**Predicción de Demanda Térmica y Flexibilidad**

# Resumen

Contenido

[Resumen 2](#_Toc520093850)

[Figuras 3](#_Toc520093851)

[Tablas 4](#_Toc520093852)

[Introducción 5](#_Toc520093853)

[Objetivos 7](#_Toc520093854)

[Estado del Arte 7](#_Toc520093855)

[Metodología 10](#_Toc520093856)

[Desarrollo 15](#_Toc520093857)

[Predicción Demanda Térmica por Series Temporales 15](#_Toc520093858)

[Creación del modelo 15](#_Toc520093859)

[Aplicación 27](#_Toc520093860)

[Apéndices 30](#_Toc520093861)

# Figuras

[Figura 1: Diagrama de principio de la instalación de Bomba de Calor 5](#_Toc519701945)

[Figura 2: Flujo del proceso de modelado y predicción 8](file:///F:\TFM\TFM-ConsumoPredictivo-BTL.docx#_Toc519701946)

[Figura 3: Definición analítica de un ARIMA de orden p,d,q. 9](#_Toc519701947)

[Figura 4: Interpretación de lag temporal de orden *d* 9](#_Toc519701948)

[Figura 5: Listado de valores Akaike obtenidos 10](#_Toc519701949)

[Figura 6: Modelo ARIMA (p=8,d=0,q=2) para el residuo de la serie temporal 11](#_Toc519701950)

[Figura 7: Modelos ARIMAX ,(p=8,d=0,q=2) 12](#_Toc519701951)

[Figura 8: Modelo ARIMAX tras suavizar y normalizar. 13](#_Toc519701952)

[Figura 9: Componente de tendencia, periódica y de residuo o resto de la serie original. 14](#_Toc519701953)

[Figura 10: Diagramas de auto correlación y autocorrelación parcial de la serie residuo grado de significancia de 95% 14](#_Toc519701954)

[Figura 11: modelo ARIMAX sobre el residuo de la serie temporal (p=8,d=0,q=2) 15](#_Toc519701955)

[Figura 12: Modelo ARIMAX sobre el residuo normalizado y normalizado 16](#_Toc519701956)

[Figura 13: Predicción original in-bound y out-bound a partir del modelo suavizado de residuos 17](#_Toc519701957)

[Figura 15: Evolución de la tendencia y aproximaciones por regresión lineal y polinómica (grado 9) 18](#_Toc519701958)

[Figura 16: Predicción total tomando aproximado la tendencia mediante regresión lineal y un polinomio de grado 9 18](#_Toc519701959)

# Tablas

[Tabla 1: Estadísticos de error obtenidos para el modelo ARIMA de la Figura 6 12](#_Toc519701960)

[Tabla 2: Coeficientes pdq para modelo ARIMAX sobre el residuo 12](#_Toc519701961)

[Tabla 3: Estadísticos de error para los modelos anteriores. 13](#_Toc519701962)

[Tabla 4: Coeficientes ARIMAX para la series tras aplicar el suavizado y normalizado 13](#_Toc519701963)

[Tabla 5: Estadisticos de error sobre el modelo ARIMAX de la serie temporal Pbld 13](#_Toc519701964)

[Tabla 6: Coeficientes pdq para modelo ARIMAX sobre el residuo 15](#_Toc519701965)

[Tabla 7: Estadísticos de error para los casos anteriores 15](#_Toc519701966)

[Tabla 8: Coeficientes ARIMAX para la series tras aplicar el suavizado y normalizado 16](#_Toc519701967)

[Tabla 9: Estadísticos de error para el modelo sobre el residuo tras aplicar el suavizado y normalizado 16](#_Toc519701968)

# Introducción

El problema planteado en este trabajo, trata de optimizar el consumo energético asociado a la climatización en una vivienda. Los datos de laboratorio se corresponden a lo de una vivienda unifamiliar que básicamente consta de 8 estancias. La tecnología utilizada para climatizar la vivienda está basada en una bomba de calor (normalmente referida como HP, Heat-Pump) que además proporciona agua caliente sanitaria, que es una fuente de aleatoriedad.

Se trata de una vivienda unifamiliar de energía casi nula disponible cuya monitorización incluye a nivel de detalle las diferentes instalaciones técnicas, componentes del sistema de climatización y confort. La villa está construida con una especial consideración por tener una estructura de construcción bien aislada, hermética y resistente a la humedad, así como por una tecnología de instalación que ahorra energía.

Actualmente el edificio consta de:

* Standard de aislamiento “nZEB” (Nearly Zero Energy Building)
* Bomba de calor geotermal
* Unidad de tratamiento de aire de recuperación de calor (FTX)
* Células solares con inversor
* Suelo radiante
* Sistema de calefacción por radiadores
* Sistema de adquisición de datos

La siguiente figura muestra el esquema hidráulico del conjunto bomba de calor, tanque dentro del conjunto del suministro de climatización y agua caliente. En la figura se han representado también las variables pertenecientes al dataset disponible

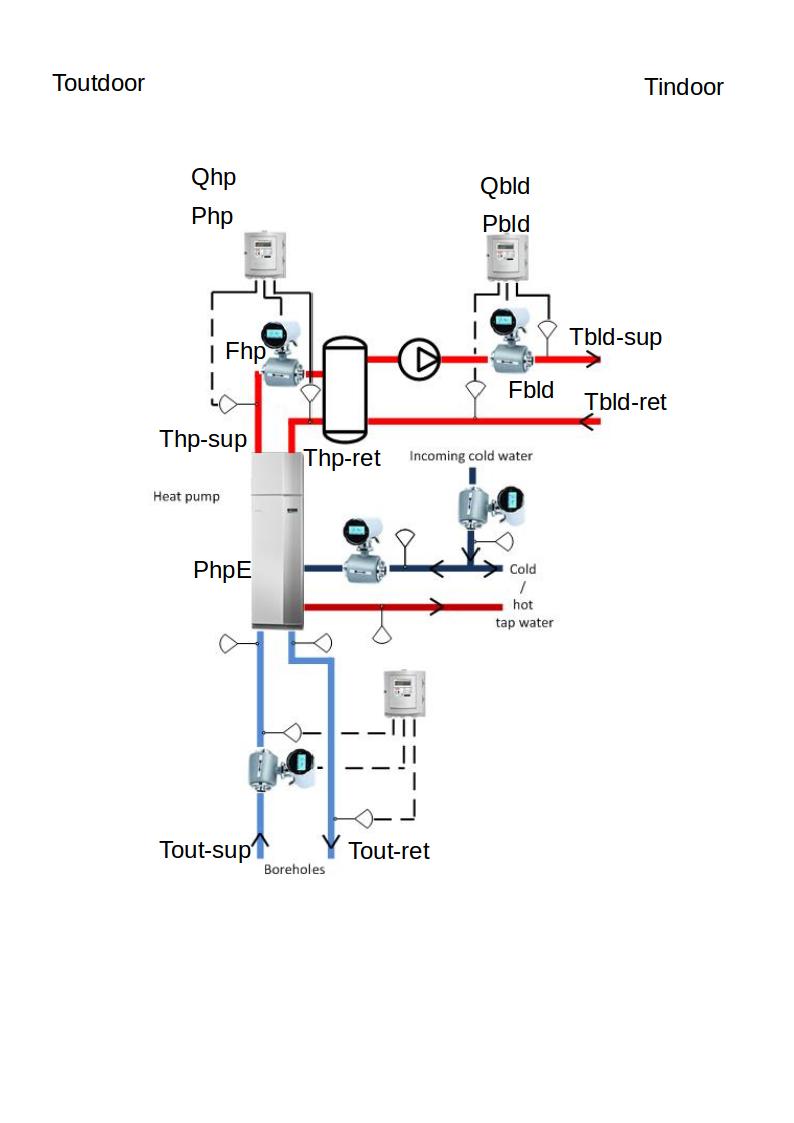


Figura : Diagrama de principio de la instalación de Bomba de Calor

Listado de features relevantes disponible en el data set.

