INDICE

1. Introducción 2
2. Mercado Eléctrico Español y Demanda de Electricidad 3
3. Modelos De Predicción. Conceptos Teóricos 13
4. Estudio Aplicado 19
4.1. Modelos de predicción sin variables exógenas 19
4.1.1. Método NAIVE 19
4.1.2. Modelo ARIMA: el modelo autorregresivo integrado de media móvil 22
4.1.2.1. ARIMA: un solo modelo 22
4.1.2.2. ARIMA: siete modelos 30
4.1.3. Modelo PLRM: Modelo de Regresión Parcialmente Lineal 50
4.1.3.1. PLRM: un solo modelo 51
4.1.3.2. PLRM: siete modelos 53
4.2. Modelos de predicción con variables exógenas 54
4.2.1. Modelo ARIMA con temperatura 57
4.2.1.1. ARIMA: un solo modelo con temperatura 57
4.2.1.2. ARIMA: siete modelos con temperatura 61
4.2.2. Modelo PLRM con Temperatura 75
4.2.2.1. PLRM: un solo modelo con temperatura 76
4.2.2.2. PLRM: siete modelos con temperatura 77
4.3. Comparación de los Modelos de Predicción 80
5. Conclusiones y Ampliación 82
6. Bibliografía 84
Apéndice 86
Código R utilizado 88

1. INTRODUCCIÓN

A partir de la liberalización que existe del mercado eléctrico a finales del siglo XX, muchos agentes se interesaron por la previsión de la demanda de electricidad. Por ello, estos últimos años se ha incrementado la relevancia de este tema y se ha extendido su estudio por parte de los agentes de la industria eléctrica.

La previsión de la demanda eléctrica a corto plazo es un problema económico importante porque la electricidad no puede ser almacenada y por tanto la demanda debe ser satisfecha de forma instantánea. Esto implica que los productores deben anticiparse a la demanda futura de forma precisa, con el objetivo de cumplir con ella y evitar además la sobreproducción. Por este motivo, desde un punto de vista práctico, las predicciones de demanda a corto plazo son una tarea muy importante para las empresas eléctricas a la hora de gestionar la producción, la transmisión y la distribución de la electricidad de la forma más segura y eficiente.

Como vemos la predicción del consumo eléctrico es la clave en la evaluación de numerosos proyectos, sin embargo la evidencia nos muestra que a la hora de predecir se comenten errores y es nuestro deber intentar que éstos sean mínimos, ya que un pequeño incremento en el porcentaje de error de una predicción, puede aumentar los costes anuales de una empresa en millones de euros.

En este trabajo nosotros vamos a utilizar el consumo eléctrico en 365 días consecutivos para predecir la demanda del día siguiente. Para realizar las predicciones utilizaremos distintos métodos como el Naive, el Modelo de regresión parcialmente lineal o el Modelo autorregresivo integrado de media móvil. Dentro de estos modelos tendremos en cuenta los ciclos estacionales que existan en nuestra serie de datos, así como el comportamiento de algunos días especiales tales como festivos o períodos vacacionales. Además al final de nuestro análisis observaremos cómo influye la incorporación de una variable exógena a los modelos, es decir, vamos a ver si el efecto de la temperatura en el consumo de electricidad en estos años es relevante, y por consiguiente, ver si nuestros modelos mejoran o no al incorporarle esta variable a la hora de predecir.

El objetivo de nuestro trabajo es realizar predicciones de la demanda eléctrica para el año 2012 con los modelos arriba citados, para posteriormente poder comparar los resultados obtenidos con cada uno de ellos y observar cuál es el mejor. De esta forma, en predicciones futuras de la demanda eléctrica a corto plazo utilizaremos el modelo que presente menores errores de predicción y evitaremos incrementos en los costes de las empresas eléctricas.

El documento está dividido en 6 epígrafes principalmente, en primer lugar tenemos una breve introducción; Luego, en el segundo epígrafe, damos una explicación a grandes rasgos sobre el mercado eléctrico español y cómo se comporta la demanda de electricidad en nuestro país, fundamentalmente en los días especiales. En el epígrafe siguiente, el tercero, comenzamos a hablar de los principales modelos que suelen emplearse a la hora de hacer predicciones, explicando en mayor profundidad los modelos que nosotros utilizaremos en nuestro posterior análisis. En el epígrafe cuatro presentamos el análisis aplicado a nuestros datos, en él se explican los pasos a seguir para realizar las predicciones con cada modelo, posteriormente se hallan los errores cometidos, comparando los valores reales del consumo eléctrico con las predicciones realizadas, y finalmente hacemos la comparación de los modelos observando los errores cometidos en cada uno de ellos. En los últimos dos epígrafes, 5 y 6, hacemos una presentación de las conclusiones fundamentales obtenidas en este estudio y finalmente expondremos la bibliografía consultada para la realización de este trabajo.

2. MERCADO ELÉCTRICO ESPAÑOL Y DEMANDA DE ELECTRICIDAD

Hasta el año 1997, el sistema de electricidad en España era un sistema regulado por el Gobierno, éste fijaba el precio de la electricidad, que retribuía el total de los costes en que incurrían las compañías eléctricas.

En 1997 se produjo una liberalización en el sector eléctrico, que dio lugar a que se estableciera lo que conocemos hoy en día como mercado eléctrico español, y que está formado por el conjunto de mercados donde se negocia la compra y venta de energía eléctrica. La liberalización del sector eléctrico se basa en la teoría de que la división vertical de actividades y su reglamentación específica pueden conseguir introducir competitividad y aumentar la eficiencia de un sector; lo que dio como resultado que el sector eléctrico quedase finalmente dividido en generación, transporte, distribución y comercialización.

La estructura legal y comercial de nuestro sistema de electricidad se basa en dos tipos de actividades, las que son parcialmente liberalizadas (generación y comercialización) y las reguladas (transporte y distribución). Pero la situación de los territorios que se encuentran fuera de la península es diferente, siendo la generación una actividad regulada, mientras que la comercialización es una actividad libre. Los comercializadores adquieren su energía al precio del mercado peninsular.

Las compañías comercializadoras que tenemos en el mercado eléctrico español por excelencia son: Iberdrola, Gas Natural Fenosa, Endesa, EDP y E.ON y reúnen sobre el 90% de las ventas a clientes finales y sobre un 60% de ventas en el mercado mayorista.

Las actividades libres pueden ser llevadas a cabo por cualquier agente como una actividad comercial más, sin embargo las actividades reguladas, se crean de la existencia de un monopolio natural y requieren una autorización y supervisión administrativas determinadas.

Lo que conocemos generalmente como mercado eléctrico se refiere a actividades libres y se constituye por dos sectores, el mercado de electricidad minorista y el mayorista.

En el mercado minorista, la mayoría de los consumidores adquieren su energía a través de empresas comercializadoras, que reflejan en las facturas de éstos, dos conceptos. Por un lado tenemos el coste de la energía, que incluye además de éste, el coste de los servicios de ajuste, pagos por capacidad, etc. Y por otro lado los costes regulados, es decir, la tarifa de acceso, en la que se incluyen los costes de las redes de distribución y transporte, así como los subsidios a energías renovables entre otros, es decir, cubren los costes establecidos por la administración. La competencia entonces en el mercado minorista se limita al primer concepto de coste de la energía, ya que la tarifa de acceso viene determinada por la administración.

El mercado mayorista abarca toda la península ibérica, integrando a España y Portugal en el mismo mercado eléctrico mayorista, que conocemos como MIBEL y que a si mismo se denomina *pool*. Dentro de todos los mercados que lo forman destaca el mercado gestionado por OMEL (Operador del Mercado de Electricidad).

Cada día OMEL efectúa la subasta para las 24h del siguiente día, entrando en ella las energías más baratas y siguiendo un orden, hasta que se cubre toda la demanda. Las primeras en acceder son las centrales nucleares por tener una energía poco flexible, y las energías renovables ya que la normativa promueve su crecimiento y desarrollo. Ambas fuentes de energía se ofrecen a precio cero para darles preferencia, y posteriormente les siguen las

energías más caras como el gas y el carbón. La última que cubre la demanda es la que marca el precio marginal de la energía en ese día, pero todas son remuneradas a ese mismo precio.

En cuanto a la demanda de electricidad, sabemos que es un indicador del estado de salud de una economía y tanto en 2011 como en 2012 tal como era de esperar por la mala situación de España, el consumo de electricidad en nuestro país descendió con respecto a los años anteriores. Tenemos que tener en cuenta además, que los precios finales para los consumidores de electricidad españoles, domésticos-residenciales, registran puestos entre los más elevados del ránking europeo. Este hecho también es importante para explicar la reducción del consumo eléctrico en estos años.

Para explicar mejor el comportamiento de la demanda eléctrica en España vamos a hacer un análisis más detallado de la misma. Para ello, cogemos una serie de datos del consumo eléctrico en España de los años 2011 y 2012 que fueron previstos por el OMIE, Operador del Mercado Ibérico de Electricidad. Este Operador del Mercado gestiona de forma integrada los mercados tanto diarios como intradiarios para toda la Península Ibérica y funciona de la misma manera que otros mercados europeos. La participación en el mercado se hace a través de una plataforma electrónica de internet, para que puedan participar un gran número de agentes y se puedan llevar a cabo numerosas ofertas de compra y venta de electricidad, en un tiempo reducido. También se encarga de realizar la facturación y liquidación de la energía comprada y vendida en los mercados.

La serie de datos con la que vamos a trabajar, abarca dos años como dijimos, con observaciones de la demanda de electricidad en España de cada día, desde el 1 de enero de 2011, hasta el 31 de diciembre de 2012. Es una serie completa, ya que consta de 731 observaciones tal como se muestra en la Figura 1 y no tiene ninguna observación perdida.

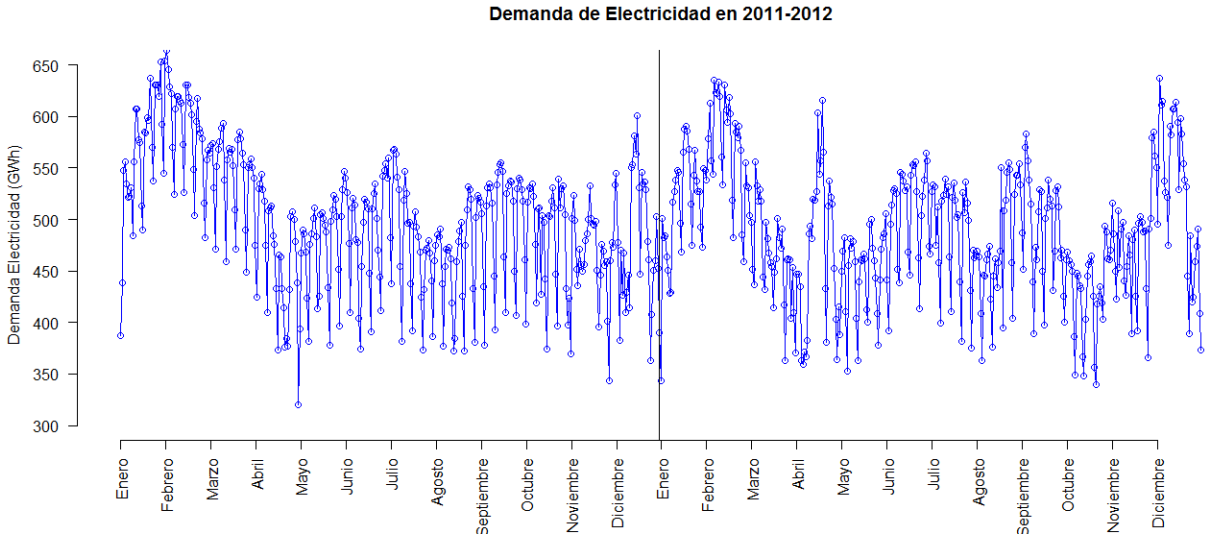

Figura 1. Demanda de Electricidad (2011 - 2012)

Tal como indican numerosas publicaciones a lo largo de los años, existen unas características comunes en las series de demanda como son, la tendencia, la estacionalidad, los efectos de los días especiales o las variables meteorológicas. Esto lo reflejan por ejemplo publicaciones como la de Espasa, Revuelta, y Cancelo (1996) que estudian series de tiempo de índole económica, centrándose en este caso en una serie de tiempo diaria.

Observando la Figura 1 desde un punto de vista general, se aprecia que la demanda se eleva tanto en las primeras semanas del año como en las finales, es decir, parece que en los meses de invierno se consume mayor cantidad de electricidad que en los meses de verano, por tanto, se puede afirmar que existe estacionalidad anual, aunque no tenemos suficiente historia como para introducirla en el modelo, por tanto no vamos a tenerla en cuenta.

Para profundizar un poco más en el análisis de nuestros datos y poder explicar cuáles de las características comunes de las series de demanda se cumplen durante el 2011 y el 2012 en el consumo de electricidad en España, vamos a representar la demanda diaria durante el año 2011 en la Figura 2.

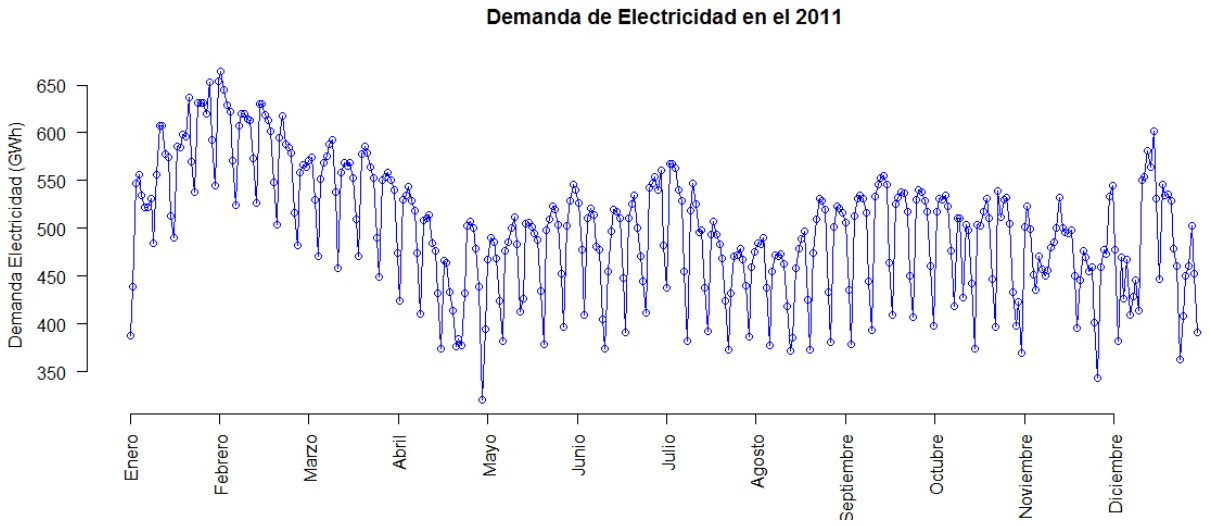

Figura 2. Demanda de Electricidad en el 2011 (GWh)

Observando la Figura 2 se puede ver la característica de la que hablábamos, que es la estacionalidad. En nuestro caso la estacionalidad es semanal, lo que quiere decir que la demanda de electricidad independientemente de la época del año en la que estemos, muestra patrones similares cada semana una vez eliminada la tendencia. Observamos que, mientras de lunes a jueves la demanda se mantiene más o menos constante, el viernes ésta comienza a caer y el sábado siempre sufre una fuerte caída, aunque ésta siempre es menor que la que sufre el domingo, que es el momento de la semana donde se alcanzan los mínimos valores de consumo de electricidad.

Pero bien, qué podemos decir de las estaciones, ¿afectan a nuestro consumo de electricidad?, pues para verlo vamos a ver cómo se comportan nuestros datos en dos meses de invierno y en dos meses de verano.

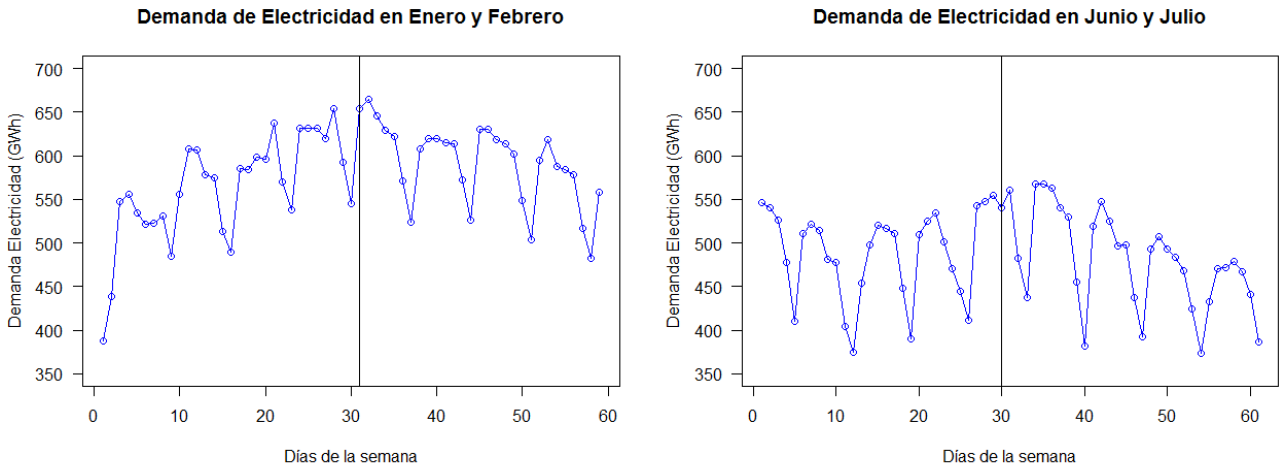

Figura 3. Demanda de Electricidad en invierno y en verano (2011)

A la vista de estos gráficos no apreciamos grandes diferencias en cuanto al comportamiento de los datos, sólo podemos decir que como parecía lógico viendo la Figura 1, en los meses de verano la demanda de electricidad es en general menor que en los meses de invierno, ya que oscilan entre 360 y 560 en verano, mientras que en invierno la demanda se sitúa en términos generales entre 500 y 650. Esto según parece, se debe entre otros muchos factores al descenso de las temperaturas en invierno y al empeoramiento de las condiciones meteorológicas, además de que en invierno hay menos horas de luz que en verano y, por tanto, es necesario un mayor uso de la electricidad para iluminar nuestras casas, fábricas y oficinas. Por otro lado se produce también un aumento de la demanda de electricidad por el mayor uso de aparatos electrónicos, como por ejemplo calentadores para mantener nuestro hogar a una temperatura confortable, y otros muchos aparatos electrónicos que no utilizamos el resto del año o que utilizamos mucho menos, como por ejemplo la secadora de ropa (ya que en verano la ropa seca al aire libre y se estropea menos) o el deshumidificador. También podemos pensar que contribuyen al aumento de la demanda de electricidad, incluso las tradicionales fiestas navideñas, que implican desde hace muchos años, llenar calles y edificios de luces. En verano también se utilizan algunos aparatos eléctricos para mantenernos frescos y evitar el calor, como climatizadores, pero su consumo energético es mucho menor que los calentadores. Por estos motivos es que generalmente el aumento de la demanda de electricidad se produce casi un mes antes del comienzo oficial del invierno.

Vamos a desglosar un poco más nuestros datos, representando un mes de invierno (enero) y un mes de verano (julio), para ver si podemos apreciar algún detalle más que los vistos hasta el momento.

En los siguientes gráficos se pueden ver dos particularidades, la primera es que parece que en las semanas de invierno existen mayores picos en el aumento de la demanda que en las semanas de verano, aunque en la primera semana de enero no se aprecie, debido a la alteración de la demanda que se produce por ser una semana con “días especiales”.

Y en segundo lugar, se puede observar que existe mayor dispersión de los datos en verano que en invierno. Esto lo vemos fijándonos en los datos en invierno, ya que en una misma semana la demanda varía por ejemplo desde 500 a 600, desde 500 hasta 640, o desde 550 a 650, es decir, en general varían en 100; mientras que en verano, pasan en una semana, de 570 a 370, o de 370 a 550, excepto en la última que va desde 400 a 500, es decir varían en una media de 150 y 200.

Esto indica que debería ser un poco más difícil predecir los cambios en la demanda en el verano que en invierno, por sus mayores cambios, ya que vemos que se alcanzan valores más extremos.

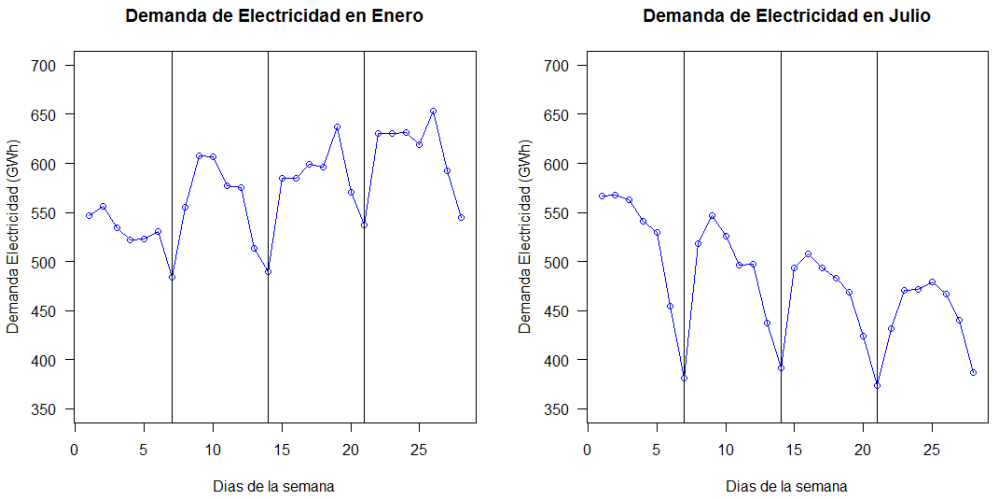

Figura 4. Demanda de Electricidad en Enero y en Julio (2011)

Ahora bien, debemos tener en cuenta que en los dos años que utilizamos para nuestro estudio, se dan días "especiales" como: festivos nacionales, festivos locales, huelgas generales, o períodos de vacaciones.

Haciendo un breve recuento, tenemos 9 días especiales en el 2011 y 9 días especiales en el año 2012, sin contar los días especiales que coinciden con domingo (que actúan como un domingo cualquiera). Observando nuestros datos, se puede apreciar que en estos días especiales se produce una demanda de electricidad inusual, ya que tanto colegios como algunas fábricas y empresas dejan de funcionar y esto se refleja en el consumo de la misma. Nosotros estos días especiales los trataremos de forma similar a como lo hacen Cancelo, Espasa y Grafe (2008).

Para realizar el estudio que tratamos en este trabajo pensamos, en crear variables dummy que tengan en cuenta que existen días diferentes en el año y que por tanto necesitan ser diferenciados de un día normal, ya que su comportamiento en términos de demanda eléctrica va a ser distinta. Por un lado, queremos tener en cuenta los efectos de los festivos nacionales y para ello creamos dos variables, una de ellas representa los festivos que coinciden en un día de entre semana, y otra de las variables representará los festivos que caigan en sábado. Los festivos que coinciden en domingo no vemos necesidad de tenerlos en cuenta ya que pensamos que no va a tener un efecto sobre la demanda, muy diferente de un domingo cualquiera.

Por otra parte creamos una variable que incorpora el efecto del período más habitual de vacaciones, que suele ser el mes de agosto.

Y finalmente, también pensamos que es importante considerar que los festivos Autonómicos de las grandes ciudades pueden tener un efecto importante en la demanda de electricidad de España. Para dictaminar qué Autonomías escogemos, pensamos en elegir las Comunidades Autónomas con un porcentaje de población superior al 10% del total de toda España, y por ello creamos una variable para los festivos de Andalucía, Cataluña, Comunidad de Madrid y Comunidad de Valencia.

Las variables las creamos de forma que tendrán valor 1 si el suceso ocurre y valor 0 en otro caso. Por lo que tenemos 7 variables dummy con las que trabajaremos posteriormente en los modelos de predicción.

Una vez explicado esto, vamos a realizar un breve análisis observando algún gráfico como hemos hecho hasta ahora, para saber en qué medida la demanda se ve afectada en estos días inusuales y qué efectos pueden tener en la demanda de esa semana o de ese mes.

No vamos a analizar todos los días especiales del año puesto que podría resultar demasiado repetitivo, además, por norma general el efecto de estos días en la demanda suele ser similar. Vamos entonces a representar únicamente 3 situaciones en las que se dan los días especiales más significativos, en cuanto al efecto que éstos tienen sobre la demanda.

Comenzamos representando los días de jueves y viernes santo, aunque no coinciden en los mismos días en los años 2011 y 2012. En el 2011 estos festivos caen en los días 21 y 22 de abril mientras que en 2012 caen en los días 5 y 6 de abril, por eso nos parece interesante mostrar en los gráficos el efecto de los festivos en el año 2011, en comparación de cómo sería el comportamiento de la demanda en esas mismas fechas en 2012. No obstante, en este primer gráfico no mostramos exactamente las mismas fechas para el año 2012 por unos acontecimientos excepcionales que tuvieron lugar en esa misma semana, que hacen que no sea una semana representativa para la comparación (utilizamos unas semanas próximas a la fecha). En cambio en el segundo gráfico si, hacemos una comparación entre lo ocurrido en 2012 con estos dos días festivos, con el 2011 en esos mismos días (sin festivos).

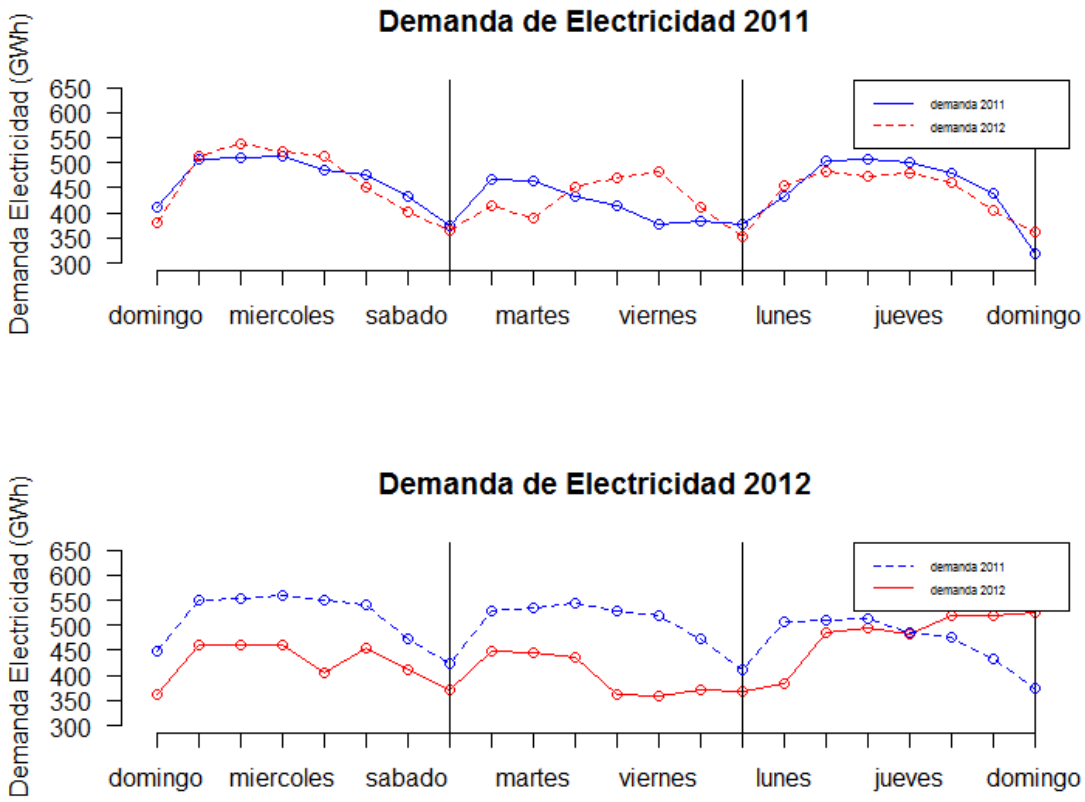

Figura 5. Demanda de Electricidad en Jueves y Viernes Santo

Estos gráficos pueden resultarnos muy interesantes, ya que si nos fijamos en el primero se puede ver que la demanda de electricidad del 2011, representada en azul, se comporta tal y como es de esperar, produciéndose una bajada de la demanda en el jueves y viernes santo e incluso se ve un poco afectado el Sábado posterior también, llegando a los valores que serían típicos de un domingo. Sin embargo en el año 2012 la demanda, se comporta siguiendo el patrón de todas las semanas que no tienen ningún acontecimiento extraordinario, excepto en el segundo martes que vemos que la demanda baja un poco porque coincide con un festivo, el 1 de mayo.

En cuanto al segundo gráfico, vemos representada la demanda de 2012, en rojo de nuevo, en la que se dan grandes variaciones que merecen ser explicadas.

En primer lugar, tenemos que en la semana anterior a la de jueves y viernes santo, el jueves, se produce una bajada considerable de la demanda de electricidad, hecho que se debe a que ese día se produjo una huelga general en toda España, en contra de la séptima reforma laboral.

En segundo lugar, vemos que se produce un efecto idéntico al que se daba el año anterior, se reduce la demanda de electricidad el jueves, el viernes y el sábado, aunque parece que la reducción en este caso se hace de forma más drástica que en el año anterior. A continuación, nos llama un poco la atención que en la siguiente semana, el lunes, la demanda no se eleva como debería si no que tiene un valor similar a la del domingo, esto es debido a que es lunes de Pascua y en la mayor parte de España en este año, se cogió el día como festivo.

Generalmente sólo vamos a analizar tres semanas en los días especiales, para ver el efecto de los días festivos en los días anteriores y posteriores, o como mucho en la semana anterior y posterior. Pero en esta Figura 5, en el segundo gráfico no hemos podido evitar fijarnos en que la demanda en la semana siguiente a la de semana santa se comporta de forma totalmente inusual, ya que en el fin de semana (viernes, sábado y domingo) ésta, se eleva aún más que en el resto de la semana (que en este caso ya había sufrido un gran aumento). Este hecho nos hace preguntarnos qué sucede en estos días para que la demanda se eleve de esta manera, y si quizás sucede algo excepcional también la semana siguiente o si la demanda vuelve ya a su comportamiento habitual. Para analizar lo sucedido, hemos decidido representar varias semanas e intentar ver lo qué ocurre.

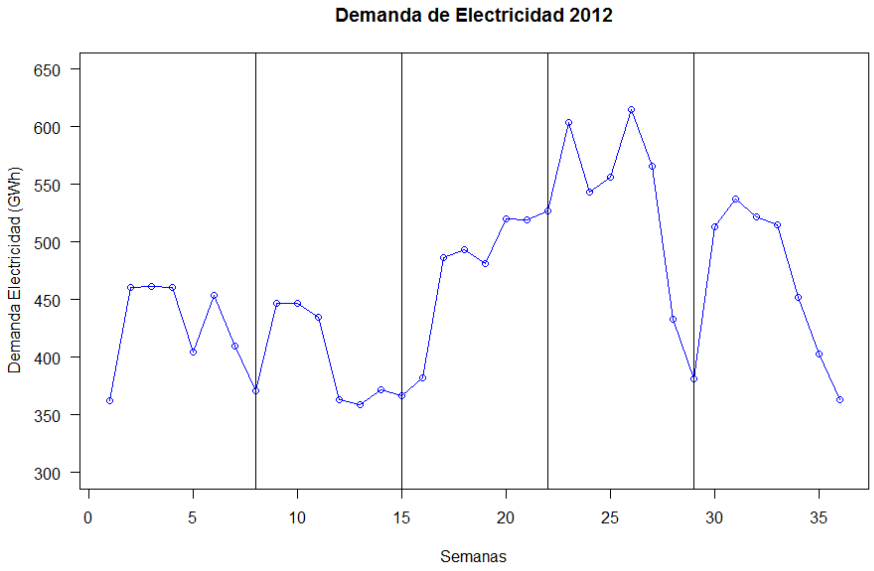

Figura 6. Demanda de Electricidad en 2012 (Días Especiales)

Observando el gráfico anterior vemos todos los efectos de los que antes hablábamos, la huelga, los dos festivos y el lunes de Pascua, así como el fuerte incremento de la demanda eléctrica que se da en el fin de semana. Si nos fijamos en la semana siguiente se ve que la demanda se mantiene elevada hasta el viernes, aunque hay dos días, el 16 y el 19 de abril, en los que se produce un mayor aumento en ésta. En la siguiente semana a ésta, el comportamiento de la demanda vuelve a la normalidad siguiendo el patrón que había seguido cada semana.

Entonces, ¿qué hace que la demanda sea tan elevada en estos días? Indagando un poco, hemos podido saber, que la causa fue que en esos días así como en el domingo 15 de abril entre otros, se obtuvieron máximos en la producción de energía eólica, lo que provocó que el precio de la electricidad fuese 0 durante algunas horas, y esto a su vez originó un fuerte incremento en la demanda de electricidad.

Una vez hemos analizado los efectos conjuntos de una huelga y dos festivos, vamos a proceder al estudio de un día festivo de forma separada para ver el efecto que tiene.

