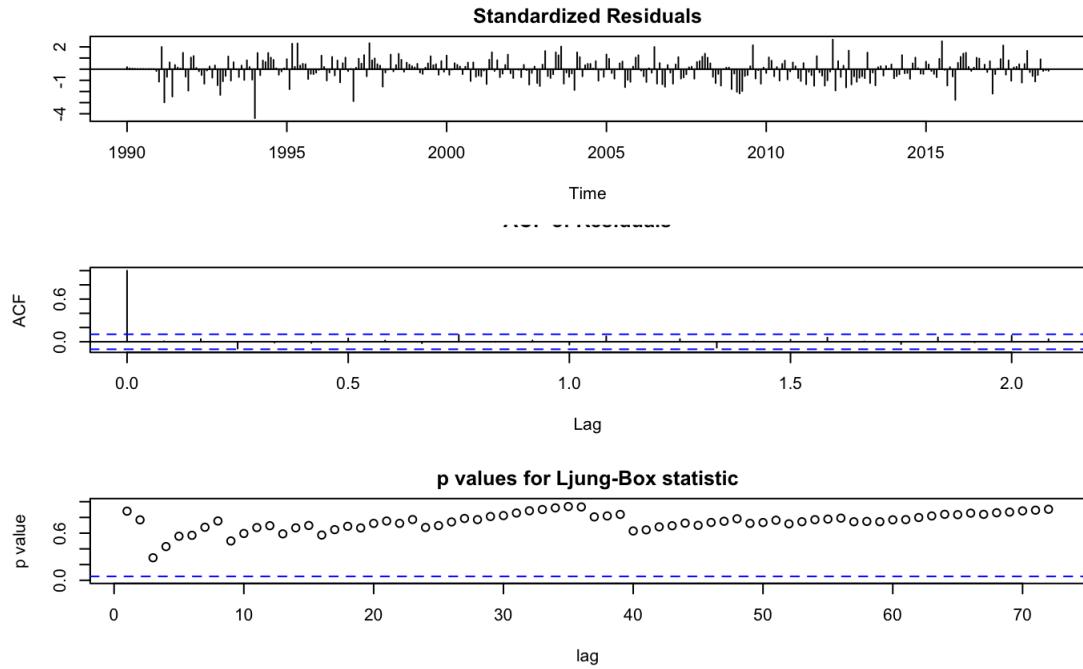
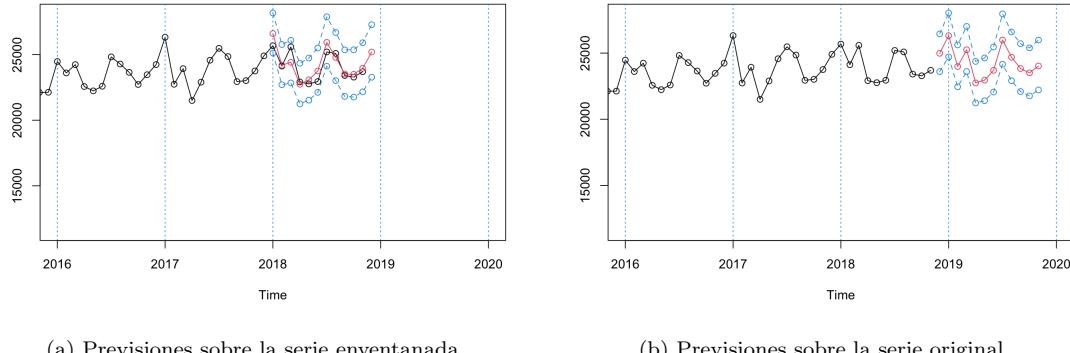


Figura 18: ACF y PACF de los residuos de *modEC*

Figura 19: Test de Llung-Box para *modEC*

Todos los módulos de las raíces de los polinomios característicos de este modelo, al igual que los del anterior, son mayores que 1. Por lo tanto, *modEC* es causal e invertible.

Como se está comprobando, el modelo tiene las mismas propiedades tanto para la serie original como para la linealizada. Por consecuencia, también compartirá la estabilidad y la capacidad de previsión al enventanar la serie sin la última observación. De modo que se pasará directamente a las predicciones, tal como se ha hecho anteriormente.

