Modelos predictivos del consumo energético para un centro comercial

Programa de experto en Data Science U-TAD

Autor:
Raúl Amarelle Valera

Director:
Carlos Gil Bellosta

03/09/2018

Índice

1. Resumen ejecutivo

El presente proyecto trata del análisis de un dataset que contiene, entre sus variables, datos del consumo energético de un centro comercial a lo largo de un periodo completo de dos años.

El objeto de este proyecto es analizar dicha serie temporal y desarrollar un nuevo modelo de predicción de consumo energético, para compararlo luego con su modelo predictivo actual y ver cuál de las dos opciones es mejor.

El proyecto se estructura en tres bloques principales:

1. Lectura y tratamiento de datos.
2. Visualización de datos y otras variables.
3. Análisis de varios modelos predictivos de la serie temporal para escoger la mejor opción.
4. Comparativa de modelo y conclusiones.

La serie y el cálculo de predicciones se han realizado aplicando cuatro métodos:

1. Impacto causal utilizando modelos bayesianos de series de tiempo
2. Utilización del método Holt-Winters
3. Método ARIMA de predicción
4. Árboles de clasificación y regresión, RPART Tree.

El empleo de diferentes técnicas ha permitido comparar diferentes resultados de predicción y cómo hemos tratado de mejorarlo, los cuales se explicarán en el presente documento.

El resultado final es que el nuevo modelo predictivo es mejor que el utilizado actualmente.

2. Problemática y objetivos

Como sabemos, la electricidad es un bien que no puede ser almacenado, por lo que es de mucha utilidad conocer la previsión de consumo, aunque sea a corto plazo (un día) con respecto el consumo real.

El objetivo principal sería desarrollar un modelo predictivo que estime nuevas predicciones, compararla con el modelo predictivo actual y decidir cuál de las dos opciones predice mejor, con mejor error.

Otros objetivos específicos:

- Análisis y tratamiento del dataset
- Generación de otra información que pudiera ser relevante.
- Analizar diferentes modelos estadísticos para la previsión de la demanda así como un análisis comparativo de la calidad predictiva de cada modelo.

3. Metodología

Tal y como hemos avanzado, el proyecto se estructura en los siguientes bloques:

1. Lectura, análisis y tratamiento de datos.
2. Visualización de datos y otras relaciones entre variables.
3. Casos de análisis de las series temporales y predicciones.
4. Conclusiones

Para el cálculo de predicciones he utilizado cuatro métodos diferentes:

1. Impacto causal utilizando modelos bayesianos de series de tiempo
2. Utilización del método Holt-Winters
3. Método ARIMA de predicción
4. Árboles de clasificación y regresión, RPART Tree.

4. Tecnologías utilizadas

En la parte técnica, se ha utilizado R como lenguaje de programación.

Análisis y tratamiento de datos: uso de algunas librerías representativas como ggplot2, dplyr, data.table, zoo, etc.

Machine Learning: librerías bsts, Holt-Winters, forecast, tseries, MLmetrics, etc.

5. Análisis del dataset inicial

Este dataset es una serie temporal con los siguientes datos relativos a un centro comercial.

Breve explicación de las variables:

1. Fecha: fecha de la medición, sólo hay una medición diaria
2. Estimado: sólo aparece "No", no nos aporta información
3. Kwh: potencia real consumida
4. LB: Línea base, es decir, la predicción de potencia consumida
5. De 5 a 10. CCDD o CHDD: Estas seis columnas son de gradientes de temperaturas, 3 de Cooling-Degree Day y 3 de Cooling Heating-Degree Day. Es la diferencia entre el promedio diario de temperatura y una determinada temperatura base de referencia, que suele ser la exterior.
   CHDD18 = 8.5, quiere decir que sobre la temperatura de 18º, el centro comercial ha tenido que climatizar 8.5º. Por eso CHDD19=9.5º y CHDD20=10.5º, porque si la temperatura de referencia sube un grado, hay que calentar un grado más, la diferencia aumenta ese grado.
   El centro comercial tiene dos modos de climatización: calefacción o refrigeración y en la práctica o funciona en un modo o en otro, por eso de las 6 columnas, son siempre 3 ceros vs 3 números.
11. Afluencia: Número de asistentes al centro comercial.

6. Tratamiento de datos

De las 11 variables iniciales, podemos eliminar la 2 ("Estimado") y 4 columnas de temperaturas. De las 3+3 columnas de temperaturas que tenemos, 2+2 son redundantes. Con quedarnos una columna de cada modo de climatización, es suficiente.

Escogemos CCDD20 y CHDD18, porque son las columnas más ventajosas. Es decir, porque si hay que calentar, calentar a 18º (CHDD18) es la línea de referencia mínima. Y si hay que refrigerar, refrigerar a 20º (CCDD20) es la referencia mínima.

De la primera exploración de las 6 variables resultantes llama la atención un aspecto importante que conviene resaltar. La columna FECHA la reconoce como cadena de caracteres (chr) y conviene tratarla como fecha, para que pase a ser Date. Esto es importante para posteriores representaciones gráficas de la serie, porque normalmente el eje X reflejará las fechas, para ver la evolución de los valores.

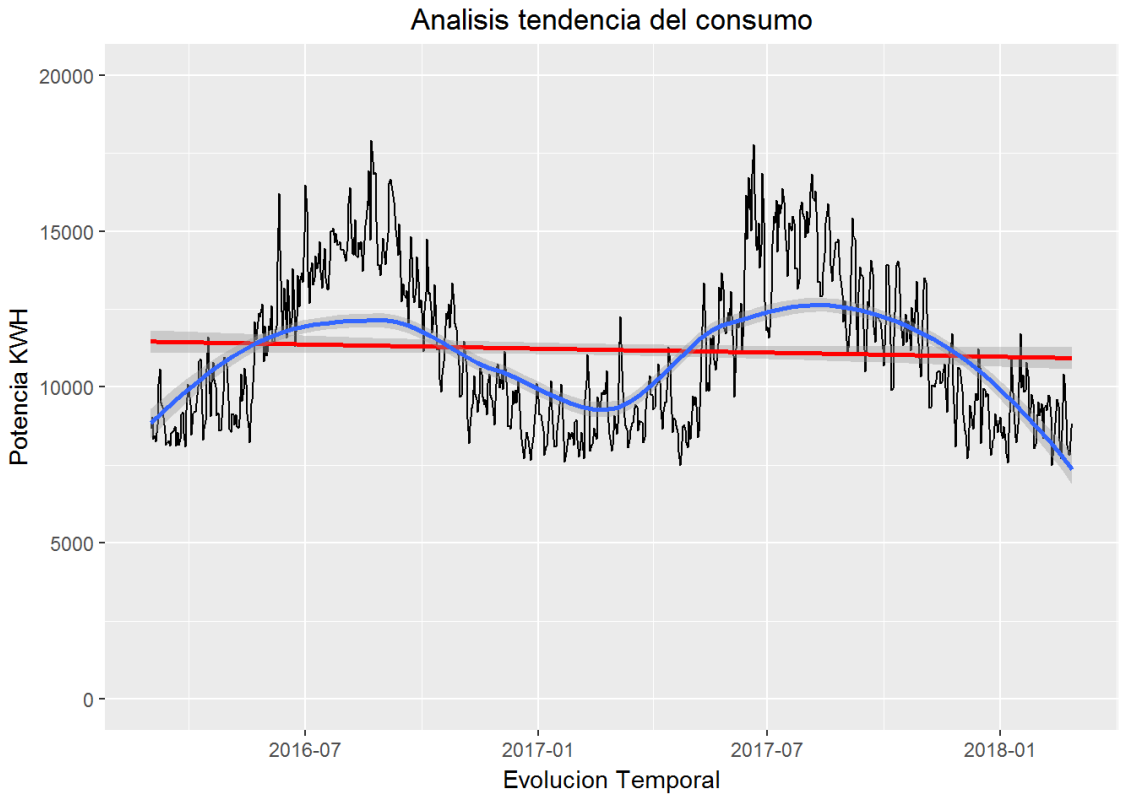
Dentro de este apartado de Tratamiento de Datos, se incluyen una serie de variables creadas ad-hoc y de las que se puede extraer información relevante. Hemos realizado algunos análisis de la información disponible, pero más con la idea de ilustrar las posibilidades que ofrece el dataset. Lógicamente se podría haber profundizado más, pero el objetivo del proyecto es analizar una serie de modelos predictivos y no tanto en detallar la casuística que nos ofrece el dataset.

Caso aparte también merece el Tratamiento de Missing Values. De 729 muestras, sólo hay 4 NA, por lo que son tan pocas, que lo más práctico es eliminarlas. Además, en nuestra serie tampoco tiene sentido realizar alguna simulación para rellenar los huecos con media por columna o similar, porque las mediciones de consumo no guardan relación entre sí.

7. Visualización de datos, variables y otras relaciones

Voy a recopilar en este punto algunas gráficas que he generado y considero de interés para explicar información relevante.