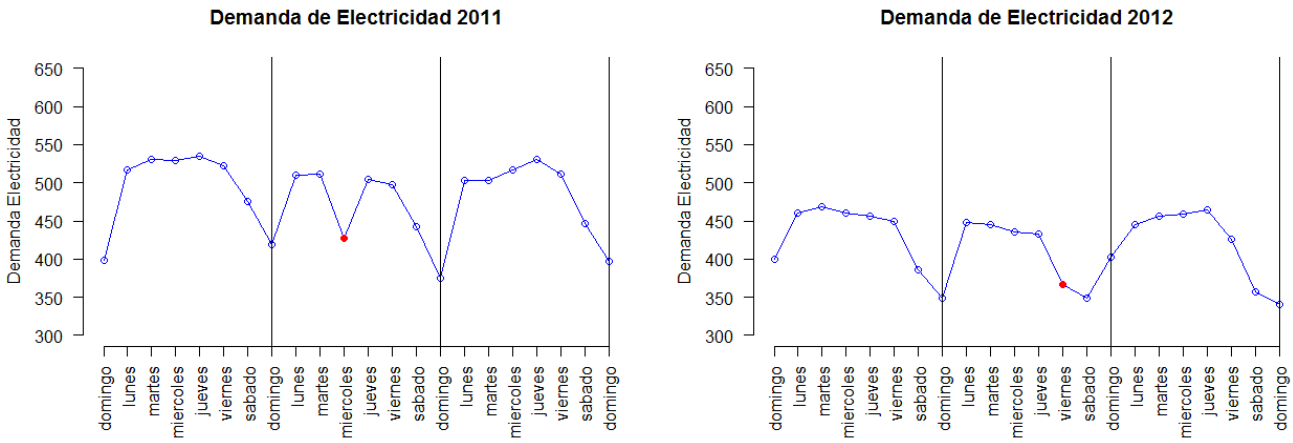

Figura 7. Demanda de Electricidad en el Día de la Hispanidad

En la Figura 7 tenemos representada la demanda en el Día de la Hispanidad con un punto rojo; En 2011 cae en miércoles y se ve que el efecto que causa es el esperado en un festivo, es decir, una reducción importante de la demanda, pero ésta no afecta a los demás días.

Sin embargo este mismo festivo en el 2012 cae viernes, y sí que parece tener cierto efecto en los días posteriores ya que el sábado disminuye mucho también la demanda, llegando a comportarse como un domingo, mientras que el domingo hay una pequeña subida de demanda, con respecto al sábado, hecho que no suele darse en condiciones normales.

Ahora vamos a analizar el efecto de una huelga general sobre la demanda, que se dio el 14 de noviembre de 2012. Fue considerada la primera Huelga Internacional y la primera Huelga General Europea y en ella se luchaba en contra de las políticas de ajuste del Gobierno.

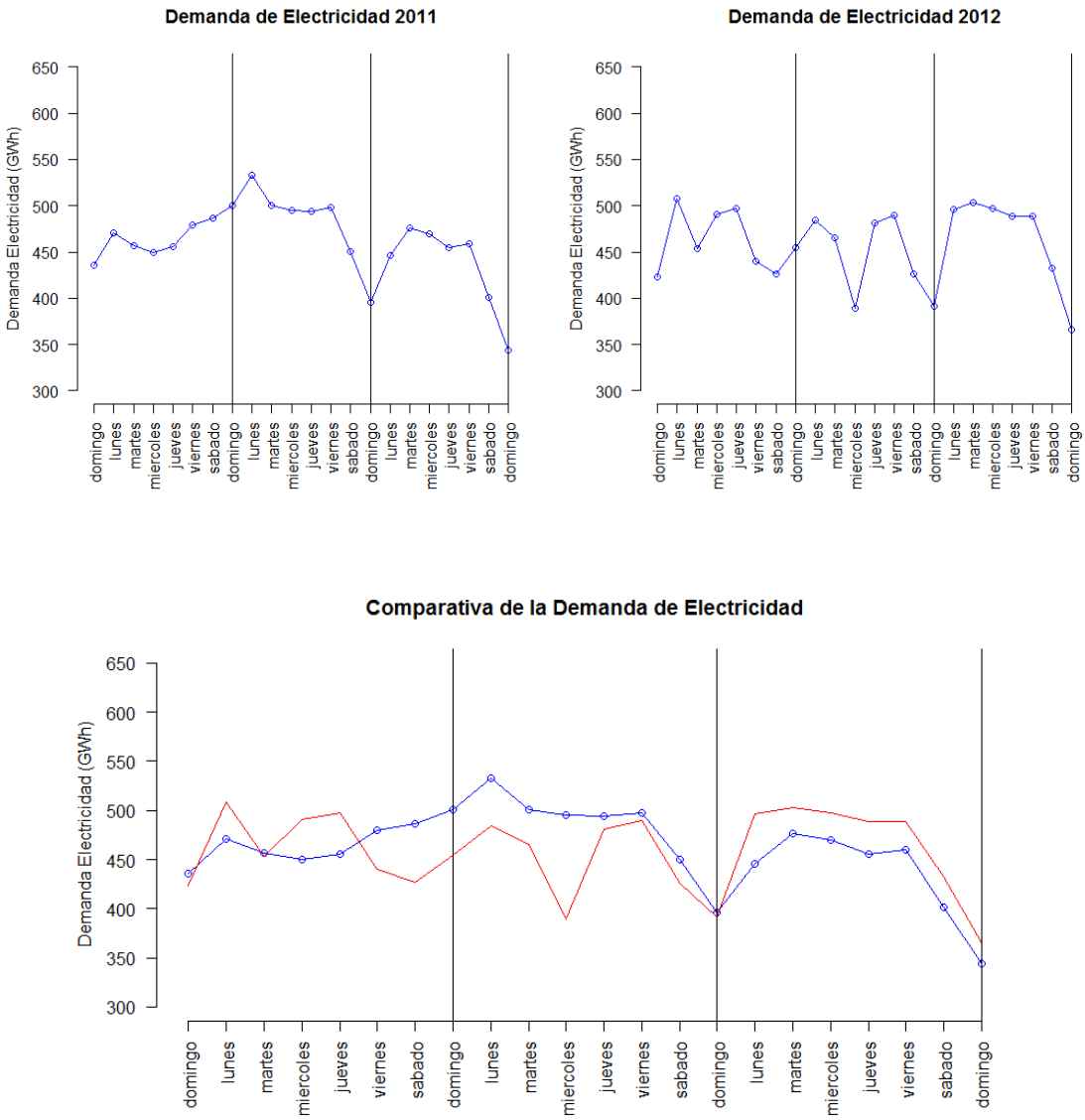

Figura 8. Demanda de Electricidad en Huelga General

Esta huelga se dio el miércoles y tuvo un efecto significativo, reduciendo fuertemente la demanda de electricidad en ese día, quizás se aprecie mejor en el gráfico donde se muestra la demanda de los dos años conjuntamente.

Entonces como podemos observar, el efecto de una huelga es muy similar al que tiene un día festivo, siempre reduce la demanda, sin embargo en el caso de la huelga, la reducción de la demanda de electricidad va a verse influenciada directamente por el número de personas que hagan la huelga.

Finalmente debemos tener en cuenta, que tanto los días con huelgas como los días festivos, a pesar de que provocan una reducción en la demanda eléctrica, su efecto no es independiente del factor de estacionalidad semanal. Es decir, el efecto que tiene un día especial sobre la demanda eléctrica va a depender de en qué día de la semana coincida, vamos a verlo con un ejemplo. En este caso hemos decidido representar los días de la Constitución y de la Inmaculada Concepción.

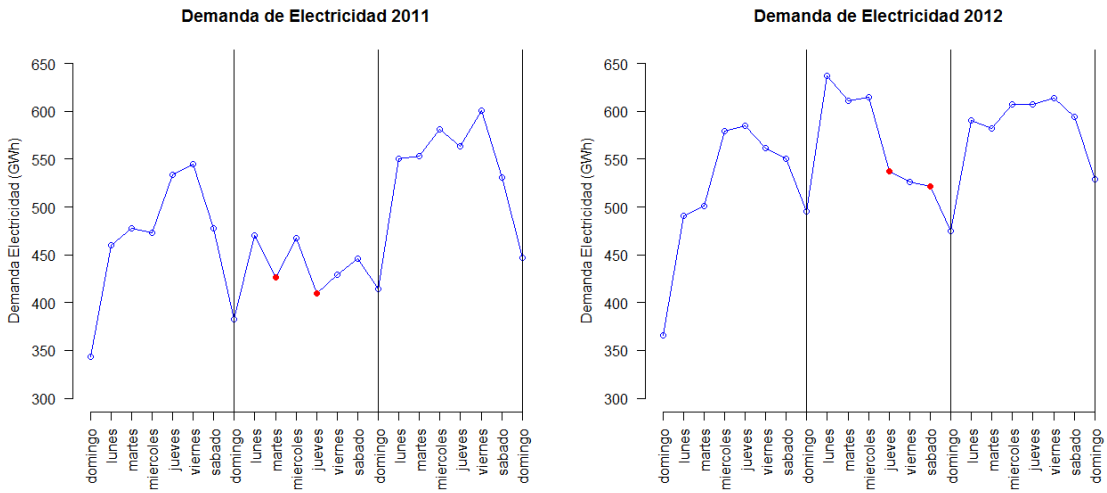

Figura 9. Demanda de Electricidad en el Día de la Constitución y de la Inmaculada Concepción

Observando lo que reflejan los gráficos anteriores se puede observar una clara diferencia en el comportamiento de la demanda en los distintos años.

En el 2011 el día de la Constitución, que representamos con un punto rojo, cae en martes, donde se aprecia una reducción considerable de la demanda eléctrica. Por otra parte el día de la Inmaculada Concepción, que es el segundo punto rojo de ambos gráficos, cae en jueves y la demanda vuelve a reducirse notablemente, incluso más que en el anterior festivo. En muchas Comunidades Autónomas, aprovechando la situación de los días festivos, cogieron libre el lunes, otras el miércoles, y otras el viernes, así como otras cogieron la semana completa libre; estos motivos afectaron a la demanda de electricidad de toda la semana que se ve reducida de forma importante.

En cambio en el año 2012 el día de la Constitución cae en jueves y el de la Inmaculada Concepción es un sábado provocando esto que la demanda se vea reducida desde el jueves hasta el sábado, sin embargo los demás días no se ven afectados.

Por tanto vemos que la estacionalidad semanal va a afectar a la demanda de los días especiales y a los días anteriores y posteriores por el "efecto puente", dependiendo del día en que estos cuadren, aunque ésta es simplemente una puntualización; nosotros en este trabajo no vamos incorporar este efecto puente en nuestro modelo de estudio.

3. MODELOS DE PREDICCIÓN. CONCEPTOS TEÓRICOS

Desde hace años se vienen utilizando para realizar predicciones los modelos de regresión. Estos modelos nos permiten saber de qué forma influyen una o varias variables explicativas *X* en una variable respuesta *Y*.

Sin embargo dentro de los modelos de regresión existen dos tipos de modelos, los paramétricos y los no paramétricos.

Nosotros en este trabajo compararemos las predicciones obtenidas a través de modelos paramétricos y no paramétricos con las resultantes del método NAIVE.

El clasificador probabilístico Naive Bayes o Bayesiano Ingenuo, tiene su fundamento en el teorema de Bayes y en algunas hipótesis simplificadoras adicionales. Es por estas simplificaciones, que se suelen resumir en la independencia entre las variables predictoras, que recibe la denominación de ingenuo. Por ejemplo una fruta puede ser considerada una naranja si es redonda, de color naranja y tiene la piel rugosa, pues bien, el método Naive considera que cada una de las características citadas contribuye independientemente a la probabilidad de que una fruta sea una naranja, independientemente de la presencia o ausencia de las demás características.

El método de predicción Naive es una técnica de predicción supervisada puesto que necesita de ejemplos previos que le ayuden a predecir los datos futuros, sin embargo, es uno de los algoritmos de aprendizaje práctico más utilizados por su sencillez a la hora de hacer predicciones. En este método las predicciones se generan mediante un mecanismo automático establecido a priori, es decir, es un procedimiento de predicción que repite de forma mecánica un comportamiento pasado. Realiza las predicciones asumiendo que el valor futuro en el instante t + 1 de una variable, coincide con el valor actual de la variable en el momento t.

$$\hat{y}_{t+1} = y_t$$

El método Naive presenta dos ventajas y es que su implementación es muy fácil, además de que requiere de una cantidad muy pequeña de datos para estimar los parámetros (medias y varianzas de las variables) necesarios para la clasificación.

La desventaja que presenta este tipo de modelos, es que su capacidad es muy limitada por lo que no suelen utilizarse más que como referencia para evaluar la calidad de métodos más complejos. Puesto que un método más complejo que no logre reducciones importantes del error de predicción respecto al Naive sería un mal método y por tanto no deberíamos utilizarlo.

En cuanto al otro tipo de modelos paramétricos, los **modelos lineales puros,** presentan una gran ventaja frente a otros modelos, que es la facilidad de interpretar el efecto que tiene cada variable explicativa sobre la variable respuesta. Para ello basta con interpretar la estimación del parámetro asociado. Además, nos permite realizar predicciones de la variable respuesta de forma muy sencilla siempre que conozcamos el valor de las variables explicativas. No obstante, este tipo de modelos presentan el inconveniente de la rigidez, es decir son muy poco flexibles, por eso en muchos casos las relaciones entre variables no pueden ser modelizadas mediante este modelo.

Para corregir el problema de la rigidez se comenzaron a utilizar los **modelos no paramétricos**, sin embargo éstos presentan otros problemas. Por una parte la interpretación

del efecto de las variables explicativas en la variable respuesta ya no es tan sencilla como en el caso del modelo de regresión lineal puro. Y por otra, los modelos no paramétricos incurren en la maldición de la dimensionalidad, lo cual implica que cuando incrementamos el número de variables regresoras en nuestro modelo aumenta la dificultad de realizar cálculos y el hecho de hacer una presentación correcta de los resultados. Además, a medida que aumentamos las variables que introducimos en el modelo las propiedades de los estimadores se ven cada vez más perjudicadas, entonces, si aumentamos mucho el número de variables en nuestro modelo necesitaríamos una gran cantidad de datos para que las estimaciones que obtenemos sean fiables.

Nosotros en este trabajo vamos a intentar solucionar estos problemas utilizando **modelos de regresión parcialmente lineal.** Estos modelos fueron propuestos por primera vez por Engle et al. (1986), con el fin de poder estudiar el efecto provocado por las condiciones meteorológicas en la demanda de la electricidad. Posteriormente fueron estudiados en la literatura por numerosos autores (Speckman, 1988; Schamalensee y Stocker, 1999 y Härdle et al., 2000).

El modelo de regresión parcialmente lineal, nos permite describir una variable respuesta como la suma de una componente paramétrica (lineal) y una componente no paramétrica caracterizada por una función suave m. Este tipo de modelos son flexibles por su parte lineal y permiten interpretar el efecto de cada variable sobre la variable respuesta de forma sencilla. Además soluciona o atenúa el problema de la maldición de la dimensionalidad. Esto hace que en muchas situaciones sea el modelo más adecuado para realizar predicciones futuras en comparación con el lineal o el no paramétrico, y por este motivo lo hemos escogido para nuestro estudio.

El modelo parcialmente lineal viene definido generalmente como:

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \ldots + X_{ip}\beta_p + m(t_i) + \varepsilon_{i,} \qquad i = 1, \ldots, n$$

Aunque otras veces lo podemos encontrar de forma simplificada:

$$Y_i = X_i^T\beta + m(t_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

En esta expresión tenemos que Y_i es la variable respuesta; $X_i = (X_{i1}, \ldots, X_{ip})^T$ y T_i son las variables explicativas; $\beta = (\beta_1, \ldots, \beta_p)^T$ es un vector de parámetros desconocidos; m es una función desconocida y ε_i los términos de error aleatorio.

A la hora de estimar β y m existes diferentes métodos que han sido extensamente estudiados a lo largo de los años. Uno de los más utilizados se fundamenta en la combinación del método de estimación por mínimos cuadrados ordinarios y la estimación tipo núcleo.

Para realizar la estimación mediante este método, tenemos que obtener en primer lugar el estimador no paramétrico de la función m, bajo la hipótesis de que β es conocido. El estimador lo obtenemos de la siguiente forma, suponiendo que $w_{n,h}(.,.)$es la función de pesos caraterizada por una función de kernel *K* y un parámetro de suavizado $h > 0$.

$$\widehat{m}_h(t, \beta) = \sum_{j=1}^{n} w_{n,h}(t, T_j)(Y_j - X_j^T\beta) \qquad (1.1)$$

Una vez tenemos el estimador de la función m, estimamos β por mínimos cuadrados ordinarios.

$$\hat{\beta}_b = (\bar{X}_b^T \bar{X}_b)^{-1} \bar{X}_b^T \bar{Y}_b \qquad (1.2)$$

Siendo $\bar{X}_b = (I - w_b)X$ e $\bar{Y}_b = (I - w_b)Y$ donde $w_b = (w_{n,b}(T_i, T_j))_{i,j}$ es una matriz de suavización y $b > 0$ es el parámetro ventana que controla el nivel de suavización tal que $nb \to \infty$ y $b \to 0$ cuando $n \to \infty$.

Robinson (1988) y Speckman (1988) demuestran que bajo ciertas circunstancias como que los errores sean independientes e idénticamente distribuidos, el estimador $\hat{\beta}_b$ es $\sqrt{n} - consistente$ para β y asintóticamente normal.

También Aneiros Pérez y Quintela del Rio (2001) probaron que bajo errores tipo $\alpha - mixing$, este estimador era asintóticamente normal.

Partiendo de la expresión (1.1) si en lugar de los β supuestamente conocidos insertamos su estimador (1.2), obtenemos el estimador de m propuesto por Robinson (1988) y Speckman (1988).

$$\hat{m}_h(t, \hat{\beta}_b) = \sum_{j=1}^{n} w_{n,h}(t, T_j)(Y_j - X_j^T \hat{\beta}_b)$$

En esta expresión b y h son los parámetros de suavización tal que $nb \to \infty$, $nh \to \infty$, $b \to 0$, $h \to 0$, cuando $n \to \infty$.

En la estimación tipo núcleo es realmente importante la elección de los parámetros de suavizado b y h. Esta importancia reside en que si el parámetro de suavizado es muy pequeño, el estimador es propenso a infrasuavizar por lo que interpola los datos, mientras que si el parámetro es grande el estimador tiende a sobresuavizar por lo que tiende a una función constante. Para ello debemos escoger un buen método que nos permita seleccionar el parámetro de ventana óptimo mediante la minimización de algún criterio de error.

En el artículo de Aneiros Pérez y Quintela del Rio (2001) se propone utilizar el método de validación cruzada modificado para seleccionar el parámetro de suavizado. Esto sería minimizando la expresión siguiente:

$$CV_{l_n}(h) = \frac{1}{n} \sum_{j=1}^{n} (Y_j - X_j^T \hat{\beta}_h - \hat{m}_{h,i,l_n}(T_i))^2$$

Donde $\hat{m}_{n,i,ln}$ es la estimación de m una vez hemos eliminado los datos altamente correlacionados con Y_j, es decir, aquellos Y_i tal que $|j - i| \leq ln$.

En el trabajo de Rice (1984) se introduce un nuevo método para la suavización tipo núcleo que es la validación cruzada generalizada. Este método requiere de la utilización de la *hat.matriz* $[A(b)]$ tal que $\hat{Y}_b = A(b)Y$, y siendo $\hat{Y}_b = X\hat{\beta}_b + \hat{m}_b = X\hat{\beta}_b + W_b(Y - X\hat{\beta}_b) = W_b Y + \tilde{X}_b \hat{\beta}_b$.

El método de validación cruzada generalizada consiste en escoger la banda b que minimice la siguiente expresión:

$$GCV(b) = \frac{RSS(b)}{[1-n^{-1}tr(A(b))]^2}$$

Donde $RSS(b)$ es la suma de cuadrados residual, que se define como:

$$RSS(b) = n^{-1} \|(I - A(b))Y\|^2$$

Otro de los métodos más empleados a la hora de hacer predicciones es el **modelo autorregresivo integrado de media móvil o ARIMA**. Es un modelo estadístico que emplea variaciones y regresiones de datos estadísticos, con el objetivo de encontrar un patrón para poder realizar predicciones futuras.

El modelo ARIMA fue desarrollado por Box y Jenkins a finales de los sesenta del siglo XX para analizar series temporales, que tienen en cuenta la dependencia existente entre los datos; el lector puede consultar la monografía dedicada a tal modelo Box, G.E.P and Jenkins,G.M (1976). Es decir, es un modelo en el que las estimaciones futuras vienen explicadas por los datos del pasado y no por variables independientes. Este modelo fue ampliamente aplicado a modelos de previsión de demanda, Fan and McDonald (1994).

El ARIMA es considerado uno de los métodos más flexibles para analizar series de tiempo, puesto que permite trabajar con patrones de datos muy variados, éstos pueden ser autorregresivos (AR), de media móvil (MA) o de promedio móvil autorregresivo mixto (ARMA) tanto simples como estacionales. Peña (2005) y Hernández (2007). Cuando hablamos de modelo ARIMA además de estar ante un caso de promedio móvil autorregresivo, que combina ambos patrones (ARMA), tenemos que tener en cuenta que éste es Integrado.

Un modelo autorregresivo tiene la siguiente forma:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_j Y_{t-j} + a_t$$

Donde tenemos que $Y_t$es la variable dependiente, c y ϕ_p son constantes y Y_{t-j} son las variables independientes. Estas variables independientes son como podemos intuir, observaciones de la variable dependiente tomadas en períodos anteriores. Por último tenemos también el a_t que es el error donde se incluyen todas las alteraciones para las que el modelo no tiene explicación.

Un modelo de promedio móvil viene representado mediante la siguiente expresión:

$$Y_t = c + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_j a_{t-j}$$

Siendo de nuevo Y_t la variable dependiente, c y θ_j constantes, a_t el error, y a_{t-j} valores anteriores del error. Es decir, en este tipo de modelos la variable dependiente depende principalmente de los valores previos del error.

Y finalmente, para el modelo que combina ambos modelos, ARMA, vemos en la siguiente expresión que la variable dependiente Y_t depende tanto de valores pasados de la misma, como de los errores pasados entre los valores reales y las predicciones.

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_j Y_{t-j} + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + \cdots + \theta_j a_{t-j}$$

Sin embargo en numerosas ocasiones estos modelos tal como lo hemos explicado no son suficientes para explicar una serie temporal y es necesario incluir un componente cíclico. Para ello debemos completar las expresiones anteriores con los parámetros estacionales. Nosotros en esta ocasión sólo mostraremos como quedaría un modelo ARMA con componente estacional, ya que incluye a los dos restantes. La componente estacional viene representada por *s*, entonces, por ejemplo si la estacionalidad es mensual $s = 12$ o si fuese semanal $s = 7$, que son los sucesos más usuales.

$$Y_t = c + \Phi_1 Y_{t-s} + \Phi_2 Y_{t-2*s} + \cdots + \Phi_p Y_{t-p*s} + a_t + \ominus_1 a_{t-s} + \ominus_2 a_{t-2*s} + \cdots + \ominus_q a_{t-q*s}$$

La metodología de Box-Jenkins se basa en un método de predicción por etapas. La primera de estas etapas consiste en identificar el modelo ARIMA que sigue la serie. Para ello en primer lugar debemos hacer las transformaciones y diferenciaciones necesarias para que la serie sea estacionaria (eliminando la tendencia existente y la estacionalidad), y posteriormente analizando las autocorrelaciones tanto simples como parciales identificaremos el orden de los parámetros del modelo ARIMA.

En la segunda fase se estiman los parámetros AR y MA del modelo por máxima verosimilitud o mediante la estimación por mínimos cuadrados ordinarios, obteniendo de esta forma sus errores estándar y los residuos del modelo. En esta etapa teniendo en cuenta si se han hecho o no diferenciaciones en la etapa anterior, estimaremos los valores de los siguientes parámetros indicando qué modelo ARIMA pudo generar la serie temporal. Serán siempre de la forma:

ARIMA_s (p, d, q) (P, D, Q)

p: Orden AR del proceso regular	P: Orden AR del proceso estacional
D: nº de diferenciaciones regulares	D: nº de diferenciaciones estacionales
q: Orden MA del proceso regular	Q: Orden MA del proceso estacional

Para ejemplificarlo vamos a representar como quedaría un modelo ARIMA_s (1, 1, 2) $(1, 1, 1)_7$

$$(1 - \phi_1 B)(1 - \Phi_1 B^7)(1 - B)(1 - B^7) Y_t = c + (1 + \theta_1 B + \theta_2 B)(1 + \ominus_1 B^7) a_t$$

En la tercera fase se lleva a cabo el diagnóstico del modelo utilizando distintos procedimientos. Lo más usual es observar el coeficiente de determinación (que indicará que el modelo es mejor cuanto más se acerce a 1), y también podemos observar el AIC y el BIC (que indicarán que un modelo es mejor cuando menor sea su valor). Por otra parte, en esta etapa también se analiza el comportamiento de los residuos, ya que es necesario comprobar que éstos cumplen las características de toda variable ruido blanco. Deben ser independientes y tener media y varianza constantes; es conveniente que cumplan normalidad también, pero no es una condición necesaria. En caso de que los residuos presenten dependencia sería necesario modificar el modelo y volver a repetir las etapas anteriores hasta que obtengamos un modelo adecuado.

Y la cuarta y última fase consiste en realizar predicciones utilizando el modelo que obtuvimos en la etapa anterior.

4. ESTUDIO APLICADO

En esta sección 4 vamos a centrarnos en tres tipos de predicción: predicción a través del método NAIVE (ingenuo), el modelo autorregresivo integrado de media móvil o ARIMA y el modelo de regresión parcialmente lineal o PLRM. Utilizaremos cada uno de los métodos para realizar predicciones sobre el año 2012, con el paquete *TSA* y el PLRModels, y poder hacer posteriormente una comparación de los resultados obtenidos con cada uno de ellos. La forma que utilizaremos para evaluar estos resultados será comparando la media del porcentaje de error absoluto de los mismos.

4.1. Modelos de predicción sin variables exógenas

En este primer subapartado consideraremos variables dummy, pero no variables exógenas continuas, a la hora de realizar las predicciones.

4.1.1. Método NAIVE:

Como ya hemos explicado, el método NAIVE es uno de los métodos más sencillos que existen para hacer predicciones, se basa en predecir un valor futuro en el instante t +1 asumiendo que éste coincide con el valor actual, es decir el valor de la variable en el tiempo t. Por tanto, si queremos predecir lo que ocurrirá con la demanda de la electricidad de un martes, asumiremos que este valor será el mismo que se dio el lunes, si queremos predecir el miércoles, el valor será el del martes y así sucesivamente. Exceptuando el sábado, el domingo y el lunes, para los cuales, cogeremos el valor anterior de los mismos, es decir, para predecir un lunes, asumiremos que tiene el mismo valor que se dio en el lunes anterior, y así con el resto de variables. Esto es porque el comportamiento de un lunes no podemos compararlo o asimilarlo como igual al comportamiento de un domingo, y lo mismo sucede con el sábado y domingo.

Para evaluar los resultados de los distintos modelos, nos centraremos en el porcentaje de error que se da en cada día de la semana y en cada uno de los trimestres. Y para calcular este porcentaje de error lo haremos mediante la fórmula siguiente:

$$\left|\frac{demanda\ real - prediccion\ de\ demanda}{demanda\ real}\right| \times 100$$

Lo que se puede observar a la vista de esta tabla es que el método NAIVE, aun siendo un método muy simple para realizar predicción, ha hecho unas predicciones bastante buenas ya que los errores no son muy elevados, además estos errores al compararlos con los errores de otros métodos de predicción nos van a dar una idea de lo bueno que es un método u otro para predecir.

Tabla 1

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	6.381480	12.500377	6.180765	9.785769	8.712098
Martes	3.201869	5.429005	2.068105	3.785159	3.621034
Miércoles	1.815097	3.081566	3.834962	4.731443	3.365767
Jueves	3.782276	3.951348	4.275552	4.879057	4.222058
Viernes	3.095406	4.234602	2.353998	5.028481	3.678122
Sábado	5.874412	6.879946	6.810000	11.566267	7.782656
Domingo	7.689165	9.356185	8.264449	15.679706	10.24738
Total	4.548529	6.490433	4.864198	7.942524	

Fijándonos detenidamente en la tabla, podemos observar que las peores predicciones se dieron en los lunes del segundo trimestre con un 12.5 % de media de error, y en el cuarto trimestre el sábado y el domingo, con un 11.56% y un 15.67%. En cambio las mejores predicciones se dieron los miércoles del primer trimestre y los martes del tercero con errores del 1.8% y 2%.

Si observamos la media total de cada día de la semana, vemos que las mejores predicciones se dieron en los días intrasemanales, de martes a viernes, mientras que los demás días para los que utilizamos una "forma distinta" de predecir, las predicciones son considerablemente peor, porque es más difícil que teniendo 7 días "por medio" la actuación de los días se asemeje tanto como si sólo hay 1 día de diferencia entre uno y otro.

También hemos querido ver si se apreciaba cierto comportamiento en los errores según el trimestre del año, pero como se puede ver no es un patrón tan claro. En este caso vemos que con nuestros datos, las mejores predicciones se dan en el primer y tercer trimestre, es decir, en invierno y en verano, por lo que podemos pensar que quizás la temperatura o las condiciones meteorológicas tengan algo que ver a la hora de hacer predicciones.

Vemos que las mejores predicciones se dan en los días intrasemanales y nos parece que puede ser interesante analizar el promedio de este grupo de días en los que se obtienen las mejores prediccioines. En la siguiente tabla, se puede observar que dentro de este grupo de días (desde Martes hasta Viernes), las mejores predicciones se dan en el primer trimestre mientras que las peores se dan en el último, con casi dos puntos porcentuales de diferencia. En cuanto al promedio total, es de un 3.72%; estos datos los tendremos en cuenta a la hora de hacer comparaciones con los demás modelos de predicción.

Tabla 2

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Ma-Vi	2.973662	4.17413	3.133154	4.606035	3.721745

Vamos a representar las predicciones y los valores reales en un gráfico para ver si se aprecia mejor en qué épocas del año las predicciones son mejores.

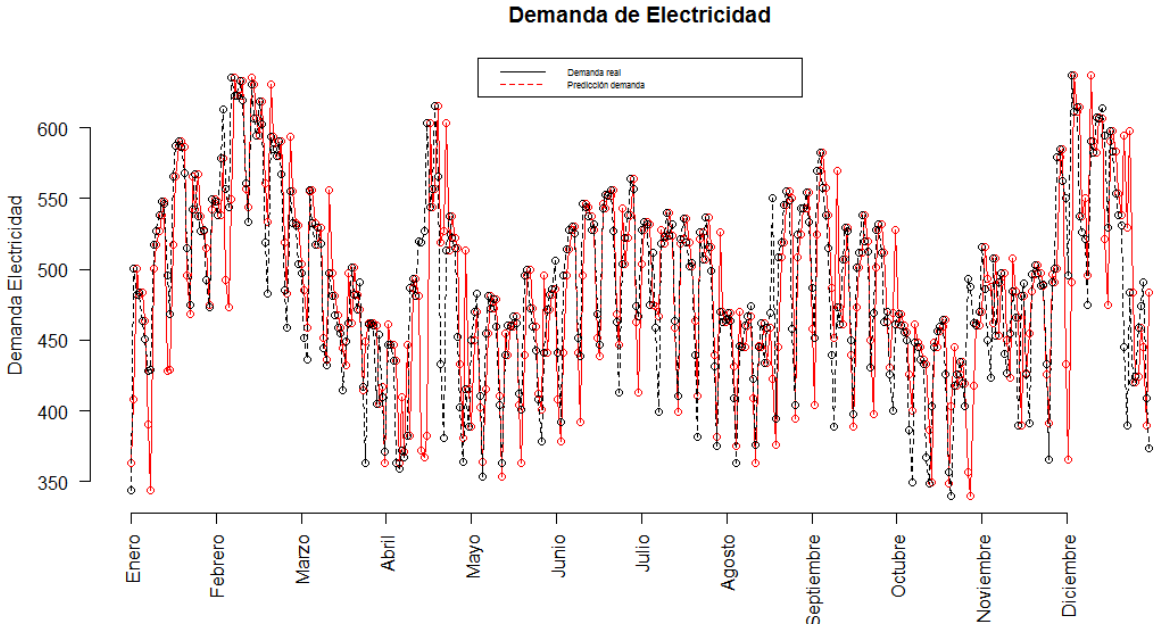

Figura 10. Comparación de la demanda real de electricidad y las predicciones de 2012

Como son muchos datos no se aprecia con claridad donde es mayor y menor el error de estimación, entonces vamos a representar directamente los errores de predicción de cada día del 2012.

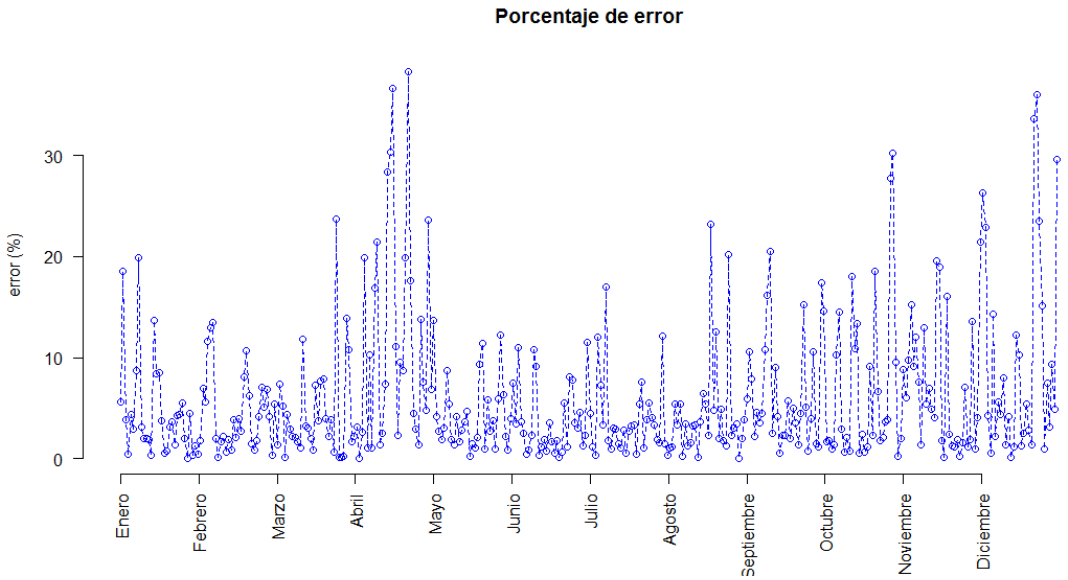

Figura 11. Errores de predicción de 2012

Tal como nos lo presentaban los datos numéricos, en el segundo trimestre empiezan a aumentar los errores de predicción, concentrándose los más elevados, principalmente en el mes de abril. En el cuarto trimestre la mayor parte de errores máximos se dan a principios de noviembre y a finales de diciembre.