Figura 20: Predicción de la serie a un año vista a partir de *modEC*

La predicción de la serie modelada con el modelo sobre la serie recortada, representada en rojo, se encuentra dentro de su intervalo de confianza, representado en azul, y se ajusta de una manera muy precisa al gráfico de la serie original, representado en negro. En el caso de la Figura 20b no se puede comprobar la precisión de esta predicción, pues no tenemos datos para ello, pero teniendo en cuenta el buen ajuste del gráfico anterior, se puede confiar en que la predicción será precisa.

	RMSE	MAE	RMSPE	MAPE	MEAN(ull-l)
<i>modEC</i>	585.6177	461.5682	0.02357383	0.01882707	3384.447

Se puede comprobar como, comparando las medidas de error con el modelo inicial sin efectos de calendario, las primeras eran más pequeñas que las obtenidas ahora, indicador de que el error del modelo es mayor cuando incluye los efectos asociados al calendario.

2.4 Tratamiento de atípicos

En estadística y análisis de datos, un outlier es un punto que se encuentra significativamente alejado del resto de los datos en una muestra. Es importante tratar los outliers en el análisis de series temporales porque pueden tener un impacto significativo en los resultados de cualquier modelo que se ajuste a los datos. Los outliers pueden distorsionar las estimaciones de los parámetros, aumentar la varianza del modelo y afectar la precisión de las predicciones.

Podemos encontrar tres tipos de outliers:

- **AO** (Additive Outlier): Solo afecta a un período.
- **TC** (Transitory Change): Afecta a un período y su efecto decrece en los siguientes.
- **LS** (Level Shift): Afecta a un período y su efecto se mantiene en los siguientes.

A partir de la detección automática de atípicos, obtenemos los siguientes:

	Obs	type_detected	W_coeff	ABS_L_Ratio	Fecha	PercVar
3	14	AO	0.07804566	3.436857	Feb 1991	108.11720
14	18	AO	-0.05667709	2.889440	Jun 1991	94.48991
10	25	TC	0.06288620	3.041845	Ene 1992	106.49056
1	49	AO	-0.11658634	4.947707	Ene 1994	88.99533
13	63	TC	0.05825792	2.925502	Mar 1995	105.99884
8	86	TC	-0.06399438	3.015613	Feb 1997	93.80103
12	97	TC	-0.05948991	2.951189	Ene 1998	94.22450
17	164	AO	0.05443965	2.874542	Ago 2003	105.59487
7	169	AO	-0.06760858	3.154946	Ene 2004	93.46262
6	199	AO	-0.06747490	3.104201	Jul 2006	106.98034
18	215	LS	0.04308703	2.861133	Nov 2007	104.40288
11	224	AO	-0.06122103	3.008113	Ago 2008	94.06153
5	230	LS	-0.05687798	3.207436	Feb 2009	94.47093
2	266	AO	0.08766443	3.785555	Feb 2012	109.16217
9	307	AO	0.06385094	3.019785	Jul 2015	106.59335
4	312	AO	-0.07564278	3.314606	Dic 2015	92.71473
15	326	TC	-0.05764212	2.878505	Feb 2017	94.39877
16	330	AO	0.05652395	2.851037	Jun 2017	105.81520

Una vez detectados las observaciones atípicas, el tratamiento de estas se basa en linealizar la serie. Para poder comparar ambas y visualizar el efecto de los atípicos, lo haremos gráficamente a continuación.