* Potencia eléctrica consumida por la bomba de calor
* Energía térmica suministrada desde la HP al tanque de agua
* Energía térmica suministrada desde el tanque de agua a la vivienda
* Temperatura exterior durante el periodo de monitorización
* Temperaturas de las habitiaciones individuales y media de la vivienda.

Anadir descripción extra

# Objetivos

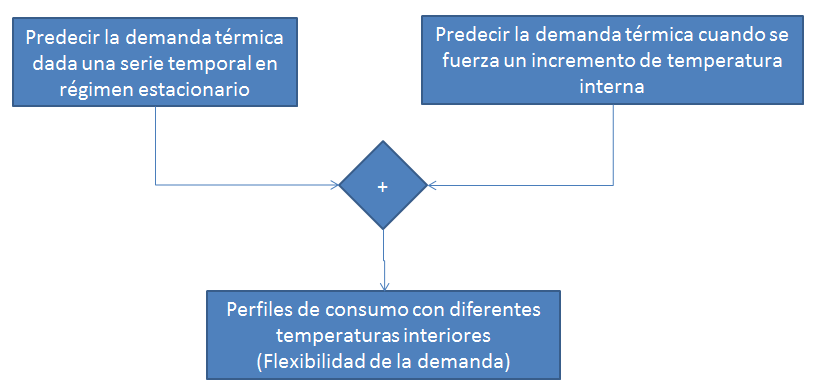
El objetivo de este trabajo es desarrollar un módulo de predicción de la demanda térmica del edificio impulsado por modelos de caja negra y pronósticos de flexibilidad energética que permitan operar en mercados de energía.

La estructura de inercia térmica permite utilizar el edificio como una herramienta de flexibilidad: los perfiles de potencia pueden modificarse y modificarse sin afectar de forma significativa el nivel de comodidad de los inquilinos mediante el uso de modelos que caracterizan los edificios comportamiento térmico dinámico.

* El primero de los objetivos elaborará algoritmos que crean dichos modelos a partir de mediciones de datos de sensores utilizando técnicas series temporales. El aspecto clave es la capacidad de crear modelos que permitan una generalización y escalabilidad sin necesidad de conocimiento experto.
* El segundo de los objetivos aplicará un algoritmo de optimización sobre los modelos térmicos evaluados, para determinar la flexibilidad en términos de demanda térmica sobre los que se puede operar el edificio.

La inercia térmica de los edificios en conjunto con las tecnologías de bomba de calor constituye un enorme potencial de flexibilidad eléctrica utilizable para absorber el exceso de generación renovable y evitar la restricción al administrar la congestión local de energía, estabilidad de voltaje en la red y evitar flujos de energía inversa, así como minimizar el desequilibrio de producción de energía renovable. Directivas Europeas relativas a la eficiencia energética [1][2] ponen especial atención en estos aspectos.

La siguiente figura describe como se unifican los objetivos anteriormente citados para concluir ofreciendo una estimación de la flexibilidad en formato de perfiles de demanda térmica



Desde el punto de vista conceptual, se puede decir que el presente documento describe como se anida dos modelos diferentes uno basado en series temporales y otro en arboles de regresión para predecir cuál será la demanda de térmica de una vivienda/edificio bien en régimen estacionario y bien cuando dicho régimen estacionario quiere verse alterado para ofreces servicios de flexibilidad.

Estado del Arte

Una predicción es un pronóstico de algún evento o evento futuro que no siempre es fácil. Existen numerosos ejemplos de predicciones fallidas por tres de las más llamativas puedas ser:

* "La población es constante en tamaño y seguirá siéndolo hasta el final de la humanidad". L'Enciclopedie, 1756.
* "1930 será un espléndido año de empleo". Departamento de Trabajo de EE. UU., Pronóstico de Año Nuevo en 1929, justo antes del colapso del mercado el 29 de octubre.
* "Las computadoras se están multiplicando a un ritmo acelerado. Hacia el cambio de siglo habrá 220,000 en los Estados Unidos". Wall Street Journal, 1966.

En la actualidad no hay ningún ámbito de aplicación en el que las predicciones no jueguen un aspecto relevante. Las predicciones son elemento fundamental en el proceso de toma de decisiones en áreas tan diversas como el marketing, la demografía, la gestión empresarial o la economía. Independientemente del ámbito de aplicación se puede decir que existen solo dos tipos de predicciones, las cualitativas y las cuantitativas.

* Predicciones cualitativas: Los pronósticos cualitativos a menudo se usan en situaciones donde hay poca o ninguna información histórica sobre la cual basar el pronóstico. Tal vez la técnica de predicción cualitativa más formal y standard de facto es el Método Delphi [3] ideado por el instituto RAND. La capacidad de predicción de la Delphi se basa en la utilización sistemática de un juicio intuitivo emitido por un grupo de expertos. El objetivo de los cuestionarios sucesivos, es disminuir el desvío de la opinión del experto de la opinión del conjunto de las respuestas obtenidas.
* Predicciones cuantitativas: Las predicciones cuantitativas formalizan valores históricos en un modelo que relaciona de forma estadística relaciones de los valores actuales de las variables con valores pasados. En otras palabras. el modelo de pronóstico se usa para extrapolar el comportamiento pasado y actual en el futuro. Las técnicas más utilizadas a la hora de hacer predicciones cuantitativas son los modelos de regresión y los modelos de series temporales generales.

Los modelos de predicción térmica, como la mayoría de los casos de modelado, se puede abordar desde diferentes enfoques o perspectivas. Con respecto al nivel en el que el enfoque de modelado caracteriza los detalles físicos del objeto o proceso para reproducir, los enfoques de modelado se clasifican en tres grupos.

Las predicciones basadas en modelos físicos intentan establecer todas las ecuaciones físicas que caracterizan la dinámica térmica y el comportamiento energético, en función de los parámetros y un conjunto de múltiples variables relacionadas. Este tipo de modelos son esenciales en la fase de diseño del sistema; suponen que la característica térmica del edificio se conoce a priori. Una de sus principales desventajas es que este conocimiento es difícil de obtener y muy incierto, no solo para la simulación de elementos individuales, sino también para sus relaciones, que se reflejarán en grandes incertidumbres en las magnitudes simuladas, lo que limitará las capacidades de ahorro de energía. Algunas de las herramientas de simulación de caja blanca más conocidas son FLUENT, TRNSYS, EnergyPlus, BLAST, ESP-r

Los modelos de caja negra son una forma de superar estos inconvenientes. Las técnicas de aprendizaje de máquina de regresión se basan en datos supervisados, deduciendo la relación no lineal de múltiples magnitudes no relacionadas con una o varias salidas independientes. En una primera etapa, los algoritmos de regresión necesitan aprender de un conjunto de datos de entrenamiento (conjunto de datos históricos supervisados). Este conjunto de datos inicial debe ser representativo de los procesos a modelar y, a veces, se necesitan procesos experimentales para generar datos. Los principales inconvenientes de esta aproximación son que las técnicas de regresión no pueden aprender ninguna situación no observada durante el proceso de capacitación y que, a veces, es una tarea difícil monitorear todas las magnitudes relacionadas con el proceso, lo que limita el rendimiento del procedimiento propuesto.

Las diferentes opciones de modelado se agrupan en tres familias, tal y como hemos anticipado anteriormente.