7.1 Análisis tendencia del consumo



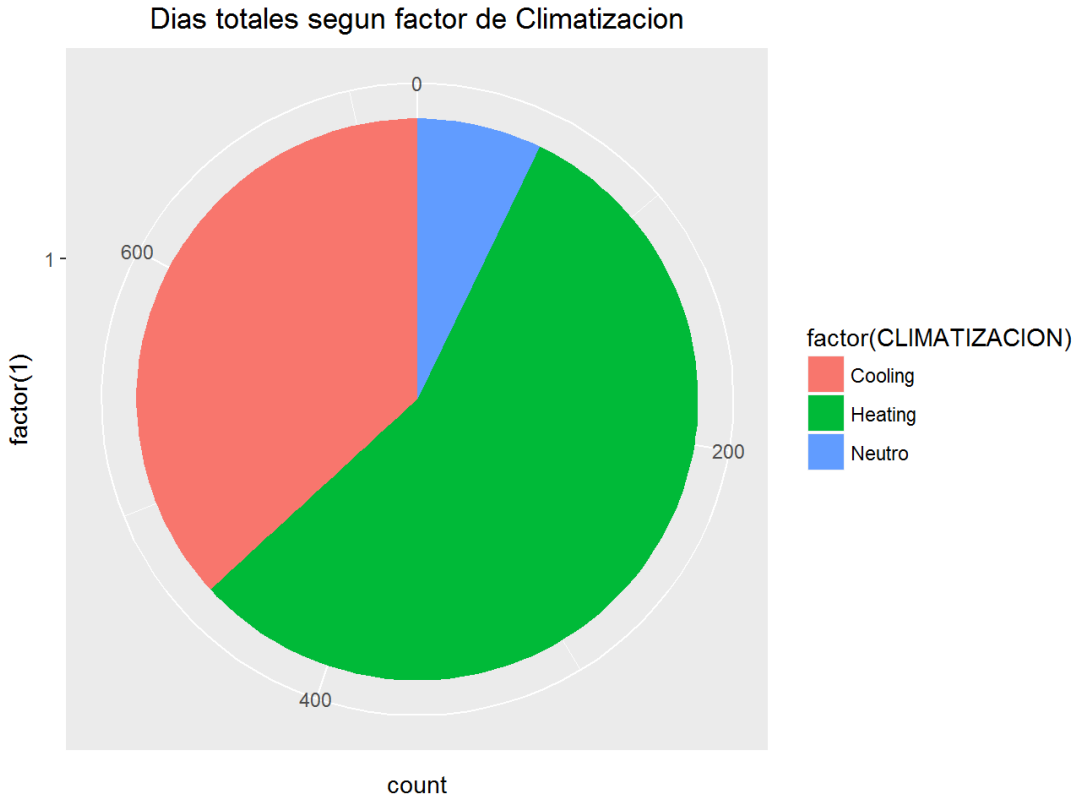
Si nos fijamos en la línea roja, se aprecia que hay una ligera tendencia de disminución del consumo a lo largo de estos dos años, motivada por la bajada de consumo de los últimos 3-4 meses de nuestros datos.

Para extraer alguna conclusión más solida habría que analizar un periodo más amplio, pues al ser la disminución más pronunciada al final, podría ser por un motivo coyuntural en una época del año con extraordinario buen tiempo. Nuestra serie está compuesta justo por dos años completos, que es el requisito mínimo para que se pueda afirmar.

7.2 Análisis de la variable Climatización

La variable Climatización es una variable categórica y tiene tres opciones:

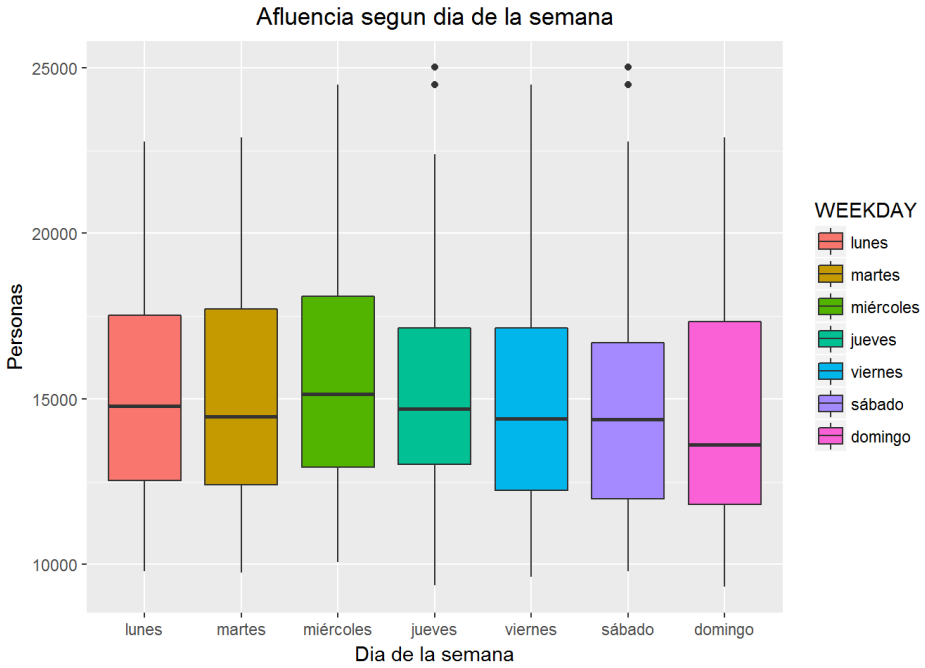
- Heating: CCDD20 es 0 y CHDD18 tiene valor.
- Cooling: CCDD20 tiene valor y CHDD18 es 0.
- Neutro: tanto CCDD20 como CHDD18 son 0.



Podemos ver que un poco más del 50% de los días el sistema de climatización funciona en Régimen de Heating (calefacción) y poco más de un tercio en Cooling (Refrigeracion). El resto, un 10% aprox. tiene gradiente Neutro.

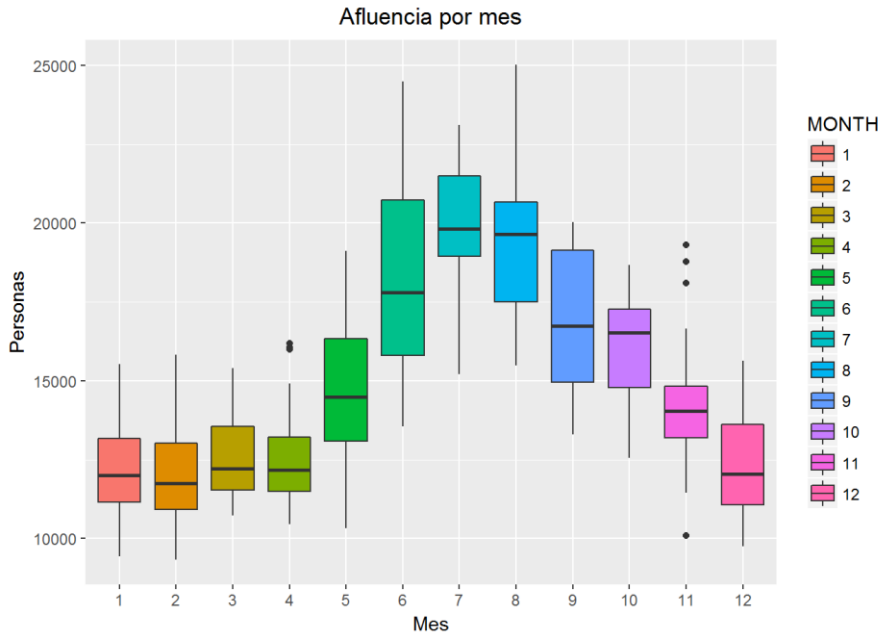
7.3 Análisis de la variable Afluencia

En primer lugar analizamos afluencia según el día de la semana.



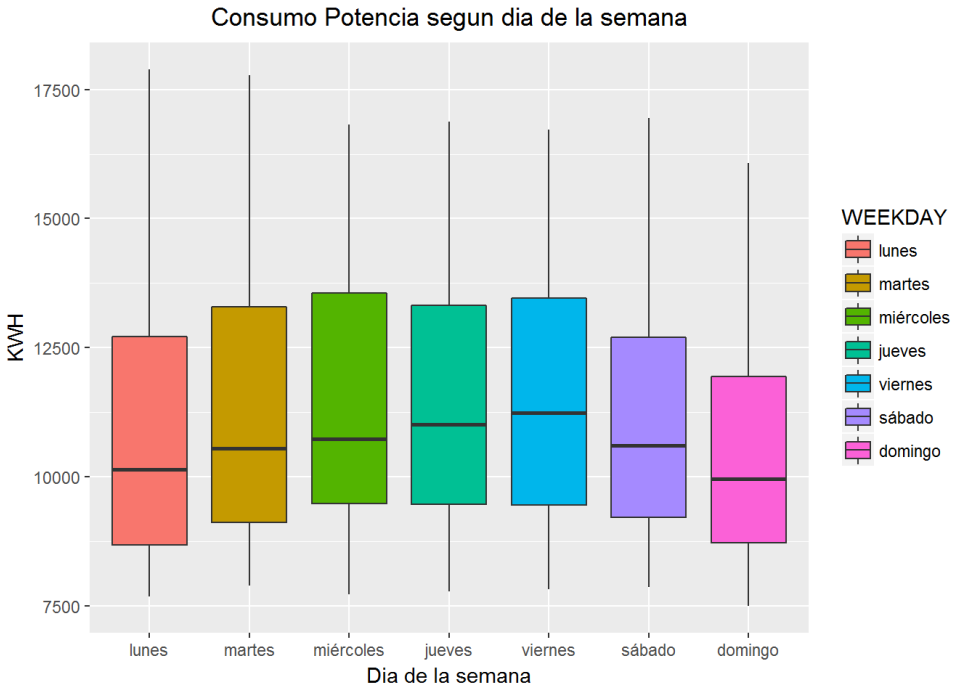
En mi opinión, sorprendentemente, el día de mayor asistencia es el miércoles, cuando esperaba que fuera el sábado.