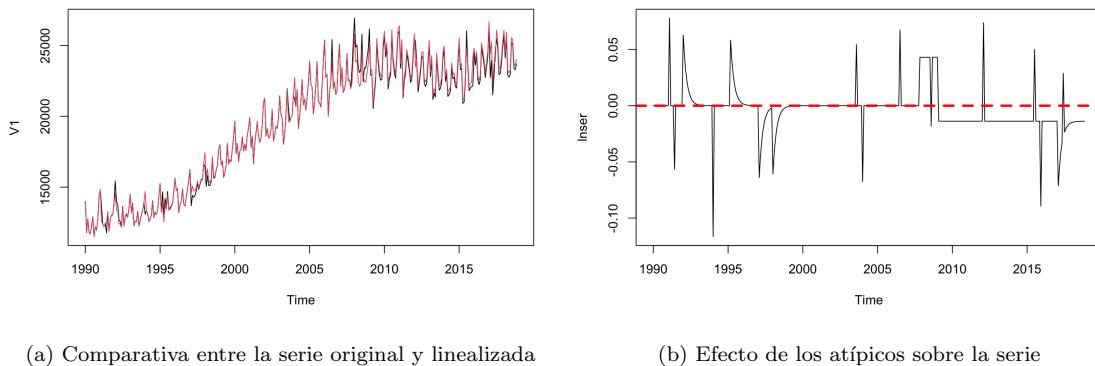
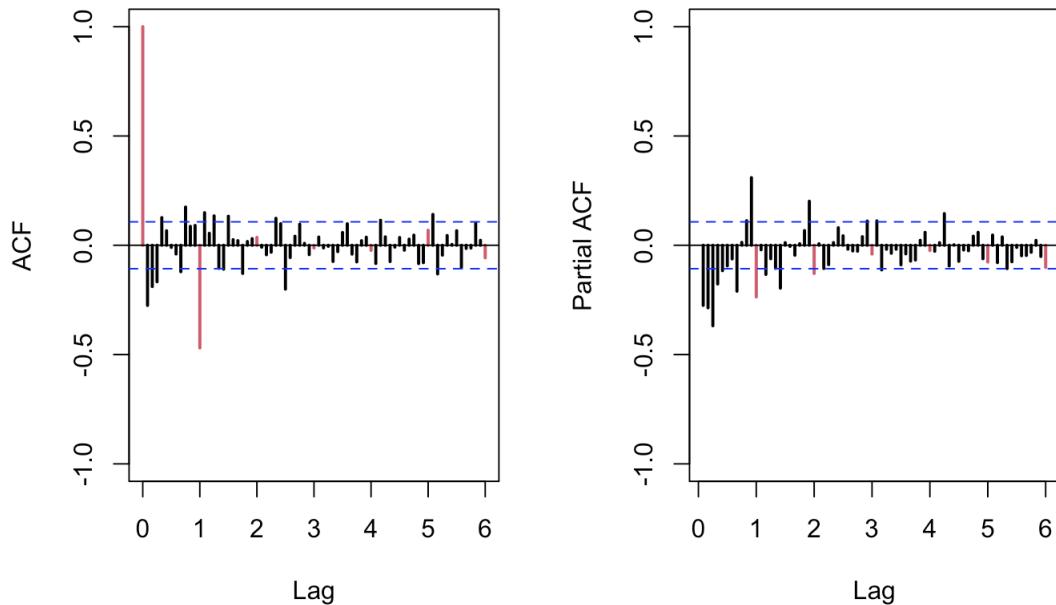


Figura 21: Comparación entre la serie y el efecto sobre ella de los outliers

En la Figura 21b puede verse como la mayoría de las observaciones atípicas son picos triangulares, correspondientes a outliers del tipo 'AO'. Si nos fijamos en las fechas asociadas a estas observaciones podemos encontrar que todas tienen algo en común. Son fechas en las que, comúnmente, se llega a las temperaturas máximas y mínimas a lo largo de un año. Es decir, los meses centrales de las estaciones extremas. La mayoría de los outliers se concentran en los meses de Enero, Febrero y Julio. Los picos más redondeados se asocian a los cambios transitorios, mientras que los picos rectangulares se asocian a outliers de tipo LS. De estos últimos encontramos dos, uno en Noviembre del año 2007 y otro en Febrero del año 2009, ambos picos más altos.

2.4.1 Identificación y estimación

A partir de la serie linealizada, deberemos aplicar de nuevo la metodología Box-Jenkins, de igual manera que con la serie original, empezando con el análisis del ACF y PACF de la serie.

Figura 22: ACF y PACF de X_{lin_t}

Por la parte estacional, se puede proponer un MA(1) estacional o se puede considerar un AR(2) estacional. Por la parte regular, puede considerarse un MA(4) regular o bien un AR(5) regular.