* White-Box: se lo conoce como modelo de caja blanca cuando se puede acceder a la arquitectura y a todos los datos del sistema. Los modelos de caja blanca son modelos de fase de diseño ampliamente reconocidos como inexactos, la mayoría de las veces, debido al uso impredecible de la construcción. Los modelos de caja blanca, basados ​​en la descripción de los detalles físicos del elemento modelado, reducen considerablemente la cantidad de sensores necesarios para producir un resultado. Por otro lado, los enfoques de caja blanca son muy sensibles a las imprecisiones en el modelado. Las imprecisiones pueden deberse a errores o simplificaciones hechas durante el modelado o debido a desviaciones en los valores dados a los parámetros de modelado.
* Grey-Box: se refiere a un sistema mientras tiene algún conocimiento del sistema interno. Por lo general, los enfoques de Gray-Box describen el modelo / proceso a modelar mediante expresiones analíticas simplificadas. Después de una fase de calibración, basada en las lecturas del sensor, se calculan los coeficientes de las expresiones analíticas. Una vez que se completa la descripción analítica, el modelo resuelve la (s) ecuación (es) para evaluar el modelo. Los modelos de Gray-Box no caracterizan la composición del artículo para modelar, sino su comportamiento a las excitaciones en sus límites.
* Black-Box: el modelo de caja negra se basa en la idea de probar un sistema sin tener ningún conocimiento del funcionamiento interno del sistema o de su arquitectura. Estos tipos de modelado son bastante útiles en la configuración de reconocimiento de patrones. A diferencia de las pruebas de caja blanca, las pruebas de caja negra facilitan la comunicación de prueba entre los módulos. Los modelos Black-Box tienen una gran dependencia de los datos de entrada disponibles y la capacidad de manipularlos, en este contexto, la potencia computacional disponible es una restricción. En los últimos años, la mejora de los marcos tecnológicos (software y hardware) ha facilitado la aplicación de enfoques Black-Box.

Particularizando para el caso de las que abordamos en el presente trabajo de fin de master, las series temporales, mediante modelos ARIMA también conocidos como modelos Bob-Jenkins. La principal característica de que una serie temporal debe presentar para poder ser modelada mediante ARIMA es la estacionaridad.

Dentro del ámbito que aplica en este trabajo de fin de master, la predicción de la demanda energética, bien sea energía térmica o eléctrica, podemos decir que es una de las áreas con mayor actividad investigadora debido a su relevancia económica como medioambiental así como los problemas de gestión de la red eléctrica que se derivan de la mayor demanda del conjunto de la población tal y como explican en su trabajo Bennett, Stewart y Ju[]. En el mismo trabajo los autores comparan diferentes aproximaciones (modelos de regresión multivariante, redes neuronales, series temporales) a la hora de realizar la predicción de la demanda.

La predicción de la demanda, independientemente de la técnica que se aplique puede considerarse, es su definición más purista como una serie temporal univariante, de todas forma existen multitud (Engle et al [] , Darbellay and Slama [] entre otros ) de trabajos que hacen referencia a la virtudes de aplicar la meteorología como variable adicional o moduladora de la demanda energética.

Aquí añadir nuevas referencias a ejemplos de la literatura que describe como se utilizan diferentes técnicas a la hora de predecir demandas termcas. ARIMAX vs NN

Metodología

Ya descrito en este documento, la mayor limitación para los enfoques de caja negra es sus dificultades para manejar escenarios no entrenados. Si la variabilidad de datos en el conjunto de datos de entrenamiento no cubre todo el rango de variabilidad de entrada, un algoritmo de predicción basado en técnicas de caja negra no funcionará bien. Una simple analogía sería que una caja negra entrenada solo con imágenes de retina sanas no es capaz de detectar enfermedades en las retinas.

Para evitar el escenario descrito anteriormente, se plantea una hibridación del modelo de caja negra de extracción de energía utilizando el modelo de caja gris de energía para generar un conjunto de datos virtual que puede complementar el conjunto de datos contenidos en el dataset.

Tal y como se ha comentado anteriormente la primera de las actividades descritas en este trabajo de fin de master tiene como objetivo predecir la demanda térmica en régimen estacionario, considerando dicha demanda como una serie temporal.

Las técnicas de series temporales, son un conjunto de técnicas de modelado que implica construir modelos de pronóstico con parámetros basados ​​en permutaciones de la variable que el modelo debe pronosticar. El promedio móvil autorregresivo integrado (ARIMA) p, d, q, es el modelo general de las técnicas de series de tiempo que encapsula el modelo autorregresivo, la diferenciación no estacional y el modelo de promedio móvil. El término "p" representa la cantidad de parámetros rezagados en el tiempo; el término "d" representa el número de diferencias discretas que han sufrido los datos de la variable de pronóstico para eliminar la estacionalidad; y el término "q" representa la cantidad de parámetros de error de pronóstico rezagados en el modelo para dar cuenta de un promedio móvil observado en los datos de la variable de previsión. El orden de los términos "p" y "q" puede identificarse mediante el uso de la función de autocorrelación parcial. El nivel de diferencia puede determinarse mediante el uso de una gráfica de autocorrelación basada en la naturaleza de la disminución o el uso de la estadística Durbin-Watson (DW) para identificar la autocorrelación en el error de pronóstico (correlación serial). Los coeficientes de los modelos de series de tiempo se pueden estimar mediante regresión o estimadores de máxima verosimilitud.

En pocas palabras, un modelo ARIMA se puede ver como un modelo de regresión múltiple con uno o más términos autorregresivos (AR) y / o uno o más términos de promedio móvil (MA). Los términos autorregresivos para una variable dependiente son simplemente valores en instantes anteriores de esa variable dependiente que tienen una relación estadísticamente significativa con su valor más reciente. Los términos promedio móviles no son más que residuos (es decir, errores en instantes anteriores) resultantes de estimaciones hechas previamente.

Por ejemplo, una serie temporal podría ser estimada por una combinación adecuadamente ponderada de las siguientes cuatro variables

1. *xt* = el valor de la variable independiente en el tiempo

2. *yt-1* = el valor inmediato anterior de la variable dependiente en el tiempo

3. *yt-2* = el valor anterior inmediato de la variable dependiente en el tiempo

4. *et-1* = el error de estimación producido por el modelo at time Este modelo de variable única e independiente, de regresión múltiple para estimar la variable dependiente se basa en el valor predictivo de la variable independiente, la variable dependiente misma (en el instante anterior #1), la variable dependiente nuevamente (en el instante anterior #2) y un término de error previamente producido (en el instante anterior # 4). Es decir,



Donde XXXX son los coeficientes a acalcula

La metodología seguida para modelar y predecir la serie temporal de estudio que se sigue en el presente trabajo de fin de master puede resumirse de la siguiente manera.

1. Trazar la serie temporal y determinar sus características básicas. Buscar posibles valores atípicos o cualquier indicación de que la serie de tiempo ha cambiado con respecto a sus características básicas (como tendencias o estacionalidad) durante el período de datos históricos disponibles.

2. Eliminar cualquier tendencia o componentes estacionales, ya sea mediante la diferenciación o ajustando un modelo apropiado a los datos. El objetivo de estas operaciones es producir un conjunto de residuos estacionarios.

3. Desarrollar un modelo de pronóstico para los residuos. No es inusual encontrar que hay varios modelos plausibles y se deberá realizar un análisis adicional para determinar cuál es el mejor para implementar. A veces, los modelos potenciales pueden eliminarse sobre la base de su ajuste a los datos históricos.

4. Validar el rendimiento del modelo (o modelos) del paso anterior. Esto probablemente involucre algún tipo de procedimiento de validación cruzada o de muestra dividida. El objetivo de este paso es seleccionar un modelo para usar en la previsión.

5. También resultan interesantes las diferencias entre la serie de tiempo original y, y los valores que predeciría el modelo en la escala original. Para pronosticar valores en la escala de la serie de tiempo original Yr. invierta las transformaciones y los ajustes de diferenciación realizados para eliminar tendencias o efectos estacionales.

6. Para pronósticos de valores futuros en el período T + r en la escala original. si se usó una transformación, digamos, x In y ,, entonces el pronóstico hecho al final del período 1 T para T + r se obtendría invirtiendo la transformación.

Es importante definir cuidadosamente el significado del rendimiento. Es tentador evaluar el rendimiento sobre la base del ajuste del modelo de pronóstico o serie de tiempo a los datos históricos. Hay muchas medidas estadísticas que describen qué tan bien un modelo se ajusta a una muestra de datos dada, y varias de ellas se describirán en capítulos posteriores. Este enfoque de bondad de ajuste a menudo usa los residuos y no refleja realmente la capacidad de la técnica de pronóstico para predecir con éxito las observaciones futuras. El usuario de los pronósticos está muy preocupado por la precisión de los pronósticos futuros, no por la bondad de ajuste del modelo, por lo que es importante evaluar este aspecto de cualquier técnica recomendada. En ocasiones, la precisión del pronóstico se denomina error de pronóstico "fuera de muestra", para distinguirla de los residuos que surgen de un proceso de ajuste del modelo (punto #3 del listado anterior).