¿Podemos decir que el NAIVE es un buen modelo de predicción? La respuesta es no, ya que haciendo un repaso del porcentaje de errores, se puede ver que de 366 observaciones, 133 de las predicciones hechas tienen un porcentaje de error mayor del 5%, que es el límite sugerido usualmente como punto de referencia en la literatura, como hacen Ranaweera, Karady & Farmer (1997) y además suele ser utilizado también por la Red Eléctrica Española (REE) a la hora de evaluar si un modelo de predicción es bueno o no.

Por tanto, en nuestro caso tenemos que un 36.33 % de los errores de nuestras predicciones superan el 5% de error, y esto es un dato muy negativo. Además tenemos que un 69.17%, es decir, casi un 70% de estos errores se dan entre los sábados, domingos y lunes. Podemos afirmar entonces que predecir estos días con el método NAIVE es totalmente imprudente, porque el error que vamos a cometer es demasiado grande y esto puede traducirse en altos costes para las personas que tienen que tomar decisiones a partir de nuestras predicciones.

4.1.2. Modelo ARIMA: el modelo autorregresivo integrado de media móvil

El modelo autorregresivo integrado de media móvil utiliza variaciones y regresiones de datos estadísticos, con el objetivo de encontrar ciertos patrones y así hacer una predicción futura. Para ello necesita determinar los coeficientes y el número de regresiones que se van a utilizar.

Nosotros vamos a realizar dos tipos de predicción a través del modelo ARIMA, en primer lugar lo haremos con un solo modelo para toda la serie y posteriormente lo haremos con siete modelos, es decir haremos lo mismo pero creando un modelo distinto para cada día de la semana.

4.1.2.1. ARIMA: un solo modelo

Como ya hemos explicado el modelo de forma general y en mayor profundidad en el apartado anterior no vamos a entrar en más detalle ahora; por ello, nos vamos a centrar en explicar los pasos a seguir para estudiar el caso concreto de nuestros datos, e ir viendo y explicando los resultados que vamos obteniendo.

En primer lugar, representamos nuestra serie para ver si existe algún aumento en la variabilidad de los datos con el paso del tiempo. Si observamos el gráfico no parece que la variabilidad aumente, no obstante, decidimos aplicar una transformación logarítmica a la serie y observar los resultados que obtenemos con la serie transformada. Al comparar estos resultados con los que obtenemos con la serie sin transformar vemos que los primeros son peores por lo que finalmente nos quedamos con la serie sin transformar para realizar nuestro análisis.

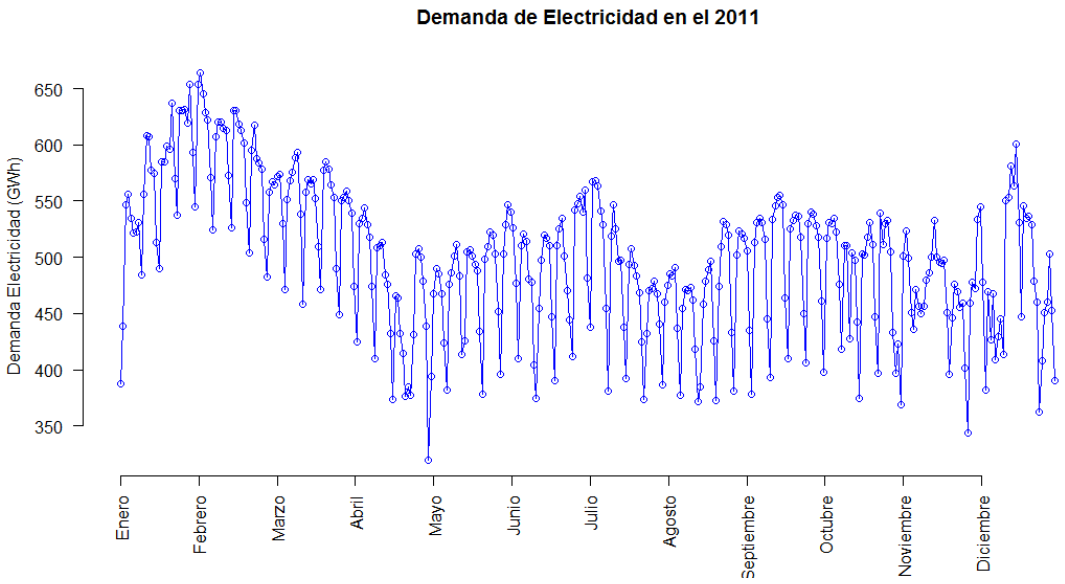

Figura 12. Demanda de Electricidad en 2011

La aproximación a los procesos estocásticos con modelos AR o MA está limitada, generalmente a los procesos estocásticos que cumplan la condición de estacionariedad. Entonces debemos analizar la estacionariedad de las series temporales, antes de iniciar la identificación de la estructura del proceso estocástico AR o MA.

Identificar si una serie es estacionaria en media o no, es una tarea sencilla por norma general, ya que suele ser suficiente con observar el gráfico de la serie para determinar si el valor medio se mantiene constante, o si con el tiempo crece o decrece.

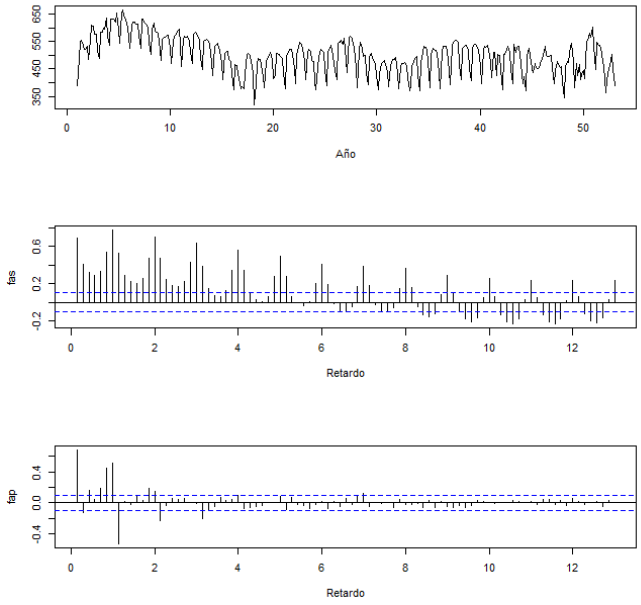

Figura 13. Serie original y correlaciones

Tal como se puede ver en el gráfico anterior, nuestros datos no son estacionarios en media, porque el gráfico nos muestra que existe cierta tendencia lineal decreciente en el primer tramo.

Lo que nos interesa es poder conservar estas variaciones eliminando el componente de cambio en la media, es decir, debemos hacer un "filtro de tendencia". De esta manera, la serie corregida va a representar exactamente las mismas variaciones que la serie original pero sin tendencia. Nosotros trabajaremos con la serie diferenciada regularmente, que no mostrará tendencia y podremos afirmar que es estacionaria en media.

Para hacer el filtro de tendencia haremos una diferenciación en la parte regular y nuestro gráfico quedará de la siguiente forma.

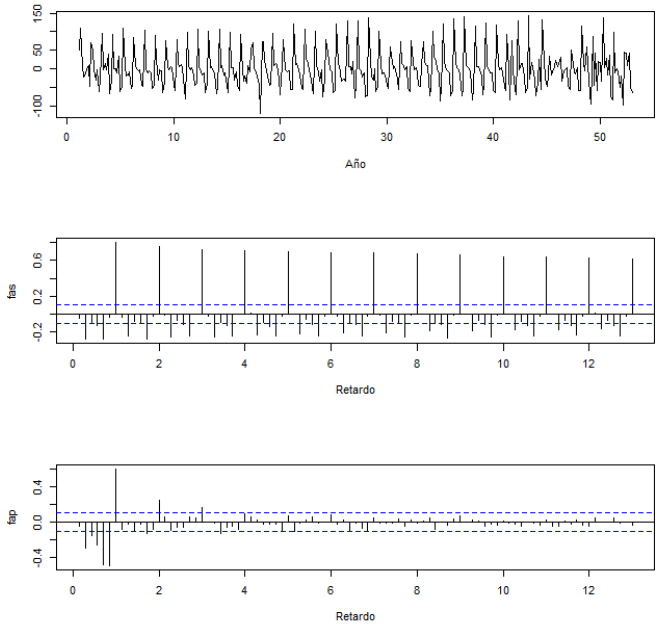

Figura 14. Serie original diferenciada en la parte regular y correlaciones

En él podemos observar que ya no hay tendencia porque las correlaciones son generalmente bajas, salvo en los múltiplos de 7 donde son elevadas, hecho que nos indica que existe estacionalidad, sin embargo al no tener tendencia no es necesario volver a diferenciar la parte regular.

Una vez eliminada la tendencia, el siguiente paso es analizar la estacionalidad de la serie estacionaria. Es fácil ver en el gráfico anterior, que sí existe una parte estacional en la serie ya que las correlaciones, tanto mirando en las "fas" como en las "fap", sobresalen en el número 7 de forma significativa y en todos sus múltiplos. Debemos hacer entonces una diferenciación estacional y observar de nuevo las correlaciones.

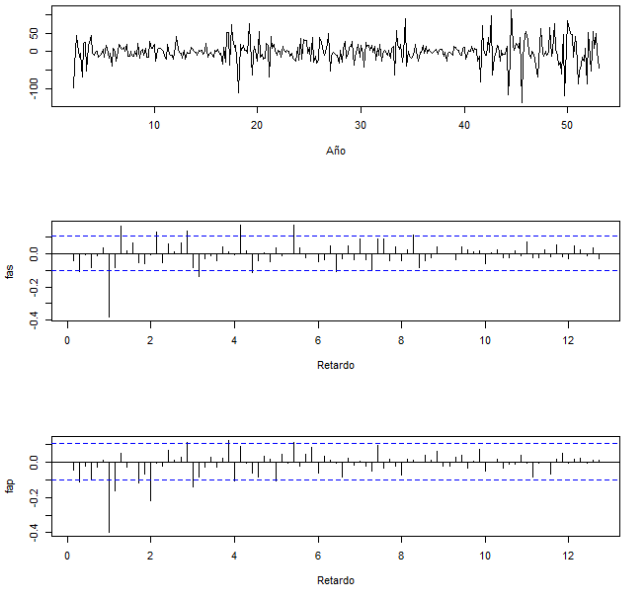

Figura 15. Serie original diferenciada en la parte estacional y correlaciones

A partir de estas correlaciones podemos tratar de identificar un modelo para ajustar la serie utilizando la información disponible sobre cómo ésta ha sido generada.

Fijándonos en las fas, podríamos proponer un ARIMA (0, 1,2) $(0, 1,1)_7$. Pero fijándonos en las fap propondríamos un ARIMA (1, 1,0) $(2, 1,0)_7$. Incluso podríamos combinarlos como un ARIMA (1, 1,2) $(2, 1,1)_7$.

Nosotros vamos a trabajar con la estructura del ARIMA que sea mejor desde el punto de vista BIC. Para ello utilizamos la función *best.arima.TSA*, ya que nos permite incorporar las 7 variables dummy (festivos, período de vacaciones de Agosto y huelgas), teniendo en cuenta de esta forma todas las variables que nos parecen importantes para nuestro estudio y que por tanto podrían influir de forma decisiva en el ajuste del modelo.

Con la función de mejor arima obtenemos que el mejor modelo sería un ARIMA $(1,1,1)(1,1,1)_7$. Y con este modelo será con el que realicemos el ajuste.

$$(1-\phi_1 B)(1-\Phi_1 B^7)(1-B)(1-B^7)Y_t = c + (1+\theta_1 B)(1+\ominus_1 B^7)a_t$$

En la siguiente salida de R, podemos observar los valores de los coeficientes después del ajuste así como sus errores estándar. Como podemos apreciar tenemos 4 coeficientes que no son significativamente distintos de cero, ya que $1.96 * s.e > |coef|$; por lo tanto debemos ajustar el modelo sin el coeficiente que tenga un menor valor cuando hacemos este cálculo, ya que será la variable menos influyente en nuestro modelo, y de esta forma además hacemos el modelo más simple. En nuestro caso la variable con un valor menor es la 3, que corresponde con el efecto que tiene el mes de Agosto (efecto vacacional) sobre la demanda eléctrica, como vemos es una variable irrelevante y la demanda no se ve afectada por este hecho, porque al incrementar en una unidad esta variable, la demanda se reduce en 9.58 unidades.

Tabla 3

arimax (x=demanda1.ts, order= c(1, 1, 1), seasonal= list (order= c (1, 1, 1)), xreg=xreg)

Coefficients	ar1	ma1	sar1	sma1	f.intra	f.sab	Agst	f.And	f.Cat	f.Mad	f.Val
	0.73	-0.92	0.16	-1.00	-51.09	-108.04	-9.58	-11.23	-38.92	-20.41	-9.20
s.e.	0.07	0.05	0.06	0.04	5.61	20.24	13.36	13.68	13.18	10.55	10.72
$1.96 * s.e > \|coef\|$	5.01	9.35	1.35	12.42	4.63	2.72	0.36	0.41	1.50	0.98	0.43

Sigma^2 estimated as 395.2: log likelihood = -1547.86, aic = 3117.71

Haciendo el ajuste de nuevo sin esta variable, obtenemos el siguiente resultado:

Tabla 4

arimax (x=demanda1.ts, order= c (1, 1, 1), seasonal= list (order= c (1, 1, 1)), xreg=xreg [,-3])

Coefficients	ar1	ma1	sar1	sma1	f.intra	f.sab	f.And	f.Cat	f.Mad	f.Val
	0.73	-0.91	0.16	-1.00	-51.13	-107.76	-11.48	-38.71	-19.98	-9.26
s.e.	0.07	0.05	0.06	0.04	5.63	20.26	13.69	13.18	10.56	10.73
$1.96 * s.e > \|coef\|$	4.77	8.85	1.36	12.62	4.63	2.71	0.42	1.49	0.96	0.44

Sigma^2 estimated as 395.9: log likelihood = -1548.11, aic = 3116.22

Seguimos viendo que 3 de nuestras variables no son significativas, repetimos el proceso de fijarnos en cuál de ellas es menos relevante para eliminarla y volver a realizar el ajuste de nuestro modelo. En este caso la menos relevante es la 4, que corresponde con los días festivos en Andalucía; la eliminamos y realizamos el ajuste de nuevo.

Tabla 5

arimax (x=demanda1.ts, order= c (1, 1, 1), seasonal= list (order= c (1, 1, 1)), xreg=xreg [,-c (3, 4)])

Coefficients	ar1	ma1	sar1	sma1	f.intra	f.sab	f.Cat	f.Mad	f.Val
	0.73	-0.91	0.16	-1.00	-51.18	-107.79	-43.26	-23.78	-10.19
s.e.	0.07	0.05	0.06	0.03	5.63	20.28	12.03	9.55	10.68
$1.96 * s.e > \|coef\|$	4.83	8.94	1.34	12.85	4.63	2.71	1.83	1.27	0.48

Sigma^2 estimated as 396.6: log likelihood = -1548.46, aic = 3114.93

Observando la última tabla obtenida, podemos apreciar que una de las variables que antes no era significativamente distinta de cero ahora si lo es. Es por este motivo que tenemos que ir eliminando las variables no significativas de una en una, ya que si lo hacemos de forma conjunta podemos estar omitiendo información relevante. Ahora bien, vemos que la única variable no importante a efectos de ajustar el modelo es la 7, que corresponde con los festivos

en Valencia, por tanto la eliminamos y deberíamos de llegar ya a un modelo con todos los coeficientes significativamente distintos de cero.

Tabla 6

arimax (x=demanda1.ts, order= c (1, 1, 1), seasonal= list (order= c (1, 1, 1)), xreg=xreg [,-c (3, 4, 7)])

Coefficients	ar1	ma1	sar1	sma1	f.intra	f.sab	f.Cat	f.Mad
	0.73	-0.91	0.16	-1.00	-51.26	-107.59	-47.83	-26.88
s.e.	0.07	0.05	0.06	0.04	5.65	20.30	11.08	9.01
$1.96 * s.e > \|coef\|$	4.77	8.85	1.33	12.40	4.62	2.70	2.20	1.52

Sigma^2 estimated as 397.7: log likelihood = -1548.92, aic = 3113.84

Después de realizar este ajuste, podemos afirmar que todos nuestros coeficientes son significativos. Además si observamos de nuevo las tablas, se puede ver que el AIC se ha ido reduciendo a medida que eliminábamos variables no significativas del ajuste, por lo tanto el ajuste se ha ido mejorando.

Vemos que nuestro ajuste va a incluir como variables significativas, los festivos tanto intrasemanales como de los sábados, así como los festivos de Cataluña y Madrid. Las variables dummy con un efecto más relevante son la de los festivos intrasemanales y la de festivos de sábado, aunque destaca en mayor medida esta última.

Nuestro modelo ARIMA quedaría de la siguiente forma después del ajuste:

$$(1 - 0.73 * B)(1 - (-0.91) * B^7)(1 - B)(1 - B^7)Y_t = c + (1 + 0.16 * B)(1 + (-1) * B^7)a_t + (-51.26 * f.intra_t) + (-107.59 * f.sab_t) + (-47.83 * f.Cat_t) + (-26.88 * f.Mad_t)$$

Procedemos entonces a analizar los residuos porque es conveniente saber si cumplen normalidad, para ello vamos a fijarnos en los siguientes gráficos.

En el Q-Q plot podemos ver que existe cierta relación lineal entre los cuantiles esperados bajo normalidad y los observados, pero no estamos seguros de que se cumpla la hipótesis de normalidad. Observando las correlaciones de los residuos vemos que todas entran en los límites.

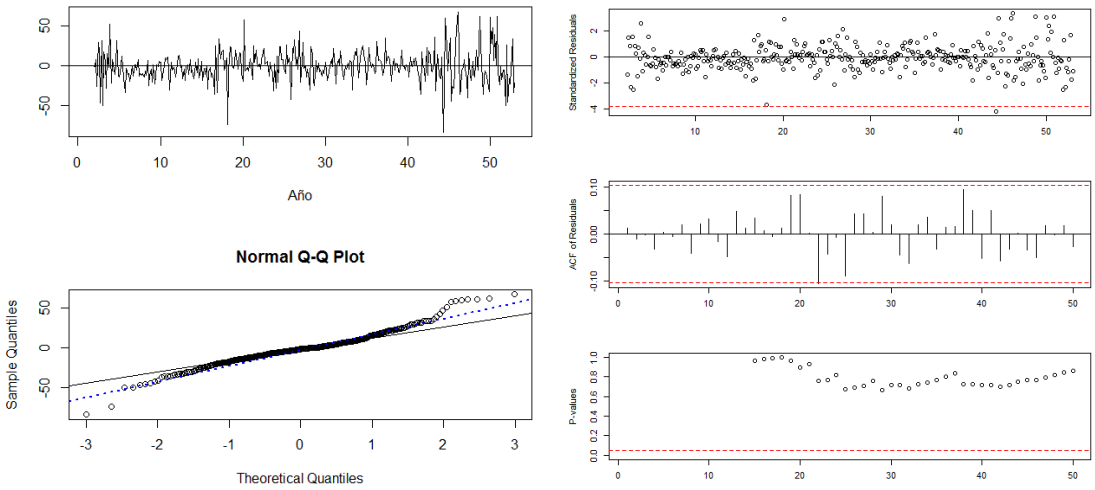

Figura 16. Q-Q plot y Ljung-Box

Para poder realizar los contrastes de normalidad de los residuos debemos hacer primero un contraste de independencia de los mismos, el Ljung-Box, debido a que si nos falla la independencia no podríamos realizar los test de que la media sea igual a cero ni los test de normalidad.

Observando los gráficos anteriores podemos ver que se cumple la hipótesis de la independencia ya que en el segundo gráfico apenas sobre sale una correlación de 50, y en el último gráfico todos los p-valores son superiores a 0.05, por tanto aceptamos independencia y pasamos a realizar los siguientes test.

Cuando hacemos el test de que la media es igual a cero obtenemos un p-valor elevado, por lo que no podemos rechazar la hipótesis nula de que si lo es.

Realizando los contrastes de normalidad obtenemos que se rechaza la normalidad en ambos test para cualquier nivel de significación. Al rechazarse la normalidad con ambos test, no calcularemos los intervalos de predicción porque no tenemos base teórica para creer que nos van a aportar una información verídica, pero sí podemos calcular predicciones puntuales para el 2012. De esta forma podremos hacer una comparación con los datos reales para saber si nuestro modelo hace buenas predicciones o no.

Tabla 7

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.3683	2.2 e-16	2.48 e-08

Observando ahora la Figura 17 podemos apreciar que nuestras predicciones reflejan bastante bien los valores reales del 2012, aunque en algunos casos la predicción podría haber sido mejor. Al representar únicamente la media absoluta del porcentaje de errores nos podemos fijar donde se dan los errores más elevados.

Principalmente los mayores errores se dan en el mes de agosto y en los meses de invierno, noviembre y diciembre. Verlo en el gráfico puede ser una tarea un poco difícil por

lo que mostraremos la media absoluta de los errores en una tabla, tal como hicimos para evaluar el método NAIVE.

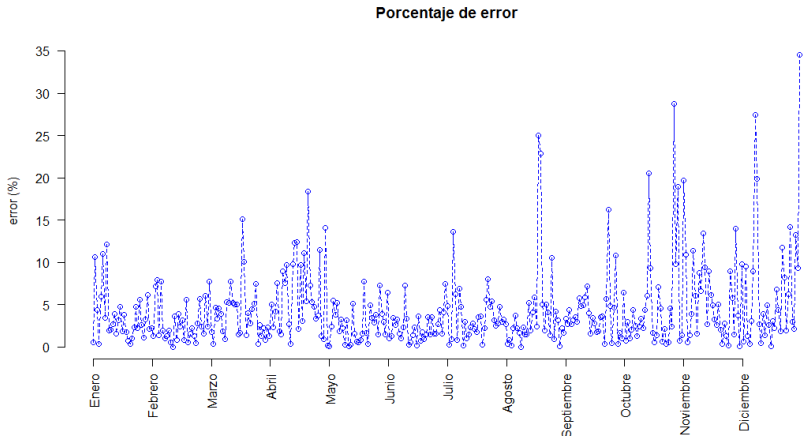

Figura 17. Errores de predicción de 2012

Observando la tabla detenidamente, podemos ver que las peores predicciones se dan el sábado y domingo del último trimestre con un 9.16% y un 7.53% de error, y el lunes del cuarto trimestre con un 8.14%, que coincide con lo que veíamos en el gráfico, es decir, las peores predicciones se dan en el invierno; Por otro lado, las mejores predicciones y por tanto las que más se aproximan a la realidad, se dan los martes y viernes del tercer trimestre y los miércoles del segundo trimestre con un 2.34%, un 1.89%, y un 2.43% respectivamente.

Analizando ahora cómo se comportan los errores trimestralmente se aprecia que los mayores errores se dan en el último trimestre y son casi el doble que en el resto. En cambio si tenemos en cuenta la media de los errores de cada día de la semana, vemos que el patrón que ya viéramos en el método de NAIVE se repite, puesto que los peores errores se dan los lunes, sábados y domingos, mientras que por la semana los errores son más pequeños, especialmente los martes, que tienen una media de un 2.84%. Aunque el mayor error se produce como dijimos el sábado del cuarto trimestre, si tenemos en cuenta la media por día de la semana, el día que tiene un error mayor es el domingo.

Tabla 8

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	3.744789	4.563803	4.377928	8.145212	5.207933
Martes	3.222814	3.227241	2.341107	2.605156	2.84908
Miércoles	2.583775	2.438362	3.465321	4.884401	3.342965
Jueves	3.515843	3.059131	4.317137	4.232315	3.781106
Viernes	2.752142	4.296760	1.891145	4.599386	3.384858
Sábado	2.954418	3.988394	5.036287	9.160447	5.284886
Domingo	5.601929	4.734544	5.894334	7.534118	5.941231
Total	3.482244	3.758319	3.924964	5.904768	

Haciendo un análisis similar al hecho en el modelo anterior, vemos en la siguiente tabla, las medias para cada trimestre para el grupo de días intrasemanales. En esta ocasión, los resultados difieren un poco de lo visto en el NAIVE, puesto que las mejores predicciones ya no se dan en el primer trimestre si no en el tercero (aunque por una diferencia mínima). Por otra parte, fijándonos en qué trimestre se obtienen las peores predicciones para los días de entre semana, vemos que vuelve a ser el cuarto. La media total para estos días refleja que s hicieron unas buenas predicciones ya que tienen únicamente un 3.34% de error.

Tabla 9

Media del porcentaje de error absoluto (MAPE) de Martes a Viernes

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Ma-Vi	3.018644	3.255374	3.003677	4.080315	3.339503

4.1.2.2. ARIMA: siete modelos

Para realizar las predicciones sobre el 2012 mediante modelos ARIMA, en esta ocasión, vamos a tener en cuenta que tenemos 7 días de la semana y por eso decidimos que vamos a modelizar cada día de la semana por separado, para después poder compararlo con los modelos anteriores y saber si de esta forma estamos mejorando las predicciones. Estamos haciendo de este modo, predicciones independientes para cada día de la semana.

4.1.2.2.1 ARIMA para los Lunes

Como los pasos que vamos a ir siguiendo son análogos a los seguidos en el ARIMA de un solo modelo, nos limitaremos a ir exponiendo los gráficos y los demás resultados dando una breve explicación, ya que lo que nos interesa es llegar a obtener las predicciones del 2012 con cada método para poder compararlas al final.

En primer lugar, de todos nuestros datos nos quedaremos con los valores que corresponden a los lunes, para trabajar con ellos. Una vez hecho ya podemos representarla.

Lo que se puede ver en el siguiente gráfico es que el comportamiento de los lunes varía mucho durante todo el año y no sigue una pauta concreta. Especialmente en días que son festivos o posteriores a festivos, como el 1 de mayo o el 15 de agosto, en los que se producen grandes reducciones de la demanda de electricidad.

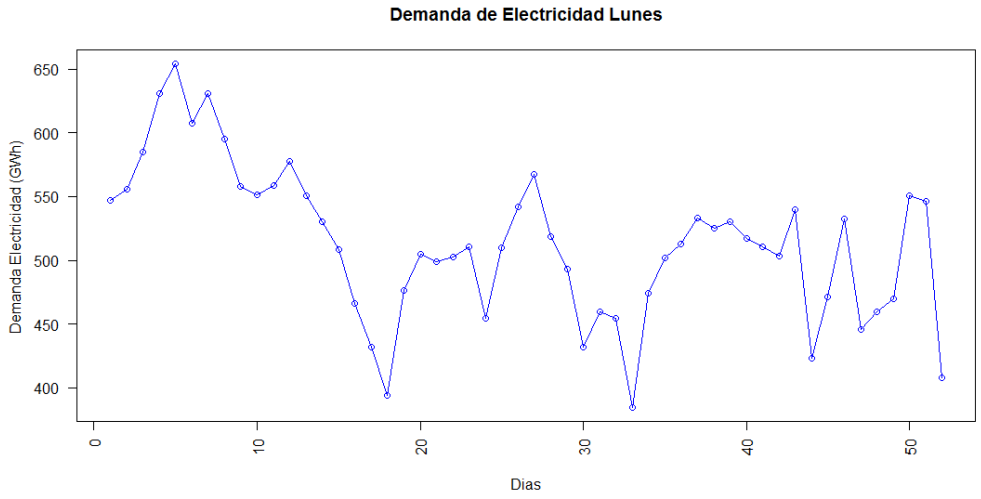

Figura 18. Demanda de electricidad de los Lunes de 2011

El siguiente paso que debemos dar es analizar la estacionariedad en media de nuestros datos. Fijándonos en la Figura 19 se ve claramente que existe cierta tendencia negativa y aunque ésta sólo se da en los primeros datos, debemos hacer una diferenciación regular. Una vez realizada la diferenciación, las correlaciones obtenidas son las que se representan en la Figura 20, en la que podemos ver que la serie no tiene componente estacional y que además es ruido blanco, es decir, son variables aleatorias ya que sus valores, en tiempos diferentes, no guardan correlación estadística.

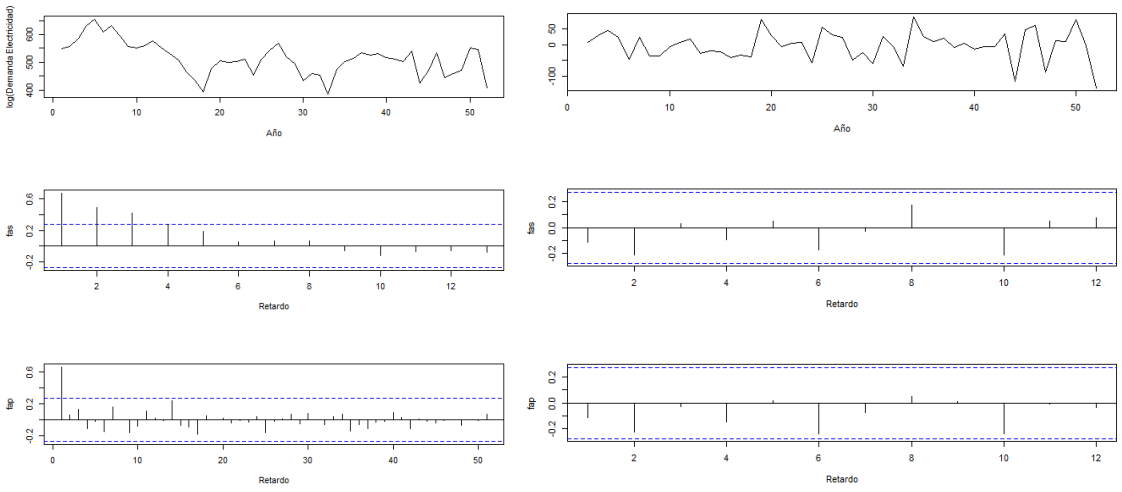

Figura 19. Serie original y correlaciones

Figura 20. Serie original (diferenciada) y correlaciones

A partir de esta última figura podemos tratar de identificar un modelo para ajustar la serie, y fijándonos en las correlaciones, el modelo que propondríamos, teniendo en cuenta que no tiene parte estacional, sería un ARIMA (0, 1, 0), que se podría presentar como:

$$(1 - B)Y_t = c + a_t$$

Utilizando la función que ya empleamos en la ocasión anterior, *best.arima.TSA,* vemos que nos devuelve el mismo valor para nuestro modelo que el que nosotros proponemos, aunque para ello tuvimos que eliminar una de nuestras variables dummy, porque para los lunes era nula, y es obvio ya que se trata de los festivos que caen en sábado.

Una vez hecho esto, procedemos al ajuste de nuestro modelo con la función *arimax*, incluyendo las variables dummy que corresponden al lunes exclusivamente. Haciendo el primer ajuste nos sale que tenemos tres coeficientes que no son significativamente distintos de cero, siendo el menos significativo el efecto de los días de agosto. Vamos realizando el ajuste de nuevo, suprimiendo en cada caso la variable correspondiente a cada coeficiente de los menos significativos, tal como hicimos en el caso del ARIMA único y la tabla final que obtenemos es la siguiente:

Tabla 10

arimax (x=lunes11.ts, order= c (0, 1, 0), xreg = xreg [,-c (2, 3, 7, 6)])

Coefficients	f.intra	f.And	f.Cat
	-79.6775	-85.9687	-44.7457
s.e.	25.7867	21.0548	21.0548
$1.96 * s.e > \|coef\|$	1.5764	2.0832	1.0842

Sigma^2 estimated as 1330: log likelihood = -250.27, aic = 506.55

Después de eliminar todas las variables que no influían de forma significativa en la demanda de electricidad, los días de Agosto, los festivos de Valencia y los festivos de Madrid, ya nos queda un modelo con todos los coeficientes significativos, que es de la siguiente forma:

$$(1-B)Y_t = c + a_t + (-79.67 * f.intra_t) + (-85.96 * f.And_t) + (-44.74 * f.Cat_t)$$

Los coeficientes que tienen relevancia en nuestro modelo son por tanto, los festivos intrasemanales, los festivos de Andalucía y los festivos de Cataluña, siendo los festivos de Andalucía los que disminuyen en mayor medida la demanda de electricidad de España. Cuando se da un día festivo en Andalucía la demanda se reduce en 85.97 unidades y el que menos efecto tiene sobre ésta son los festivos de Cataluña, ya que sólo disminuye la demanda en 44.75 unidades.

Siguiendo con el análisis procedemos a estudiar los residuos. Nos conviene que los residuos cumplan normalidad y observando la Figura 21 parece que si se cumple, puesto que en el Q-Q plot vemos una relación lineal clara entre los cuantiles esperados suponiendo normalidad y los cuantiles observados.

Pero para analizar la normalidad de residuos, tenemos que ver primero si cumplen independencia, ya que si no la cumplen no se podría realizar los test de media igual a cero ni los test de normalidad de los datos. En este caso se ve que la independencia se cumple

también de forma clara. En el último gráfico se puede ver que todos los p-valores son superiores a 0.05 por lo que podemos proceder a elaborar los siguientes test.

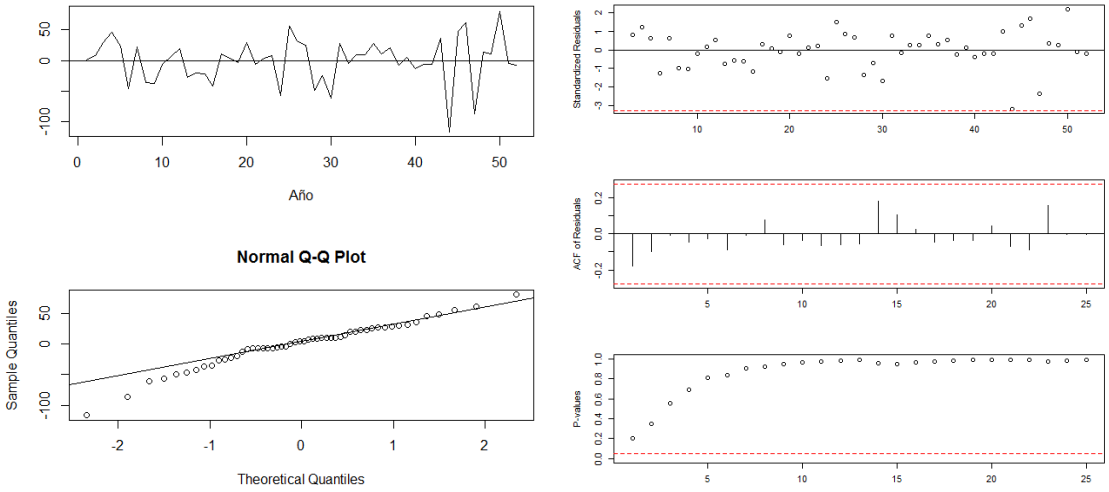

Figura 21. Q-Q plot y Ljung Box

Hacemos el test de que la media es igual a cero y con el p-valor que obtenemos, un 0.97, debemos aceptar que la hipótesis se cumple.