A continuación, analizamos afluencia según el mes del año. También en mi opinión de manera sorprendente, los meses de mayor afluencia son julio y agosto de manera muy destacada, época de vacaciones y de buen tiempo. Y es que casi dobla, de unas 12.000 personas durante los meses de invierno a casi 20.000 en verano.



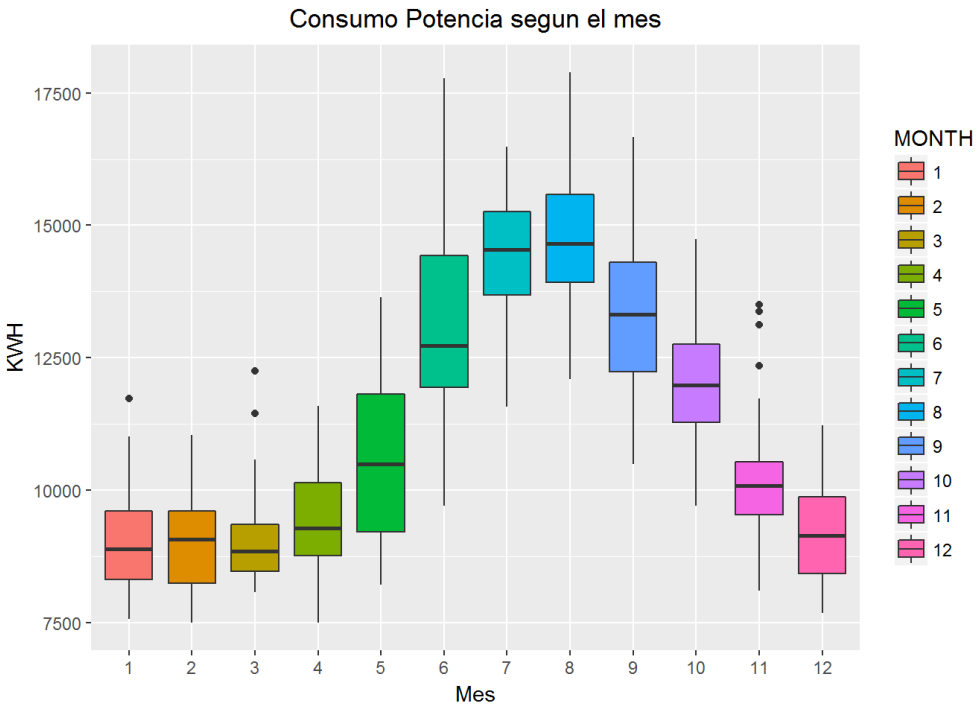
7.4 Análisis de la variable potencia

En primer lugar analizamos el consumo de potencia según el día de la semana.



Hay un contraste entre el día de la semana que más consumo hay (viernes) con el de mayor asistencia (miércoles). También vemos que hay un aumento progresivo de desde el lunes (min) a viernes (max).

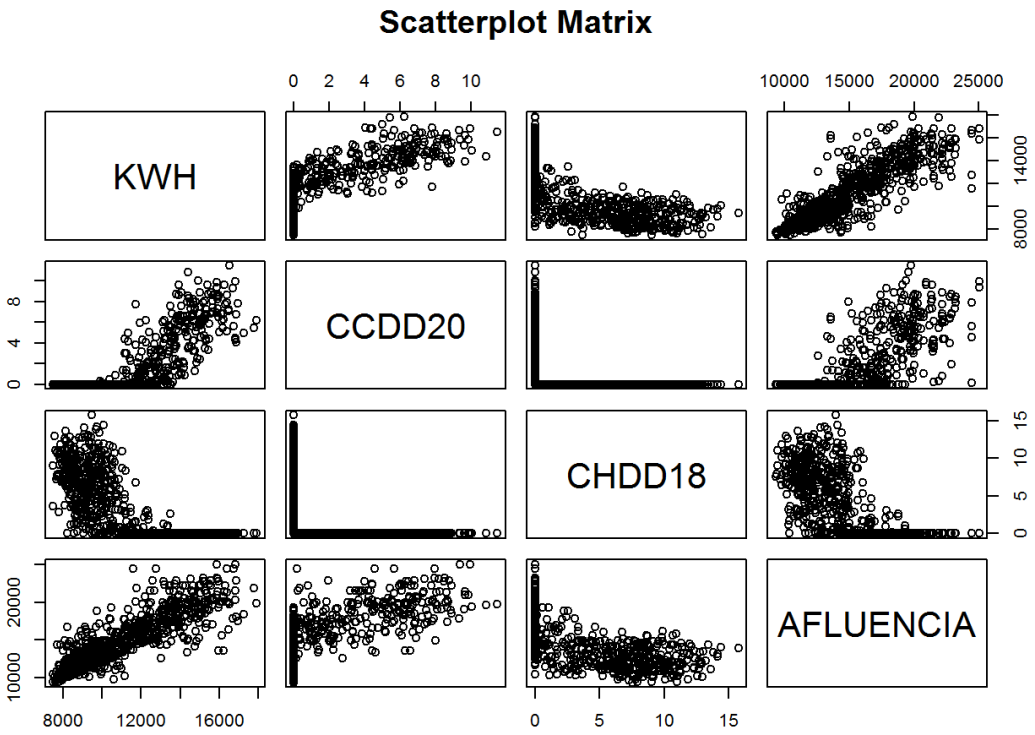
Y ahora el consumo de potencia según el mes del año.



Los meses de mayor consumo son julio y agosto, que coinciden con los meses de mayor afluencia. Hay una relación directa entre la temperatura de confort en su interior y la afluencia al centro comercial.

7.5 Análisis de correlaciones

Vamos a ver la matriz de correlaciones entre algunas de las variables más interesantes: KWH (2), CCDD20(4), CHDD18(5) y AFLUENCIA (6).



Aunque hay bastante dispersión en los datos, en primer lugar destaca una fuerte dependencia entre Consumo vs Afluencia. También llama la atención la relación entre Consumo y CCDD20 (Refrigeración). Esto es coherente, cuanta mayor diferencia de temperatura con el exterior, mayor consumo. Ya vimos anteriormente que en los meses de verano hay mucho más consumo.

En cambio, en modo Calefacción (CHDD18), al ser una línea plana significa que para el mismo consumo hay días de poca diferencia de temperatura y otros días con mayor gradiente térmico.

Con la tabla de correlaciones numéricas para confirmar toda esta información:

```
##                  KWH     CCDD20     CHDD18  AFLUENCIA
## KWH        1.0000000  0.8072247 -0.7207035  0.8513208
## CCDD20     0.8072247  1.0000000 -0.5228964  0.7284759
## CHDD18    -0.7207035 -0.5228964  1.0000000 -0.6668167
## AFLUENCIA  0.8513208  0.7284759 -0.6668167  1.0000000
```

8. Análisis de la serie temporal según 4 casos diferentes

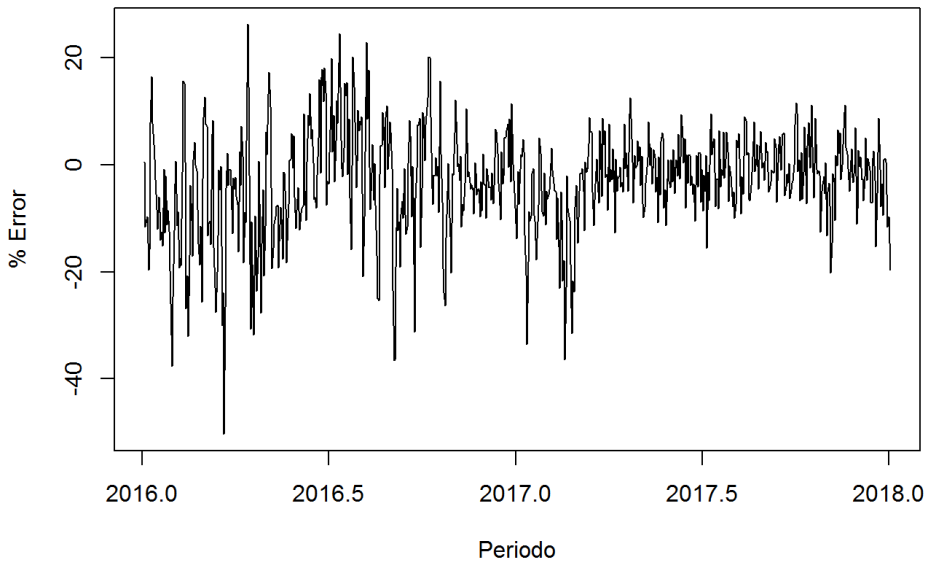
El objetivo principal es desarrollar un modelo predictivo que prediga nuestro Consumo Real (KWH) y mejorar en la medida de lo posible, los resultados de la predicción actual (LB).

Para conseguirlo vamos a definir nuestro plan:

1. Aplicar diferentes métodos de análisis a nuestra serie.
2. Analizar la calidad predictiva de cada modelo.
3. Escoger el modelo que mejor prediga.
4. Comparar el error entre su modelo predictivo y nuestra mejor propuesta.
5. Comprobar cuál de las dos opciones es mejor.