Es costumbre evaluar el rendimiento del modelo de predicción utilizando los errores de pronóstico de un paso adelante, donde *ŷt* (t - 1) es la predicción de *y* que se hizo un período anterior.

*e*t(*1*)=*y*t-*ŷt*(*t-1*)

Supongamos que hay n observaciones para las cuales se han hecho pronósticos *yn*errores de predicción de un paso adelante, e (1). t = I, 2, ..., n. Las medidas estándar 1 de precisión la predicción son el error promedio o error medio la desviación media absoluta (o error absoluto medio) y el error cuadrático medio.



El error promedio de pronóstico en Eq. (2.30) es una estimación del valor esperado del error de pronóstico, que esperamos sea cero; es decir, la técnica de pronóstico produce predicciones sesgadas. Si el error promedio de pronóstico difiere apreciablemente de cero, se indica un sesgo en el pronóstico. Si el error de pronóstico medio se aleja de cero cuando la técnica de predicción está en uso, esto puede ser una indicación de que la serie de tiempo subyacente ha cambiado de alguna manera, la técnica de pronóstico no ha capturado este cambio. Tanto la desviación absoluta media (MAD) en Eq. (2.31) y el error cuadrático medio (MSE) en la Eq. (2.32) miden la variabilidad en los errores de predicción, es deseable que la variabilidad en los errores de pronóstico sea pequeña.

Las medidas de precisión que dependen de la escala no facilitan las comparaciones de una única técnica de pronóstico a través de diferentes períodos de tiempo. Para lograr esto, necesitamos una medida de error de pronóstico relativo.

El error medio de previsión porcentual (MPE) se define como:



y el error medio de pronóstico porcentual absoluto (MAPE) es:



Un aspecto importante a la hora de aplicar MAPE es que solo tiene sentido si la serie de tiempo y no contiene el valor cero.

Una vez definido como evaluar la bondad de un modelo a través de diversos estadísticos de error debemos atender al modo de seleccionar entre varios modelos competitivos que se pueden usar para pronosticar una serie temporal particular. Por ejemplo, hay varias formas de modelar y pronosticar tendencias. En consecuencia, la selección de un modelo de pronóstico apropiado es de considerable importancia práctica.

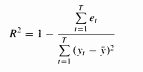
La selección del modelo que mejor se ajusta a los datos históricos generalmente no da como resultado un método de pronóstico que produce las mejores previsiones de datos nuevos. Concentrarse demasiado en el modelo que produce el mejor ajuste histórico a menudo resulta en sobreajuste, o incluir demasiados parámetros o términos en el modelo solo porque estos términos adicionales mejoran el ajuste del modelo. En general, el mejor enfoque es seleccionar el modelo que dé como resultado la desviación estándar más pequeña (o error cuadrático medio) de los errores de pronóstico de un paso adelante cuando el modelo se aplica a datos que no se usaron en el proceso de ajuste.

Al evaluar el ajuste del modelo a los datos históricos, existen varios criterios que pueden ser de valor. El error cuadrático medio de los residuos es



donde T períodos de datos se han utilizado. para ajustarse a un modelo con parámetros p y e 1 2 el residuo del proceso de ajuste del modelo en el período t.

Otro criterio es la estadística R-cuadrado,



Por lo tanto, seleccionando el modelo que maximiza R2 el modelo que minimiza la suma de los residuos al cuadrado.

Otros dos criterios importantes son el Criterio de información de Akaike (AIC) y el Criterio de información de Schwarz (SIC). En ambos casos un valor bajo de AIC o SIC indican mejora en la bondad del modelo. Posterior volveremos a incidir sobre el criterio de Akaike que es el seleccionado para elegir entre las diferentes opciones de modelado.

Una vez una serie de tiempo se juzga estadísticamente como estacionaria y se tienen un abanico de modelos tentativos, se examinar la función del coeficiente de autocorrelación (ACF) y la función del coeficiente de autocorrelación parcial (PACF) para conocer la naturaleza de los modelos propuestos y evaluar que no tratemos de modelar ruido blanco en el sentido que lo indica el test de Ljung-Box .

Una vez introducido los modelos ARIMA se puede complejizar un poco el modelado considerando una aproximación ARIMAX. ARIMAX es un acrónimo de media móvil integrada autorregresiva con variables exógenas. Es una extensión lógica del modelado ARIMA puro que incorpora variables independientes que agregan valor explicativo. Cuando los términos AR y MA en un modelo ARIMA puro no son suficientes para proporcionar una aceptable alta (o alguna otra medida del poder explicativo general de un modelo), es natural buscar otros fenómenos de conducción cuya influencia con el tiempo no esté suficientemente integrada en los valores históricos de la serie de tiempo dependiente.

La construcción de un modelo ARIMAX requiere combinar el valor predictivo de los valores de la serie temporal final (*yt*) y los errores del modelo final (*et*) con el valor predictivo de las variables exógenas.

Las asunciones estadísticas a considerar sobre la variable exógena son:

* El coeficiente estimado para una variable exógena debe ser significativamente diferente de 0.
* Una variable exógena no debe mostrar evidencia de recibir comentarios de la variable dependiente. Es decir, un candidato atractivo de variable exógena debe mostrar una relación causal significativa con la variable dependiente sin que la variable dependiente muestre una relación causal con ella. Las direcciones de la causalidad significativa entre una variable exógena y la variable dependiente se pueden probar usando la/causalidad de Granger []

Dadas todas las consideraciones técnicas y metodológicas consideradas anteriormente estamos en condiciones de comenzar a aplicar el modelado ARIMA/ARIMAX a nuestro caso de uso.

Aquí Explicar la hibridación de Grey-Box y Black-Box para evaluar al demanda ante escenario de cambio de condiciones interiores.

## Desarrollo

## Predicción Demanda Térmica por Series Temporales

### Creación del modelo

El predicción de la demanda basada en series numéricas y la metodología de aplicación seguirá la siguiente secuencia.

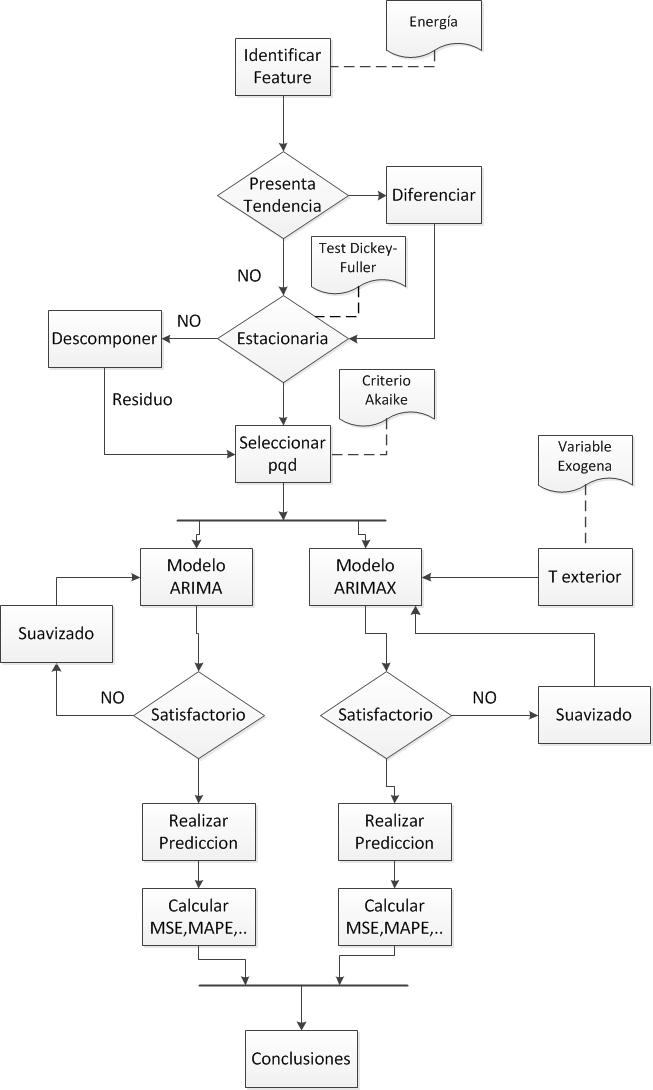


Figura : Flujo del proceso de modelado y predicción

El primero de los métodos considerados va a ser el tratamiento por series temporales univariantes. Se propone tratar la serie numérica Pbld (energía térmica suministrada)como si se tratar de una serie temporal, que de hecho lo es, obviado el efecto distorsionador que puede tener una temperatura exterior muy dispar a lo estadísticamente habitual al periodo que aplique.