De los cuatro posibles modelos que se podrían proponer, el que menor AIC tiene y, por tanto, mejor

adecuación a los datos, es el formado por un MA(1) estacional y un MA(4) regular. Por lo tanto, estimamos un $ARIMA(0, 1, 4)(0, 1, 1)_{12}$, *modEClin*, con el cual obtenemos los siguientes coeficientes:

Coefficients:

	ma1	ma2	ma3	ma4	sma1	vEa	vTD
	-0.5760	-0.2914	0	0.1735	-0.714	-0.0197	1e-03
s.e.	0.0531	0.0584	0	0.0478	0.041	0.0045	3e-04

`sigma^2 estimated as 0.0004873: log likelihood = 795.02, aic = -1576.03`

Según el test de significancia para cada coeficiente, todos son significativos en el modelo ajustado, excepto *ma3*, por tanto este es fijado a 0. A continuación, procedemos a validarla de igual manera que hemos hecho en secciones anteriores.

2.4.2 Validación

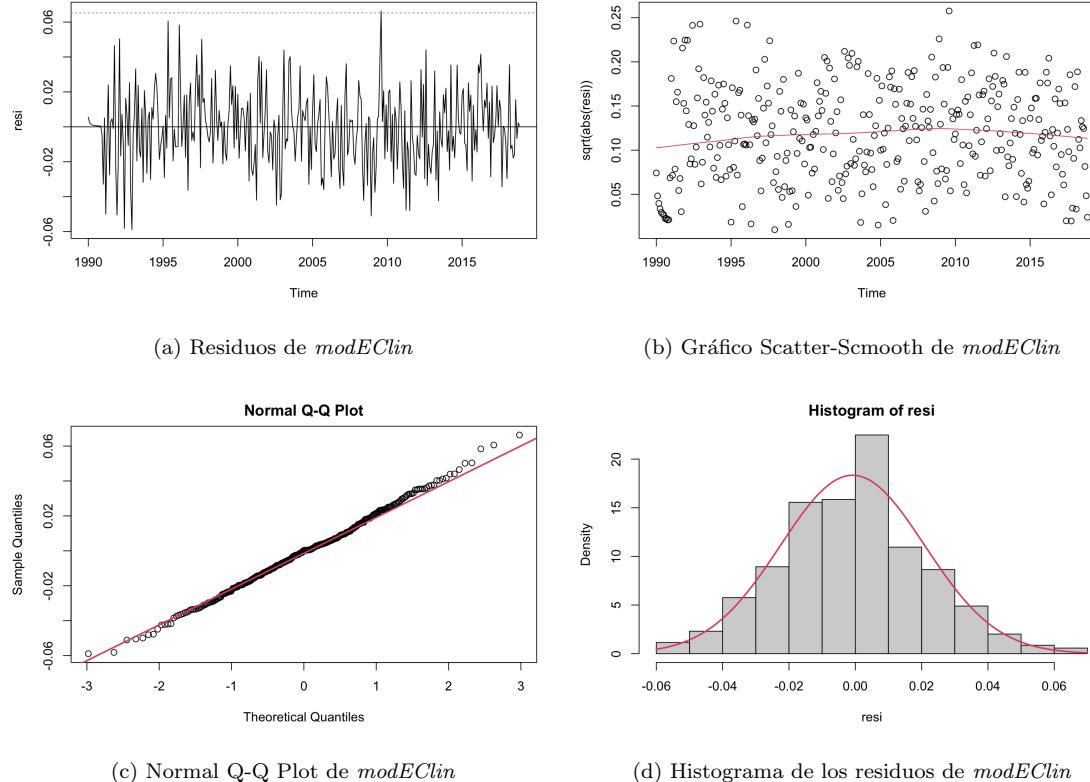
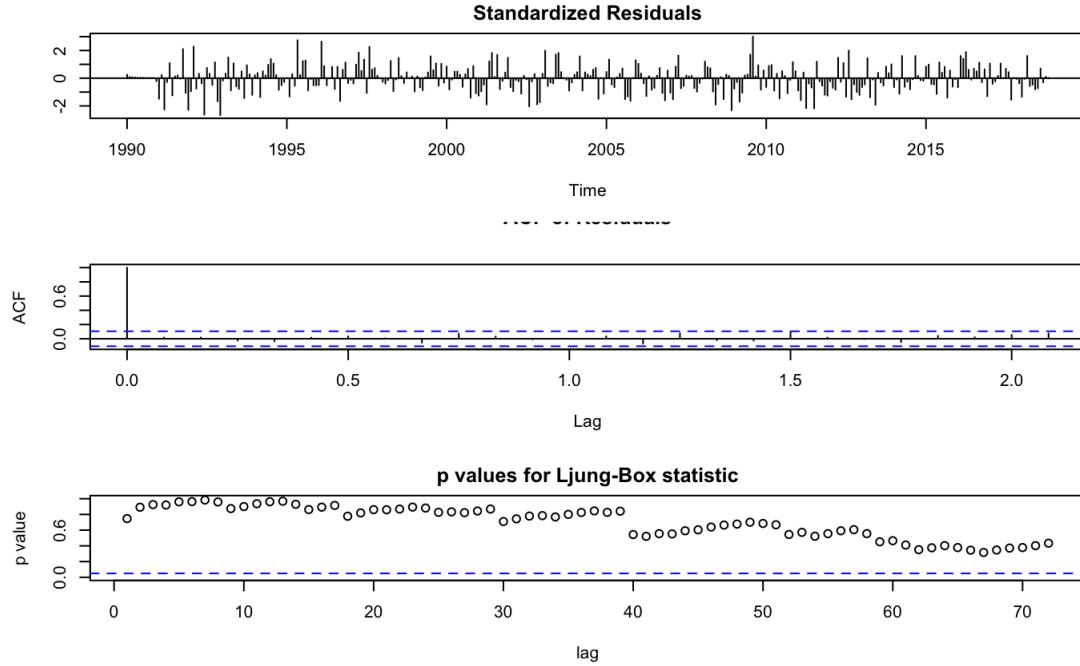
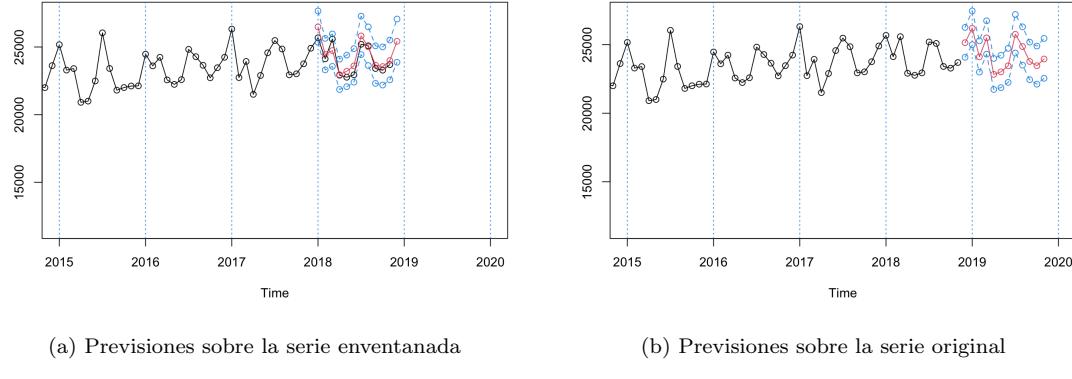