La dificultad de analizar esta serie viene dada porque está compuesta por muestras diarias en un ciclo estacional de dos años. Al tener justo el ciclo de dos años, se observa claramente una estacionalidad anual. Por un lado, tampoco es una serie de amplio rango lo suficientemente larga (ej. diez años). Por otro, la estacionalidad para periodos más cortos no resulta evidente.

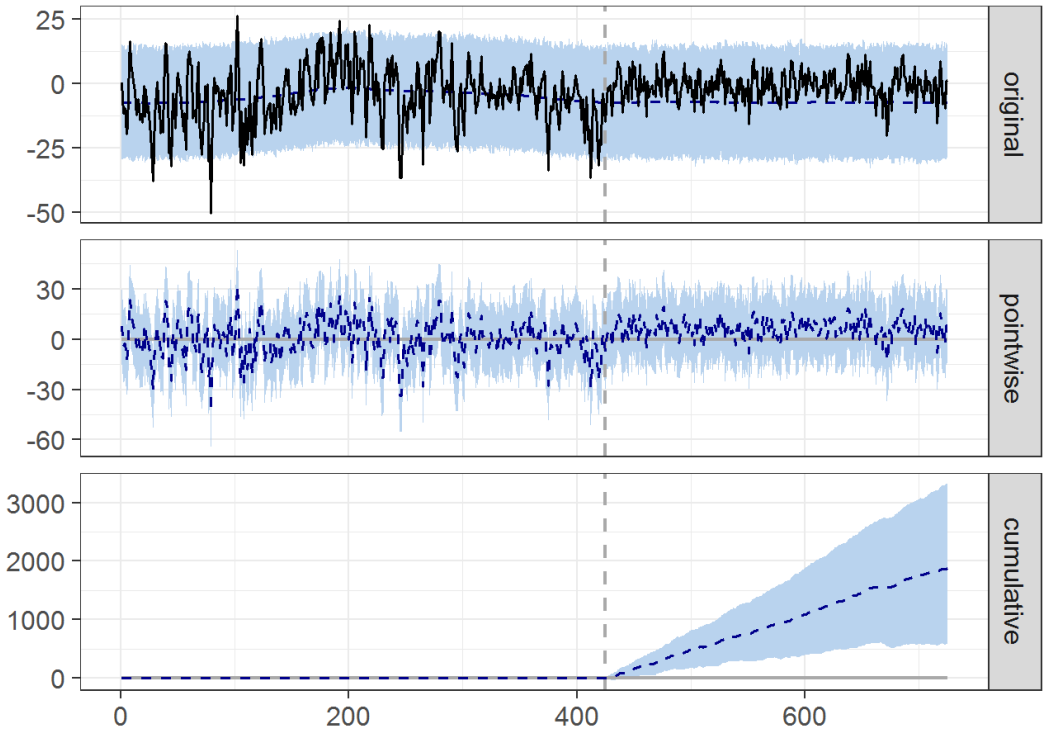
Empezamos analizamos la evolución de la variable Error:



Como primera aproximación, podemos observar que el modelo de predicción ha mejorado a partir de la segunda mitad de 2017, ya que el % Error es mucho menor, tiende a concentrarse en la horquilla de -10% a +10%.

CASO 1: Impacto causal utilizando modelos bayesianos de series

Vamos a suponer que a partir de la medición 426, coincidiendo con el 01/05/2017, se han introducido mejoras en el algoritmo de predicción y vamos a comparar la mejora que ha supuesto aplicando Causal Impact.



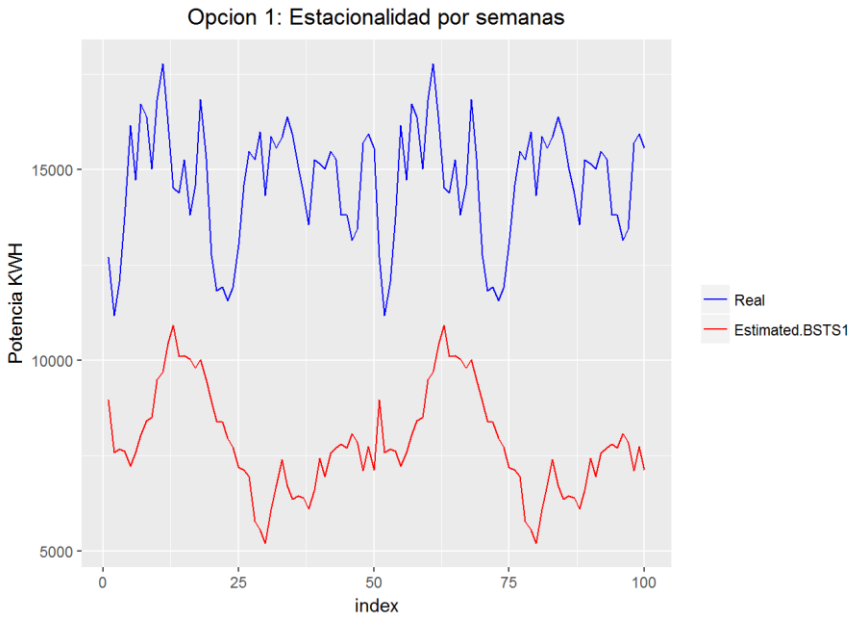

```
## Posterior inference {CausalImpact}
##
##                          Average          Cumulative
## Actual                   -0.91            -273.72
## Prediction (s.d.)        -7.2 (2.4)       -2153.3 (722.7)
## 95% CI                   [-12, -2.9]      [-3609, -866.9]
##
## Absolute effect (s.d.)   6.3 (2.4)        1879.6 (722.7)
## 95% CI                   [2, 11]          [593, 3335]
##
## Relative effect (s.d.)   -87% (-34%)      -87% (-34%)
## 95% CI                   [-28%, -155%]    [-28%, -155%]
##
## Posterior tail-area probability p:   0.00235
## Posterior prob. of a causal effect:  99.76471%
##
## For more details, type: summary(impact, "report")
```

El resultado del análisis nos dice que la intervención en el modelo redujo la previsión de ERROR en un 87%.

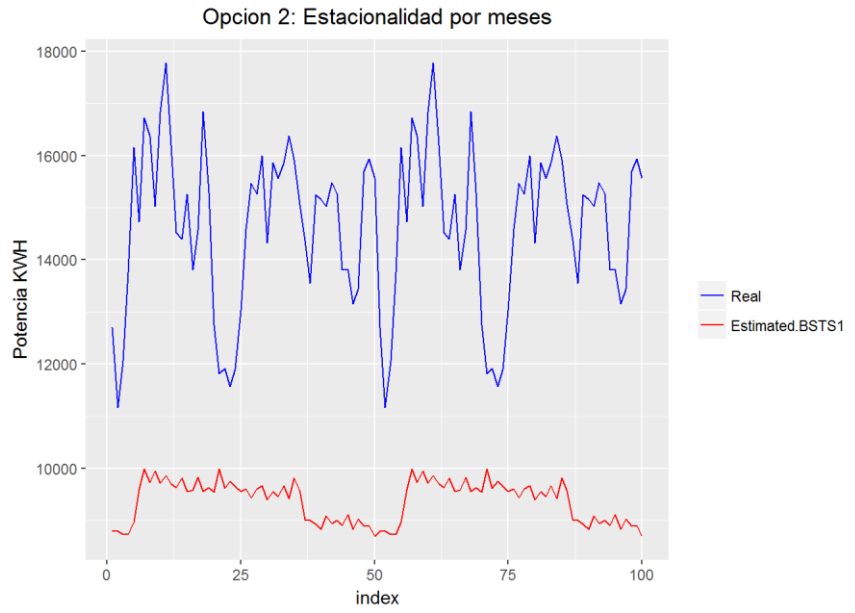
Ahora vamos a calcular la predicción utilizando un modelo personalizado con la función bsts.model ().

Una serie tiene dos componentes principales, una parte lineal y otra estacional. La parte lineal está clara, pero para la estacional vamos a analizar cuál de los supuestos encaja mejor: semanal o mensual.