#### Aproximación ARIMA

Se aplicará el modelo ARIMA a la serie temporal Pbld. Los modelos ARIMA son en los que una variable es explicada utilizando exclusivamente una "endogena": su propio pasado. Podemos decir que la consideración exclusiva de los valores pasados de una determinada variable para explicar su evolución presente y futura supone, al mismo tiempo, una ventaja y un inconveniente:

* La ventaja radica en el hecho de no necesitar distintas series de datos (distintas variables) referidas al mismo período de tiempo (característica común a todos los modelos univariantes) y, al mismo tiempo, ahorrarnos la identificación y especificación del modelo en el sentido tradicional.
* El inconveniente es que, al renunciar a la inclusión de un conjunto más amplio de variables explicativas, no atendemos a las relaciones que pueden existir entre ellas.

Dentro de estos modelos univariantes se desarrollarán suficientemente los conocidos con el nombre de ARIMA. Posteriormente se complementará esta perspectiva univariante añadiéndose a la

especificación una o más variables exógenas al modelo "tradicional" aproximándonos al estudio de los conocidos como modelos de transferencia.

La palabra ARIMA significa Modelos Autorregresivos Integrados de Medias Móviles.

Definimos un modelo como autorregresivo si la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores añadiéndose, como en los modelos estructurales, un término de error. En el caso de procesos estacionarios con distribución normal, la teoría estadística de los procesos estocásticos dice que, bajo determinadas condiciones previas, toda Y t puede expresarse como una combinación lineal de sus valores pasados (parte sistemática) más un término de error (innovación).

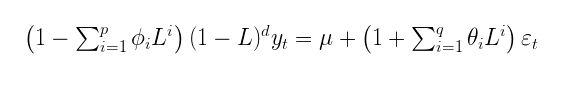


Figura : Definición analítica de un ARIMA de orden p,d,q.

Donde L denota el operador de ***lag*** que realiza la trasformación de lag temporal de la siguiente manera:



Figura : Interpretación de lag temporal de orden *d*

Los modelos autorregresivos se abrevian con la palabra AR tras la que se indica el orden del modelo: AR(1), AR(2),....etc. El orden del modelo expresa el número de observaciones retasadas de la series temporal analizada que intervienen en la ecuación.

Queda clara que la aproximación con modelos AR o MA está restringida, en términos generales, a aquellos procesos que cumplan, al menos de forma débil, la restricción de estacionariedad es decir, series en las que la tendencia y las estacionalidad has sido substraídas.

Para eliminar o al menos minimizar la tendencias existen varios metodos, como pueden ser, la aplicación de logaritmos, la diferenciación o el cálculo de medias “flotantes” ( en ingles, moving avearage)

El test más habitual a la hora de determinar la estacionariedad de una serie temporal, consiste en la aplicación del conocido como test de Dickey–Fuller (Test DF) o Dickey-Fuller Ampliado (Test ADF). Éste es un contraste de “No estacionariedad” (hipotesis nula) ya que la hipótesis nula es precisamente la presencia de una raíz unitaria en el proceso generador de datos de la serie analizada.

El código tras realizar varias comprobaciones de la estacionalidad de la serie resultante tras aplicar bien el logaritmo y la diferenciación, hace uso de las librería de Python para descomponer la serie en su parte “tendencia” , “periódica” y “residual”, una vez obtenida la parte residual se busca la mejor tripleta de parámetro p-q-d para el modelo ARIMA completo. Esta parte del código requiere de mucho tiempo computacional.

En el caso que nos aplica, es decir la predicción de la cantidad de energía a suministrar para mantener ciertas condiciones interiores de confort la serie temporal es la energía, más en concreto energía térmica ( feature Pbld del dataset)

El tratamiento de la serie temporal comenzará por verificar si es estacionaria o no mediante el test aumentado de Dickey-Fuller (ADF), en caso de que no lo sea, se descompondrá en sus componentes de “tendencia”, “periodica” y “residual”, tras verificar si la parte residual es estacionaria se procederá a identificar el modelo ARIMA que mejor represente la serie.

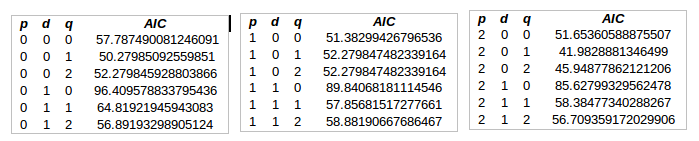
Los resultados optenidos son los siguientes:

* Test de ADF sobre la feature Pbld:
  + Test Statistic -2.302028
  + p-value 0.171315
  + Critical Value (1%) -3.466201
  + Critical Value (5%) -2.877293
  + Critical Value (10%) -2.575168

De los resultados anteriores se concluye que para al feature Pbld al ser el valor de test mayor que los umbrales 10%, 5% y 1% no se puede rechazar la hipotesis nula (H0 = No Estacionaria). Al considerar la componente residual, ocurre lo contrario, se puede rechazar H0, por tanto la serie es estacionaria. Del valor ***p-value*** se concluye lo mismo, en el primer caso un valor 0,17 es superior al valor de significación mínimo exigible 10% (0.1) al contrario de lo que ocurre para el caso de la serie residual

Una vez identificada la componente sobre la que se aplicará el método ARIMA se debe proceder a identificar los mejores valores para *p,d,q* para dicha serie. Se ha implementado un método ***doSelectBestArima*** que de forma iterativa busca la mejor combinación de parámetros pdq para lo que se han utilizado el criterio de información de Akaike (AIC). El AIC o Criterio de información de Akaike evalúa la bondad de la solución en función de la entropía. Su idea clave es la de penalizar un exceso de parámetros ajustados.

Los valores obtenidos para el coeficiente AIC son:



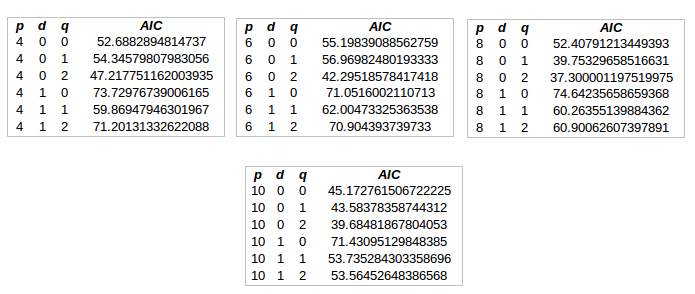


Figura : Listado de valores Akaike obtenidos

En una primera instancia se evaluarán tres estadísticos (error absoluto medio -MAE- y el error cuadrático medio -MSE- y error porcentual absoluto medio -MAPE-) para determina la bondad de cada uno de los modelos que se implementen.

El error absoluto medio (MAE) y el error cuadrático medio (MSE) son dos de las medidas más comunes utilizadas para medir la precisión de las variables continuas. Tanto MAE como RMSE expresan el error de predicción promedio del modelo en unidades de la variable de interés. Ambas métricas pueden variar de 0 a ∞ y son indiferentes a la dirección de los errores. Son puntuaciones negativamente orientadas, lo que significa que los valores más bajos son mejores. Tomlos errores cuadrados promedio tiene algunas implicaciones interesantes para MSE. Dado que los errores se cuadran antes de promediarlos, el MSE otorga un peso relativamente alto a los grandes errores. Esto significa que el MSE debería ser más útil cuando los grandes errores son particularmente indeseables.