Al realizar los test de normalidad obtenemos dos valores que nos aportan información diferente. Para un nivel de significación del 5% el test de Jarque Bera nos da que se rechazaría la hipótesis de normalidad, mientras que el Shapiro Wilk la acepta. Sin embargo para un nivel de significación del 10% o más, la hipótesis de normalidad se rechazaría en ambos test, por tanto no podemos construir los intervalos de predicción, pero si hacer las predicciones del 2012.

Tabla 11

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.9764	0.02025	0.09962

La media absoluta del porcentaje de errores tanto de los lunes como el resto de días lo pondremos al final del apartado para tener una tabla completa con todos los errores de todos los días de la semana y así poder hacer las comparaciones correspondientes como en los métodos anteriores.

4.1.2.2.2 ARIMA para los Martes

Separando los datos correspondientes a los martes del 2011 y representándolos, podemos ver que tienen un comportamiento bastante irregular aunque no tanto como los lunes, ya que solo tienen una reducción drástica, debido al día 1 de noviembre en el que se celebra el día de todos los santos, que es un festivo Nacional.

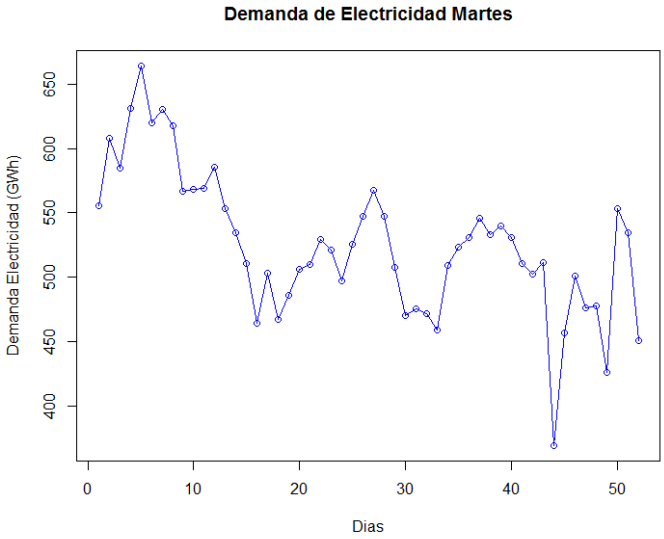

Figura 22. Demanda de electricidad de los Martes de 2011

Si nos detenemos en la Figura 23 en el primer gráfico para observar si hay algún tipo de tendencia en nuestros datos, vemos que si la hay y es decreciente. La eliminamos mediante un filtro de tendencia, es decir, diferenciamos la parte regular de nuestros datos. Una vez hecha la diferencia obtenemos las correlaciones representadas en el segundo gráfico, en el que además de tener corregido el efecto de la tendencia, podemos ver que no existe parte estacional y por tanto no necesitamos hacer ninguna diferencia en esta parte para corregirla.

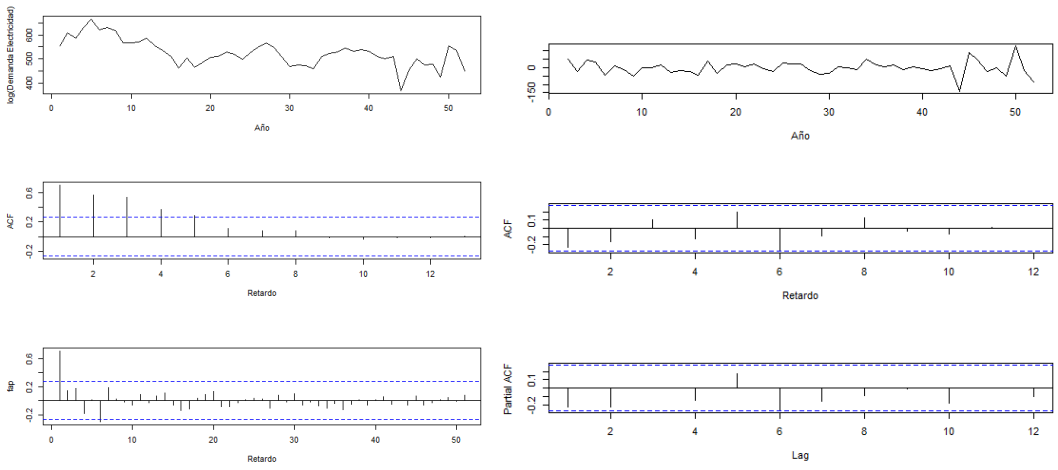

Figura 23. Serie original y serie diferenciada en la parte regular

Observando este último gráfico de las correlaciones, la estructura de modelo ARIMA que proponemos es un ARIMA (0, 1, 0), que es el mismo que posteriormente nos da la función de *best.arima.TSA.*

$$(1 - B)Y_t = c + a_t$$

Realizamos el ajuste que procede en ese caso, es decir, ajustamos un ARIMA (0, 1, 0) y solo incluimos en las variables dummy los festivos intrasemanales y los días de agosto, ya que las demás variables son nulas para los martes. Haciendo el ajuste vemos que tenemos uno de los dos coeficientes que no es significativamente distinto de cero y es el coeficiente 3, los días de agosto, por tanto debemos eliminarla para mejorar nuestro modelo. Y una vez hecho el ajuste obtenemos los siguientes resultados:

Tabla 12

arimax (x=martes11.ts, order= c (0, 1, 0), xreg = xreg [,-c (2, 4, 5, 6, 7, 3)])

Coefficients	f.intra
	-101.94
s.e.	14.88
$1.96 * s.e > \|coef\|$	3.49

Sigma^2 estimated as 886.3: log likelihood = -240.13, aic = 482.25

Ahora que nuestro coeficiente es significativo, los festivos entre semana, vemos que realmente va a tener un efecto importante en nuestro modelo, ya que cada vez que se dé un festivo los martes la demanda se va a reducir en 101.94 unidades. Afectando a nuestro modelo del siguiente modo:

$$(1 - B)Y_t = c + a_t + (-101.94 * f.intra_t)$$

Realizamos el análisis de residuos observando el Q-Q plot y vemos que parece que existe claramente normalidad en los residuos, además viendo el segundo gráfico de la Figura 24 podemos afirmar que existe independencia ya que todos los p-valores superan el límite de 0.05. Podemos entonces realizar los test de media igual a cero y de normalidad de los datos para confirmar nuestra afirmación.

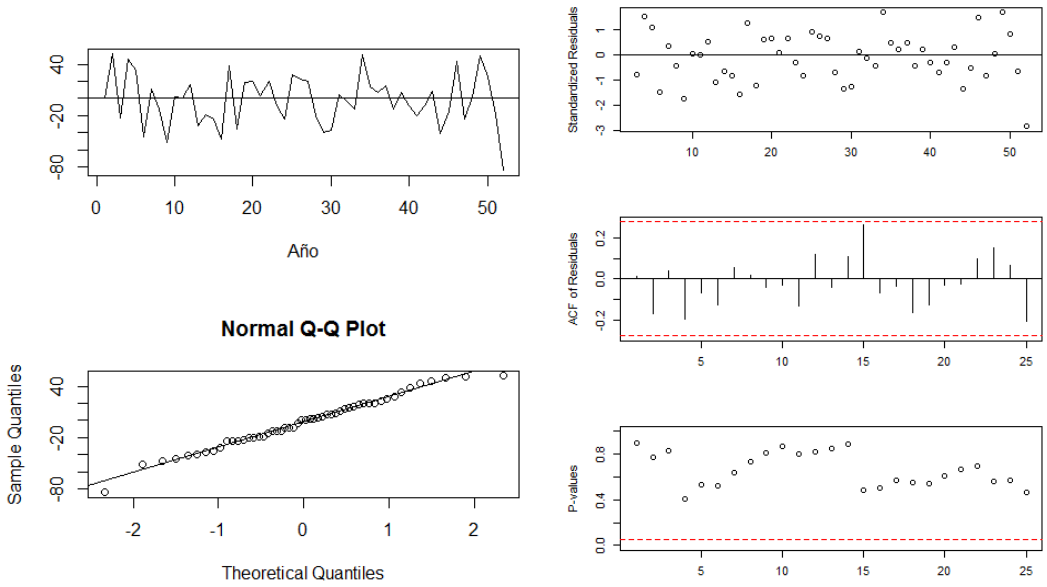

Figura 24. Q-Q plot y Ljung Box

Con los resultados obtenidos y presentados en la tabla que tenemos a continuación, no debemos rechazar la hipótesis de que la media sea igual a cero; y en cuanto a la normalidad se cumple para cualquier nivel de significación y para ambos test.

Tabla 13

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.623	0.8625	0.6946

4.1.2.2.3 ARIMA para los Miércoles

El panorama que nos presenta el miércoles es muy similar al del lunes ya que su comportamiento a lo largo del año es bastante cambiante. De nuevo nos fijamos que es una serie con tendencia y por tanto es necesario realizarle una diferenciación en la parte regular, de esta forma nos quedan unas correlaciones que como en los casos anteriores no tienen parte estacional y además son ruido blanco.

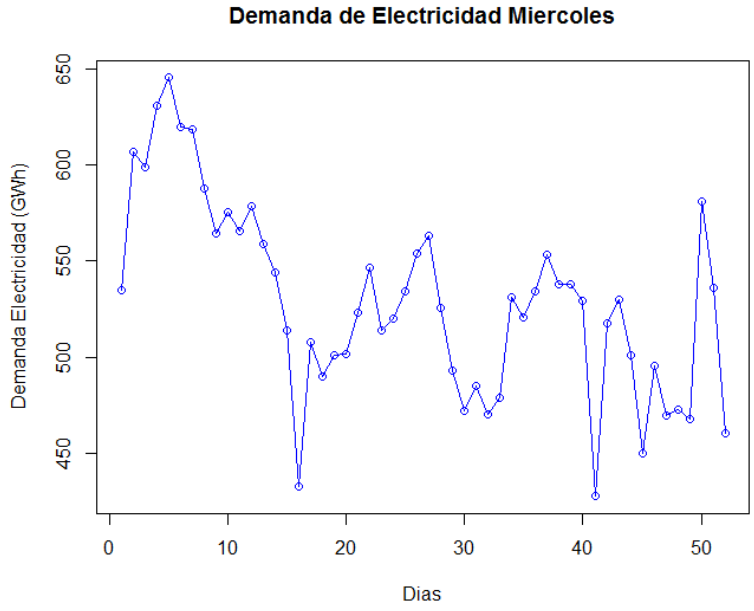

Figura 25. Demanda de Electricidad de los Miércoles de 2011

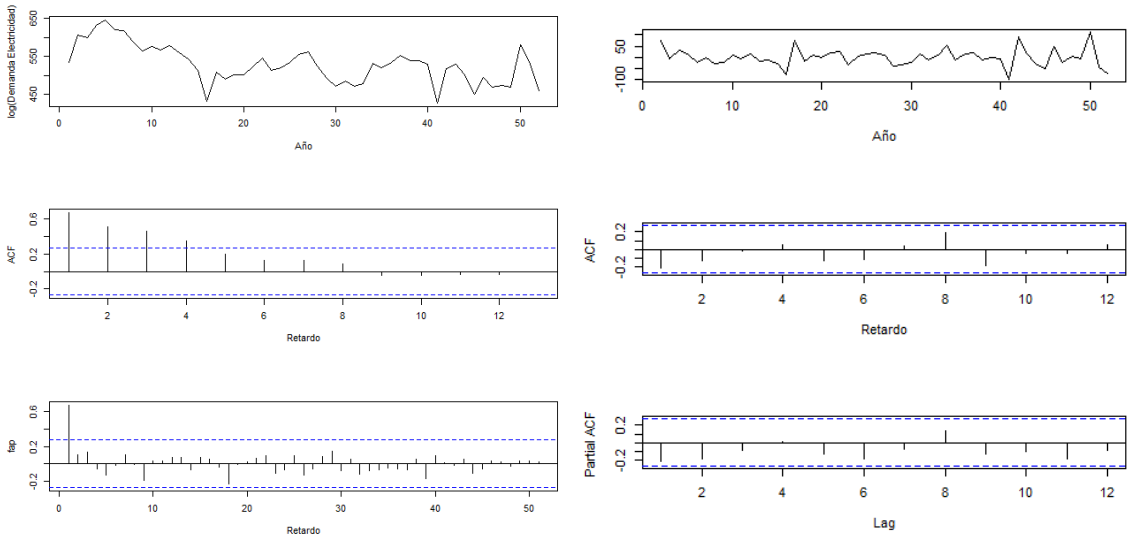

Figura 26. Serie original y serie diferenciada en la parte regular

Por tanto a la vista de las correlaciones obtenidas en el último gráfico, nuestro modelo propuesto será el ARIMA (0, 1, 0) que es el que obtenemos con la función que utilizamos hasta ahora.

$$(1 - B)Y_t = c + a_t$$

Hacemos el ajuste utilizando esta estructura del ARIMA, eliminando las variables que son nulas en este caso. Quedándonos con las mismas que para el martes, es decir, los festivos de entre semana y los días de agosto.

La situación se vuelve a repetir, y tenemos que eliminar la variable de agosto ya que su coeficiente no es significativo. De esta forma nos quedamos con un ajuste que nos muestra que un día festivo que caiga en miércoles va a provocar una reducción de 96.10 unidades en la demanda eléctrica total.

Tabla 14

arimax (x=miercoles11.ts, order= c (0, 1, 0), xreg = xreg [,-c (2, 4, 5, 6, 7, 3)])

Coefficients	f.intra
	-96.1040
s.e.	24.5119
$1.96 * s.e > \|coef\|$	2.0003

Sigma^2 estimated as 1202: log likelihood = -247.74, aic = 497.48

Con referencia a los datos de la tabla anterior, podemos afirmar que nuestro modelo quedaría de la siguiente forma:

$$(1 - B)Y_t = c + a_t + (-96.10 * f.intra_t)$$

Procedemos al análisis de los residuos, y vemos en el Q-Q plot que no está claro del todo que vayan a cumplir normalidad, sin embargo si cumplen independencia. Podemos por tanto realizar los test de media igual a cero y de normalidad de los datos, es decir, podríamos afirmar entonces que el modelo ARIMA (0, 1, 0) puede utilizarse para generar la serie de demanda de los miércoles.

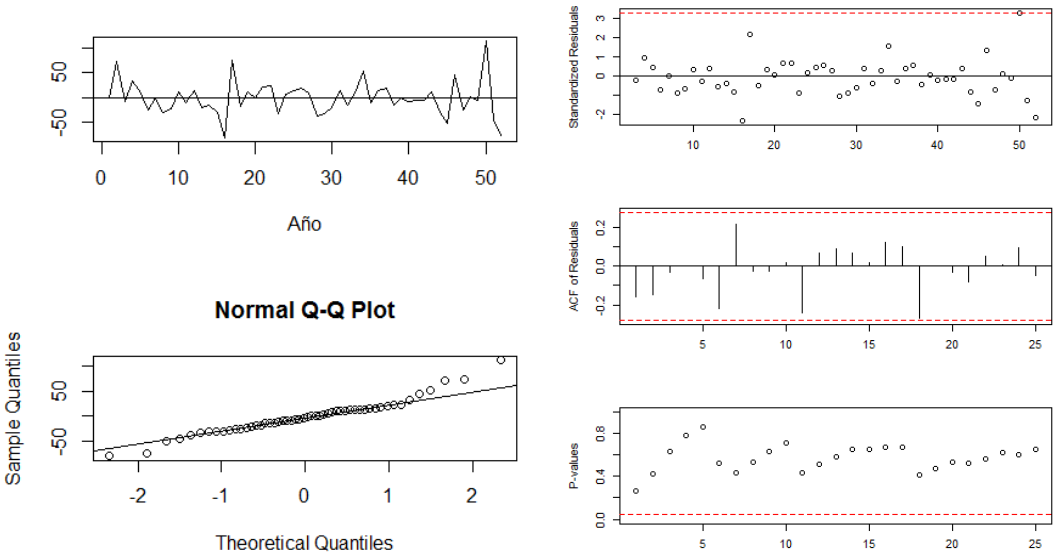

Figura 27. Q-Q plot y Ljung Box

En cuanto al test de la media igual a cero, debemos aceptar que se cumple la hipótesis, no obstante, no podemos decir que los residuos cumplan normalidad ya que se rechaza la hipótesis de normalidad para ambos test y para cualquier nivel de significación, excepto en el test de Shapiro Wilk en el que podríamos aceptar normalidad para un 1% significación.

Tabla 15

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.7671	0.002962	0.02329

4.1.2.2.4 ARIMA para los Jueves

Analizando lo que sucede los jueves del 2011, podemos ver que hay dos grandes reducciones en la demanda de electricidad, una se da el 21 de abril, que es jueves santo, y otro se da el día de la Inmaculada Concepción, el 8 de diciembre; los demás días se mantienen bastante constantes, exceptuando también los primeros jueves del año.

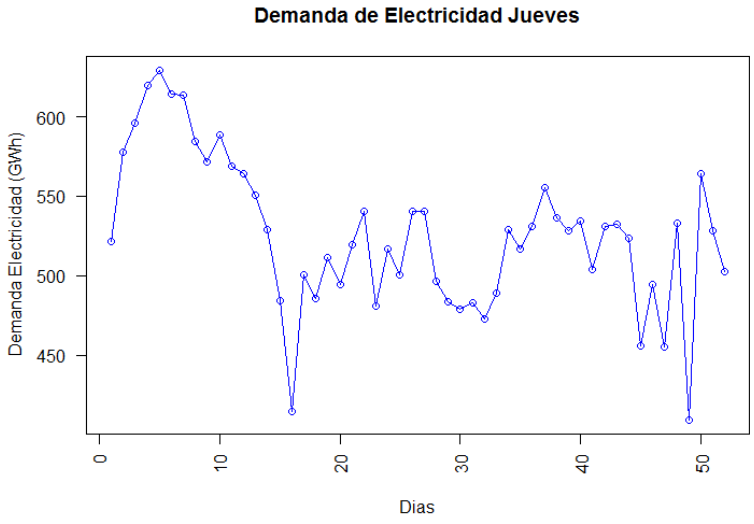

Figura 28. Demanda de Electricidad de los Jueves de 2011

Esta serie tiene una tendencia decreciente tal como se puede observar en el primer gráfico de la Figura 29. Ésta la eliminamos haciendo una diferenciación regular y obtenemos las correlaciones que presentamos en el segundo gráfico de esta figura, en el cual podemos ver que no hay parte estacional y es ruido blanco.

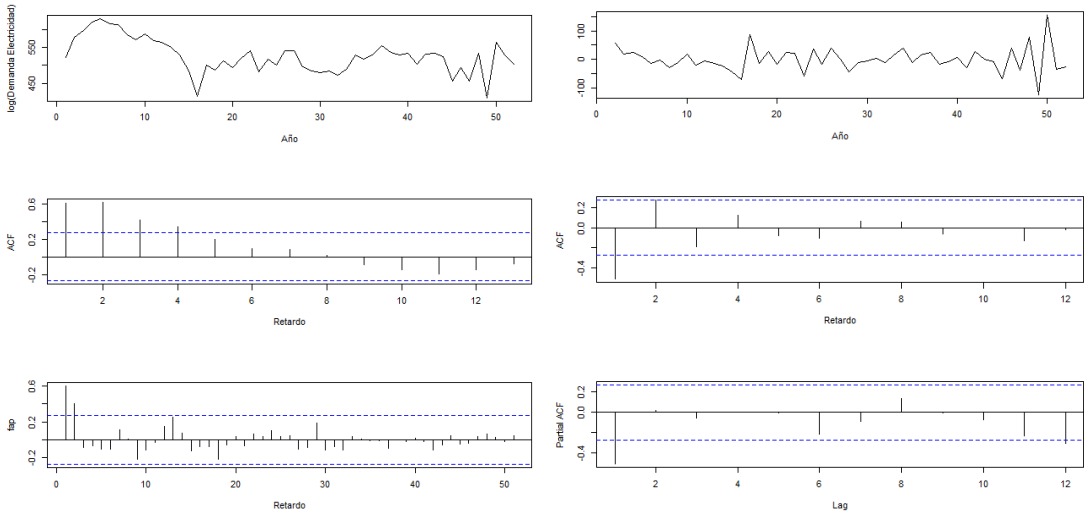

Figura 29. Serie original y serie diferenciada en la parte regular

Si tenemos que hacer un modelo tentativo a partir de este gráfico diremos que nuestro modelo podría ser un ARIMA (1, 1, 0) o un ARIMA (0, 1, 1) y estos son exactamente los dos modelos que obtenemos utilizando la función *best.arima.TSA.* El modelo que nos ofrece un BIC menor es el ARIMA (1, 1, 0), por tanto será el que utilicemos para realizar el ajuste del modelo.

$$(1 - \phi_1 B)(1 - B)Y_t = c + a_t$$

Para realizar el ajuste, eliminamos las variables dummy nulas y nos quedamos en este caso con los festivos intrasemanales, los días de agosto, y los festivos de Madrid.

En este caso, cuando hacemos el primer ajuste vemos que existen dos coeficientes que no son significativos el 3 y el 6, y con los valores que obtenemos sabemos que vamos a tener que eliminar los dos, por tanto eliminaremos los días de agosto y los festivos de Madrid.

Tabla 16

arimax (x=jueves11.ts, order= c (1, 1, 0), xreg = xreg [,-c (2, 4, 5, 7, 3)])

Coefficients	ar1	f.intra
	-0.44	-59.57
s.e.	0.12	13.39
$1.96 * s.e > \|coef\|$	1.78	2.26

Sigma^2 estimated as 958: log likelihood = -242.18, aic = 488.36

Llegamos al punto en que nuestros coeficientes son significativos y vuelve a ser únicamente la variable de los festivos de entre semana, con lo cual, si se da un festivo el jueves la demanda de electricidad se reducirá en 59.57 unidades. Esto podemos representarlo en nuestro modelo como sigue:

$$(1-(-0.44)*B)(1-B)Y_t = c + a_t + (-59.57 * f.intra_t)$$

Podemos comenzar a analizar entonces los residuos, para ver cómo actúan.

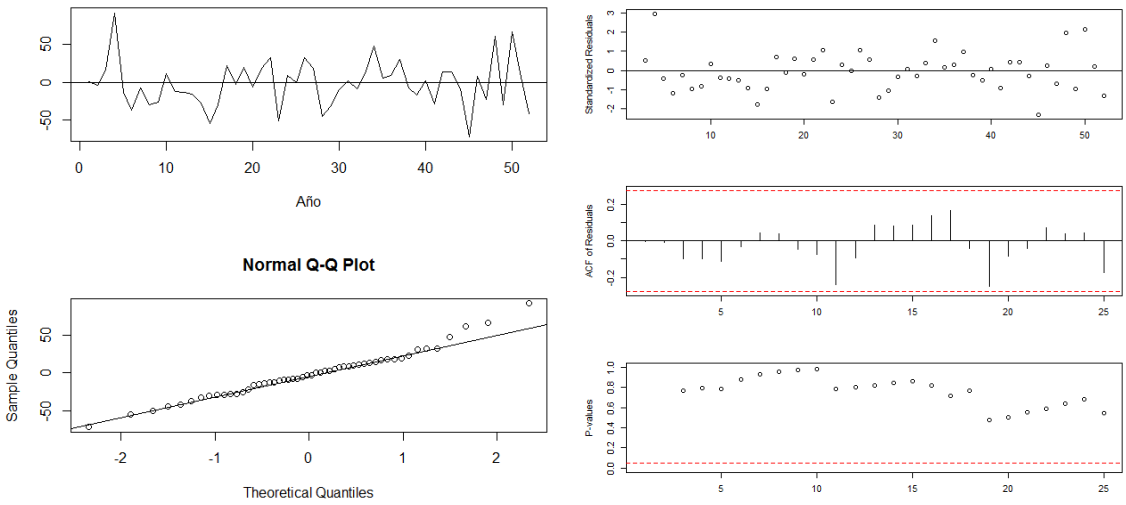

Figura 30. Q-Q plot y Ljung box

En primer lugar como siempre observamos el primer gráfico de la Figura 30, y vemos el Q-Q plot en el que parece indicarnos que va a existir normalidad, pero antes de realizar los test correspondientes debemos saber si se cumple independencia.

A la vista del segundo gráfico de la figura anterior, se puede afirmar que sí se cumple la hipótesis de independencia, por lo que se puede decir que el ARIMA (1, 1, 0) es un buen modelo para generar la serie de la demanda de los jueves.

Para los jueves, la hipótesis de que la media es igual a cero se cumple así como la de normalidad que se cumple para los dos test y para cualquier nivel de significación.

Tabla 17

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.6485	0.1277	0.3406

4.1.2.2.5 ARIMA para los Viernes

Analizando los viernes del 2011 podemos observar que su comportamiento es uno de los más estables de la semana, exceptuando una fuerte bajada de la demanda de electricidad que se da el día de viernes santo, el 22 de abril.

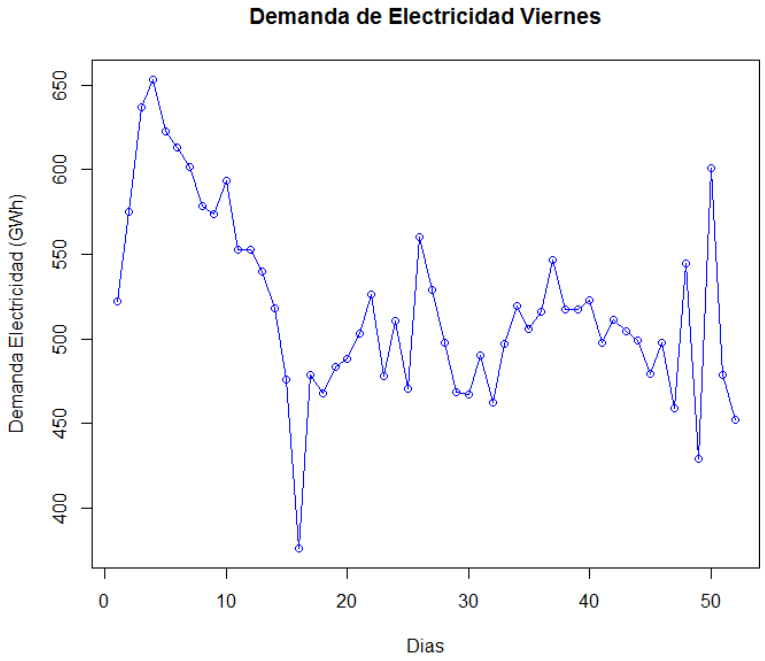

Figura 31. Demanda de Electricidad de los Viernes de 2011

Representando las correlaciones parciales en la Figura 32 vemos que tiene tendencia decreciente y por tanto debemos hacer una diferenciación en la parte regular para eliminarla, de esta manera obtendremos el último gráfico.

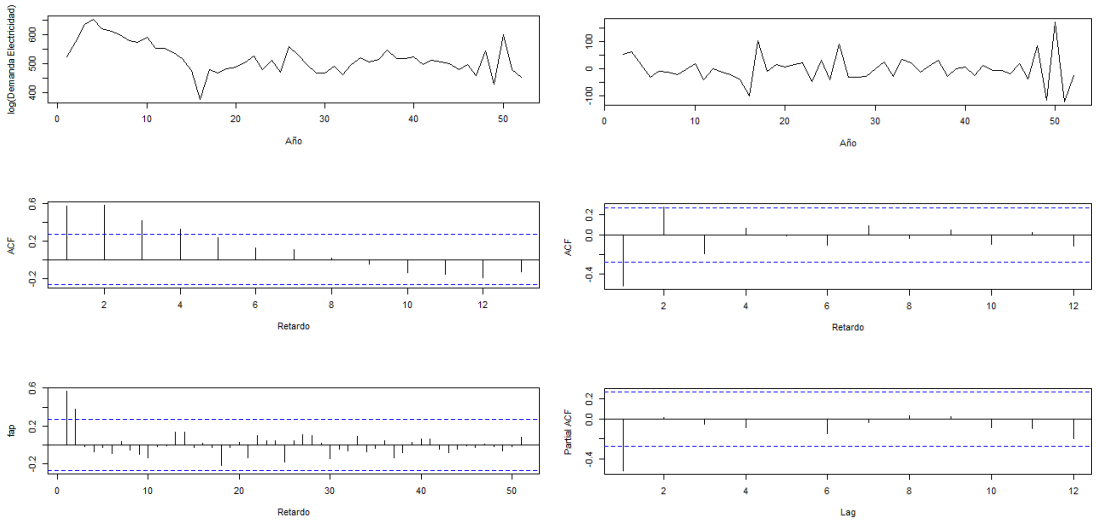

Figura 32. Serie original y serie diferenciada en la parte regular

A la vista de este último gráfico podemos proponer dos estructuras para nuestro modelo de ajuste. Un ARIMA (1, 1, 0) o un ARIMA (0, 1, 1). Utilizando la función *best.arima.TSA* y eliminando todas las dummy nulas, obtenemos los mismos modelos tentativos que los que nosotros propusimos a partir de las fas y las fap, sin embargo, debemos coger el modelo con el menor BIC, que es el ARIMA (1, 1, 0).

$$(1 - \phi_1 B)(1 - B)Y_t = c + a_t$$

Realizamos el ajuste y éstos son los valores que obtenemos de los coeficientes y de los errores estándar una vez hemos eliminado la variable de Agosto, ya que al principio del ajuste vemos que esta variable no resulta significativa en nuestro estudio.

Tabla 18

arimax (x=viernes11.ts, order= c (1, 1, 0), xreg = xreg [,-c (2, 4, 5, 6, 7, 3)])

Coefficients	ar1	f.intra
	-0.56	-112.76
s.e.	0.11	30.76
$1.96 * s.e > \|coef\|$	2.46	1.87

Sigma^2 estimated as 1420: log likelihood = -252.1, aic = 508.2

Al realizar el ajuste dos veces vemos que los coeficientes son significativos y de nuevo sólo nos quedamos con uno de ellos. Sabemos que cuando un viernes sea festivo nuestra demanda de electricidad se reducirá en 112.76 unidades.

Procedemos a analizar los residuos para saber si el modelo que hemos propuesto es apropiado para generar la serie de la demanda de electricidad de los viernes. Pero no sin antes mostrar cómo quedaría formulado nuestro modelo para este caso:

$$(1 - (-0.56) * B)(1 - B)Y_t = c + a_t + (-112.76 * f.intra_t)$$

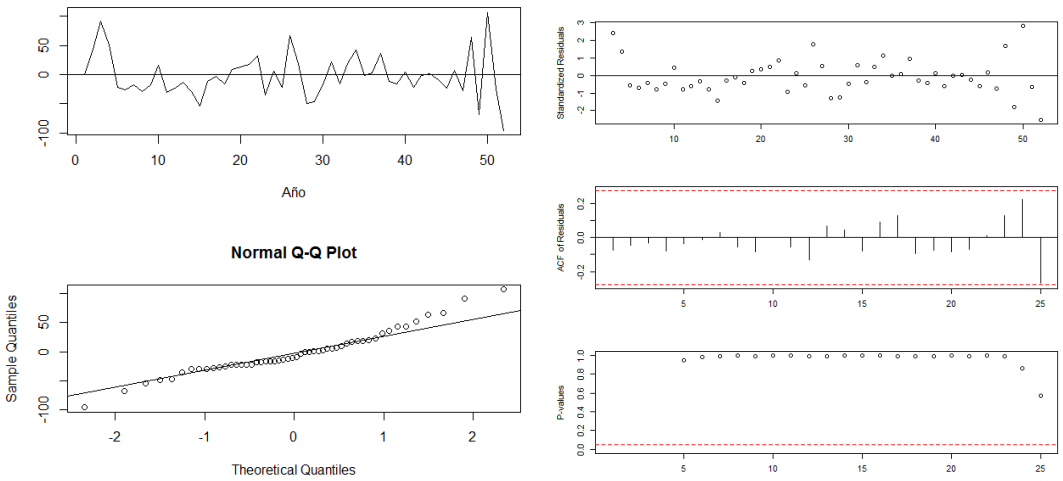

Figura 33. Q-Q plot y Ljung Box

A partir de los gráficos obtenidos, parece que los residuos cumplen normalidad e independencia por tanto podemos afirmar que el ARIMA (1, 1, 0) es un buen modelo para generar la serie de la demanda de los viernes. Además podemos realizar los siguientes test, de que la media es igual a cero y de normalidad de nuestros datos.

La hipótesis de que la media es igual a cero se cumple, como era de esperar. En cuanto a la normalidad podemos decir que se cumple para un 1% y un 5% de significación, en cambio para un 10% ya no se cumpliría para ninguno de los test.

Tabla 19

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.7024	0.05817	0.05916

4.1.2.2.6 ARIMA para los Sábados

En el gráfico siguiente podemos observar cómo actúan los sábados de 2011 y en general podemos decir que se comporta de una forma un poco aleatoria, ya que existen fuertes variaciones a lo largo del año y sin motivo aparente, exceptuando uno de los valores más bajos en la demanda que se da el 19 de marzo, en San José.