Figura 23: Representación gráfica de los residuos de *modEClin*

Figura 24: Test de Llung-Box para *modEClin*

Tal como se puede observar en los gráficos de la Figura 23 y en el test de Llung-Box [Figura 24], las hipótesis de varianza constante e independencia se cumplen. Si realizamos el test de Shapiro-Wilks, el p-valor es de 0.9009, por encima de 0.05. Por tanto, también cumple la de normalidad residual. El modelo es causal e invertible, y tiene un AIC de -1540.034 y un BIC de -1444.068. Estos valores son más pequeños que cualquiera de los obtenidos al ajustar anteriormente.

2.4.3 Previsiones

Figura 25: Predicción de la serie a un año vista a partir de *modEClin*

	RMSE	MAE	RMSPE	MAPE	MEAN(ull-ll)
<i>modEClin</i>	729.864	681.6363	0.03018659	0.02830442	2541.563

A partir de la buena precisión de las predicciones y la disminución en las medidas de error respecto a las anteriores, se puede pensar, en un principio, en este modelo como el que mejor previsiones proporciona de todos los planteados.

3 CONCLUSIONES

	par	σ^2	AIC	BIC	RMSE	MAE	RMSPE	MAPE	CIml
<i>ARIMA</i>	3	0.0009084550	-1370.715	-1355.471	517.8877	434.0531	0.02119971	0.01784540	3441.743
<i>ARIMA+EC</i>	5	0.0008518283	-1388.719	-1365.852	585.6177	461.5682	0.02357383	0.01882707	3384.447
<i>ARIMA+EC+OutTreat</i>	25	0.0004872871	-1540.034	-1444.068	729.8640	681.6363	0.03018659	0.02830442	2541.563

En conclusión, para estudiar el consumo eléctrico en España se ha optado por utilizar un modelo $ARIMA(0, 1, 2)(0, 1, 1)_{12}$. A la hora de lidiar con los efectos de calendario, el modelo propuesto es también el mismo. Este segundo mejora al anterior, tanto aumentando la capacidad de previsión como ajustándose mejor a los datos. Sin embargo, se han podido encontrar varios atípicos, principalmente en los meses centrales de estaciones extremas en determinados años, en los cuales la demanda eléctrica se vio disparada. A la hora de proponer un modelo en el cual estas se trataran, se ha optado por un $ARIMA(0, 1, 4)(0, 1, 1)_{12}$.

Ambos son modelos válidos que dan buenas previsiones. Aún así, tal y como se puede ver en la tabla, este último se adecua mejor a los datos, proporcionando un mejor AIC y reduciendo a la mitad la varianza estimada, aunque comporta también unas medidas de error mayores.

Esto indica que, por un lado, al actuar sobre la serie linealizada, este último modelo trata todas las observaciones. Esa es la razón por la cual las medidas de adecuación a los datos se ven mejoradas. Mientras que, por otro lado, el modelo *ARIMA+EC* proporciona unas previsiones más precisas y exactas. Por lo tanto, si el objetivo es predecir el consumo eléctrico en España para un año concreto a partir de la serie temporal, el modelo siguiente es el que mejor pronósticos concede:

$$(1 - B)(1 - B^{12})(X_t + 0.0201TD_t - 0.0014Ea_t) = (1 - 0.5531\theta_1B - 0.1877\theta_2B^2)(1 - 0.7830\Theta_1B^{12})Z_t$$

Bibliografía

- [Sán23] Josep Anton Sánchez. *Apuntes de la asignatura Análisis de Datos*. Atenea, 2023.
- [365] Calendario 365.es. *Calendario 1990*. URL: <https://www.calendario-365.es/calendario-1990.html>.
- [Ene] Plena Energía. *Red Eléctrica de España: Qué es, funciones, precio de la luz y más*. URL: <https://www.plena-energia.com/post/red-electrica-espana>. (23 de Marzo de 2022).
- [Espa] Gobierno de España. Ministerio de Industria Comercio y Turismo. *Series Estadísticas*. URL: <https://sedeaplicaciones.minetur.gob.es/Badase/BadasiUI/lstSeriesInformesPostBack.aspx>.
- [Espb] Gobierno de España. Ministerio para la Transición Ecológica y el Reto Demográfico. *Plan de Energías Renovables 2005-2010*. URL: <https://energia.gob.es/desarrollo/EnergiaRenovable/Plan/Paginas/planRenovables.aspx>. (21 de Julio de 2005).
- [For] Aleasoft Energy Forecasting. *Historia del ‘pool’ (Parte II): de 2008 a 2014, la crisis económica cambia por completo el mercado eléctrico*. URL: <https://elperiodicodelaenergia.com/historia-del-pool-parte-ii-de-2008-a-2014-la-crisis-economica-cambia-por-completo-el-mercado-electrico/>. (24 de Octubre de 2019).
- [Ide] Ideal. *Fechas Semana Santa 1990*. URL: <https://calendarios.ideal.es/semana-santa/1990>.
- [New] B2B News. *La Liberalización Del Mercado Eléctrico Español*. URL: <https://www.totalenergies.es/es/pymes/blog/liberalizacion-mercado-electrico-espanol>. (20 de Noviembre de 2018).
- [Red] Redeia. *Red Eléctrica de España*. URL: <https://www.ree.es/es/conocenos/el-grupo>.