Opción 1: Tantos periodos como semanas del año



Opción 2: Tantos periodos como meses del año y le incluimos MonthlyAnnualCycle

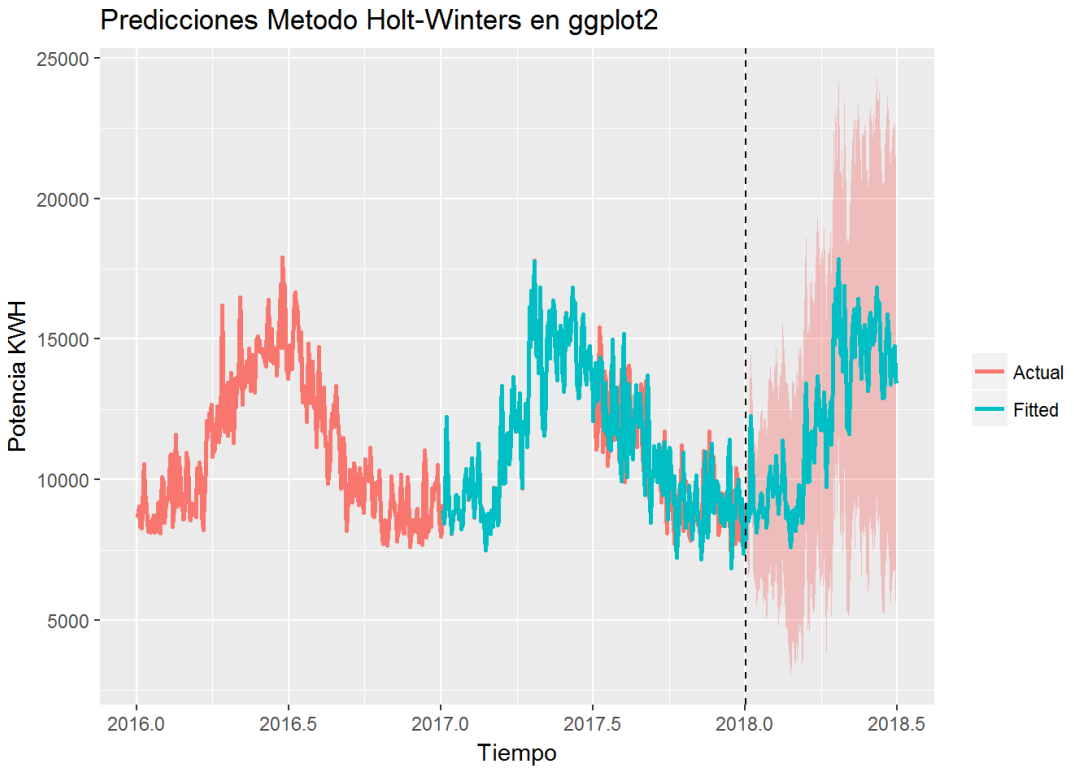


Con la Opción 2, la de periodos por meses, se mejora la predicción y baja sensiblemente el MAPE% del 92% al 56%. No obstante, las predicciones siguen lejos de la potencia consumida y podemos concluir que no son de buena calidad.

CASO 2: Método Holt-Winters

Vamos a hacer predicciones según el Método Holt-Winters, que se caracteriza por hacer pronósticos utilizando un suavizado exponencial con un componente de tendencia y un componente estacional, es decir, suavizado exponencial triple.

Calculamos la previsión para los próximos 6 meses con un intervalo de confianza de 0,95 y representamos el pronóstico junto con los valores reales y ajustados. En el código se puede ver como transformamos el plot para hacerlo más atractivo.

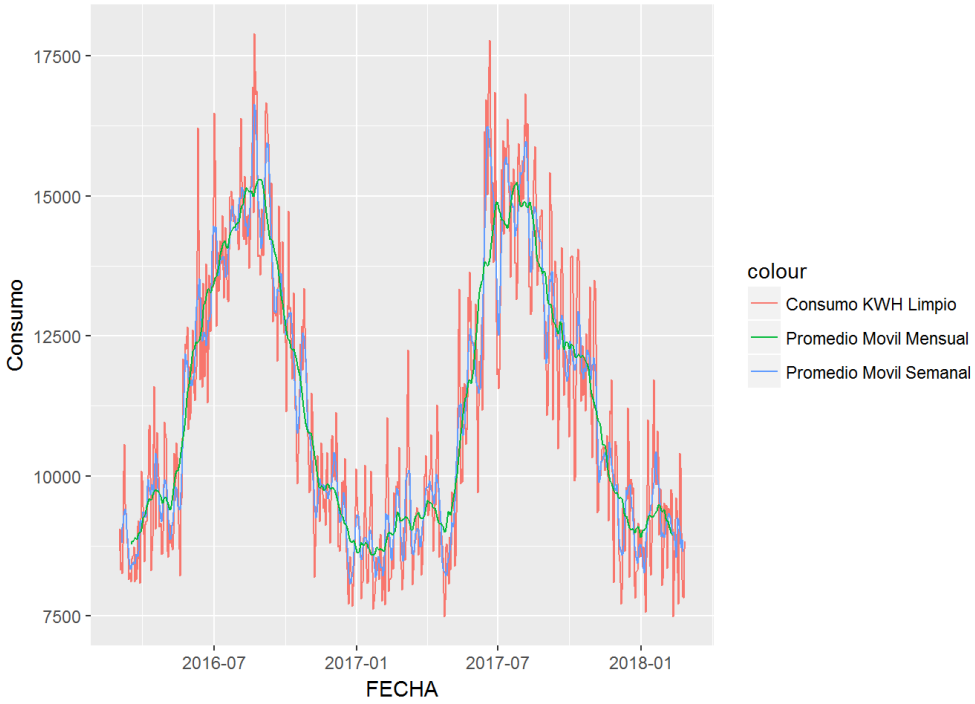


El modelo predictivo parece bastante bueno, aunque ofrece dudas por el amplio intervalo de confianza.

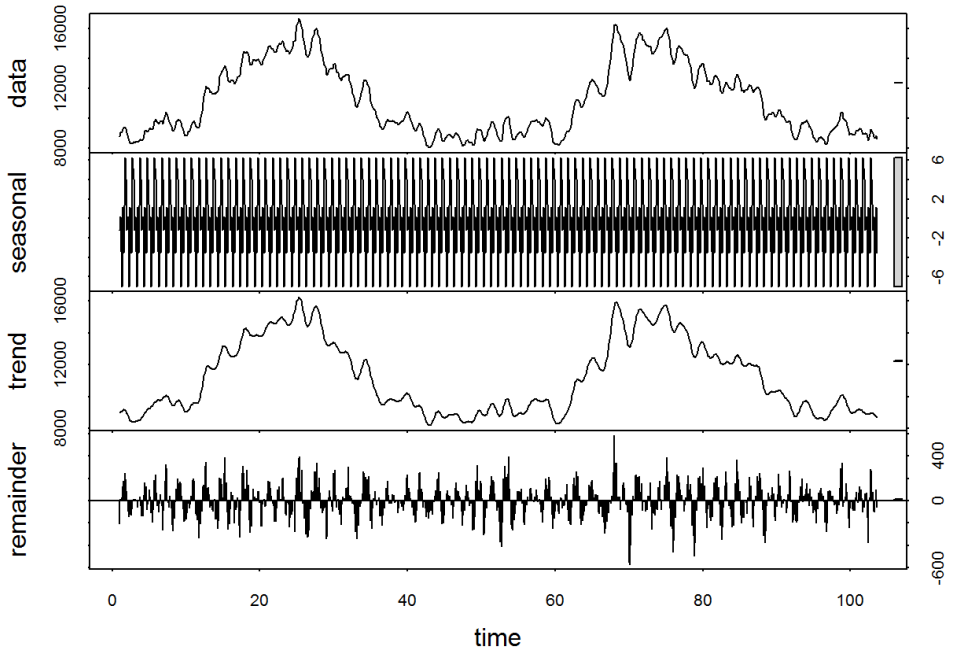
Caso 3: Método ARIMA

Los modelos ARIMA son una clase de modelo de pronóstico que utiliza información histórica para hacer predicciones y está pensado para suavizar las fluctuaciones de la curva utilizamos el promedio móvil.

Cuanto más ancha es la ventana de la media móvil, más suave se vuelve la serie original. En nuestro caso, podemos tomar la media móvil semanal o mensual, suavizando la serie en algo más estable y, por lo tanto, más predecible.



Los componentes básicos de un análisis de series de tiempo son la estacionalidad, la tendencia y el ciclo. Calculamos el doble componente estacional de los datos (semanal, anual) usando stl.



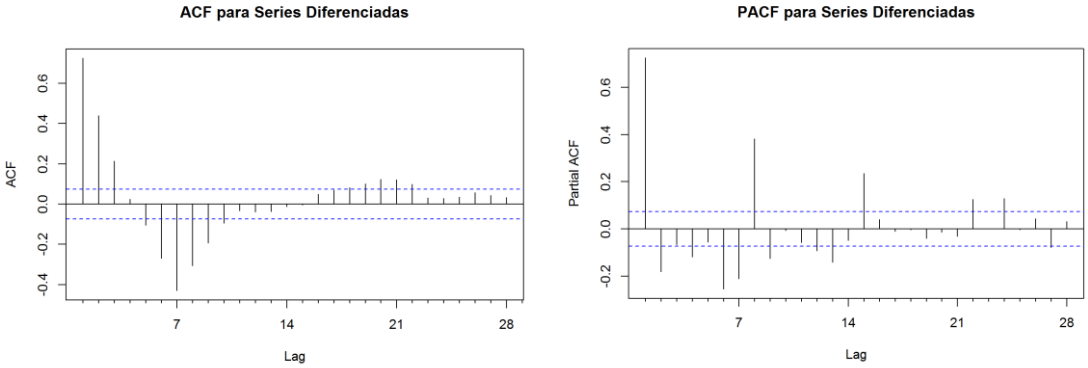
La aplicación de un modelo ARIMA requiere que la serie sea estacionaria. La prueba aumentada Dickey-Fuller (ADF) es una prueba estadistica formal para comprobar la estacionalidad, donde:

- Ho: La serie es no estacionaria: tiene raíz unitaria.
- H1: La serie es estacionaria: no tiene raíz unitaria.

La serie original no es estacionaria porque tiene raíz unitaria (-2.05) y el valor p es 0.55 (supera la referencia de 0.05). En cambio, la serie en primera diferencia es estacionaria. El valor del estadístico es -9.5, con un valor p de 0.01, por lo que la hipótesis nula ya se puede rechazar.