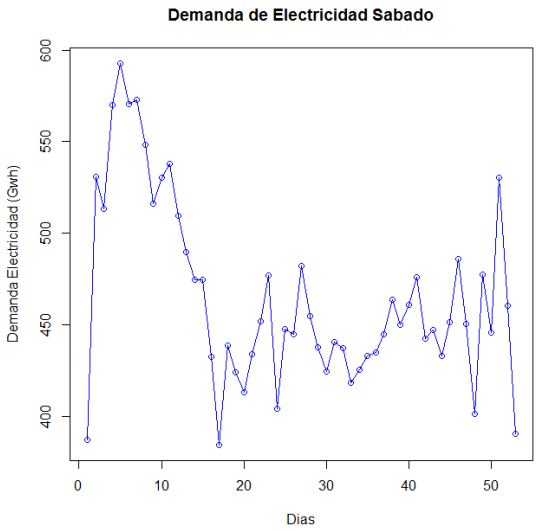

Figura 34. Demanda de Electricidad de los Sábado de 2011

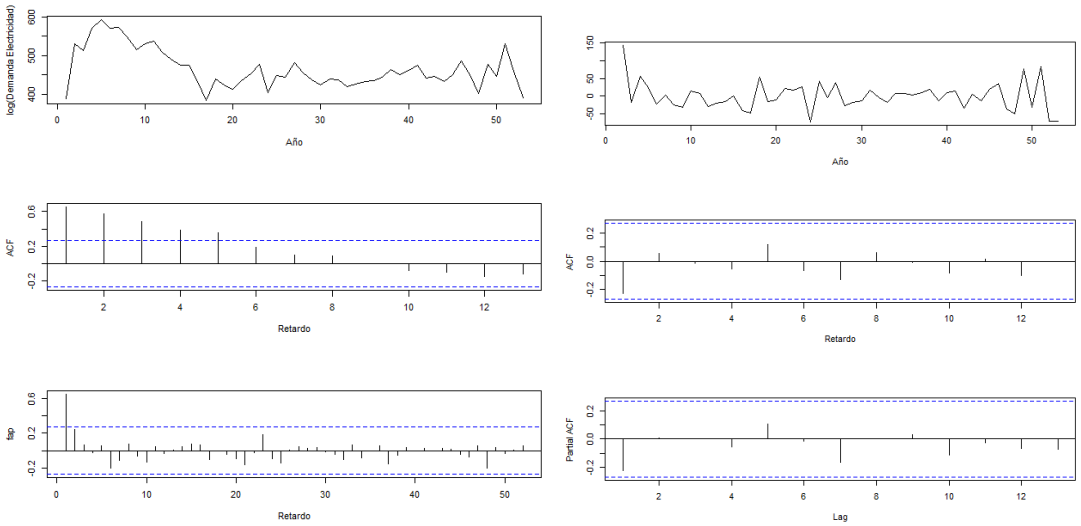

Figura 35. Serie original y serie diferenciada en la parte regular

En los gráficos anteriores también vemos la existencia de la tendencia que caracteriza a todos los modelos que hemos analizado anteriormente y que debemos corregir diferenciando la parte regular de nuestros datos. De esta forma obtenemos las correlaciones del segundo gráfico de la Figura 35, con las que podemos afirmar que no existe parte estacional y podemos proponer el modelo tentativo ARIMA (0, 1, 0).

Utilizando la función *best.arima.TSA*, y eliminando las variables dummy nulas de esta ocasión, nos quedamos simplemente con los festivos en sábado, los días de agosto y los festivos de Valencia. El único modelo que nos propone la función es el ARIMA (0, 1, 1), por tanto será este el modelo que utilicemos para realizar el ajuste.

$$(1 - B)Y_t = c + (1 + \theta_1 B)a_t$$

Cuando realizamos el ajuste apreciamos que hay dos coeficientes que no son significativos y por tanto que vamos a tener que eliminar. Nosotros a la hora de hacer el estudio lo vamos a ir haciendo de uno en uno como explicamos anteriormente, sin embargo ahora presentaremos la tabla de coeficientes y valores estándar que nos saldría al final (como en los días anteriores), en la que vemos que solo serán significativos los festivos del sábado. Tal como podemos observar en la tabla y recordando valores pasados, podríamos afirmar que hasta ahora es uno de los coeficientes que presenta mayor significación, puesto que un festivo en sábado hará disminuir la demanda eléctrica en 148.08 unidades.

Tabla 20

arimax (x=viernes11.ts, order= c (0, 1, 1), xreg = xreg [,-c (1, 4, 5, 6, 7)])

Coefficients	ma1	f.sab
	-0.45	-148.04
s.e.	0.14	32.21
$1.96 * s.e > \|coef\|$	1.54	2.34

Sigma^2 estimated as 1010: log likelihood = -248.38, aic = 500.77

Como vemos, los coeficientes son significativos por tanto si presentamos el modelo ajustado final nos quedaría de la siguiente forma:

$$(1 - B)Y_t = c + (1 + (-0.45)B)a_t + (-148.04 * f.sab_t)$$

En los gráficos siguientes vemos el comportamiento de los residuos, que parece que cumplen los contrastes de normalidad e independencia, por lo que podemos afirmar que el ARIMA (0, 1, 1) es un modelo adecuado para generar la serie de demanda de electricidad de los sábados.

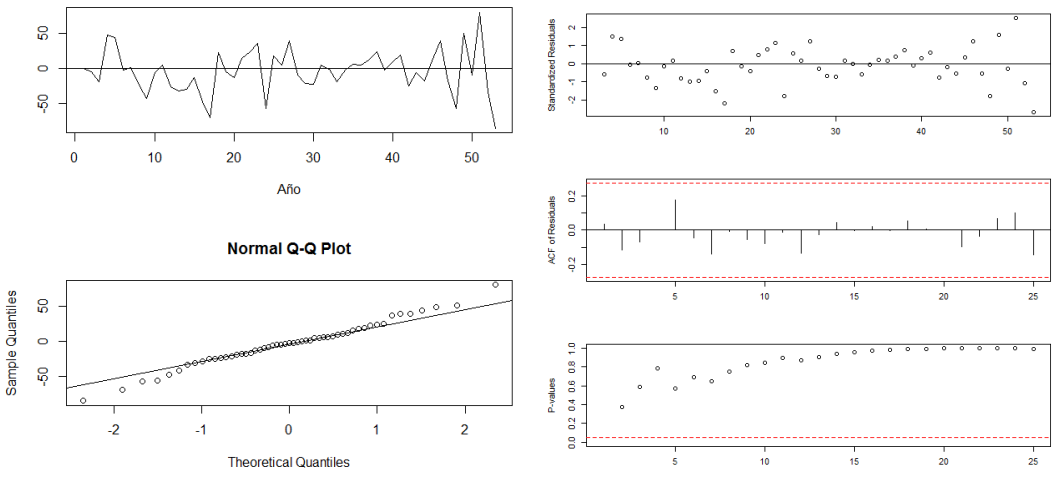

Figura 36. Q-Q plot y Ljung Box

Al cumplirse la hipótesis de independencia de los residuos podemos realizar los siguientes test, en los que tenemos como hipótesis nula que la media es cero y que nuestros datos son normales.

Observando la siguiente tabla, vemos que se cumplen ambas hipótesis para cualquier nivel de significación.

Tabla 21

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.4031	0.7154	0.8364

4.1.2.2.7 ARIMA para los Domingos

Y finalmente tenemos representado el comportamiento de los domingos, que según lo que podemos apreciar en la gráfica es muy similar a lo que ocurre los sábados.

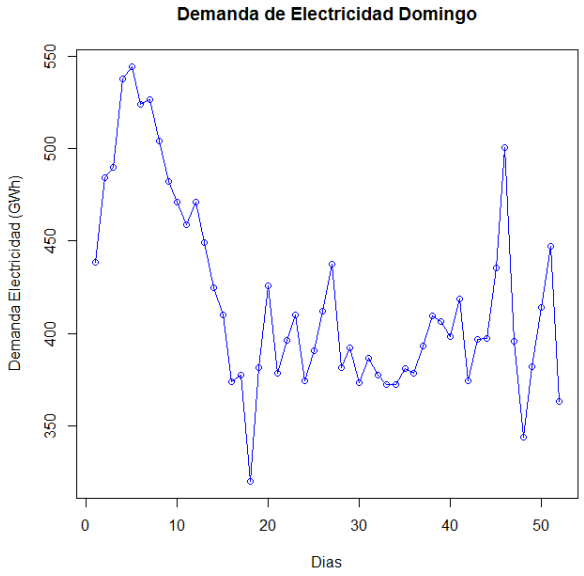

Figura 37. Demanda de Electricidad de los Domingo de 2011

De nuevo observando la Figura 38 vemos que existe tendencia como en los demás días de la semana, con lo cual es necesario diferenciar la parte regular para eliminar su efecto. De esta forma tenemos el segundo gráfico de esta figura en el que tampoco se observa que exista parte estacional ya que es ruido blanco. La estructura del modelo que podríamos proponer es un ARIMA (0, 1, 0) para realizar el ajuste del modelo.

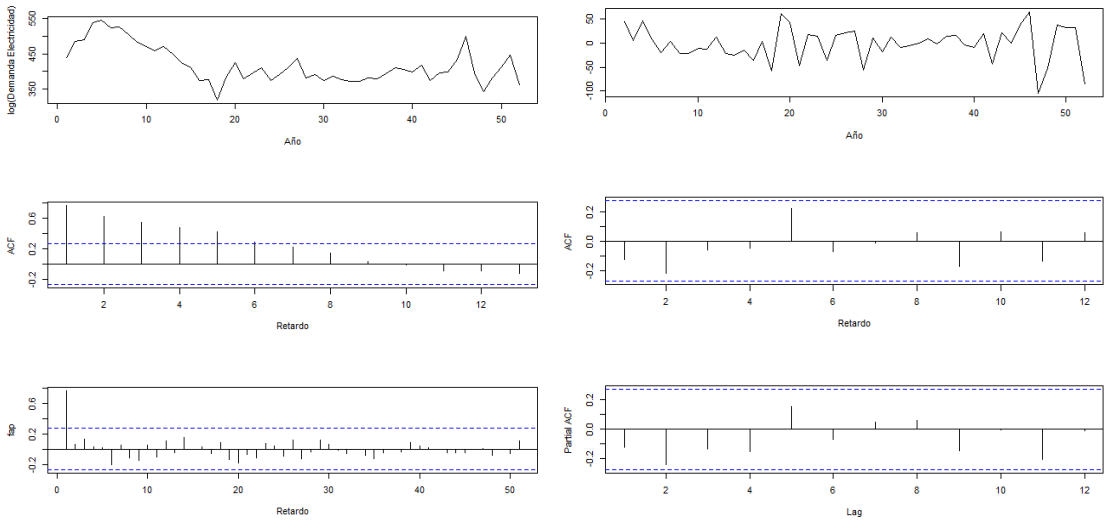

Figura 38. Serie original y serie diferenciada en la parte regular

Para comprobar si nuestro modelo tentativo es acertado utilizamos la función *best.arima.TSA*, después de eliminar todas las variables dummy que eran nulas, y quedarnos solamente con la de los días de agosto. El modelo tentativo que nos propone la función es ARIMA (0, 1, 0), por tanto procedemos a realizar el ajuste.

$$(1 - B)Y_t = c + a_t$$

Tabla 22

arimax (x=domingo11.ts, order= c (0, 1, 0), xreg = xreg [,-c (1, 2, 4, 5, 6, 7)])

Coefficients	Agst
	-3.34
s.e.	24.68
$1.96 * s.e > \|coef\|$	0.06

Sigma^2 estimated as 1219: log likelihood = -248.09, aic = 498.19

Como ya ocurría en el resto de días de la semana, esta variable de los días de agosto, no es significativamente distinta de cero por lo tanto debemos eliminarla del modelo, quedándonos el ajuste sin coeficientes. Esto quiere decir que ninguna de las variables dummy que hemos creado va a influir en la demanda de los domingos, por lo que el modelo ajustado quedaría de la misma forma que veíamos antes:

$$(1 - B)Y_t = c + a_t$$

Observando los siguientes gráficos vemos que parece que existe un claro cumplimiento de normalidad de los residuos y podemos confirmar que se cumple

independencia. Por tanto el modelo ARIMA (0, 1, 0) es un modelo apropiado para generar la serie de demanda de los domingos.

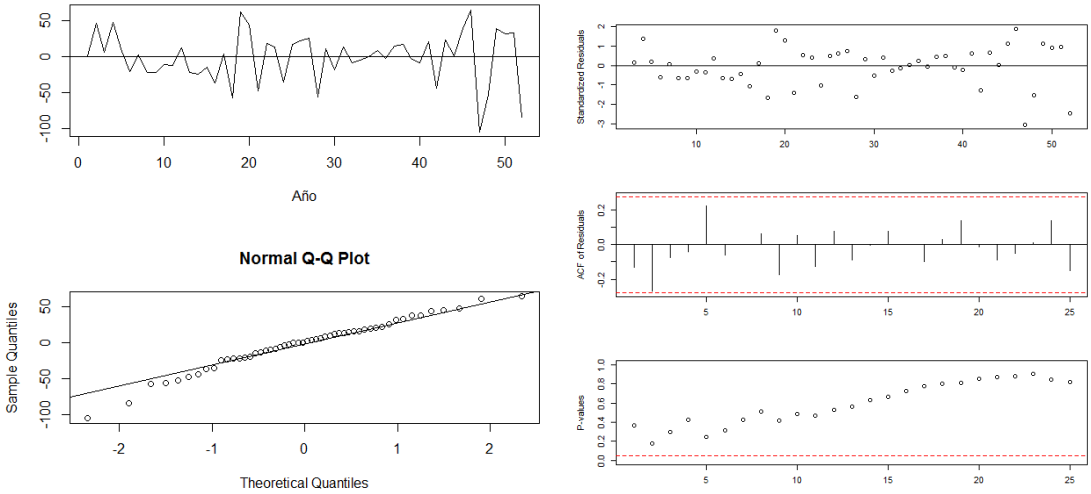

Figura 39. Q-Q plot y Ljung Box

Como la hipótesis de independencia anterior se ha aceptado, podemos realizar los test de la media igual a cero y de normalidad de nuestros datos.

Observando la tabla que sigue, manifestamos que la media es igual a cero y que nuestros datos son normales para cualquier nivel de significación y para ambos test.

Tabla 23

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.7643	0.1213	0.2583

Media del porcentaje de error absoluto (MAPE)

A la vista de la tabla siguiente podemos decir que el ARIMA de siete modelos no es un buen modelo para realizar las predicciones del 2012.

Dando una visión general vemos que los errores oscilan entre valores del 5.57% registrado el sábado del tercer trimestre y el 19.01% registrado el domingo del cuarto trimestre. Podemos afirmar entonces que además de que el ARIMA para cada día de la semana no hace buenas predicciones en general, éstas tienen valores muy dispersos. Las mejores predicciones se dan el sábado del tercer trimestre y el lunes también de dicho trimestre. Mientras que los mayores errores y por tanto las peores predicciones se dan el domingo y el miércoles del cuarto trimestre.

Observando la tabla desde un punto de vista trimestral nos llama la atención que, exceptuando el tercer trimestre, la media de error en todos ellos sobrepasa el 10% que es el doble del umbral que habíamos establecido como máximo para considerar una predicción como buena. Dándose las peores predicciones en el último trimestre.

Analizando ahora los datos por día de la semana debemos destacar que el comportamiento de los datos que habíamos visto hasta ahora cambia de forma drástica, ya que las mejores predicciones se producen los lunes, miércoles y sábados, cuando los lunes y sábados estaban siempre entre los días con mayores errores de predicción. En cambio, en los días intrasemanales salvo el miércoles (que acabamos de mencionar) se dan las mayores medias de error de la semana.

Podemos calificar los resultados obtenidos por este método de muy deficientes.

Tabla 24

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	6.488114	11.029704	6.180765	9.785769	8.371088
Martes	9.833596	10.296150	9.423031	10.476727	10.007376
Miércoles	9.986361	8.548664	7.041598	14.164745	9.935342
Jueves	12.264701	11.594508	8.746307	13.336433	11.48548725
Viernes	13.470336	12.475604	6.252036	13.229479	11.35686375
Sábado	8.567597	6.539871	5.574540	12.474112	8.28903
Domingo	14.005371	13.032597	7.465162	19.009092	13.3780555
Total	10.65943943	10.50244257	7.240491286	13.21090814	

4.1.3. Modelo PLRM: Modelo de Regresión Parcialmente Lineal

El modelo PLRM tal como se explica en la sección anterior está formado de una componente paramétrica y otra no paramétrica. Se caracteriza por su flexibilidad y por la sencilla interpretación del efecto de cada variable lineal, además mitiga el problema de la dimensionalidad, por tanto es un modelo adecuado para evaluar datos en numerosas situaciones y por eso lo hemos escogido para estudiar nuestros datos.

En este trabajo vamos a realizar dos tipos de predicción a través del modelo de regresión parcialmente lineal, en primer lugar lo vamos a hacer con un solo modelo para nuestras 366 observaciones, y en segundo lugar lo haremos con siete modelos, uno para cada día de la semana.

Sin embargo lo primero que tenemos que hacer antes de empezar a hacer los ajustes oportunos para realizar las predicciones, es observar que nuestros datos tienen tendencia y por tanto debemos eliminar su efecto al principio, para evitar problemas posteriores. Tenemos que hacer entonces un filtro de tendencia, es decir, debemos diferenciar los datos (incluidas las variables dummy que utilizamos en el ARIMA).

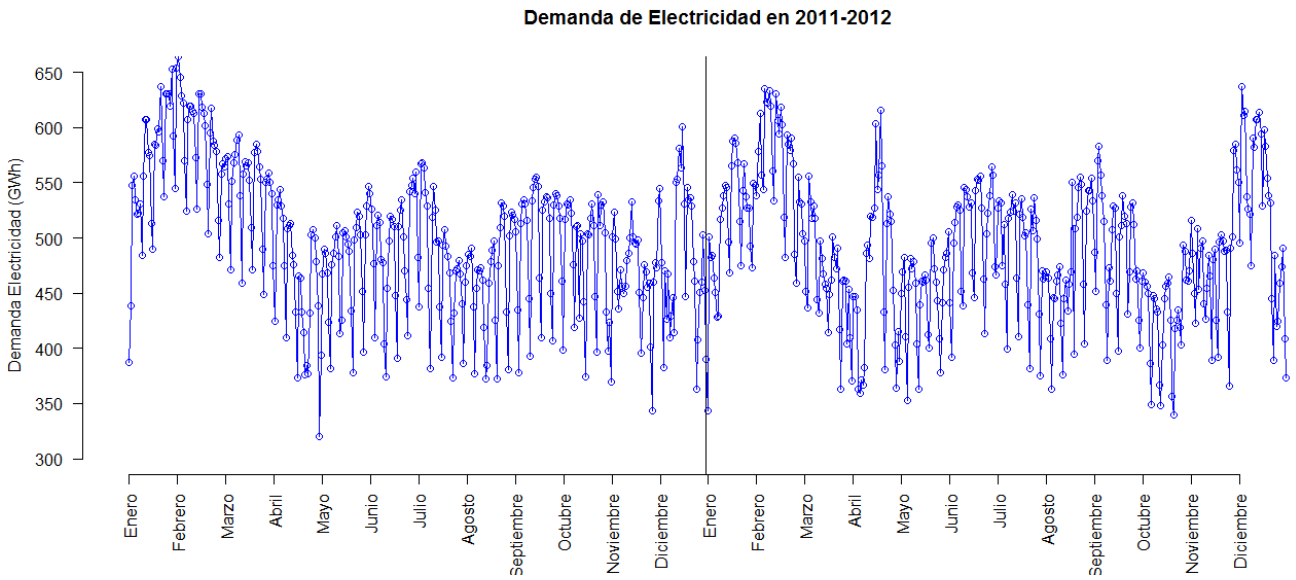

Figura 40. Demanda de Electricidad (2011 - 2012)

Es importante organizar los datos para comenzar a realizar las predicciones que necesitamos.

4.1.3.1. PLRM: un solo modelo

Tenemos los datos diferenciados y por tanto ya hemos eliminado el efecto de la tendencia, pero debemos corregir también el efecto de la estacionalidad. En los modelos ARIMA para corregirlo bastaba con hacer una diferenciación en la parte estacional, sin embargo en este modelo es necesario crear variables dummy para indicar los períodos con características específicas. Las variables que vamos a crear serán de la siguiente forma:

Variable D1: Esta variable tomará valor 1 si la diferencia es entre un lunes y un domingo, y cero en otro caso.

Variable D2: Esta variable tomará valor 1 si la diferencia es entre un sábado y un viernes, y cero en otro caso.

Variable D3: Esta variable tomará valor 1 si la diferencia es entre un domingo y un sábado, y cero en otro caso.

Estas variables D que hemos creado lo que hacen es diferenciar las demandas en días en los que la demanda se comporta de forma distinta.

Una vez que ya hemos creado las variables, las incorporamos sin diferenciar en nuestra matriz de datos totales para poder trabajar con ellas. Estas variables se introducen en nuestra matriz sin diferenciar porque cuando las creamos ya tuvimos en cuenta la diferenciación hecha anteriormente. También introducimos en la matriz de datos totales la variable T (en la última columna), que es la demanda diferenciada retardada un instante.

En la matriz de datos con la que vamos a trabajar tenemos además las variables dummy que utilizamos en el ARIMA, nuestra intención era utilizar únicamente las que fueron más influyentes, aunque al realizar el ajuste vemos que sólo podemos utilizar 3 de ellas, ya que con las demás nos da error la función. Por tanto de las primeras variables dummy que hemos creado vamos a poder utilizar exclusivamente, la variable de los festivos intrasemanales, la de los festivos en Cataluña y la de los festivos en Madrid.

Para hacer el ajuste y obtener los β y m con los que posteriormente haremos las predicciones, utilizamos la función *plrm.est*. Vamos a ir cogiendo siempre los 363 datos anteriores al día que queremos predecir. Es decir, si queremos hallar el dato del 1 de enero de 2012, cogeremos los datos desde el 1 hasta el 363, que corresponde con los datos desde el 3 de enero de 2011 (diferenciado), hasta el 31 de diciembre de 2011 (también diferenciado).

Cuando tenemos ya los valores de β y m hallados para cada uno de los datos, hacemos las predicciones utilizando los valores de las variables dummy del año 2012, que es lo que queremos predecir. Entonces el procedimiento es multiplicar los β por los valores de las dummy en 2012, siempre que el valor de estás sea 1 y no 0 (de esta forma la variable dummy no tendría ningún efecto en la predicción), y sumarle el m que corresponda en cada caso. Sin embargo debemos recordar deshacer la diferenciación que realizamos al principio para obtener realmente las predicciones, y hallar luego los errores cometidos en cada una de ellas.

Los errores que hemos obtenido, utilizando la misma fórmula descrita en el modelo NAIVE, son los de la tabla siguiente:

Tabla 25

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	3.785505	4.014922	4.504886	8.553456	5.214692
Martes	3.390196	4.581161	2.053885	2.924693	3.237484
Miércoles	1.756138	1.948655	3.982152	4.657190	3.086034
Jueves	2.745969	2.777426	4.353995	4.480062	3.589363
Viernes	2.880451	4.306825	2.392444	5.273870	3.713397
Sábado	3.031510	4.545335	4.849077	7.623738	5.012415
Domingo	6.301880	4.557819	6.470294	6.997874	6.081967
Total	3.413093	3.818878	4.112585	5.817336	

Observando esta tabla de errores no apreciamos grandes diferencias con los modelos anteriores (excepto con el ARIMA de siete modelos), pero si hay algún que otro matiz que comentar.

En primer lugar como siempre nos fijamos en la parte más desglosada de la tabla y aquí podemos ver, que la media absoluta de errores oscila entre valores del 1.77% y 8.51%. Siendo las mejores predicciones las de los miércoles tanto en el primer como en el segundo trimestre con valores del 1.77% y 1.97% respectivamente. En cuanto a las peores predicciones coinciden perfectamente con lo que ya veíamos también en el estudio anterior, se dan el sábado y el lunes del último trimestre.

Es relevante destacar que el comportamiento de los datos por trimestres sufren una pequeña variación, puesto que generalmente las peores predicciones se daban siempre en el segundo y el cuarto trimestre, mientras que en este caso se dan en los dos últimos trimestres; no podemos decir que exista una gran diferencia entre los errores del segundo y del tercer trimestre pero si la hay con respecto al cuarto trimestre, hecho que debemos tener en cuenta porque al final queremos comparar los resultados de forma óptima y ver qué modelo predice mejor cada momento.

Por último viendo los errores totales de cada día de la semana, sólo podemos comentar que vuelve a cumplirse el patrón de que los días intrasemanales tienen menor porcentaje de error que el resto de días y en este caso hay otro aspecto importante, ya que el domingo tiene peores predicciones que el sábado y el lunes, que más o menos tienen el mismo porcentaje de error.

Fijándonos ahora en la tabla siguiente, donde tenemos la media por trimestres de los días intrasemanales, podemos observar que se vuelve a repetir el patrón que veíamos en el NAIVE, es decir, las peores predicciones se dan en el último trimestre, mientras que las mejores se dan en el primero, y con bastante diferencia. En cuanto a la media total de este grupo de días, tenemos que admitir que las predicciones son buenas, puesto que sólo se comete un error del 3.4%.

Tabla 26

Media del porcentaje de error absoluto (MAPE) de Martes a Viernes

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Ma-Vi	2.693188	3.403517	3.195619	4.333954	3.40657

4.1.3.2. PLRM: siete modelos

En esta ocasión vamos a utilizar un modelo de regresión parcialmente lineal para cada día de la semana. Sin embargo hay un gran cambio con respecto al modelo único de regresión parcialmente lineal, ya que al utilizar un modelo para cada día de la semana, dentro de nuestros modelos no va a existir la componente estacional que antes teníamos al analizar toda la semana en su conjunto. Esto quiere decir que no vamos a utilizar las variables dummy (D1, D2 y D3) creadas antes, para este tipo de modelo.

En cuanto al procedimiento para hacer el ajuste, hallar las predicciones y posteriormente los errores, va a ser el mismo que utilizábamos en el modelo anterior.

Las variables que podemos introducir en cada modelo no son siempre las más significativas, ya que en alguna de las variables más significativas dan error, como nos ocurría en el modelo anterior. Por ejemplo el lunes no se puede introducir la variable de festivos intrasemanales porque el modelo no funciona, por eso introducimos los festivos de Cataluña y Madrid, en cambio el martes por ejemplo, sí que ya podemos introducir las 3 variables sin ocasionar ningún problema. Lo más destacable que nos ocurre haciendo estos análisis es que el sábado y el domingo no podemos introducir ninguna de las variables dummy del ARIMA, por tanto el ajuste se hará sin variables dummy.

En estos análisis, exceptuando el modelo de los sábados y domingos, los ajustes se harán mediante la función *plrm.est* ya que tenemos variables explicativas de la variable

respuesta. En el caso de los sábados y domingos, al no tener variables explicativas, el ajuste se hará mediante la función *np.est*.

La tabla de media de errores que obtenemos es la que se muestra a continuación:

Tabla 27

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	6.002032	9.442155	7.796669	9.958455	8.299827
Martes	6.643600	6.009589	7.362390	6.310143	6.581430
Miércoles	7.691458	6.211748	7.732196	9.213954	7.712339
Jueves	8.106958	5.864724	5.999326	11.735859	7.926716
Viernes	10.193586	5.336611	4.289699	9.497549	7.329361
Sábado	8.18830	5.69567	5.67619	11.23550	7.698915
Domingo	8.927217	6.699880	6.988820	15.952326	9.642060
Total	7.964735	6.465768	6.549327	10.557683	

Centrándonos ahora en los datos reflejados arriba en la tabla 34, vemos que el porcentaje de error es considerablemente alto puesto que supera en todos los casos (con la excepción del viernes del tercer trimestre) el umbral del 5% que habíamos establecido para considerar una predicción como buena.

Por otra parte vemos que el dato sobre el que se concentra de nuevo el mayor porcentaje de error vuelve a ser en los domingos del 4º trimestre. No obstante se refleja que el valor más próximo a la realidad se registra los viernes del tercer trimestre, hecho que no habíamos visto hasta ahora.

Otro dato curioso que debemos destacar es que observando la media diaria, las mejores predicciones se dan los martes, viernes y sábados; por lo que se rompe el patrón que se venía estableciendo hasta el momento.

Observando la tabla desde un punto de vista trimestral vemos que las peores predicciones se siguen dando con bastante diferencia en el cuarto trimestre, ya que se alcanza un error del 10.55%.

4.2. Modelos de predicción con variables exógenas

Cuando pensamos en la demanda eléctrica y en los factores que pueden afectar a la misma, es inevitable considerar la temperatura como uno de los más influyentes. Esto es porque tanto en verano, con el aumento de demanda de electricidad generada por los climatizadores, como en invierno, con el aumento de demanda eléctrica provocada por los aparatos calefactores, la demanda se ve afectada y aumentada por los efectos de la temperatura.

Por este motivo vamos a añadirle el efecto que produce la temperatura a nuestro análisis, en cuanto a la variación de la demanda eléctrica, para poder observar si el efecto que ésta tiene sobre la demanda es la esperada.

Podemos encontrar dos características principales en cuanto a la relación que existe entre la temperatura y la demanda de electricidad.

La primera es que es una relación no lineal. Tenemos dos valores críticos entre los cuales se encuentra la zona de confort de temperatura o zona neutral (Cancelo, Espasa & Grafe, 2008). El primer valor crítico es de 20C (68F), ya que si la temperatura es menor que este punto entraríamos en una zona fría, mientras que el otro valor crítico es de 24C (75.2F), por encima del cual estaríamos en una zona de calor. Por tanto nuestra función de respuesta va a ser no lineal en las zonas frías y cálidas.

Y en segundo lugar cuando tenemos datos diarios, como en nuestro caso, la demanda para un día *t* va a depender de la influencia que tenga la temperatura observada en *t, t-1, t-2,…,t-h.* Ver por ejemplo Gross and Galiana (1987).

También se podría tener en cuenta que va a existir una temperatura con un valor suficientemente alto (bajo) para obligar a todos los sistemas de refrigeración (calefacción) a trabajar a su completa capacidad, por lo que en esos momentos cualquier incremento (decrecimiento) de la temperatura no tendría efectos adicionales de la demanda. Sin embargo este efecto no vamos a incluirlo en nuestro modelo porque lo complicaría y no nos interesa.

Otro factor que debemos señalar es que la relación que existe entre la temperatura y la demanda eléctrica de un día va a ser diferente dependiendo de si se trata de un día de trabajo, un fin de semana o un festivo, ya que los sistemas de refrigeración o calefacción por ejemplo son muy diferentes en los lugares de trabajo y en la residencias privadas (Smith, 2000).

Vamos a hacer un breve análisis general observando el comportamiento de nuestros datos de la demanda junto con la evolución de la temperatura en el mismo año, 2011. Los datos de la temperatura son resultados obtenidos a partir de la información cedida por la Agencia Estatal de Meteorología; Ministerio de Agricultura, Alimentación y Medio Ambiente.

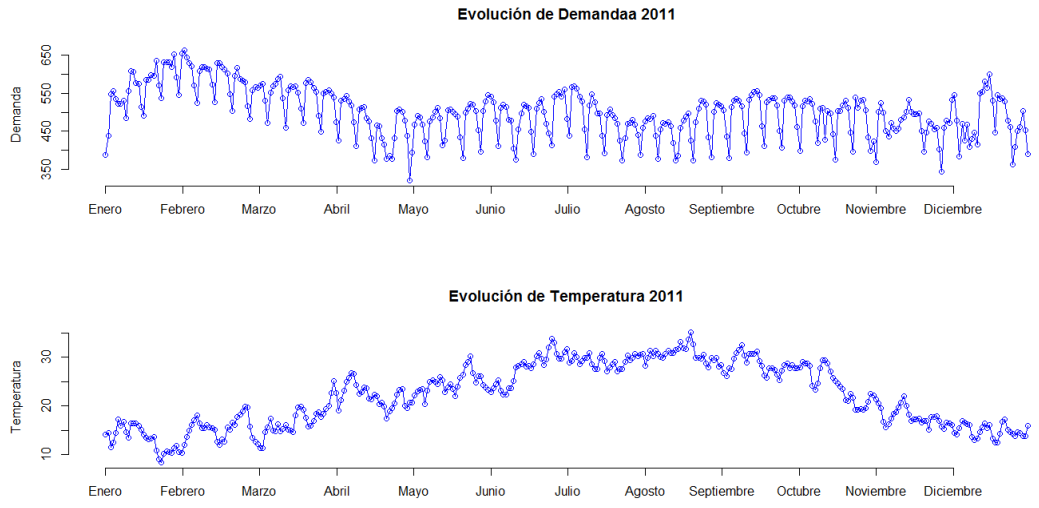

Figura 41. Evolución de la demanda y la temperatura en el 2011

Al observar el gráfico detenidamente se puede apreciar que especialmente en los primeros meses del año, se observa que la reducción en la temperatura conlleva un incremento en la demanda eléctrica.

Por tanto, podemos sacar una primera conclusión diciendo que tienen mayor impacto en la demanda de electricidad las bajas temperaturas que las altas, ya que éstas apenas tienen efecto en la misma. Este hecho podemos explicarlo en parte, reiterándonos en una primera suposición que hicimos anteriormente en este mismo trabajo y es que los aparatos calefactores provocan una mayor demanda de electricidad que los climatizadores.

Esta relación entre la demanda de electricidad y la temperatura se puede ver en el gráfico siguiente. Pero debemos matizar que el efecto de las nuevas variables HDDS y CDDS va a ser modelizado de forma lineal.