MSE tiene el beneficio de penalizar más los errores grandes, por lo que puede ser más apropiado en algunos casos, por ejemplo, si estar apagado 10 es más del doble que estar fuera de 5, pero estar 10 veces fuera de tan solo es el doble de malo que estando fuera por 5, entonces MAE es más apropiado.  
  
Desde el punto de vista de la interpretación, MAE es claramente el ganador. MSE no describe solo el error promedio y tiene otras implicaciones que son más difíciles de interpretar.

Por otro lado, el error porcentual absoluto medio (MAPE) es una de las medidas de previsión más utilizadas, debido a sus ventajas de independencia de escala e interpretabilidad. Sin embargo, MAPE tiene la desventaja significativa de que produce valores infinitos o indefinidos para valores reales cero o cercanos a cero tal y como se podrá ver en los modelos creados a partir del análisis del residuo, es por ello que no será considerado para medir la bondad del modelos, pero se dará su valor como parte del estudio.

La fórmula de cálculo del valor MAPE es la siguiente:



Donde *At* son los valores observados y Ft los valores predichas

Por lo tanto se considerarán el MSE y el MAE para la elección del mejor de los modelos.

La medida de la bondad mediante AIC ha dado como mejor opción (p=8,d=0,q=2).

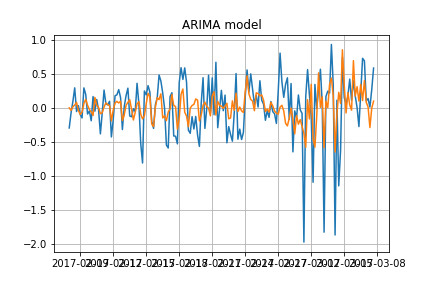


Figura : Modelo ARIMA (p=8,d=0,q=2) para el residuo de la serie temporal

Los coeficientes obtenidos para el modelo ARIMA son los siguientes:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ar.L1 | ar.L2 | ar.L3 | ar.L4 | ar.L5 | ar.L6 | ar.L7 | ar.L8 | ma.L1 | ma.L2 |
| 0.0862 | 0.8264 | -0.1629 | -0.1409 | 0.0351 | 0.3744 | -0.0168 | -0.3609 | -0.0121 | -0.9864 |

|  |  |  |
| --- | --- | --- |
| Error Cuadrático Medio  (MSE) | Error absoluto Medio  (MAE) | Error absoluto Medio Porcentual  (MAPE) |
| 0.138 | 0.266 | 197.003 |

Tabla : Estadísticos de error obtenidos para el modelo ARIMA de la Figura 6

#### Aproximación ARIMAX

Como siguiente paso se plantea la introducción de una variable más al modelo de regresión, se trataría de lo que se llama, variable exógena, que en este caso es la temperatura exterior. El motivo de incluir una variable exógena es la de tratar de explicar las variaciones bruscas en la serie temporal por la influencia de dichas variables. Dichos modelos son conocidos como ARIMAX.

Modelo ARIMAX calculado sobre el logaritmo de la serie temporal de energía (Pbld) y considerando la temperatua exterior como variable exogena. La serie temporal Pbld, tal y como se ha visto anteriormente no pasa el test de aumentado de Dickey-Fuller (ADF), pero debido a que queda en un valor relativamente razonable se realiza el estudio.

Las siguientes figuras los se describen:

* Estudio de la serie temporal Pbld considerando como variable exógena la temperatura exterior (Izquierda)
* Estudio de la serie temporal Pbld considerando como variable exógena la temperatura exterior, habiendo normalizado ambas variables (Centro)
* Estudio de la serie temporal Pbld considerando como variable exógena la temperatura exterior y realizando un suavizado por convolución de la serie (Derecha).

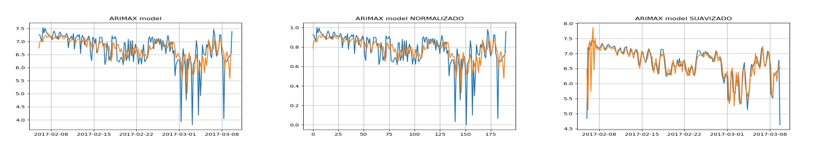


Figura : Modelos ARIMAX ,(p=8,d=0,q=2)

Los coeficiente de la serie ARIMAX de orden p=8,d=0,q=2 para los casos anteriores son los siguientes:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ar.L1 | ar.L2 | ar.L3 | ar.L4 | ar.L5 | ar.L6 | ar.L7 | ar.L8 | ma.L1 | ma.L2 |
| 1.4229 | -0.3437 | -0.0916 | -0.0428 | 0.1994 | 0.2421 | -0.5209 | 0.1298 | -1.1424 | 0.1424 |
| 1.6791 | -0.6451 | -0.0620 | -0.0258 | 0.1937 | 0.2163 | -0.5900 | 0.2329 | -1.4006 | 0.4006 |
| 0.4051 | 0.1340 | -0.0232 | 0.5141 | -0.2088 | 0.5143 | -0.1639 | -0.1853 | 0.9221 | 0.9999 |

Tabla : Coeficientes pdq para modelo ARIMAX sobre el residuo

Los diferentes estadísticos de error obtenidos son los siguientes:

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo | Error Cuadrático Medio  (MSE) | Error absoluto Medio  (MAE) | Error absoluto Medio Porcentual  (MAPE) |
| ARIMAX | 0.22119934852 | 0.31558113716 | 5.236 |
| Normalizado | 0.015 | 0.084 | nan |
| Suavizado | 0.103 | 0.165 | 2.619 |

Tabla : Estadísticos de error para los modelos anteriores.

Finalmente se realiza el estudio de la serie temporal Pbld considerando como variable exógena la temperatura exterior y realizando un suavizado tras normalizar ambas series.

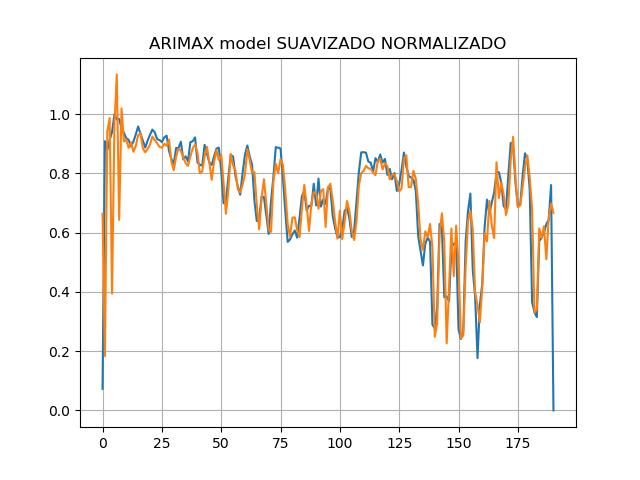


Figura : Modelo ARIMAX tras suavizar y normalizar.

Los coeficiente de la serie ARIMAX, tras realizar el suavizado y normalizado, de orden p=8,d=0,q=2 son:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ar.L1 | ar.L2 | ar.L3 | ar.L4 | ar.L5 | ar.L6 | ar.L7 | ar.L8 | ma.L1 | ma.L2 |
| 0.4078 | 0.0213 | 0.0290 | 0.4236 | -0.1751 | 0.4242 | -0.1259 | -0.2347 | 0.9165 | 1.0000 |

Tabla : Coeficientes ARIMAX para la series tras aplicar el suavizado y normalizado

Los diferentes estadísticos de error obtenidos son los siguientes:

|  |  |  |
| --- | --- | --- |
| Error Cuadrático Medio  (MSE) | Error absoluto Medio  (MAE) | Error absoluto Medio Porcentual  (MAPE) |
| 0.012 | 0.058 | nan |

Tabla : Estadisticos de error sobre el modelo ARIMAX de la serie temporal Pbld

El estudio anterior nos anticipa que el modelo ARIMAX suavizado puede dar resultados aceptables ante lo cual evaluamos el mismo escenario tras descomponer la serie temporal en sus componentes, periódica, tendencia y residuo o resto, considerando en este caso el residuo no como la desviación con relación a la predicción sino que como el “resto” de la serie temporal tras extraer las componentes periódicas y de tendencia anteriormente citadas.