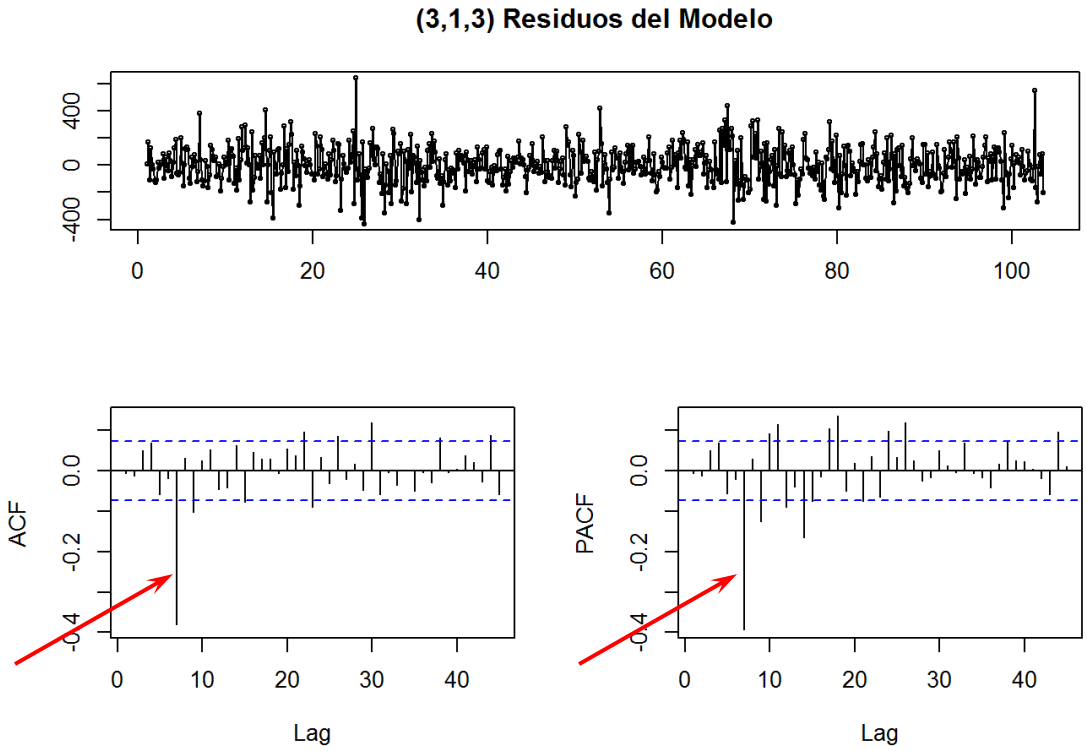
Para aplicar el Modelo ARIMA hay que conocer tres coeficientes que nos permitan modelarlo, que son (p, d, q). Como la serie en primera diferencia es estacionaria, significa que d=1.

Ploteamos la Función de Autocorrelación Simple y Parcial para series diferenciadas.



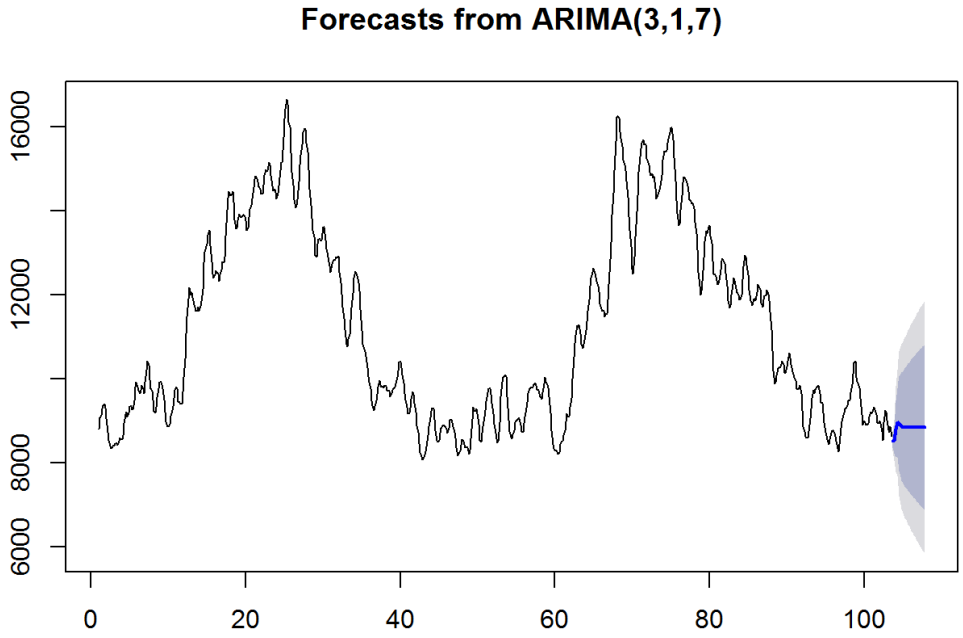
Podemos observar que en gráfico ACF hay retardos importantes en los picos 1, 2 y 7, lo que significa que son nuestros candidatos para q. En cambio, la componente parcial PACF vemos que tiene claramente el pico en el 1, lo que significa que el coeficiente p=1.

En cualquier caso, al aplicar el ajuste automático auto.arima() obtenemos un modelo ARIMA(3,1,3). Recalculamos de nuevo nuestros gráficos ACF, PACF y Residuos con la autoconfiguración y se confirma el retardo en 7. Esto indica que nuestro modelo puede mejorarse con un nuevo parámetro q = 7.

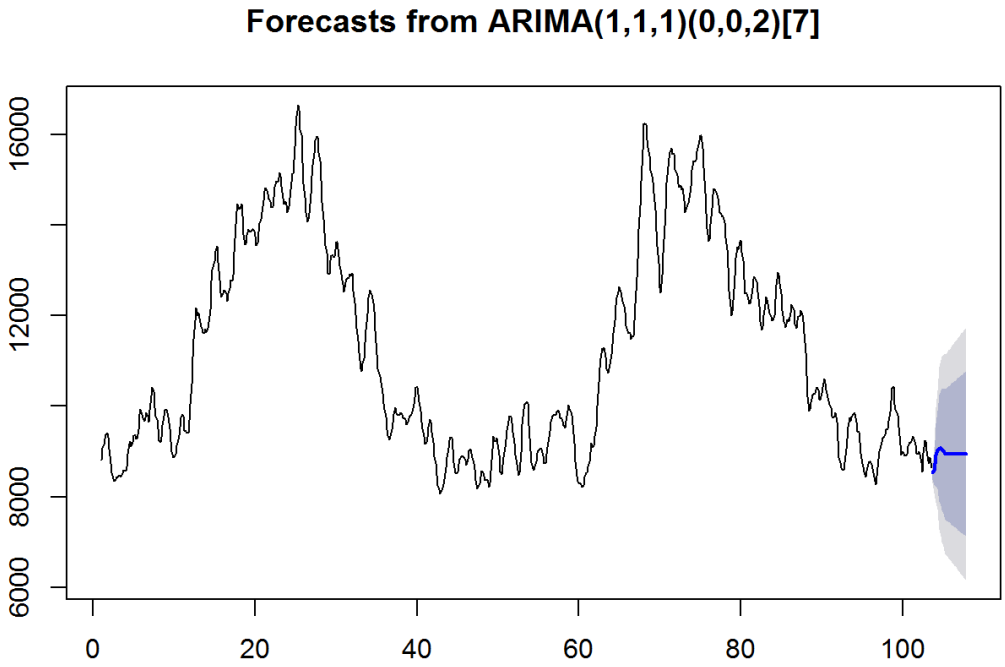


Así que volvemos a ajustar con ARIMA(3,1,7). También sería posible con ARIMA(7,1,7), pero he comprobado que no afecta, así que mi criterio es dejar p=3, tal y como recomienda la autoconfiguración.

Una vez evaluado e iterado el modelo óptimo, pasamos a la parte de Predicciones, especificando el pronóstico h de unidades de tiempo por delante (días).



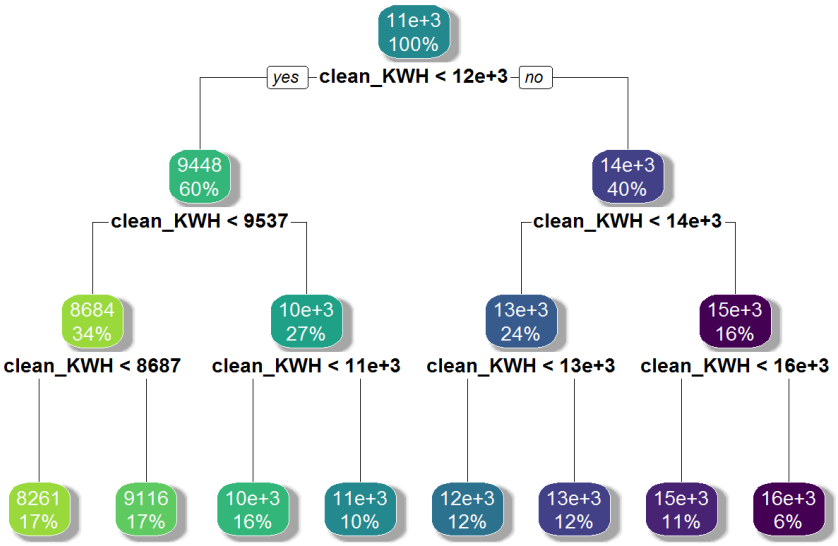
También probamos la función auto.arima con estacionalidad:



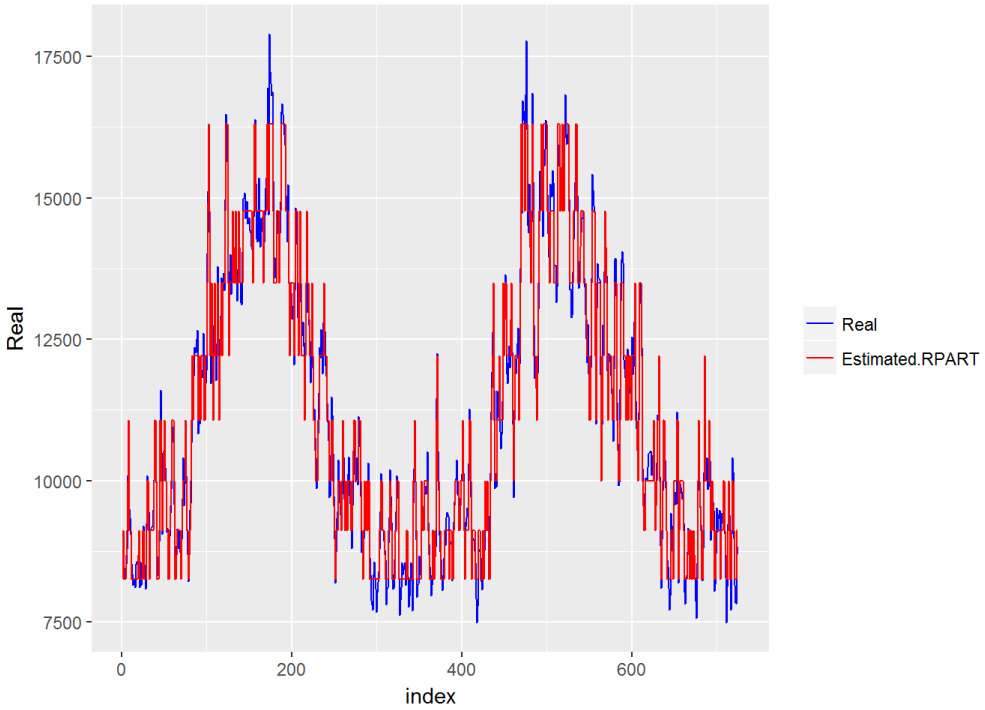
Tampoco mejora. Así que podemos concluir que los resultados predictivos bajo este modelo son decepcionantes. He realizado diferentes pruebas de simulación pero en ningún caso parece reaccionar la curva.

CASO 4: Arboles de clasificación y regresión, RPART (CART) Tree

Planteamos un último modelo predictivo, RPART. Entrenamos el primer árbol con rpart, variable KWH.



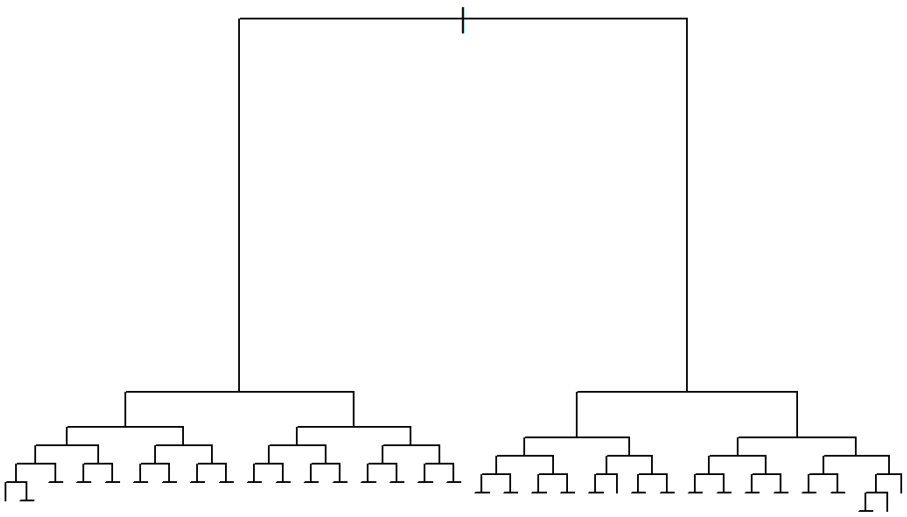
Hacemos la predicción de la serie y vemos que es bastante buena con respecto la real, pero en los picos no es capaz de predecir bien. Para evaluar la predicción de la serie utilizamos la métrica MAPE (mean absolute percentage error).



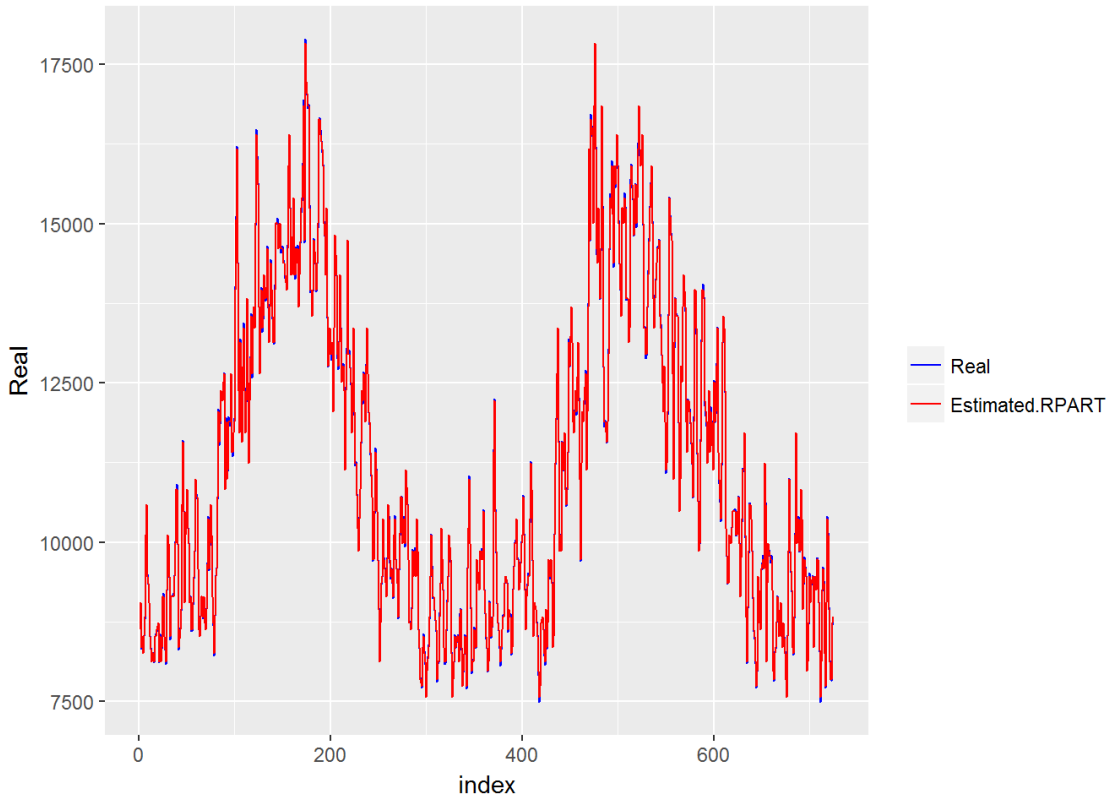
El error es razonable (MAPE 2.47%), pero para mejorar predicciones hay que tunear un poco nuestro tree1 con rpart.control.

Vamos a entrenar un segundo árbol en el que podremos especificar los siguientes parámetros: minsplit, maxdepth, cp.

El árbol 2 ya es mucho más complejo y tiene 63 splits:



Calculamos las nuevas predicciones y ploteamos consumo real vs estimado.

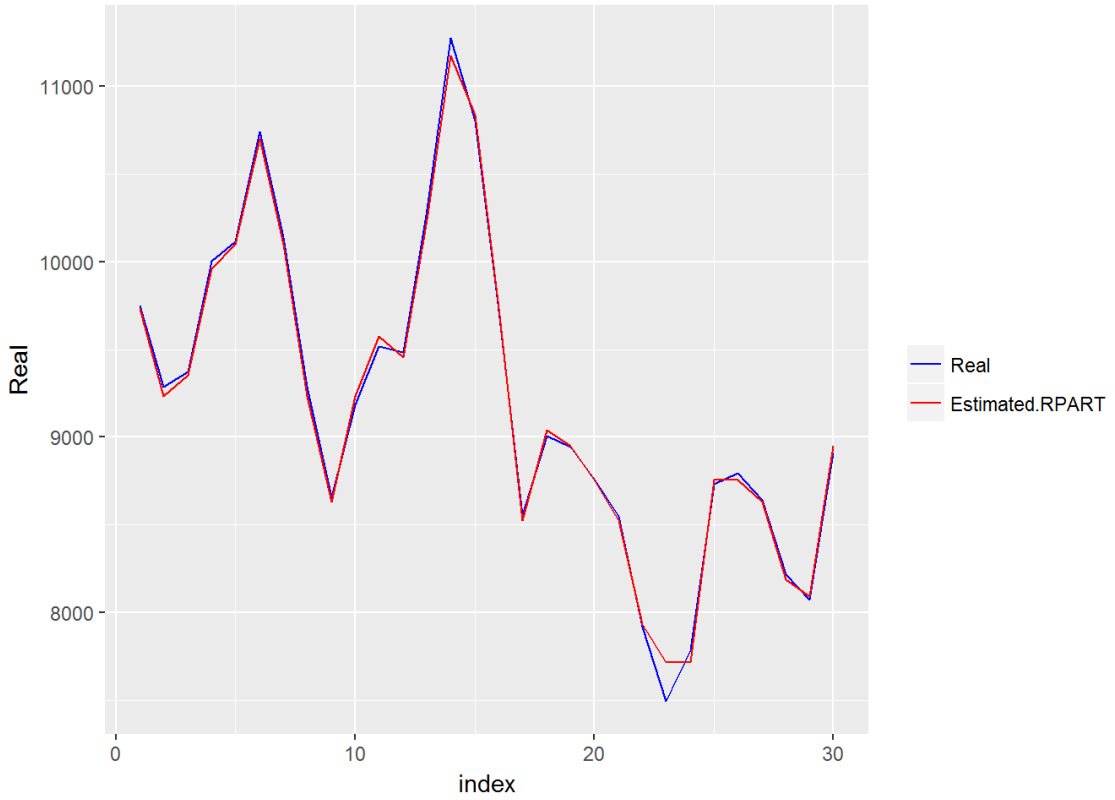


Vemos que prácticamente ambas líneas se superponen y el MAPE se reduce a un 0.28%. Aun así vamos a buscar una combinación de los hiperparámetros de control para ver cuál es la configuración que mejor predice.