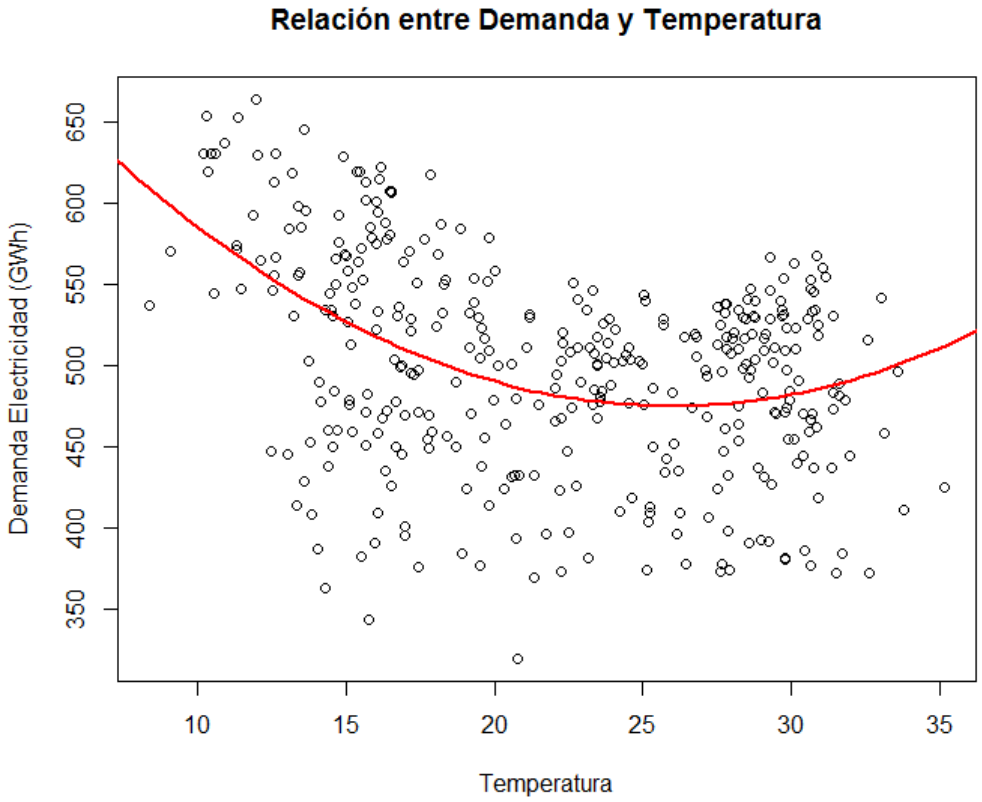

Figura 42: Relación entre la Demanda (GWh) y la Temperatura (C)

Vamos a introducir entonces el efecto de la temperatura en nuestro análisis para ver la influencia que realmente tiene en la demanda y poder así compararlo con la primera idea que sacamos viendo los gráficos.

Para incorporar el efecto de la temperatura en el análisis, lo haremos creando dos variables, la HDDS (heating degree-days) y la CDDS (cooling degree-days), de modo que CDDS_t va a tomar el valor de Tmax_t – 24 si Tmax_t – 24 es positivo y cero en otro caso, y HDDS tomará el valor 20 – Tmax_t si 20 – Tmax_t es positivo y cero en otro caso. Estas variables las creamos para tener en cuenta del mismo modo, los grados que sobrepasan la zona de confort y los que no llegan a la misma. De esta forma, la HDDS va a incluír únicamente el número de grados que superan el límite de 24C, mientras que la CDDS incluirá

los grados que faltan para alcanzar la temperatura neutral que comienza en 20C. Si la temperatura media de algún día se encuentra en la zona de confort, obviamente, estas dos variables dummy que acabamos de crear van a ser cero, ya que no necesitaríamos indicar que la temperatura tiene un comportamiento fuera de lo "común".

Estas variables serán incluidas en el modelo además de las dummy que ya habíamos construido anteriormente e irán diferenciadas.

Comenzamos de nuevo el análisis por los modelos autorregresivos integrados de media móvil y posteriormente pasaremos a realizar el mismo estudio con los modelos regresivos parcialmente lineales, aunque ya no nos vamos a centrar tanto en explicar detalladamente cada parte de los distintos análisis sino que vamos a hacer una explicación más breve que en los apartados anteriores.

4.2.1. Modelo ARIMA con temperatura

En primer lugar vamos a realizar el estudio con el ARIMA utilizando un solo modelo general para todos los días de la semana, aunque eso sí esta vez además de las variables dummy que habíamos creado al principio vamos a introducir en el modelo las dos variables que creamos para la temperatura. Y en segundo lugar repetiremos el proceso de utilizar siete modelos diferentes, uno para cada día de la semana, añadiéndole también las dos variables de la temperatura.

4.2.1.1. ARIMA: un solo modelo con temperatura

Cuando realizamos el estudio con un solo modelo del ARIMA sin temperatura ya observamos que los datos tenían variabilidad constante, es decir, nuestros datos son homocedásticos y por tanto no necesitan transformación logarítmica, por lo que vamos a seguir trabajando con los datos sin transformar como hasta ahora.

El siguiente paso que debemos dar es comprobar si la demanda eléctrica del 2011 es estacionaria, pero volviendo la vista a la Figura 13 sabemos que no lo es. En esta Figura 13 podemos observar las correlaciones parciales y fijándonos en las fas se aprecia cierta tendencia lineal decreciente. Es necesario entonces, para eliminar esta tendencia, realizar una diferenciación en la parte regular. Una vez hemos eliminado la tendencia volvemos a representar las correlaciones y el gráfico que obtenemos es el mismo que en la Figura 14. En esta figura se ve que existe presencia de estacionalidad ya que cada 7 correlaciones éstas se elevan, por lo tanto vamos a tener que diferenciar también la parte estacional para eliminarla y obtener un proceso estacionario. El resultado de hacer las dos diferenciaciones se puede ver en la Figura 15, en la que ya tenemos un proceso estacionario y según el cual podríamos proponer un modelo ARIMA adecuado para ajustar nuestra serie de tiempo.

A partir de la Figura 15 podemos proponer un ARIMA $(0,1,1)(0,1,1)_7$ fijándonos en las fas o bien un ARIMA $(1,1,0)(1,1,0)_7$fijándonos en las fap, para ajustar nuestra serie de datos.

Sin embargo, comprobaremos si estos modelos son los más apropiados para nuestra serie de tiempo con la función *best.arima.TSA*, ya que a veces existen modelos que nos proporcionan mayor información que los que podemos proponer a simple vista.

Con esta función, el mejor ARIMA que obtenemos es el (1, 1, 1) $(1, 1, 1)_7$ desde el punto de vista BIC. Por tanto éste es el modelo que emplearemos para realizar el ajuste.

El modelo podremos representarlo de la siguiente forma:

$$(1 - \phi_1 B)(1 - \Phi_1 B^7)(1 - B)(1 - B^7)Y_t = c + (1 + \theta_1 B)(1 + \ominus_1 B^7)a_t$$

Realizamos el ajuste con este ARIMA y obtenemos la siguiente salida que hemos representado en forma de tabla, en la que podemos ver los coeficientes del modelo y sus correspondientes errores estándar. Se puede apreciar que tenemos 6 coeficientes que no son significativos, por tanto debemos ir ajustando el modelo eliminando de cada vez el coeficiente menos significativo puesto que es el coeficiente que menos influye en el modelo.

Si analizamos detenidamente la siguiente tabla se puede apreciar que el coeficiente menos significativo es el que corresponde con la variable de la temperatura *HDDS* ya que si aumentamos esta variable en una unidad la demanda eléctrica se incrementaría en 0.1431 unidades. Y el siguiente coeficiente menos significativo es la otra variable que creamos de la temperatura *CDDS* ya que un incremento de una unidad en esta variable provoca un aumento de 0.8765 unidades de la demanda de electricidad. Estas dos variables son las menos significativas de todas las que hemos añadido, especialmente la de *HDDS,* sin embargo no vamos a eliminarlas aunque no sean importantes porque queremos introducirlas en el ajuste para ver el efecto que tienen en los resultados finales, aunque ya suponemos que no habrá grandes cambios.

Los siguientes coeficientes menos significativos son en primer lugar el efecto del mes de agosto, como ocurría en los casos anteriores, los festivos de Andalucía y de la Comunidad de Valencia. Vamos a realizar el ajuste sin estas variables y aunque las vamos a ir eliminando una por una, pondremos la tabla del ajuste final para analizar el efecto de los coeficientes.

Tabla 28

arimax (x=demanda1.ts, order= c(1, 1, 1), seasonal= list (order= c (1, 1, 1)), xreg=xreg)

Coefficients	ar1	ma1	sar1	sma1	f.intra	f.sab	Agst	f.And	f.Cat	f.Mad	f.Val	HDDS	CDDS
	0.73	-0.92	0.15	-1.00	-51.12	-108.23	-9.95	-11.36	-38.84	-20.15	-9.21	0.14	0.87
s.e.	0.07	0.04	0.06	0.04	5.61	20.27	13.37	13.70	13.18	10.56	10.72	1.26	1.23
$1.96 * s.e > \|coef\|$	5.07	9.52	1.32	12.12	4.64	2.72	0.37	0.42	1.50	0.97	0.43	0.05	0.36

Sigma^2 estimated as 394.5: log likelihood = -1547.6, aic = 3121.2

Esta es la última tabla del ajuste en la que podemos ver los coeficientes que son significativamente distintos de cero. El más relevante es el que nos refleja los festivos del sábado ya que un incremento en esta variable en una unidad, provoca una disminución de 108.23 en la demanda eléctrica. Y la siguiente variable con mayor relevancia es la de festivos intrasemanales, porque un incremento en esta variable provocará una disminución de la demanda en 51.12 unidades.

Tabla 29

arimax (x=demanda1.ts, order= c(1, 1, 1), seasonal= list (order= c (1, 1, 1)), xreg=xreg)

Coefficients	ar1	ma1	sar1	sma1	f.intra	f.sab	f.Cat	f.Mad	HDDS	CDDS
	0.73	-0.91	0.15	-1.00	-51.28	-107.75	-47.78	-26.71	0.15	0.81
s.e.	0.07	0.50	0.06	0.04	5.64	20.32	11.07	9.01	1.26	1.24
$1.96 * s.e > \|coef\|$	4.82	8.98	1.30	12.16	4.63	2.70	2.20	1.51	0.06	0.33

Sigma^2 estimated as 397.1: log likelihood = -1548.69, aic = 3117.38

Una vez que tenemos el ajuste hecho con todos los coeficientes significativos, nuestro modelo quedaría de la siguiente forma:

$$(1 - 0.73 * B)(1 - (-0.91) * B^7)(1 - B)(1 - B^7)Y_t = c + (1 + 0.15 * B)(1 + (-1) * B^7)a_t + (-51.28 * f.intra_t) + (-107.75 * f.sab_t) + (-47.78 * f.Cat_t) + (-26.71 * f.Mad_t) + (0.15 * HDDS_t) + (0.81 * CDDS_t)$$

Pasamos ahora a analizar los residuos para comprobar si cumplen normalidad e independencia. Observando los gráficos que siguen no parece que se cumpla normalidad ya que en el Q-Q plot no existe una relación lineal muy clara. En cambio, con respecto a la independencia vemos que si se cumple, puesto que todos los p-valores son superiores a 0.05, esto quiere decir que podemos realizar los test de la media igual a cero y los de normalidad para confirmar si ésta se cumple con nuestro ajuste o no.

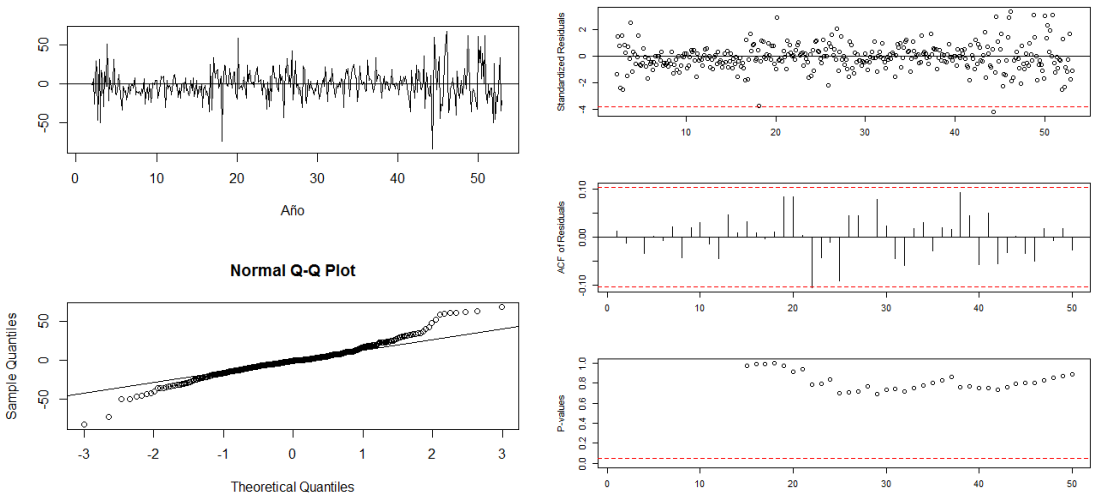

Figura 43. Q-Q plot y Ljung Box

Haciendo el primer test de que la media sea cero, obtenemos un p-valor de 0.37 por tanto no podemos rechazar la hipótesis. No obstante en los test de normalidad de los datos tenemos unos p-valor muy reducidos, por tanto no podemos decir que nuestros datos cumplan normalidad para ningún nivel de significación. Esto indica que podremos realizar las predicciones del 2012 pero no podremos construir los intervalos de predicción.

Tabla 30

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.3754	2.2 e-16	2.572e-8

Una vez hecho el análisis de los residuos podemos comenzar a realizar las predicciones del 2012, para las que utilizaremos los datos del 2011, tal como hicimos en los demás ARIMA.

Después de realizar las debidas predicciones mediante el modelo ARIMA con el efecto de la temperatura incorporado, obtenemos la siguiente tabla de errores que explicaremos a continuación de forma detallada.

Tabla 31

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	3.744690	4.624866	4.320280	8.115108	5.201236
Martes	3.216769	3.165846	2.410750	2.610682	2.851012
Miércoles	2.578698	2.453122	3.476908	4.865609	3.343584
Jueves	3.514643	3.099625	4.432127	4.208830	3.813806
Viernes	2.751424	4.334969	1.808171	4.583367	3.369483
Sábado	2.959516	3.970464	5.027472	9.141953	5.274851
Domingo	5.606284	4.709194	5.855845	7.579223	5.937637
Total	3.481718	3.765441	3.925718	5.896491	

Observando de forma general dicha tabla podemos ver que los resultados obtenidos son bastante buenos ya que, salvo contadas excepciones, la media de errores no supera el 5% que habíamos establecido al principio como límite máximo permitido para considerar que un error no era preocupante. Centrándonos ahora en las excepciones podemos apreciar un dato curioso y es que aunque las peores predicciones se dan en los mismos días que habíamos visto anteriormente, es decir, sábado, domingo y lunes del último trimestre, con un 9.14%, un 7.57% y un 8.11% respectivamente, los errores más elevados ya no se dan el domingo si no que pasan a darse los sábados y lunes del último trimestre. Por otra parte, fijándonos en los menores errores producidos vemos que también existe una pequeña

variación, puesto que hasta ahora siempre se daban en los martes y miércoles, en cambio en este caso tenemos que la media de errores de los viernes del tercer trimestre es de 1.80% y la siguiente mejor predicción se da el martes también del tercer trimestre con un 2.41%, por lo que podemos pensar que el tercer trimestre va a ser uno de los que tenga las mejores predicciones.

Estudiando ahora la tabla desde un punto de vista trimestral, nos llama la atención el dato del cuarto trimestre ya que saca dos puntos porcentuales al resto de los errores trimestrales, es decir, los datos del último trimestre son más difíciles de predecir que los de los demás trimestres, con este modelo. Además debemos mencionar que aunque los dos mejores datos se encontraban en el tercer trimestre, éste es el segundo con mayor error.

En cuanto a la media de porcentaje de error por día de la semana, se puede apreciar que no hay grandes diferencias con respecto a las anteriores tablas de error, ya que las peores predicciones se dan los sábados, domingos y lunes, siendo el peor día el domingo. De nuevo, los menores errores se dan por el medio de la semana, destacando el martes con una media de error del 2.85%.

Como hicimos en modelos anteriores, vamos a ver en la tabla 32, como se comporta la media de los días intrasemanales. Se vuelve a repetir el patrón que seguían algunos de los modelos analizados, las peores predicciones se dan en el cuarto cuadrimestre y las mejores se dan en el primero, con un 4.04% y un 3.01% respectivamente. La media total de este grupo de días es de 3.34%, por lo que tenemos que admitir que las predicciones de este grupo de días, con el ARIMA de un solo modelo con temperatura, son buenas.

Tabla 32

Media del porcentaje de error absoluto (MAPE) de Martes a Viernes

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Ma-Vi	3.015383	3.26339	3.031989	4.067122	3.344471

4.2.1.2. ARIMA: siete modelos con temperatura

En este apartado del trabajo vamos a volver a hacer el análisis del ARIMA pero modelizando cada día de la semana por separado. Tal como vimos ya en el análisis del ARIMA sin temperatura, no va a existir estacionalidad en los siguientes modelos ya que la única diferencia con el estudio realizado anteriormente es que le añadimos dos variables dummy más, las de la temperatura.

4.2.1.2.1 ARIMA para los Lunes

El primer paso que debemos dar, es examinar si existe estacionariedad en media o no y para ello tenemos que observar la representación de las correlaciones de los lunes.

Las correlaciones de los lunes ya las tenemos representadas en la Figura 19, en la que podemos apreciar que existe cierta tendencia en la primera parte de las mismas, por tanto debemos eliminarla. Para ello hacemos una diferenciación de la parte regular mediante la que obtenemos las correlaciones que se muestran la Figura 20. Como vemos nuestra serie ya es estacionaria y además es ruido blanco, es decir, su comportamiento es aleatorio y no

depende en absoluto de lo ocurrido en los lunes anteriores (hecho que explicaría los resultados obtenidos en el NAIVE).

A partir de las correlaciones obtenidas, podemos tratar de identificar el mejor ARIMA, y observándolas el modelo que proponemos es un ARIMA (0, 1, 0), sin embargo utilizando la función de *best.arima.TSA* el mejor ARIMA que obtenemos para nuestros datos es un (0, 1, 1), por lo tanto trabajaremos con un modelo de esta forma:

$$(1 - B)Y_t = c + (1 + \theta_1 B)a_t$$

Procedemos entonces a realizar el ajuste de este modelo, eliminando primero las variables dummy que son nulas, como en este caso, los festivos que caen en sábado. En el primer ajuste que hacemos nos salen que no son significativas tres de las dummy que hemos introducido en el modelo, los festivos en Andalucía, los festivos en Valencia y la variable CDDS. Esta última variable no vamos a eliminarla del modelo aunque no sea significativa, para poder estudiar el efecto que tiene en las predicciones finales, aunque suponemos que éste no va a ser relevante.

Eliminando las dos variables no significativas, nos quedaría el modelo ajustado del siguiente modo:

Tabla 33

arimax (x= lunes11.ts, order = c(0, 1, 1), xreg = xreg [, -c (2,7,4)])

Coefficients	ma1	f.intra	Agst	f.Cat	f.Mad	HDDS	CDDS
	-0.64	-92.84	-50.68	-87.96	-93.69	7.94	3.26
s.e.	0.16	31.29	20.34	22.38	21.89	2.94	3.05
$1.96 * s.e > \|coef\|$	2.01	1.51	1.27	2.00	2.18	1.37	0.54

Sigma^2 estimated as 1073: log likelihood = -245.17, aic = 504.33

Si observamos la tabla anterior donde tenemos los coeficientes de los parámetros con sus errores estándar, vemos que las variables menos significativas son las dos de la temperatura y la de los días de agosto. Aunque con mucha diferencia la menos importante es la CDDS, es decir, en el comportamiento de los lunes, que las temperaturas sean superiores a los 24C no provoca incrementos muy grandes en la demanda eléctrica, en comparación por ejemplo a que la temperatura sea más baja de los 20C.

Nuestro modelo quedaría ajustado con las variables dummy, de la siguiente forma:

$$(1 - B)Y_t = c + (1 + (-0.64) * B)a_t + (-92.84 * f.intra_t) + (-50.68 * f.And_t) + (-87.96 * f.Cat_t) + (-93.69 * f.Mad_t) + (7.94 * HDDS_t) + (3.26 * CDDS)$$

Ahora que ya tenemos el modelo ajustado, vamos a realizar el análisis de los residuos. En el Q-Q plot vemos una primera presentación de los cuantiles esperados bajo normalidad

en comparación con los cuantiles observados y parece que se va a cumplir normalidad. Pero para poder realizar los test debemos mirar si se cumple primero la hipótesis de independencia.

En el segundo gráfico, Ljung Box, podemos ver que las correlaciones se encuentran dentro de los límites y que además los p-valor superan el 0.05, por lo que podemos afirmar que sí se cumple la independencia de los residuos.

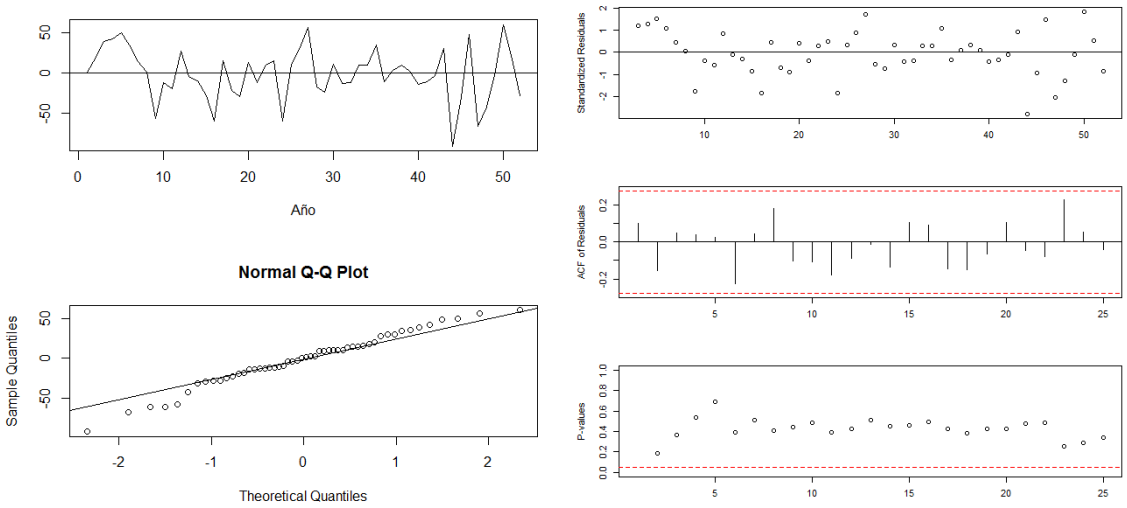

Figura 44: Q-Q plot y Ljung Box

Ya podemos realizar los siguientes test, tanto el de que la media es igual a cero como el de que existe normalidad de nuestros datos.

A la vista de los resultados, podemos confirmar que se cumplen ambas hipótesis y para cualquier nivel de significación.

Tabla 34

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.7573	0.4285	0.4307

4.2.1.2.2 ARIMA para los Martes

Observando ahora las correlaciones de nuestra demanda eléctrica de los martes, representadas en la Figura 23 en el primer gráfico, vemos que tienen tendencia, por tanto debemos eliminarla haciendo un filtro de tendencia, es decir, una diferenciación en la parte regular.

Las correlaciones pasan a ser de la forma que vemos en el segundo gráfico. De nuevo se puede observar que nuestros datos no tienen componente estacional y es ruido blanco, por lo tanto ya es una serie estacionaria.

Observando las correlaciones, el modelo tentativo que proponemos va a ser de la forma que vemos a continuación, un ARIMA (0 ,1, 0):

$$(1 - B)Y_t = c + a_t$$

Este modelo va a ser con el que realicemos el ajuste ya que es también el mejor modelo que obtenemos con la serie del mejor arima, desde el punto de vista BIC.

Para proceder a ajustar el modelo, primero debemos eliminar las variables que sean nulas en este caso, y sólo nos quedamos con 4 de las 9 que tenemos, los festivos intrasemanales, los días de agosto y las dos variables de la temperatura. Al hacer el primer ajuste, vemos que 3 de ellas no son significativas, entre las mismas se encuentran las dos de la temperatura, pero no las eliminamos por el mismo motivo que explicábamos antes. Entonces repetimos el ajuste eliminando únicamente la variable de los días de agosto y el resultado que obtenemos es el siguiente:

Tabla 35

arimax (x= martes11.ts, order= c(0, 1, 0), xreg = xreg[, -c (2, 4, 5, 6, 7, 3)])

Coefficients	f.intra	HDDS	CDDS
	-102.17	-0.28	-1.75
s.e.	14.92	2.41	2.35
$1.96 * s.e > \|coef\|$	3.49	0.06	0.38

Sigma^2 estimated as 876.3: log likelihood = -239.84, aic = 485.69

Vemos de nuevo que las dos variables de la temperatura siguen sin ser significativas, especialmente la HDDS, con lo cual si esta variable aumenta en una unidad, es decir, si la temperatura baja un grado de los 20C va afectar a la demanda eléctrica reduciéndola en 0.28 unidades. Para que se perciba que esto es un dato irrelevante, lo comparamos con que exista un festivo que caiga en martes, lo cual reduciría la demanda en 102.17 unidades. En esta ocasión el comportamiento de las variables dummy de la temperatura es justamente el contrario de lo esperado, ya que en lugar de aumentar la demanda eléctrica, la reducirían.

De todos modos vamos a tener en cuenta las variables de la temperatura aunque no tengan un coeficiente significativo, para ver si las predicciones mejoran aunque no sea en gran medida, por lo que nuestro modelo quedaría:

$$(1 - B)Y_t = c + a_t(-102.17 * f.intra_t) + (-0.28 * HDDS_t) + (-1.75 * CDDS_t)$$

Pasamos ahora a analizar los residuos de nuestra serie de datos. El Q-Q plot parece que nos muestra que nuestros datos van a ser normales, pero para asegurarnos lo comprobaremos haciendo los test de Shapiro Wilk y Jarque Bera. Sin embargo antes debemos comprobar que se cumpla independencia de residuos y vemos que sí se cumple, ya que los p-valor superan todos el 0.05.

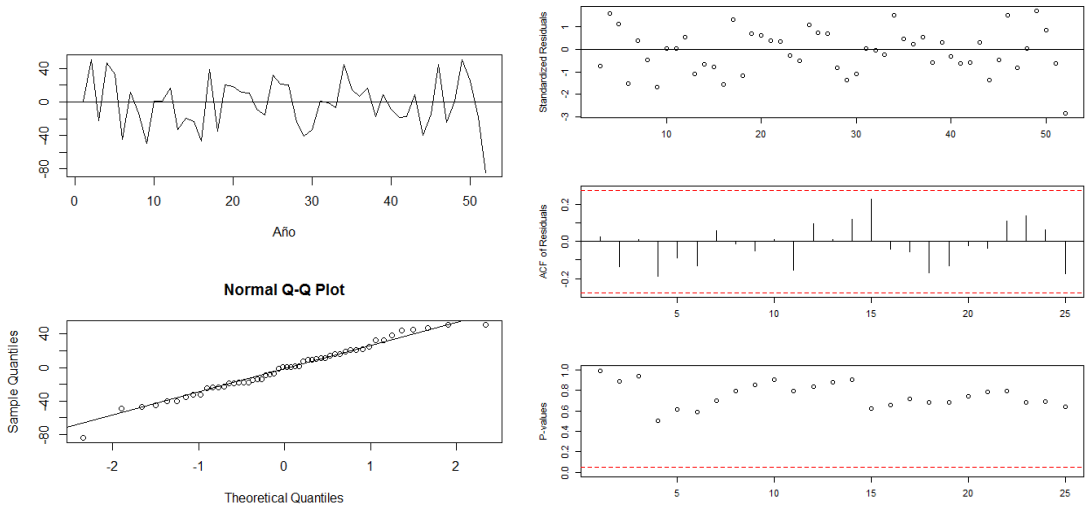

Figura 45: Q-Q plot y Ljung Box

Hacemos también el contraste de hipótesis de si la media es igual a cero y vemos que sí lo es para cualquier nivel de significación.

En cuanto a los contrastes de normalidad también se cumplen para cualquier nivel de significación.

Tabla 36

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.6192	0.8309	0.5973

4.2.1.2.3 ARIMA para los Miércoles

En esta ocasión vamos a centrarnos en los miércoles, aunque vemos en la Figura 26 en el primer gráfico, que el comportamiento de las correlaciones es similar al de los días anteriores. Como se puede apreciar existe tendencia lineal decreciente por lo tanto debemos hacer una diferenciación en la parte regular. De esta forma, nos quedan las correlaciones parciales tal como podemos ver en el segundo gráfico. Tenemos una serie sin parte estacionaria y que es ruido blanco, por tanto tenemos una serie estacionaria.

Observando las correlaciones anteriores, tanto las fas como las fap, podemos proponer una estructura para nuestro modelo y el modelo que conviene proponer es un ARIMA (0, 1, 0). Utilizando la función que nos devuelve el mejor ARIMA para un modelo obtenemos el mismo que hemos propuesto, por tanto será el que utilicemos para realizar el ajuste:

$$(1 - B)Y_t = c + a_t$$

Para comenzar a ajustar el modelo eliminamos las variables nulas, en este caso, las mismas que en los martes, por lo que nos quedamos con los festivos en miércoles, los días de agosto, y las dos de temperatura.

Como podemos ver en el primer ajuste nos sale que los días de agosto no son significativos, por tanto eliminamos esa variable y el modelo nos queda ajustado de esta forma:

$$(1 - B)Y_t = c + a_t$$
$$+(-98.39 * f.intra_t) + (-1.24 * HDDS_t) + (0.77 * CDDS_t)$$

Tabla 37

arimax (x = miercoles11.ts, order= c(0, 1, 0), xreg = xreg [, -c (2, 4, 5, 6, 7, 3)])

Coefficients	f.intra	HDDS	CDDS
	-98.39	-1.24	0.77
s.e.	25.83	2.92	2.81
$1.96 * s.e > \|coef\|$	1.94	0.21	0.14

Sigma^2 estimated as 1196: log likelihood = -247.61, aic = 501.22

Ya tenemos los coeficientes significativos, aunque sólo en parte, ya que las variables de la temperatura siguen sin ser significativas. Ahora la menos importante es la CDDS, es decir, cuando la temperatura sobrepasa los 24C el aumento que provoca en la electricidad es mínimo. Sin embargo vuelve a darse el suceso de que con una temperatura que desciende de los 20C la demanda de electricidad se reduciría.

Echando un vistazo ahora a los residuos, parece que el Q-Q plot ya no muestra normalidad, ya que los cuantiles observados difieren de los esperados bajo normalidad. Vamos a proceder a realizar los test que nos confirmen si el Q-Q plot nos da una información fiable en este caso, pero antes debemos ver si los residuos cumplen independencia y vemos que una vez más, sí la cumplen.

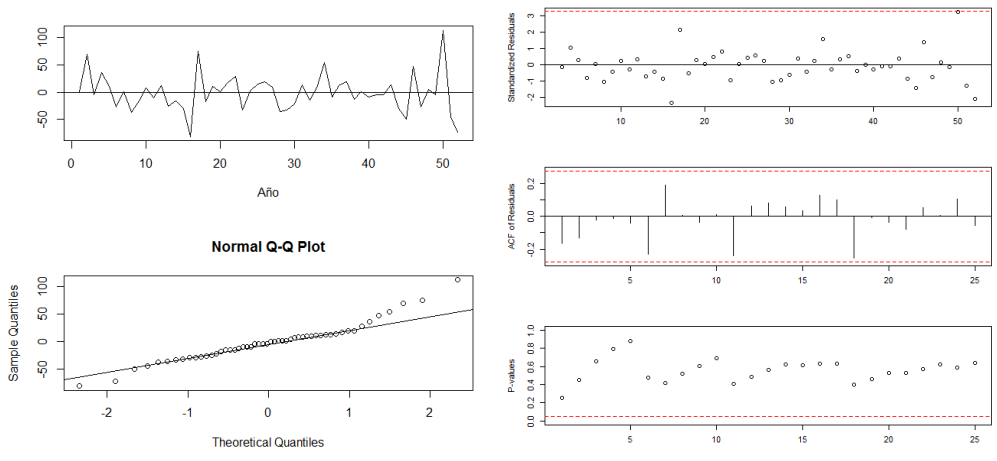

Figura 46: Q-Q plot y Ljung Box

Si observamos la tabla siguiente, vemos que la hipótesis de que la media es igual a cero sí que cumple para cualquier nivel de significación.

En cambio, en cuanto a la hipótesis de normalidad tenemos unos resultados distintos, ya que con el test de Jarque Bera no se cumple normalidad para ningún nivel de significación, mientras que con el test de Shapiro Wilk sí se cumple, aunque sólo para un nivel de significación del 1%.

Tabla 38

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.7668	0.003386	0.02573

4.2.1.2.4 ARIMA para los Jueves

Para analizar el comportamiento de la demanda eléctrica de los jueves, nos vamos a fijar en cómo se comportan sus correlaciones. Tal como se puede observar en el primer gráfico de la Figura 29 existe cierta tendencia negativa, por lo que tenemos que hacer una diferenciación de la parte regular para suprimirla. De este modo las correlaciones nos quedan como vemos en el segundo gráfico, en el que podemos ver que no existe estacionalidad y por tanto nuestro proceso ya es estacionario en media.

Observando las correlaciones de este último gráfico podemos proponer un modelo tentativo para nuestra serie como un ARIMA (1, 1, 0) o un ARIMA (0, 1, 1). En este caso, el modelo que nos devuelve la función de *best.arima.TSA* es el segundo que hemos propuesto, es decir, el ARIMA (0, 1, 1), por tanto el modelo con el que vamos a trabajar será de la siguiente forma:

$$(1 - B)Y_t = c + (1 + \theta_1 B)a_t$$

Realizamos el ajuste con el ARIMA (0, 1, 1) eliminando las variables nulas y nos quedamos únicamente con 5 variables, con los festivos intrasemanales (en este caso festivos en jueves), los días de agosto, los festivos de Madrid y las dos variables de la temperatura.

En el primero de los ajustes nos salen que no son significativos ni los días de agosto, ni los festivos de Madrid ni la CDDS. Decidimos eliminar estos coeficientes no significativos (salvo la CDDS) para quedarnos únicamente con los significativos para nuestro modelo. Y de esta forma obtenemos la siguiente tabla:

Tabla 39

arimax (x = jueves11.ts, order= c(0, 1, 1), xreg = xreg [, -c (2, 4, 5, 7, 6, 3)])

Coefficients	ma1	f.intra	HDDS	CDDS
	-0.50	-71.71	6.09	1.37
s.e.	0.14	13.87	2.59	2.98
$1.96 * s.e > \|coef\|$	1.74	2.63	1.19	0.23

Sigma^2 estimated as 875.9: log likelihood = -239.98, aic = 487.95

Si nos detenemos brevemente en su análisis vemos que mientras que los festivos intrasemanales así como la variable HDDS son significativos, la CDDS no lo es. Esto indica que es de mayor importancia en cuanto a la influencia en la demanda, que un jueves sea festivo o que las temperaturas sean inferiores a 20C, antes que las temperaturas rebasen los 24C. En este caso el comportamiento de las dummy es el esperado, ya que una disminución o incremento de un grado fuera de los límites expuestos (20C y 24C), supondría un aumento en la demanda eléctrica.