Se plantea realizar el estudio sobre la serie temporal tras extraer su parte periódica y de tendencia.

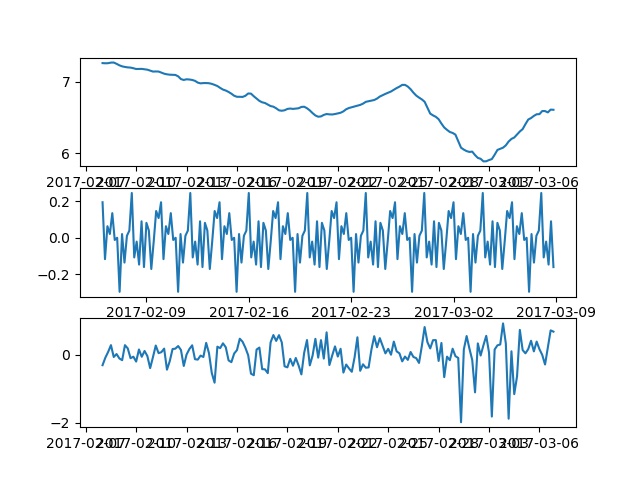


Figura : Componente de tendencia, periódica y de residuo o resto de la serie original.

Para verificar que el residuo presenta algún patrón temporal realizamos los diagramas de correlación y autocorrelación del mismo.

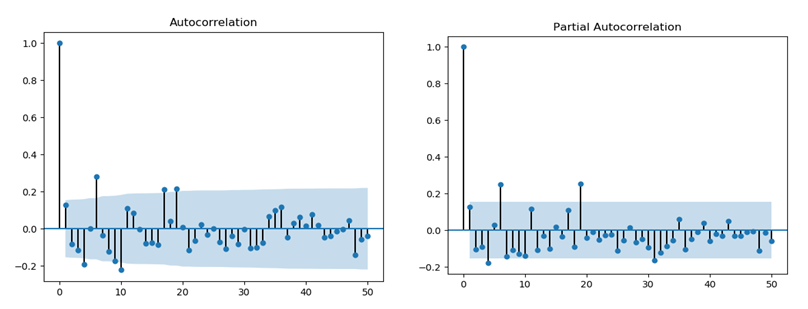


Figura : Diagramas de auto correlación y autocorrelación parcial de la serie residuo grado de significancia de 95%

Se puede observar que hay coeficientes te autocorrelación y autocorrelación parcial que presentan cierta relevancia por lo que se concluye que se puede proceder a realizar el estudio del residuo como serie temporal. Para determinar los mejores coeficientes p,d,q se realiza el estudio mediante el criterio de Akaike cuyo resultado vuelve a dar como mejor combinación p=8,q=0,d=2. Por último queda verificar el resultado de los para el test ADF, es decir, los residuos son una serie estacionaria.

Recordamos los resultados de la ejecución del test aumentado Dickey-Fuller son:

* Test Statistic -3.947550
* p-value 0.001711
* Critical Value (1%) -3.472979
* Critical Value (5%) -2.880252
* Critical Value (10%) -2.576747

De donde se ve que la serie es estacionario con una probabilidad superior al 99%.

Las siguientes figuras los se describen:

* Estudio sobre el residuo de la serie temporal Pbld considerando como variable exógena la temperatura exterior (Izquierda)
* Estudio sobre el residuo de la serie temporal Pbld considerando como variable exógena la temperatura exterior, habiendo normalizado ambas variables (Centro)
* Estudio sobre el residuo de la serie temporal Pbld considerando como variable exógena la temperatura exterior y realizando un suavizado por convolución de la serie (Derecha).

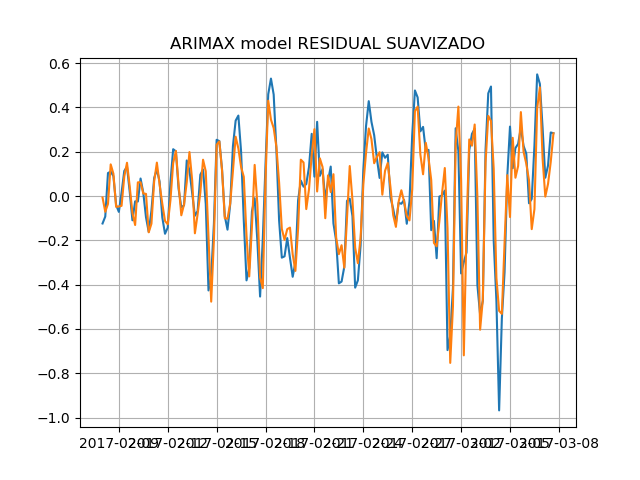
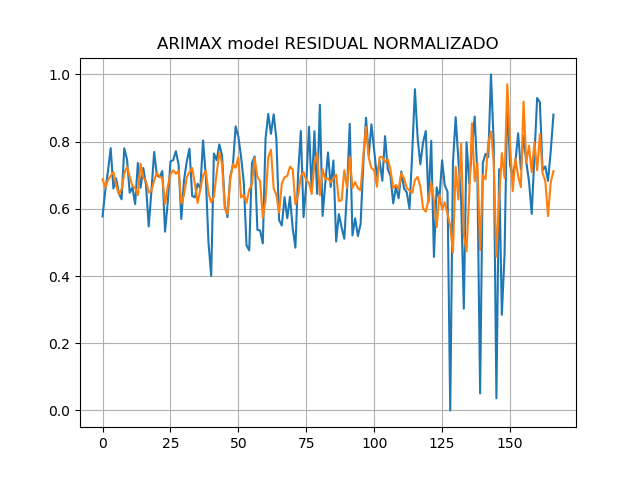
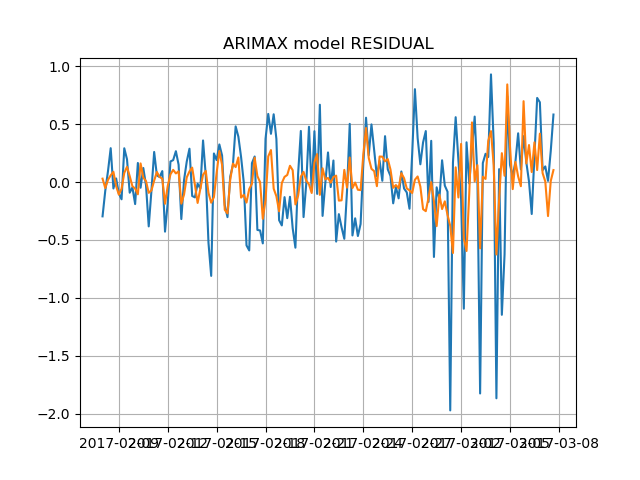


Figura : modelo ARIMAX sobre el residuo de la serie temporal (p=8,d=0,q=2)

Los coeficiente de la serie ARIMAX de orden p=8,d=0,q=2 para los casos anteriores son los siguientes:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ar.L1 | ar.L2 | ar.L3 | ar.L4 | ar.L5 | ar.L6 | ar.L7 | ar.L8 | ma.L1 | ma.L2 |
| 0.0862 | 0.8264 | -0.1629 | -0.1409 | 0.0351 | 0.3744 | -0.0168 | -0.3609 | -0.0121 | -0.9864 |
| 0.0716 | 0.8239 | -0.1559 | -0.1388 | 0.0210 | 0.3774 | -0.0024 | -0.3593 | -0.0078 | -0.9922 |
| 0.9143 | -0.9084 | 0.1045 | 0.4678 | -0.6673 | 0.3187 | 0.1183 | -0.3102 | 0.1190 | 0.9322 |

Tabla : Coeficientes pdq para modelo ARIMAX sobre el residuo

Los diferentes estadísticos de error obtenidos son los siguientes:

|  |  |  |  |
| --- | --- | --- | --- |
| Modelo | Error Cuadrático Medio  (MSE) | Error absoluto Medio  (MAE) | Error absoluto Medio Porcentual  (MAPE) |
| ARIMAX | 0.138 | 0.266 | 197.037 |
| Normalizado | 0.016 | 0.091 | Nan |
| Suavizado | 0.018 | 0.097 | 134.804 |

Tabla : Estadísticos de error para los casos anteriores

Finalmente se realiza el estudio sobre el residuo de la serie temporal Pbld considerando como variable exógena la temperatura exterior y realizando un suavizado tras normalizar ambas series.