Como tenemos una serie larga, vamos a crear 2 set de training y 2 set de test y compararemos resultados, para evitar overfitting o influencia.

- data.train.1: 1/03/2016 -> 31/03/2017 (13 meses)
- data.test.1: 1/04/2017 -> 30/04/2017 (1 mes)
- data.train.2: 1/01/2017 -> 31/01/2018 (13 meses)
- data.test.2: 1/02/2018 -> 27/02/2018 (1 mes)

Entrenamos el árbol tree.data.1 con data.train.1 y posteriormente predeciremos sobre data.test.1. Luego le tuneamos los hiperparámetros para ver si podemos mejorar. Realizamos el mismo procedimiento con el segundo set de training y test.



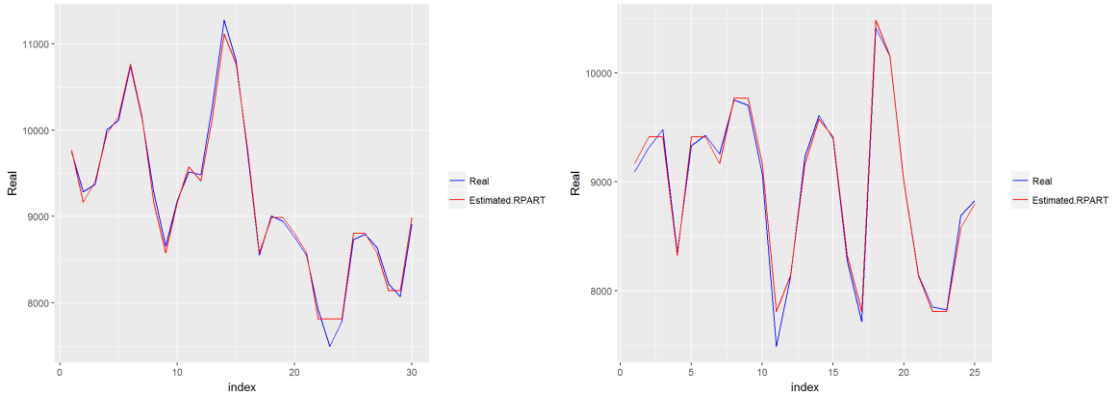
El resultado es que se obtiene la misma configuración de hiperparámetros en ambos casos, aunque en el segundo set disminuye ligeramente el MAPE.

Aun no estamos seguros si hemos cogido los valores óptimos de hiperparámetros, aunque estaremos muy cerca. Haremos un último intento al observar que:

- Por lo visto en el caso de tree1 y tree2, cuando entrenábamos y predecíamos con la misma serie de datos, a mayor profundidad de los árboles se tiende al overfitting. Nos hemos fijado que la mejor predicción (menor MAPE), tomaba como max_depth el valor mínimo posible (5). Es por esto que vamos a dirigir la búsqueda final de este hiperparámetros hacia valores menores de máxima profundidad.
- En cuanto al mínimo número de splits, y tal como se ha experimentado podemos ver que es 6 para ambos casos y que a su vez es el mayor número de la serie propuesta. Aumentaremos los valores de la serie.
- El valor de cp, de igual manera a la máxima profundidad, cuando entrenábamos y predecíamos con los mismos datos proporcionaba mejores resultados a medida que disminuíamos el parámetro (induciendo al overfitting). En las nuevas predicciones, se ha encontrado como valor el 0.001 (mínimo local) y es en sus valores adyacentes donde centraremos la búsqueda.

Los test con los nuevos parámetros sugeridos, los valores finales para el menor MAPE salen casi calcados, no afecta seleccionar ni la primera mitad de los datos ni la segunda. Así que para terminar, vamos a entrenar un árbol para las dos series con los óptimos de hiperparámetros encontrados, predecimos y ploteamos.

Pongo ambos data.test, vemos que los resultados son casi iguales y se acoplan muy bien entre predicción y real.



Podemos afirmar que la calidad predictiva de este modelo es muy alta y de los cuatro casos analizados, es la mejor propuesta.

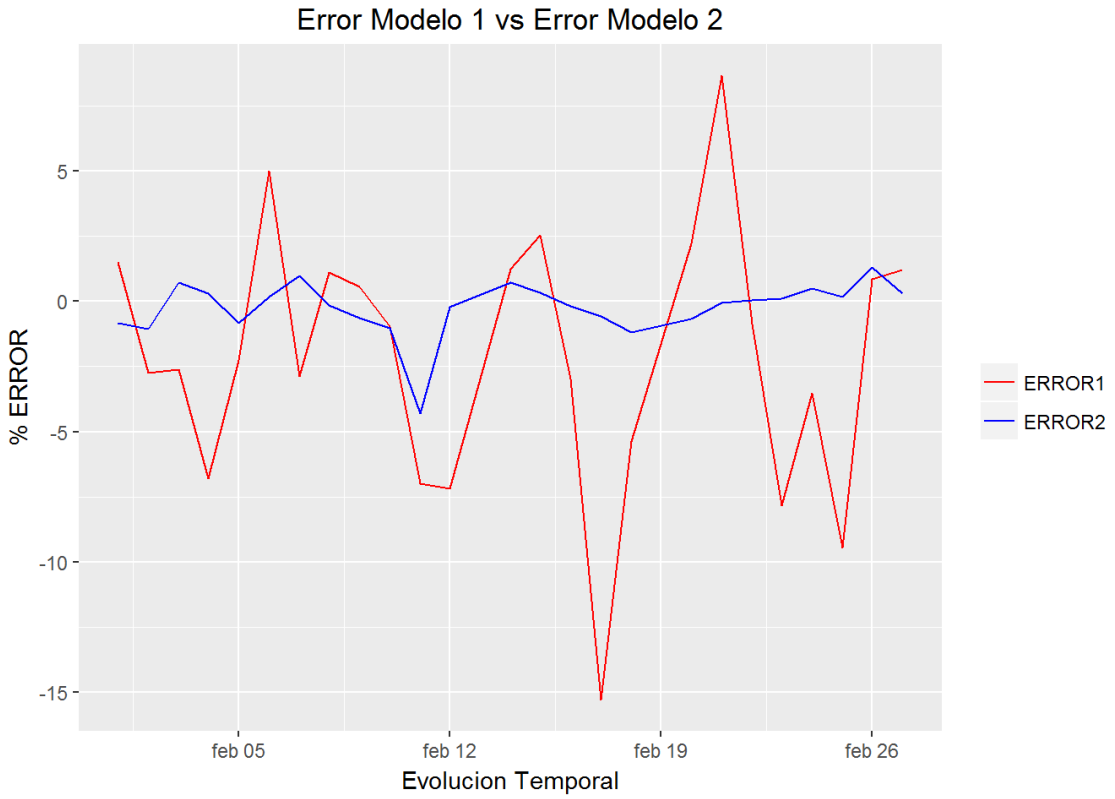
9. Conclusiones

De los 4 modelos analizados, podemos llegar a la conclusión que el modelo con el que hemos obtenido mejor resultado es el último, el RPART. Por tanto, es el seleccionado para pasar a la siguiente fase.

Tal y como planteamos inicialmente, el objeto del estudio es comparar el error existente entre el modelo predictivo actual con el error según un nuevo modelo desarrollado, para decidir cuál de los dos modelos predice mejor.

En primer lugar, calculamos el error de las nuevas predicciones: real vs predicción, creamos una nueva tabla compuesta únicamente por fecha y ambos errores. Recuerdo que los valores que se han tenido en cuenta provienen del data.test.2, bajo su columna predictions.2$Estimated.RPART y por tanto sólo se han generado las predicciones para el periodo data.test.2: 1/02/2018 -> 27/02/2018.

Vamos a plotear ambos Errores en la misma gráfica para compararlos mejor:



Según podemos observar, el nuevo modelo predictivo predice mejor porque la gráfica apenas se desvía del 0, mientras el modelo predictivo inicial tiene algunos picos del 15% o del 8% de error, vemos que claramente el Error1 es peor que el Error2.

Por tanto, concluimos este estudio afirmando que el nuevo modelo predictivo es mejor que el que utilizan actualmente.