Si representamos como queda nuestro modelo ajustado con las variables dummy, tenemos:

$$(1-B)Y_t = c + (1 + (-0.5) * B)a_t + (-71.71 * f.intra_t) + (6.09 * HDDS_t) + (1.37 * CDDS_t)$$

Pasamos ahora a analizar los residuos, que según parece en el Q-Q plot tienen un comportamiento normal. Sin embargo debemos primero comprobar que se cumple independencia de los residuos. En el segundo gráfico de esta figura vemos que sí se cumple ya que el p-valor supera en todas las ocasiones el 0.05.

Podemos pasar entonces, a realizar los siguientes test tanto de si la media es igual a cero, como los de si existe o no normalidad.

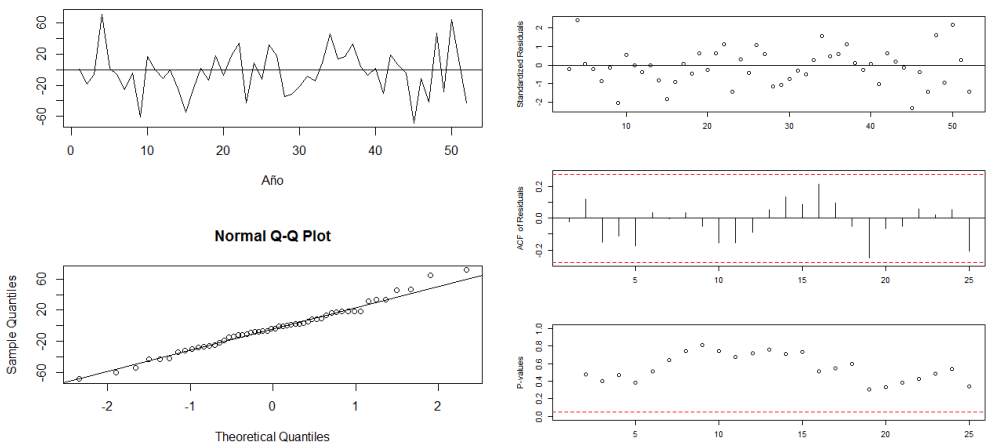

Figura 47: Q-Q plot y Ljung Box

A la vista de la siguiente tabla se puede confirmar que se cumplen las dos hipótesis, para cualquier nivel de significación. El contraste de hipótesis de que la media es igual a cero tiene un p-valor de 0.40, y en los test de normalidad el p-valor supera el 0.70 en ambos casos.

Tabla 40

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.4071	0.7437	0.7736

4.2.1.2.5 ARIMA para los Viernes

Observando los gráficos que tenemos en la Figura 32 podemos ver las correlaciones de la demanda de los viernes. En ellas se aprecia de nuevo que nuestros datos tienen tendencia negativa por lo que tenemos que hacer un filtro de tendencia para eliminarla. Una vez hemos diferenciado nuestros datos en la parte regular obtenemos las correlaciones que se ven en el segundo gráfico.

A partir de estas correlaciones podemos plantear dos estructuras del ARIMA para realizar el ajuste de nuestro modelo. Podemos sugerir un ARIMA (1, 1, 0) o un ARIMA (0, 1, 1), sin embargo, la función del mejor arima nos presenta como mejor modelo según el punto de vista del BIC, el ARIMA (1, 1, 0). Por tanto la estructura de ARIMA que emplearemos en el ajuste será la siguiente:

$$(1 - \phi_1 B)(1 - B)Y_t = c + a_t$$

Para realizar el ajuste debemos eliminar en primer lugar las variables dummy que sean nulas, por lo que nos quedamos con los festivos intrasemanales, los días de agosto y las dos de temperatura.

En el primer paso del ajuste vemos que sólo nos salen significativos los festivos intrasemanales, por tanto mantenemos esa variable y también las de la temperatura, para poder tener en cuenta su efecto por mínimo que sea.

De esta manera, obtenemos los resultados de la siguiente tabla. Podemos ver que en comparación con el efecto de los festivos de los viernes, el efecto de la temperatura es ínfimo. Ya que mientras que un festivo reduciría la demanda en 120.48 unidades, las variables de la temperatura, HDDS y CDDS la aumentarían en 3.38 y 1.26 unidades respectivamente.

Tabla 41

arimax (x = viernes11.ts, order= c(1, 1, 0), xreg = xreg [, -c (2, 4, 5, 6, 7, 3)])

Coefficients	ar1	f.intra	HDDS	CDDS
	-0.58	-120.48	3.38	1.26
s.e.	0.11	30.98	2.54	2.85
$1.96 * s.e > \|coef\|$	2.58	1.98	0.67	0.22

Sigma^2 estimated as 1367: log likelihood = -251.16, aic = 510.31

Aun siendo conscientes de que el efecto de la temperatura en los viernes es muy reducido nuestro ARIMA final estimado quedaría de la siguiente forma:

$$(1-(-0.58)*B)(1-B)Y_t = c + a_t + (-120.48 * f.intra_t) + (3.38 * HDDS_t) + (1.26 * CDDS_t)$$

Analizando ahora los residuos, podemos decir que el Q-Q plot ya no muestra que se vaya a cumplir normalidad de los mismos, puesto que los cuantiles observados se diferencian de los esperados bajo el supuesto de normalidad.

Observando el segundo de los gráficos vemos que se cumple la hipótesis de independencia de los residuos por lo tanto vamos a poder realizar los test de que la media sea nula y los de normalidad.

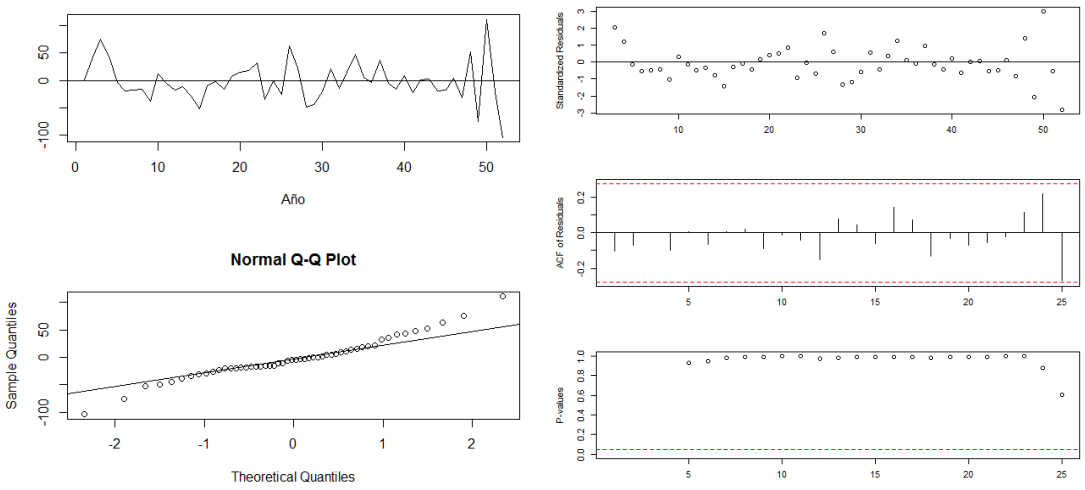

Figura 48: Q-Q plot y Ljung Box

Según los valores que hemos obtenido en el primero de los test para el p-valor, podemos afirmar que la media es igual a cero para cualquier nivel de significación.

Sin embargo cuando hablamos de los test de normalidad ya no podemos ser tan concluyentes. Para el test de Jarque Bera podemos ver que tenemos un p-valor de 0.046, por

tanto podemos aceptar que existe normalidad únicamente con un 1% de significación. En cambio, en el Shapiro Wilk ya podemos confirmar que existe normalidad para un nivel de un 10% de significación.

Tabla 42

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.667	0.04631	0.1013

4.2.1.2.6 ARIMA para los Sábados

Analizando ahora la demanda de electricidad para los sábados, vemos en la Figura 35 que debemos hacer una diferenciación regular para eliminar la tendencia que se aprecia en el primer gráfico. Hecho esto, las correlaciones de nuestra serie quedarían tal como vemos en el segundo gráfico y parece que tenemos ruido blanco, por tanto nuestra serie es estacionaria.

Observando las correlaciones anteriores podemos proponer una estructura para el ARIMA que ajuste nuestro modelo y la estructura más razonable en este caso sería un ARIMA (0, 1, 0). Sin embargo la función del mejor arima nos muestra que la mejor estructura sería un ARIMA (0, 1, 1). Entonces el modelo con el que trabajemos será de este modo:

$$(1 - B)Y_t = c + (1 + \theta_1 B)a_t$$

En el momento de realizar el ajuste debemos como siempre de eliminar las variables que sean nulas ya que si no el ajuste no funcionaria. Eliminamos entonces todas las variables excepto los festivos de sábado, los días de agosto, los festivos de Valencia y las de las temperaturas.

En el primer ajuste que hacemos nos salen que no son significativas 3 de las variables que hemos introducido, por lo que debemos eliminarlas si queremos que nuestro modelo mejore. Sin embargo una de ellas es la CDDS y ésta no queremos eliminarla del ajuste, ya que queremos analizar el efecto que tiene la temperatura sobre la demanda eléctrica.

Tabla 43

arimax (x = sabado11.ts, order= c(0, 1, 1), xreg = xreg [, -c (1, 4, 5, 6, 7, 3)])

Coefficients	ma1	f.sab	HDDS	CDDS
	-0.52	-155.82	5.42	1.04
s.e.	0.13	30.45	2.35	2.32
$1.96 * s.e > \|coef\|$	1.94	2.61	1.17	0.22

Sigma^2 estimated as 909.7: log likelihood = -245.76, aic = 499.52

Ésta sería la tabla final de nuestro ajuste. En ella podemos ver que la CDDS sigue siendo una variable no significativa, ya que si se supera en 1 grado los 24C, la demanda de electricidad sólo se incrementaría en 1.04 unidades, mientras que si la temperatura baja 1 grado de los 20C la demanda de electricidad se incrementaría en 5.42 unidades. Y la variable de mayor relevancia es que se dé un festivo el sábado ya que este hecho reduciría la demanda en 155.82 unidades.

Entonces nuestro ARIMA final ajustado con las dummy, quedaría:

$$(1-B)Y_t = c + (1 + (-0.52) * B)a_t + (-155.82 * f.intra_t) + (5.42 * HDDS_t) + (1.04 * CDDS_t)$$

En lo que se refiere a los residuos, vemos que en el Q-Q plot no está del todo claro si va a existir o no normalidad. Primero vamos a comprobar que se cumple independencia y posteriormente podemos realizar los test que correspondan.

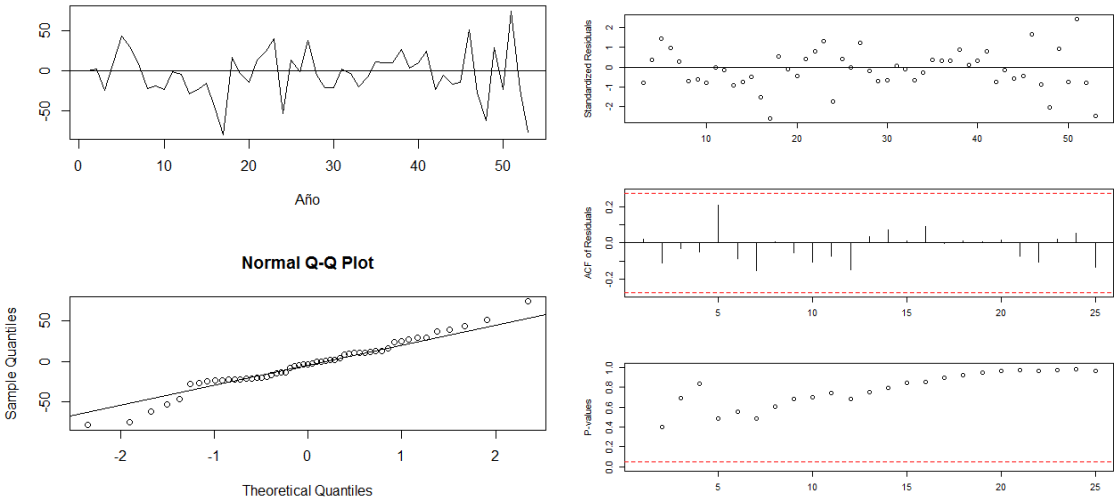

Figura 49: Q-Q plot y Ljung Box

Vemos que los p-valor del segundo gráfico superan todos el 0.05 por tanto podemos afirmar que sí se cumple la independencia de los residuos.

Podemos entonces realizar los contrastes pertinentes. En primer lugar, vemos que la media es nula para cualquier nivel de significación. Pero además también podemos afirmar que nuestros datos son normales para cualquier nivel de significación y para ambos test.

Tabla 44

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.3244	0.5891	0.3211

4.2.1.2.7 ARIMA para los Domingos

Finalmente vamos a ver cómo se comporta la demanda de electricidad de los domingos. Para ello tenemos representadas las correlaciones en la Figura 38, que nos indican que existe una tendencia lineal negativa un poco mayor que en el resto de los días de la semana. Esta tendencia debe ser eliminada, por lo que debemos hacer una diferenciación de la parte regular.

Al hacer la diferenciación regular, las correlaciones quedan tal como podemos ver en el segundo gráfico. Podemos decir que nuestra serie es estacionaria ya que es ruido blanco.

A la vista de las tablas anteriores podemos sugerir un único modelo tentativo para la estructura del ARIMA, que es un ARIMA (0, 1, 0). Este es el mismo modelo que nos plantea la función de *best.arima.TSA* según el punto de vista del BIC. Por lo tanto realizaremos el ajuste con el siguiente modelo:

$$(1 - B)Y_t = c + a_t$$

Cuando realizamos el ajuste para la demanda eléctrica de los domingos, nos ocurre algo que no nos había ocurrido en los demás días de la semana, y es que todos los coeficientes que hemos incorporado en el modelo son no significativos. Nosotros vamos a quedarnos como siempre con las variables de la temperatura para tenerlas en consideración, a la hora de hacer las comparaciones de las predicciones y sus errores.

En esta tabla donde se muestran tanto los coeficientes como sus errores estándar, se puede ver la influencia de cada una de las variables en la demanda de electricidad. La de mayor importancia es la CDDS, aunque tiene un comportamiento distinto al resto de los días de la semana.

Si miramos detenidamente los valores, vemos que un incremento de 1 grado en las temperaturas que superan los 24C supondría que se reduzca la demanda en 3.44, cuando lo lógico y razonable sería que aumentase la demanda, que era lo que sucedía en la mayor parte de los días. Pero esto puede tener una sencilla explicación y es que estas temperaturas se van a dar en verano. En verano las personas aprovechan para ir de vacaciones o aunque no tengan vacaciones los fines de semana, especialmente el domingo, van a la playa o a refrescarse a otros lugares, por lo que dejan de consumir tanta electricidad.

El ajuste de nuestro modelo nos queda del modo siguiente:

$$(1 - B)Y_t = c + a_t + (1.35 * HDDS_t) + (-3.44 * CDDS_t)$$

Tabla 45

arimax (x = domingo11.ts, order= c(0, 1, 0), xreg = xreg [, -c (1, 2, 4, 5, 6, 7, 3)])

Coefficients	HDDS	CDDS
	1.35	-3.44
s.e.	2.65	3.22
$1.96 * s.e > \|coef\|$	0.25	0.54

Sigma^2 estimated as 1186: log likelihood = -247.41, aic = 498.83

Una vez hemos analizado el comportamiento de nuestras variables, comenzamos el estudio de los residuos.

El Q-Q plot parece indicarnos que los residuos se comportan de forma normal, aunque para asegurarnos realizaremos los test de Jarque Bera y Shapiro Wilk.

Pero antes de realizar estos test así como el de que la media sea nula, es necesario ver que se cumple independencia. Y por lo que se puede ver en el segundo gráfico que tenemos a continuación, sí se cumple.

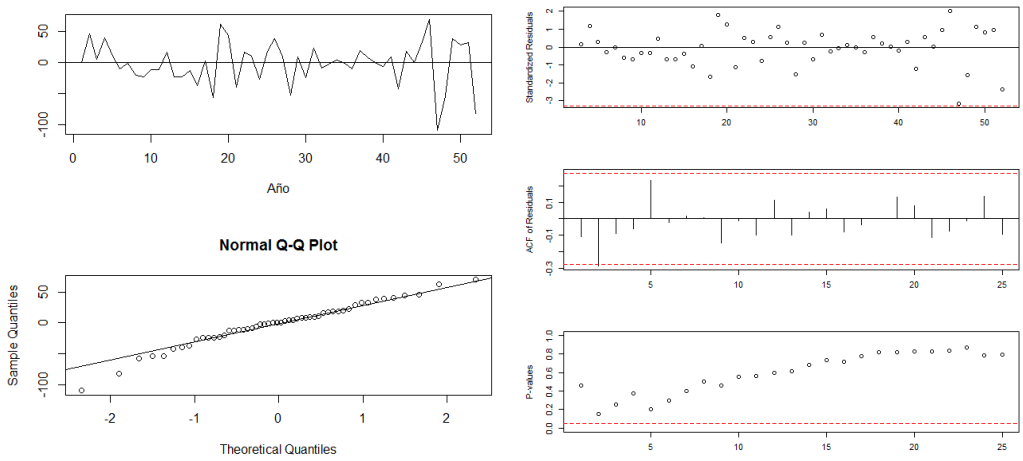

Figura 50: Q-Q plot y Ljung Box

A la vista de la tabla siguiente, lo primero que podemos afirmar es que la media es igual a cero, ya que el p-valor que nos devuelve el contraste es de 0.76.

Sin embargo en cuanto a la normalidad, los test difieren un poco en los resultados. Mientras que en el test de Shapiro Wilk se acepta normalidad de los datos para cualquier nivel de significación, en el de Jarque Bera sólo se acepta normalidad para un 5%. Como conclusión de ambos test aceptaríamos que nuestros datos son normales para un nivel de significación no mucho mayor del 5%.

Tabla 46

Test de que la media es igual a cero y test de normalidad

	test mu = 0	**jarque.bera.test (ajuste)**	**shapiro.test (ajuste)**
p-valor	0.7677	0.05788	0.2144

Media del porcentaje de error absoluto (MAPE)

Ante la lectura de la tabla abajo situada podemos destacar de forma general que dicho método no presenta los datos deseados debido al elevado porcentaje de error que ésta muestra. Así vemos por ejemplo, que la mayor parte de los datos obtenidos constan de un error superior al 5%, por lo que supera el límite que habíamos establecido para considerar que una predicción es fiable.

En segundo lugar, siguiendo en la misma visión general, es importante recalcar que los valores obtenidos fluctúan entre el 5.76% y el 18.77%. Registrándose por tanto, los valores menos reales, es decir, los que presentan los mayores porcentajes de error el domingo del primer, segundo y cuarto trimestre, seguido del sábado del cuarto trimestre. Mientras que los menores errores se dan los sábados del tercer trimestre y los lunes del primer trimestre.

Centrándonos ahora en un análisis semanal, podemos señalar que los datos más próximos a la realidad y que presentan por tanto menor porcentaje de error son los lunes y sábados, ya que se mueven entre el 7.96% y el 8.05% de error. Registrándose los mayores errores especialmente el domingo.

Estudiando ahora la tabla desde otro punto de vista, trimestralmente, observamos que el tercer trimestre es el que nos ofrece una visión de la realidad más verídica, puesto que la media del mismo es de 7.28%, en cambio los demás trimestres varían entre el 10.02% y el 13.01% de error.

Tabla 47

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	5.809679	9.743210	6.140360	10.545904	8.059788
Martes	9.855578	10.349092	9.238791	10.615702	10.014790
Miércoles	10.210769	8.819213	7.248335	14.212895	10.122803
Jueves	10.923135	10.020613	7.799565	11.807886	10.137799
Viernes	12.788446	11.875583	6.619552	12.470233	10.938453
Sábado	6.849027	6.557333	5.767403	12.692901	7.968453
Domingo	13.726008	13.402260	8.153796	18.774387	13.514112
Total	10.023234	10.109614	7.281114	13.017124	

4.2.2. Modelo PLRM con Temperatura

En este apartado elaboraremos dos tipos de predicción mediante el modelo de regresión parcialmente lineal, tal como hicimos anteriormente. En primer lugar lo haremos utilizando un solo modelo de regresión para el total de nuestros datos y en segundo lugar lo haremos empleando siete modelos diferentes, uno para cada día de la semana. Eso sí, en esta ocasión añadiremos a las variables dummy ya existentes, las dos variables que hemos creado de la temperatura.

El siguiente paso es eliminar la tendencia de nuestros datos, y para ello debemos hacer un filtro de tendencia, es decir, debemos diferenciar todos los datos excepto las variables de la temperatura (porque si no, no funciona el modelo).

Una vez hecho esto, tenemos nuestros datos en condiciones para comenzar a hacer los ajustes y las predicciones correspondientes.

4.2.2.1. PLRM: un solo modelo con temperatura

Para realizar el ajuste con un solo modelo de regresión parcialmente lineal, una vez tenemos los datos corregidos sin tendencia, debemos corregir también el efecto de la estacionalidad. En este modelo la forma de corregir la estacionalidad es crear unas variables dummy indicando cuales son los días o períodos con características distintas. Entonces tenemos que repetir los mismos pasos que dimos cuando hicimos el ajuste con un solo PLRM sin variables exógenas.

Es necesario que volvamos a crear las variables siguientes:

Variable D1: Esta variable tomará valor 1 si la diferencia es entre un lunes y un domingo, y cero en otro caso.

Variable D2: Esta variable tomará valor 1 si la diferencia es entre un sábado y un viernes, y cero en otro caso.

Variable D3: Esta variable tomará valor 1 si la diferencia es entre un domingo y un sábado, y cero en otro caso.

Una vez que hemos creado las variables, las anexionamos sin diferenciar a nuestra matriz de datos totales para poder trabajar con ellas. Estas variables como habíamos explicado, se incorporan sin diferenciar debido a que cuando las creamos ya lo hacemos considerando las debidas diferenciaciones.

En nuestra matriz de datos van a estar entonces, todas las variables dummy que hemos creado pero sólo van a estar diferenciadas las creadas para el ARIMA. Ahora bien, debemos puntualizar que no podemos utilizar todas las variables para hacer el ajuste, ya que de las creadas al principio sólo podemos introducir en el modelo, los festivos intrasemanales, los festivos de Cataluña y los festivos de Madrid (si introducimos cualquier otra variable no funcionaría el modelo).

La forma de realizar el ajuste y obtener tanto los β como los m va a ser semejante a cuando realizamos el ajuste sin variables exógenas, es decir, vamos a ir tomando siempre los 363 datos anteriores al día que queremos predecir. De esta forma vamos a ir obteniendo los valores de β y m de cada una de las observaciones y con ellos podemos hacer las predicciones para el 2012. Las predicciones también se harán de la misma forma, es decir, multiplicando los β por los valores de las dummy que empleemos en el 2012, y sumarle el m correspondiente. Al final tenemos que recordar deshacer la diferenciación del principio para obtener las predicciones, y poder hallar después el porcentaje de error de cada una y las medias de porcentaje de error tanto por días de la semana como por trimestres.

Tabla 48

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	3.817266	4.079127	4.402708	8.548855	5.211989
Martes	3.414204	4.606166	2.125158	3.014735	3.290066
Miércoles	1.721675	2.103607	4.012326	4.617318	3.113731
Jueves	2.820260	2.770298	4.475170	4.559780	3.656377
Viernes	2.923307	4.129774	2.266199	5.281589	3.650217
Sábado	2.985414	4.531391	4.866716	7.621340	5.001215
Domingo	6.341210	4.482420	6.274681	6.980459	6.019693
Total	3.431905	3.814683	4.080994	5.833281	

A la vista de los resultados anteriores podemos afirmar que este método ha hecho en general unas buenas predicciones ya que la mayor parte de los errores no superan el 5%. Esto se observa en los errores obtenidos en la tabla, puesto que oscilan entre el 1.72% en el primer trimestre del miercoles y el 8.54% en el cuarto trimestre del lunes respectivamente.

Observando ahora los datos por trimestres vemos que el comportamiento de los datos es similar al que llevamos visto hasta ahora. El trimestre en el que se siguen dando las peores predicciones es el cuarto, y con bastante diferencia del resto, el mejor es el primero.

En cuanto a los datos por semana más o menos mantienen las pautas seguidas también en los métodos anteriores, ya que entre semana los errores de predicción son menores que los sábados, domingos y lunes. El día que peor refleja la realidad es el domingo con un 6.01% de error, mientras que entre semana los errores rondan el 3%. Por tanto, debemos destacar que en este modelo existen grandes diferencias entre los errores de los días intrasemanales y el resto.

Analizando ahora únicamente el grupo de días intrasemanales, es decir, desde el Martes al Viernes vemos que las predicciones son buenas con este método, puesto que el error total cometido es de un 3.42%. Obteniendo las peores predicciones en el cuarto trimestre con un 4.36% de error y las mejores en el primer trimestre, con un 2.71%.

Tabla 49

Media del porcentaje de error absoluto (MAPE) de Martes a Viernes

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Ma-Vi	2.719861	3.402461	3.219713	4.368355	3.427598

4.2.2.2. PLRM: siete modelos con temperatura

Cuando hablamos de que vamos a realizar siete modelos de regresión nos referimos como ya explicamos anteriormente a elaborar un modelo de regresión para cada día de la semana.

En esta ocasión no podemos trabajar con nuestra tabla de datos completa, es decir, no podemos trabajar con todas las variables creadas al principio. Vamos a ir comprobando en cada día cuantas y cuáles son las variables dummy que podemos introducir sin que nos dé error la función. Así es, por ejemplo, que en el lunes y el martes podemos incorporar al ajuste las variables de festivos intrasemanales y festivos en Cataluña y Madrid. El miércoles al igual que el viernes sólo podremos incorporar la variable de los festivos intrasemanales. El jueves podremos añadir además de la variable de festivos intrasemanales, la de festivos de Madrid, y el sábado únicamente podremos introducir en el ajuste la variable de festivos en sábado. El domingo no va a permitir que se introduzca ninguna de las variables de este tipo.

Además de estas variables dummy tenemos que incorporar a nuestro modelo el efecto de la temperatura, para ver la influencia que ésta tiene en el comportamiento de la demanda de electricidad. Para ello añadimos a nuestra tabla de datos las dos variables dummy creadas con los efectos de la temperatura.

Una vez tenemos la tabla de datos que necesitamos para hacer los ajustes de cada día de la semana, debemos observar si existe tendencia y corregirla en caso afirmativo. En nuestro caso sabemos que los datos sí tienen tendencia, por tanto debemos hacer una diferenciación en la parte regular para eliminarla.

Por otra parte debemos comprobar si existe o no componente estacional, pero como sabemos, nuestros datos no tienen estacionalidad cuando analizamos cada día de la semana por separado. Por tanto, no va a ser necesario utilizar las variables dummy que hemos creado (D1, D2 y D3) para evitar la estacionalidad en el apartado anterior.

Ahora, con los datos corregidos podemos comenzar a realizar los ajustes para cada día, de igual forma que lo hacíamos en el modelo único de regresión parcialmente lineal. La función que se utiliza para este ajuste es de nuevo la plrm.est.

La forma de realizar el ajuste y obtener tanto los β como los m va a ser similar a cuando realizamos este mismo ajuste sin variables exógenas. Vamos a ir cogiendo en cada caso los 50 datos anteriores al día que queremos predecir, por ejemplo, si queremos predecir el dato 3 del 2012 (3 de enero de 2012), cogeríamos los datos para hacer el ajuste desde el 3 al 53.

Cuando tenemos los valores de β y m podemos realizar las predicciones del mismo modo que realizamos siempre en este tipo de modelos (PLRM).

Debemos recordar que una vez hechas las "predicciones" tenemos que deshacer la diferenciación de datos hecha al principio, es decir, devolver el efecto de la tendencia, para obtener las predicciones reales. Al final se hallan los porcentajes de error y las medias por trimestres y por días de la semana.

La tabla de errores que obtenemos en este estudio es la siguiente:

Tabla 50

Media del porcentaje de error absoluto (MAPE)

	Trimestre 1	Trimestre 2	Trimestre 3	Trimestre 4	Total
Lunes	4.367814	9.166995	7.373705	10.858829	7.941835
Martes	5.160948	6.339504	7.435976	6.008468	6.236224
Miércoles	6.371958	6.939796	7.776544	7.296777	7.096268
Jueves	8.747579	4.899920	6.376330	11.683424	7.926813
Viernes	11.086954	4.657699	4.459660	9.620377	7.456172
Sábado	7.994830	6.422589	5.786231	11.888421	8.023017
Domingo	8.658179	6.091291	6.568417	14.640286	8.989543
Total	7.484037	6.359684	6.539551	10.285226	

Observando de forma general la tabla anterior, podemos decir que los resultados no son tan buenos como esperábamos. Los valores de los errores oscilan entre el 4.36% el primer trimestre de los lunes, y el 14.64 el último trimestre de los domingos. Existe como podemos ver gran diferencia en la calidad de las predicciones según el día que analicemos. Las mejores predicciones se dan los lunes del primer trimestre (como ya dijimos) y los viernes del segundo y tercer trimestre. En cambio las peores predicciones y por tanto los datos que más se alejan de la realidad se dan el último trimestre de los domingos y el último trimestre también de los sábados y jueves. No parece que con estos datos podamos decir que el comportamiento de los datos siga un patrón.

Pasamos ahora a analizar nuestra tabla desde un punto de vista trimestral. Podemos ver que ahora sí que parece que se sigue un patrón ya visto en otros modelos y es que las peores predicciones se dan en el último trimestre de forma destacable. Mientras los demás trimestres se mantienen más o menos en el mismo nivel de error, aunque el primer trimestre sea el más elevado.

Y finalmente centrándonos en los errores según el tipo de día, vemos que también se mantiene una pauta ya seguida por los errores en otros modelos de predicción. Los peores errores se dan los lunes, sábados y domingos. Aunque fijándonos un poco más detenidamente vemos que los jueves tienen un porcentaje de error casi igual a los lunes, y el viernes tampoco tiene un error mucho más bajo que estos días. Podemos decir que los únicos días que tienen unas predicciones un poco mejores son los martes y los miércoles, pero aun así sobrepasan el límite del 5% que habíamos fijado para considerar que las predicciones sean suficientemente buenas.

4.3. Comparación de los Modelos de Predicción

Haciendo un breve repaso por los distintos modelos que hemos analizado y estudiado en este trabajo así como por los errores que hemos obtenido con cada uno de ellos, podemos afirmar que las predicciones que hemos realizado no han sido todo lo buenas que esperábamos. Sin embargo el objetivo de nuestro trabajo es comparar los resultados de cada uno de los métodos, con y sin variables exógenas, para poder determinar con qué modelo y de qué forma obtendríamos las mejores predicciones para cada día de la semana.

Comenzaremos descartando los peores modelos comparando dentro de cada uno de los métodos el modelo único con los siete modelos para cada día de la semana. En el método NAIVE ya nos quedamos con los resultados que obtuvimos porque sólo se puede hacer de este modo.

En cuanto a los datos obtenidos en el ARIMA sin temperatura, podemos apreciar claramente que el ARIMA de los siete modelos es un modelo deficiente en comparación con el modelo ARIMA único; ya que mientras la media de errores del ARIMA de un solo modelo oscila entre valores del 1.89% y el 9.16%, el de siete modelos oscila entre el 5.57% y el 19%.

Comparando ahora los modelos PLRM vemos que el suceso se repite, puesto que el de un solo modelo presenta menores errores con respecto al de siete modelos, eso sí, en este método los errores entre un modelo y el otro varían en menor magnitud que en el ARIMA, en el cual podíamos ver variaciones hasta del 10% entre ambos modelos.

Analizando los mismos métodos pero añadiéndoles la variable exógena, la temperatura, vemos que el comportamiento de nuestros modelos dentro de cada método es similar al que acabamos de explicar, siendo en ambos casos tanto en el ARIMA como en el PLRM mejores los modelos únicos que los siete modelos.

Ahora bien, tenemos un modelo de cada método, el NAIVE y los modelos únicos del ARIMA y del PLRM, tanto con la variable exógena como sin ella. Pero ¿Cuál es el modelo que mejor se ajusta a la realidad a la hora de realizar predicciones?

Comparando el método NAIVE con los modelos únicos del ARIMA y del PLRM, podemos afirmar que aunque el método NAIVE no hace unas malas predicciones (para su poca complejidad), ya que muchas de sus medias de error son menores al 5%, es el método con el que obtenemos los errores más elevados, tal como era de esperar.

Contrastando ahora los errores obtenidos tanto del ARIMA como del PLRM vemos que los resultados obtenidos son muy similares. Si nos fijamos más detenidamente observamos que las diferencias de error entre ambos métodos son mínimas, por ejemplo, el segundo y el tercer trimestre presentan un menor error con el ARIMA mientras los dos restantes son menores en el PLRM. En cuanto a los errores de los días de la semana tenemos que lunes, martes, viernes y domingo obtienen mejores predicciones con el ARIMA en cambio miércoles, jueves y sábado obtienen mejores predicciones con el PLRM. Observando ahora el análisis hecho de los días intrasemanales podemos apreciar que las mejores predicciones se dan con el ARIMA, con un 3.33% de error frente a un 3.4% obtenido en el PLRM, aunque las diferencias entre ambos métodos siguen siendo mínimas

Dicho esto, podemos concluir diciendo que en igualdad de condiciones, el ARIMA podría ser el recomendable puesto que es más sencillo de interpretar y más rápido el obtener predicciones a través de él.

Por otra parte, observamos las tablas de error de los mismos modelos pero añadiendo la variable exógena y vemos que la situación vuelve a repetirse, por tanto también parece el más recomendable el ARIMA de modelo único con temperatura.

Pero nos queda saber si los modelos han mejorado o no en cuanto a las predicciones al añadir la variable exógena.

La verdad es que es muy difícil ver si la situación mejora o no, ya que el cambio en la cuantía de los errores es ínfima cuando añadimos la temperatura. Además de esto, la variación de los errores no sigue ningún patrón aparente, en unos días los errores se reducen y en otros aumentan. Debemos hacer entonces un análisis más detallado en esta ocasión para poder sacar una conclusión.

Las predicciones mejoran en el primer y cuarto trimestre al añadirles a los modelos las variables de la temperatura, aunque como decíamos el error varía únicamente unas décimas. En cambio en los dos trimestres restantes la situación empeora.

Haciendo un análisis desde un punto de vista semanal y fijándonos los días en lo que las predicciones mejoran con la incorporación de la temperatura, vemos que son los lunes, los viernes, los sábados y los domingos. Recordamos que la mejora que se observa en estos días es ínfima y debemos estudiar si es productivo el aumento de cálculos y de tiempo que conlleva añadir dos variables más con respecto a la mínima mejora de resultados que conseguimos.

De todos modos, debemos hacer una pequeña puntualización y es que parece que añadiendo la variable de la temperatura las predicciones mejoran en los días que tenían unos errores mayores en todos los modelos, por su comportamiento distinto al resto de días. Ya que si recordamos siempre obteníamos los peores errores de cada modelo los lunes, viernes, sábados y domingos.

Conclusión, con nuestros datos la forma en la que obtendríamos las mejores predicciones para la demanda eléctrica de 2012 sería, utilizando el ARIMA de un solo modelo con temperatura para los lunes, viernes, sábados y domingos, y el ARIMA de un solo modelo sin temperatura para los martes, miércoles y jueves.

5. CONCLUSIONES Y AMPLIACIÓN

Al principio de este trabajo una de las primeras conclusiones que sacamos viendo los gráficos de la demanda de los años 2011 y 2012, es que la demanda de electricidad es mayor en los meses de invierno que en verano. Nuestra primera suposición es que esto se debe a la temperatura y que por tanto la demanda se va a incrementar con la bajada de las temperaturas, por eso decidimos hacer un análisis y observar la influencia de ésta en el consumo eléctrico.

Por otra parte también hemos querido ver la importancia que tenía que un día fuese festivo o se hiciese huelga así como que las personas tuviesen vacaciones. Por ello nuestro primer análisis fue sobre el efecto de los días festivos y las huelgas, y como vimos sí tienen un impacto relevante en el consumo energético, ya que cuando un día tiene alguna de estas características, la demanda sufre una fuerte bajada en ese día, e incluso dependiendo del día de la semana en qué éste ocurra, puede afectar a los días anteriores y posteriores. Sin embargo, no ocurre lo mismo con los períodos vacacionales (entendiendo nosotros que es el mes de agosto por excelencia), ya que en el estudio realizado veíamos que ésta era una variable totalmente irrelevante en todos los modelos. Esto puede deberse a que no tuvimos en cuenta que aunque gran parte de las personas tienen sus vacaciones en el mes de agosto, muchas otras no y es más, muchas personas, especialmente en la industria hostelera y de servicios, trabajan más horas y contratan a más empleados en este mes, por tanto ya se compensa el efecto vacacional.

Un dato curioso que nos sorprendió fue que en cuanto a los días festivos de las grandes ciudades que escogimos, sólo son relevantes en la demanda de electricidad dos de ellas, los días festivos de Madrid y Cataluña, siendo de mayor relevancia en cuanto a términos de la demanda éstos últimos. Los festivos de Cataluña provocan una reducción del consumo eléctrico de España en el doble que los festivos que se celebran en Madrid.

Ahora bien, al principio de este documento explicábamos que nuestro objetivo además de hacer predicciones sobre la demanda de electricidad para el año 2012 con distintos modelos, era hacer una comparación de los mismos para saber cuál de los modelos que hemos utilizado nos da unos valores más próximos a la realidad. Ahora que ya hemos realizado el análisis aplicado a nuestros datos podemos afirmar que el mejor modelo para realizar predicción (de los que hemos estudiado) es el ARIMA de modelo único. Por una parte tenemos que admitir que los errores obtenidos con este modelo son muy similares a los que obtuvimos con el PLRM de modelo único, sin embargo, mirando los resultados con detalle vemos que el ARIMA predice una mayor cantidad de datos con menor error y que además lo hace en mucho menos tiempo computacional. Por tanto nos quedamos con este modelo para predicciones futuras.

Las predicciones con los distintos modelos han sido muy similares en cuanto a cuáles son los días que más cuestan predecir o los trimestres, debido a la variabilidad en su comportamiento. En general con todos los modelos vimos que los días más difíciles de predecir son los sábados, domingos y lunes, ya que son días que tienen un comportamiento muy diferente al resto de la semana; y en cuanto a trimestres, el más difícil de predecir es el cuarto que coincide con los meses de otoño e invierno. Esto nos indica que en este trimestre

y en estos días de la semana los datos están más dispersos y son más extremos por eso su dificultad a la hora de predecirlos.

Finalmente, ¿podemos decir que nuestra primera hipótesis de que la temperatura tenía una influencia relevante en la demanda de electricidad? La respuesta es no. Una vez hecho el análisis incorporándoles a nuestros modelos los efectos de la temperatura vimos que la influencia de la misma era mínima en todos los modelos. No obstante tenemos una pequeña curiosidad en este caso, y es que cuando comparamos los modelos con y sin temperatura, vimos que generalmente las predicciones son las mismas e incluso mejores con el ARIMA sin la temperatura, aunque en los lunes, sábados y domingos (los días con los mayores errores de predicción) las predicciones mejoran al añadirle al modelo el efecto de la temperatura.

Puesto que los coeficientes que acompañan a las variables exógenas en los modelos ARIMA únicos no son significativamente distintos de cero, no podemos afirmar que la temperatura influya sobre la demanda. Después de darle vueltas pensamos en el hecho de que en los meses de "invierno" en el último trimestre, puede que aumente la demanda de electricidad debido a que los días se vuelven más cortos y tenemos menos horas de luz que en verano o en el resto del año donde necesitamos un menor uso de la electricidad. Aunque a este factor habría que añadirle muchos más. Éste sería por ejemplo un análisis interesante en el momento en que queramos ampliar nuestro trabajo.

Para ampliar nuestro trabajo tenemos varias ideas: por una parte lo que acabamos de comentar, sería buena idea estudiar el efecto de las horas de luz sobre la demanda de electricidad, ya que esto podría darnos respuesta a lo que nos preguntábamos al principio. Y también podemos incrementar el número de modelos de nuestro estudio, haciendo un análisis por ejemplo con modelos funcionales tanto lineales como no paramétricos.

6. BIBLIOGRAFÍA

Aneiros Pérez, G. y Quintela del Río, A. (2001): Modified Cross-Validation in Semiparametric regression models with dependent errors, *Comm. Statist Theory Methods*, Vol. 30, (pp. 289 – 307).

Aneiros Pérez, G y López Cheda, A. (2014): PLRModels: Statistical inference in partial linear regression models. R package version 1.1. `http://CRAN.R-project.org/package= PLRModels`

Box, G.E.P. and Jenkins, G.M. (1976): Time series analysis: forecasting and control. *Prentice Hall*.

Cancelo, J. R, Espasa, A. &Grafe, R. (2008): Forecasting the electricity load from one day to one week ahead for the Spanish system operator. *Internation Journal Forecasting*; Vol. 24, (pp. 588-602).

Engle, R. F., Granger, W. J., Rice, J. & Weiss, A. (1986): Semiparametricestimates of the relation between weather and electricitysales. *Journal of the American Statistical Association*, Vol.81, (pp. 310-320).

Espasa,A., Revuelta, J.M., & Cancelo, J.R. (1996):Automatic modeling of daily series of economic activity. In A. Prat (Ed), *Proceedings in computational statistics COMPSTAT 1996* (pp. 51-63).

Fan, J.Y. and McDonald, J.D. (1994): A real-time implementation of short-term load forecasting for distribution power systems, *IEEE Transactions on Power Systems* 9, (pp. 988-994)

Härdle, W., Liang, H. and Gao, J. (2000): Partially linear models. *Physica Verlag*.

Hernández, J. (2007): Análisis de series temporales económicas II. España. Primera Edición. *Editorial ESIC*.

Gross, G. and Galiana, F. D. (1987): "Short-term load forecasting", *Proceedings of the IEEE*, Vol. 75, (pp. 1558-1573)).

Peña, D. (2005): Análisis de series temporales. España. *Editorial alianza*.

Ranaweera, D. K., Karady, G.G., & Farmer, R. G. (1997): Economic impact analysis of load forcasting. *IEEE Transactions on Power Systems*, Vol. 12, (pp. 1388 – 1392).

Rice, J. (1984): Bandwidth choice for nonparametric regression, *Ann. Statist*. , Vol. 12, N 4, (pp. 1215 – 1230).

Robinson, P. (1988): Root-n-consistent semiparametric regression, *Econometrica*, Vol. 56, (pp. 931 – 954).

Schmalensee y Stocker, (1999): Household gasoline demand in the United States, *Econometrica*, Vol.67, (pp. 645 – 662).

Smith, M. (2000): Modeling and short-term forecasting of New South Wales electricity system load. *Journal of Business and Economic Statistics*, Vol. 18, (pp. 465-478).

Speckman, P. (1988): Kernel smoothing in partial linear models, *Journal of the Royal Statistical Society*, Series B, Vol. 50, (pp. 413 – 436).

Weron, R. (2006): Modeling and forecasting electricity loads and prices: a statistical approach. *Editorial Wiley Finance* .

Wolfgang, H., Hua, L. and Jiti, G. (2000): Partially Linear Models, *Contributions to statistics* (pp. 127 – 180).

APÉNDICE

INDICE FIGURAS

Figura 1. Demanda de Electricidad (2011 - 2012)……..…4
Figura 2. Demanda de Electricidad en el 2011 (GWh) ……..…….5
Figura 3. Demanda de Electricidad en invierno y en verano (2011) ……...…6
Figura 4. Demanda de Electricidad en Enero y en Julio (2011) ……....7
Figura 5. Demanda de Electricidad en Jueves y Viernes Santo……...…8
Figura 6. Demanda de Electricidad en 2012 (Días Especiales) ……..……9
Figura 7. Demanda de Electricidad en el Día de la Hispanidad……..…...10
Figura 8. Demanda de Electricidad en Huelga General……..…11
Figura 9. Demanda de Electricidad en el Día de la Constitución y de la Inmaculada Concepción……...12
Figura 10. Comparación de la demanda real de electricidad y las predicciones de 2012……..…….21
Figura 11. Errores de predicción de 2012……..21
Figura 12. Demanda de Electricidad en 2011……....23
Figura 13. Serie original y correlaciones……...23
Figura 14. Serie original diferenciada en la parte regular y correlaciones…….....…24
Figura 15. Serie original diferenciada en la parte estacional y correlaciones……..…25
Figura 16. Q-Q plot y Ljung-Box ……..…...28
Figura 17. Errores de predicción de 2012……..…29
Figura 18. Demanda de electricidad de los Lunes de 2011……....31
Figura 19. Serie original y correlaciones ……..…….31
Figura 20. Serie original (diferenciada) y correlaciones……..…….....31
Figura 21. Q-Q plot y Ljung Box…….....33
Figura 22. Demanda de electricidad de los Martes de 2011…….……..…34
Figura 23. Serie original y serie diferenciada en la parte regular…….....34
Figura 24. Q-Q plot y Ljung Box…….……36
Figura 25. Demanda de Electricidad de los Miércoles de 2011……..…….....37
Figura 26. Serie original y serie diferenciada en la parte regular……..……37
Figura 27. Q-Q plot y Ljung Box…….…38

Figura 28. Demanda de Electricidad de los Jueves de 2011……..39
Figura 29. Serie original y serie diferenciada en la parte regular……...40
Figura 30. Q-Q plot y Ljung box…….....41
Figura 31. Demanda de Electricidad de los Viernes de 2011……42
Figura 32. Serie original y serie diferenciada en la parte regular……..42
Figura 33. Q-Q plot y Ljung Box……..44
Figura 34. Demanda de Electricidad de los Sábado de 2011……..……...45
Figura 35. Serie original y serie diferenciada en la parte regular……...45
Figura 36. Q-Q plot y Ljung Box…….....46
Figura 37. Demanda de Electricidad de los Domingo de 2011…….....47
Figura 38. Serie original y serie diferenciada en la parte regular…….....48
Figura 39. Q-Q plot y Ljung Box…….....49
Figura 40. Demanda de Electricidad (2011 - 2012) ……..51
Figura 41. . Evolución de la demanda y la temperatura en el 2011…….....55
Figura 42: Relación entre la Demanda (GWh) y la Temperatura (C) …….....56
Figura 43. Q-Q plot y Ljung Box…….....59
Figura 44: Q-Q plot y Ljung Box…….....63
Figura 45: Q-Q plot y Ljung Box…….....65
Figura 46: Q-Q plot y Ljung Box…….....66
Figura 47: Q-Q plot y Ljung Box…….....68
Figura 48: Q-Q plot y Ljung Box…….....70
Figura 49: Q-Q plot y Ljung Box…….....72
Figura 50: Q-Q plot y Ljung Box…….....74

CÓDIGO R UTILIZADO

NAIVE

```
#cargamos datos
datos<-read.csv("demanda.csv",header=T,sep=";")
attach(datos)
A<-as.matrix(datos);A


B<-matrix(0,366,11);B   #Introduciremos los datos de 2012

### Vamos a predecir los datos de 2012 por el método NAIVE

#ponemos en la 1ª columna el tipo de dia
B[,1]<-A[A[,1]>365,2]

###posición domingos
domingos<-A[,2]==7

#demandas de los domingos
DEM.dom<-A[domingos,10]

#cantidad de domingos 2012
num.dom2012<-sum(A[366:731,2]==7);num.dom2012

#predición NAIVE domingos 2012
NAI.dom2012<-DEM.dom[-1][(105-num.dom2012-1):103];NAI.dom2012

#Insertamos datos
B[B[,1]==7,2]<-NAI.dom2012


###posición sábados
sábados<-A[,2]==6
```

```
#demandas de los domingos
DEM.sab<-A[sabados,10];DEM.sab

#cantidad de domingos 2012
num.sab2012<-sum(A[366:731,2]==6);num.dom2012

#predición NAIVE domingos 2012
NAI.sab2012<-DEM.sab[-1][(105-num.sab2012-1):103];NAI.sab2012

#Insertamos datos
B[B[,1]==6,2]<-NAI.sab2012


###posición lunes
lunes<-A[,2]==1

#demandas de los lunes
DEM.lun<-A[lunes,10];DEM.lun

#cantidad de lunes 2012
num.lun2012<-sum(A[366:731,2]==1);num.lun2012

#predición NAIVE lunes 2012
NAI.lun2012<-DEM.lun[-1][(105-num.lun2012-1):103];NAI.lun2012

#Insertamos datos
B[B[,1]==1,2]<-NAI.lun2012


###posición martes
martes<-A[,2]==2

#demandas de los martes
DEM.mar<-A[martes,10];DEM.mar
```

```
#cantidad de martes 2012
num.mar2012<-sum(A[366:731,2]==2);num.mar2012

#predición NAIVE martes 2012
NAI.mar2012<-DEM.lun[-1][(num.lun2012-1):103];NAI.mar2012

#Insertamos datos
B[B[,1]==2,2]<-NAI.mar2012


###posición miercoles
miercoles<-A[,2]==3

#demandas de los miercoles
DEM.mier<-A[miercoles,10];DEM.mier

#cantidad de miercoles 2012
num.mier2012<-sum(A[366:731,2]==3);num.mier2012

#predición NAIVE miercoles 2012
NAI.mier2012<-DEM.mar[-1][(105-num.mar2012-1):103];NAI.mier2012

#Insertamos datos
B[B[,1]==3,2]<-NAI.mier2012


###posición jueves
jueves<-A[,2]==4

#demandas de los jueves
DEM.juev<-A[jueves,10];DEM.juev

#cantidad de jueves 2012
num.juev2012<-sum(A[366:731,2]==4);num.juev2012

#predición NAIVE jueves 2012
```

```
NAI.juev2012<-DEM.mier[-1][(105-num.mier2012-1):103];NAI.juev2012

#Insertamos datos
B[B[,1]==4,2]<-NAI.juev2012


###posición viernes
viernes<-A[,2]==5

#demandas de los viernes
DEM.vier<-A[viernes,10];DEM.vier

#cantidad de viernes 2012
num.vier2012<-sum(A[366:731,2]==5);num.vier2012

#predición NAIVE viernes 2012
NAI.vier2012<-DEM.juev[-1][(105-num.juev2012-1):103];NAI.vier2012

#Insertamos datos
B[B[,1]==5,2]<-NAI.vier2012


#Incorporamos el resto de datos a la matriz (festivos y Demanda)
B[,3]<-A[A[,1]>365,10] #Demanda 2012
B[,4]<-A[A[,1]>365,3]#Festivos nacionales, huelgas
B[,5]<-A[A[,1]>365,4]#Festivos sábados, huelgas
B[,6]<-A[A[,1]>365,5]#Dias de Agosto
B[,7]<-A[A[,1]>365,6]#Festivos Andalucia
B[,8]<-A[A[,1]>365,7]#Festivos Cataluña
B[,9]<-A[A[,1]>365,8]#Festivos Madrid
B[,10]<-A[A[,1]>365,9]#Festivos Valencia


#Añadimos los errores MAPE A LA columna 11
B[,11]<-abs(((B[,3])-(B[,2]))/(B[,3]))*100
```

```
numerror<-sum(B[,11]>5);numerror
errores<-B[B[,11]>5,c(1,11)];errores
erroreslu<-sum(errores[,1]==1);erroreslu
erroressab<-sum(errores[,1]==6);erroressab
erroresdom<-sum(errores[,1]==7);erroresdom


##CREAMOS TABLA CON LA MEDIA DE LOS ERRORES MAPE PARA CADA TRIMESTRE DE 2012
############################################################

#PRIMER TRIMESTRE
C<-matrix(0,91,2);C

t1<-B[1:91,1];t1
C[,1]<-t1
errort1<-B[1:91,11]
C[,2]<-errort1

l1<-C[C[,1]==1,2];l1
medial1<-mean(l1);medial1
ma1<-C[C[,1]==2,2];ma1
mediama1<-mean(ma1);mediama1
mi1<-C[C[,1]==3,2];mi1
mediami1<-mean(mi1);mediami1
j1<-C[C[,1]==4,2];j1
mediaj1<-mean(j1);mediaj1
v1<-C[C[,1]==5,2];v1
mediav1<-mean(v1);mediav1
s1<-C[C[,1]==6,2];s1
medias1<-mean(s1);medias1
d1<-C[C[,1]==7,2];d1
mediad1<-mean(d1);mediad1
```

```
### SEGUNDO TRIMESTRE
D<-matrix(0,91,2)

t2<-B[92:182,1];t2
D[,1]<-t2
errort2<-B[92:182,11]
D[,2]<-errort2

l2<-D[D[,1]==1,2];l2
medial2<-mean(l2);medial2
ma2<-D[D[,1]==2,2];ma2
mediama2<-mean(ma2);mediama2
mi2<-D[D[,1]==3,2];mi2
mediami2<-mean(mi2);mediami2
j2<-D[D[,1]==4,2];j2
mediaj2<-mean(j2);mediaj2
v2<-D[D[,1]==5,2];v2
mediav2<-mean(v2);mediav2
s2<-D[D[,1]==6,2];s2
medias2<-mean(s2);medias2
d2<-D[D[,1]==7,2];d2
mediad2<-mean(d2);mediad2


### TERCER TRIMESTRE
E<-matrix(0,92,2)

t3<-B[183:274,1];t3
E[,1]<-t3
errort3<-B[183:274,11]
E[,2]<-errort3

l3<-E[E[,1]==1,2];l3
medial3<-mean(l3);medial3
ma3<-E[E[,1]==2,2];ma3
mediama3<-mean(ma3);mediama3
```

```
mi3<-E[E[,1]==3,2];mi3
mediami3<-mean(mi3);mediami3
j3<-E[E[,1]==4,2];j3
mediaj3<-mean(j3);mediaj3
v3<-E[E[,1]==5,2];v3
mediav3<-mean(v3);mediav3
s3<-E[E[,1]==6,2];s3
medias3<-mean(s3);medias3
d3<-E[E[,1]==7,2];d3
mediad3<-mean(d3);mediad3



### CUARTO TRIMESTRE
F<-matrix(0,92,2)

t4<-B[275:366,1];t4
F[,1]<-t4
errort4<-B[275:366,11]
F[,2]<-errort4

l4<-F[F[,1]==1,2];l4
medial4<-mean(l4);medial4
ma4<-F[F[,1]==2,2];ma4
mediama4<-mean(ma4);mediama4
mi4<-F[F[,1]==3,2];mi4
mediami4<-mean(mi4);mediami4
j4<-F[F[,1]==4,2];j4
mediaj4<-mean(j4);mediaj4
v4<-F[F[,1]==5,2];v4
mediav4<-mean(v4);mediav4
s4<-F[F[,1]==6,2];s4
medias4<-mean(s4);medias4
d4<-F[F[,1]==7,2];d4
mediad4<-mean(d4);mediad4
```

```
#Creamos una tabla con todas las medias
MediaMAPE<-matrix(0,4,7)
MediaMAPE[1,]<-c(medial1,mediama1,mediami1,mediaj1,mediav1,medias1,mediad1)
MediaMAPE[2,]<-c(medial2,mediama2,mediami2,mediaj2,mediav2,medias2,mediad2)
MediaMAPE[3,]<-c(medial3,mediama3,mediami3,mediaj3,mediav3,medias3,mediad3)
MediaMAPE[4,]<-c(medial4,mediama4,mediami4,mediaj4,mediav4,medias4,mediad4)
colnames(MediaMAPE)<-
c("Lunes","Martes","Miercoles","Jueves","Viernes","Sabado","Domingo")
rownames(MediaMAPE)<-c("Trimestre 1","Trimestre 2", "Trimestre 3", "Trimestre 4")
MediaMAPE


#media de los trimestres
t1<-C[,2];t1
mediat1<-mean(t1);mediat1
t2<-D[,2];t2
mediat2<-mean(t2);mediat2
t3<-E[,2];t3
mediat3<-mean(t3);mediat3
t4<-F[,2];t4
mediat4<-mean(t4);mediat4
trimestres<-matrix(0,4,1)
trimestres[1,1]<-mediat1
trimestres[2,1]<-mediat2
trimestres[3,1]<-mediat3
trimestres[4,1]<-mediat4
trimestres

#media para cada dia de la semana
lunes<-c(medial1,medial2,medial3,medial4)
ml<-mean(lunes)
martes<-c(mediama1,mediama2,mediama3,mediama4)
mma<-mean(martes)
miercoles<-c(mediami1,mediami2,mediami3,mediami4)
mmi<-mean(miercoles)
jueves<-c(mediaj1,mediaj2,mediaj3,mediaj4)
```

```
mj<-mean(jueves)
viernes<-c(mediav1,mediav2,mediav3,mediav4)
mv<-mean(viernes)
sabado<-c(medias1,medias2,medias3,medias4)
ms<-mean(sabado)
domingo<-c(mediad1,mediad2,mediad3,mediad4)
md<-mean(domingo)
semana<-matrix(0,1,7)
semana[1,1]<-ml
semana[1,2]<-mma
semana[1,3]<-mmi
semana[1,4]<-mj
semana[1,5]<-mv
semana[1,6]<-ms
semana[1,7]<-md
semana
```

ARIMA sin temperatura (modelo único)

```
demanda.ts <- ts(demanda,frequency=7);demanda.ts
head(demanda.ts)

D<-as.matrix(demanda.ts);D

#Separamos las valores de las variables D por años.
#para el año 2011
D1<-D[D[,1]<366,3]
D2<-D[D[,1]<366,4]
D3<-D[D[,1]<366,5]
D4<-D[D[,1]<366,6]
D5<-D[D[,1]<366,7]
D6<-D[D[,1]<366,8]
D7<-D[D[,1]<366,9]
```

```
#para el año 2012
D21<-D[D[,1]>365,3]
D22<-D[D[,1]>365,4]
D23<-D[D[,1]>365,5]
D24<-D[D[,1]>365,6]
D25<-D[D[,1]>365,7]
D26<-D[D[,1]>365,8]
D27<-D[D[,1]>365,9]

# Reservamos los datos del último año para valorar el comportamiento del modelo Box-Jenkins que ajustaremos al resto de datos.

demanda1.ts <- ts(D[D[,1]<366,10], frequency=7)
demanda2.ts <- ts(D[D[,1]>365,10],frequency=7)


#Representamos nuestros datos
par(mfrow=c(1,1))
plot(demanda1.ts, type="o", xlab="Semanas", ylab="Demanda Electricidad",main="Demanda de Electricidad", col="blue")

#miramos si hay tendendia y componente estacional con las correlaciones
windows()
par(mfrow=c(3,1))
plot(demanda1.ts, type="l", xlab="Año", ylab="")
acf(demanda1.ts, xlab="Retardo",ylab="fas", main="", lag.max=length(demanda1.ts)/4)
pacf(demanda1.ts, xlab="Retardo",ylab="fap", main="", lag.max=length(demanda1.ts)/4)

# Presencia de tendencia (y de componente estacional): necesidad de diferenciación regular
dif.demanda1.ts <- diff(demanda1.ts, lag=1)


### Gráfico secuencial, fas y fap de la serie diferenciada

windows()
par(mfrow=c(3,1))
plot(dif.demanda1.ts, type="l", xlab="Año", ylab="")
```

```
acf(dif.demanda1.ts, xlab="Retardo",ylab="fas", main="",lag.max=length(dif.demanda1.ts)/4)
pacf(dif.demanda1.ts, xlab="Retardo",ylab="fap",main="", lag.max=length(dif.demanda1.ts)/4)


# Presencia de componente estacional con periodo estacional s=7: necesidad de diferenciación estacional
dif7.dif.demanda1.ts <- diff(dif.demanda1.ts, lag=7)

### Gráfico secuencial, fas y fap de la serie diferenciada regularmente y estacionalmente
windows()
par(mfrow=c(3,1))
plot(dif7.dif.demanda1.ts, type="l", xlab="Año", ylab="")
acf(dif7.dif.demanda1.ts, xlab="Retardo",ylab="fas", main="", lag.max=length(dif7.dif.demanda1.ts)/4)
pacf(dif7.dif.demanda1.ts, xlab="Retardo",ylab="fap", main="", lag.max=length(dif7.dif.demanda1.ts)/4)

#creamos matriz xreg con las variables dummy (festivos, huelgas, etc)
xreg<-matrix(0,365,7)
xreg[,1]<-D1
xreg[,2]<-D2
xreg[,3]<-D3
xreg[,4]<-D4
xreg[,5]<-D5
xreg[,6]<-D6
xreg[,7]<-D7
xreg

#buscamos major ARIMA
best.arima.TSA(x=demanda1.ts, p.max=5, q.max=5, d=1, P.max=5, Q.max=5, D=1, xreg=xreg, criterio="BIC", dist.max.crit=10)

#  p  q  P  Q   BIC
#  1  1  1  1 3160.369
#  1  1  0  2 3160.582

#realizamos el ajuste
ajuste<-arimax(demanda1.ts,xreg=xreg,order=c(1,1,1), seasonal=list(order=c(1,1,1)))
```

```
ajuste
abs(ajuste$coef)/(1.96*sqrt(diag(ajuste$var.coef)))


ajuste<-arimax(demanda1.ts,xreg=xreg[,-3],order=c(1,1,1), seasonal=list(order=c(1,1,1)))
ajuste
abs(ajuste$coef)/(1.96*sqrt(diag(ajuste$var.coef)))
ajuste<-arimax(demanda1.ts,xreg=xreg[,-c(3,4)],order=c(1,1,1), seasonal=list(order=c(1,1,1)))
ajuste
abs(ajuste$coef)/(1.96*sqrt(diag(ajuste$var.coef)))


ajuste<-arimax(demanda1.ts,xreg=xreg[,-c(3,4,7)],order=c(1,1,1), seasonal=list(order=c(1,1,1)))
ajuste
abs(ajuste$coef)/(1.96*sqrt(diag(ajuste$var.coef)))



# ETAPA: ANÁLISIS DE RESIDUOS
################################

###### Métodos gráficos

## Gráficos secuencial y Q-Q normal de los residuos
windows()
par(mfrow=c(2,1))
plot(ajuste$resid, type="l", xlab="Año", ylab="")
abline(h=0)

qqnorm(ajuste$resid)
qqline(y=(1.4*ajuste$resid), col=4,lty=3,lwd=2)
qqline(ajuste$resid)



###### Contrastes de hipótesis
```

```
## Independencia:fas y fap de los residuos
windows()
par(mfrow=c(2,1))
acf(ajuste$resid, main="")
pacf(ajuste$resid, main="")

# fas y Ljung-Box
windows()
tsdiag(ajuste, gof.lag=50)

# mu_a = 0
t.test(residuals(ajuste), mu=0)

# Normalidad
library(tseries)
jarque.bera.test(residuals(ajuste))
shapiro.test(residuals(ajuste))




#########################################################
######PREDICCIÓN##
#########################################################

demanda2012.pr.t <- 0
for (t in 1:366){
 cat(t, " ")
 index <-  (D[,1]>=t)&(D[,1]<=(364+t))
 serie.t<- ts(D[index,10], frequency=7)
 xreg.t<- D[index,c(3,4,7,8)]
 newxreg.t <- as.data.frame(t(D[(365+t),c(3,4,7,8)]))

 ajuste.t<-arimax(serie.t,xreg=xreg.t,order=c(1,1,1),
seasonal=list(order=c(1,1,1)),fixed=ajuste$coef)
 ajuste.t
```

```
  demanda2012.pr.t[t] <- predict( ajuste.t, n.ahead=1, newxreg=newxreg.t)$pred
}

demanda2012.pr.t



#creamos matriz para de predicciones y valores reales para luego poder hallar las medias del % de error
B<-matrix(0,366,4)
B[,1]<-D[D[,1]>365,2]
B[,2]<-D[D[,1]>365,10]
B[,3]<-demanda2012.pr.t
B[,4]<-abs(((B[,2])-(B[,3]))/(B[,2]))*100
B
```

PLRM sin temperatura (modelo único)

```
datos<-as.matrix(demanda);datos

#########
#CREAMOS las variables Dummy indicando cuales van a ser lunes, sabado y domingo.
#########

#LUNES - domingo
x<-D[,2]
D1<-0
for(i in 1:731){
  if(x[i]==1) D1[i]<-1
  else {D1[i]<-0}
}
```

```
D1


#SABADO-viernes
x<-D[,2]
D2<-0
for(i in 1:731){
  if(x[i]==6) D2[i]<-1
  else {D2[i]<-0}
}
D2


#DOMINGO  - sabado
x<-D[,2]
D3<-0
for(i in 1:731){
  if(x[i]==7) D3[i]<-1
  else {D3[i]<-0}
}
D3


#####
#Añado las variables Dummy que creé a mi matriz total-> datos
#####
datos<-cbind(D,D1,D2,D3)
datos
head(datos)

datos.ts <- ts(datos[,10],frequency=7);datos.ts

#representamos los datos
par(mfrow=c(1,1))
plot(datos.ts, type="o", xlab="Semanas", ylab="Demanda Electricidad",main="Demanda de Electricidad", col="blue")
```

```
#Diferencio los datos para eliminar la tendencia de los mismos
dif.datos <- diff(datos, lag=1)
head(dif.datos)
dim(dif.datos)# en estos momento tengo una matriz de 730x13


##################
# CREAMOS LA MATRIZ DATA TOTAL
##################
#ponemos en la primera columna la variable respuesta "Y".
#desde la 2ª columna a la 8, añadimos las variables "X"
#en la columna 9 ponemos la t
#y yo le añado por último una columna contabilizando las filas

p=6
t<-dif.datos[1:729,10]
data<-matrix(0,729,9)
data[,1]<-dif.datos[2:730,10]
data[,2:4]<-datos[3:731, c(11,12,13)] #sin diferenciar
data[,5:7]<-dif.datos[2:730,c(3,7,8)] #festivos diarios, festivos en Cataluña y festivos en Madrid
data[,p+2]<-t
data[,9]<-seq(1:729)
head(data)

#Buscamos el minimo valor de t es -155.498, para hacerlas positivas sumaremos a todos los valores
160
min(data[,8])

#sumamos en la columna de las "t" 160 para hacerlo positivo
x<-data[,8]
t<-0
for(i in 1:729){
  t[i]<-x[i]+160
}
t
```

```
#y modificamos el valor de la columna "t" con los nuevos valores (todos positivos)
data[,p+2]<-t
data



######
##Hacemos el ajuste
####

ajuste<-list()
for(i in 1:366){
 index<-(data[,9]>=i)&(data[,9]<=(362+i))
 dat<-data[index,c(1:8)]#matriz que me interesa
 newt<-(data[363+i,8])
 ajuste[[i]]<-plrm.est(data=dat, newt= newt)}

ajuste1npsint<-ajuste
ajuste1npsint[[i]]$beta
ajuste1npsint[[i]]$m.newt



##############################
#  Hacemos las predicciones

y<-0
for(i in 1:366){
 index<-(data[,9]=363+i) #cogemos indices a partir de 364 a 729 (1 fila por vez)
 newx<-data[index,c(2:7)] #cogemos vector de los x de 2012
 beta<-ajuste1npsint[[i]]$beta  #cogemos los beta de cada caso
 mnewt<-ajuste1npsint[[i]]$m.newt  #cogemos lo m de cada caso
 variacion<-newx%*%beta  #hallamos la variación de cada día de 2012 (afectada por las variables
dummy)
 y[[i]]<-variacion+mnewt}  #sumamos los m nuevos correspondientes
y
```

```
demanda1<-datos[365:730,10]
y
prediccion<-demanda1+y;prediccion

demanda2<-datos[366:731,10]
error<-(demanda2-prediccion)/demanda2*100;error


#creamos la matriz con los valores reales y predichos para poder luego hallar el % de error

B<-matrix(0,366,4)
B[,1]<-demanda[demanda[,1]>365,2]  #tipo de día de 2012
B[,2]<-demanda[demanda[,1]>365,10] # demanda de 2012
B[,3]<-prediccion
B[,4]<-abs(error)
B
```