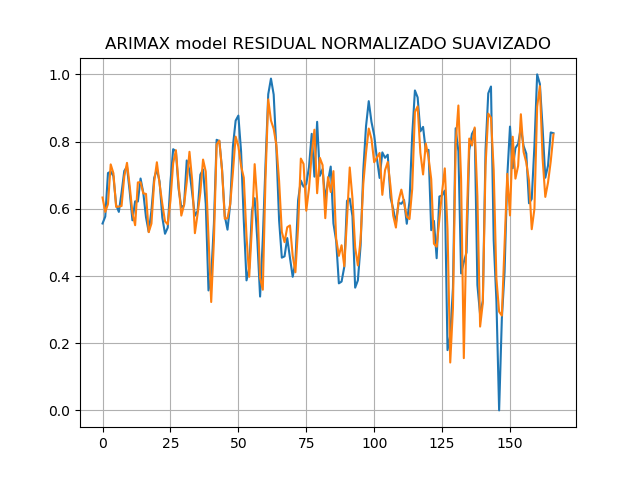


Figura : Modelo ARIMAX sobre el residuo normalizado y normalizado

Los coeficiente de la serie ARIMAX, tras realizar el suavizado y normalizado, de orden p=8,d=0,q=2 son:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ar.L1 | ar.L2 | ar.L3 | ar.L4 | ar.L5 | ar.L6 | ar.L7 | ar.L8 | ma.L1 | ma.L2 |
| 0.9143 | -0.9084 | 0.1045 | 0.4678 | -0.6673 | 0.3187 | 0.1183 | -0.3102 | 0.1190 | 0.9322 |

Tabla : Coeficientes ARIMAX para la series tras aplicar el suavizado y normalizado

Los diferentes estadísticos de error obtenidos son los siguientes:

|  |  |  |
| --- | --- | --- |
| Error Cuadrático Medio  (MSE) | Error absoluto Medio  (MAE) | Error absoluto Medio Porcentual  (MAPE) |
| 0.007 | 0.063 | Nan |

Tabla : Estadísticos de error para el modelo sobre el residuo tras aplicar el suavizado y normalizado

# Aplicación

La aplicación del modelo ARIMAX consiste en predecir los valores en un horizonte temporal de 24h. Para ello se procederá de la siguiente manera:

Evaluar el modelo. (Out-Bound)

Evaluar el modelo. (In-Bound)

Regenerar el modelo

Excluir último día del dataset

Se considera **in-bound** la predicción sobre datos “aprendidos” por el modelo, mientras que las predicción sobre datos no aprendido (futuro) se denominan **out-bound**, ambos casos se han realizado para el modelo sobre los residuos y tras aplicar el suavizado.

* **Predicción in-bound:** Las fechas utilizadas para la validación son los días comprendidos entre el 1 de marzo de 2017 y el 7 de marzo de 2017
* **Predicción out-bound:** La fecha elegida para la predicción ha sido el día 8 de marzo de 2017, para ello se ha suprimido dicha fechas del set de datos de entrenamiento.

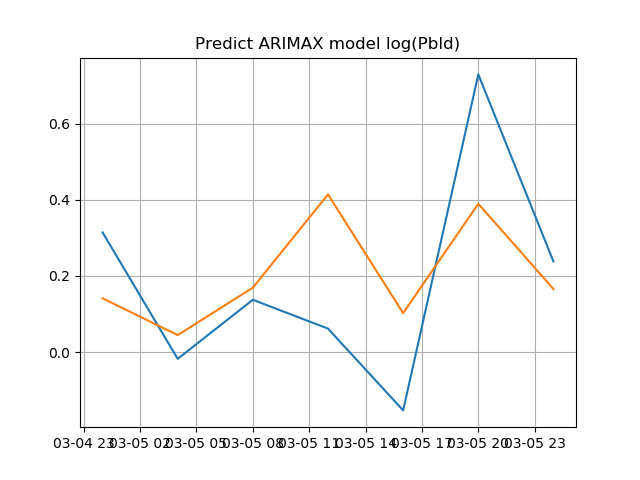
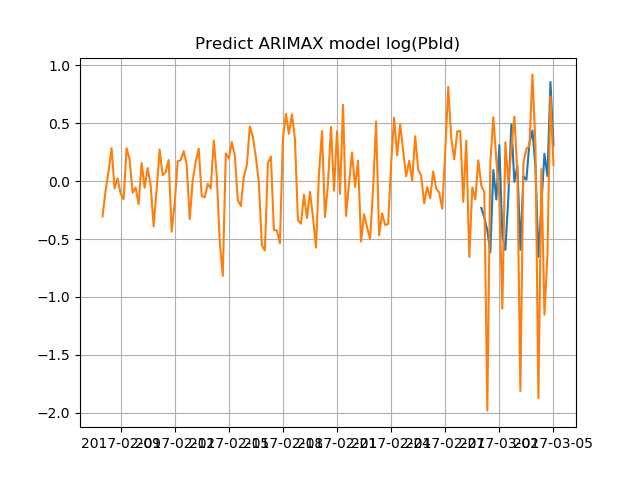


Figura 13: Predicción original in-bound y out-bound a partir del modelo suavizado de residuos

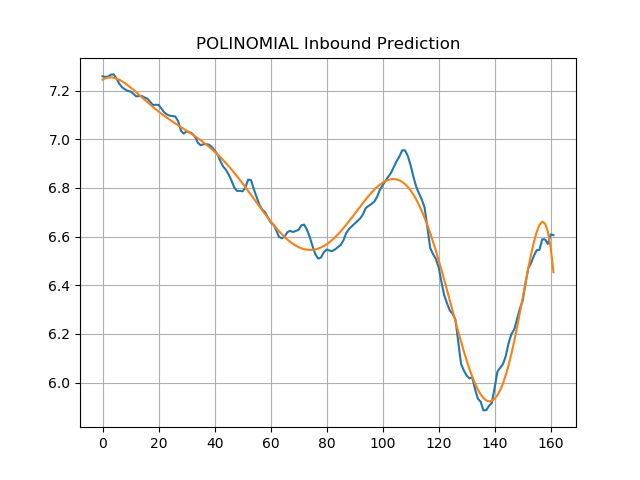
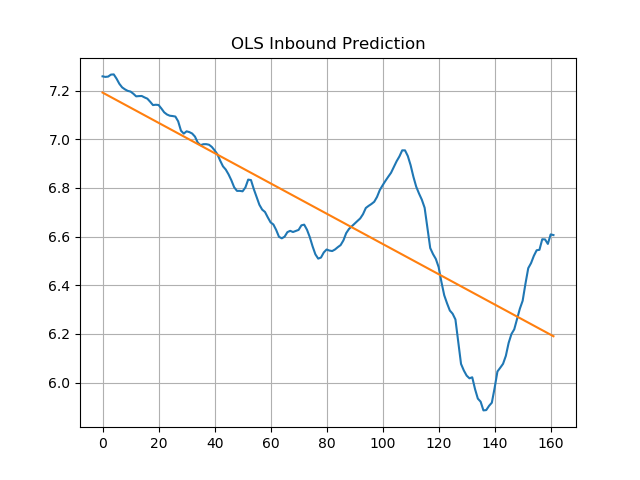


Figura : Evolución de la tendencia y aproximaciones por regresión lineal y polinómica (grado 9)

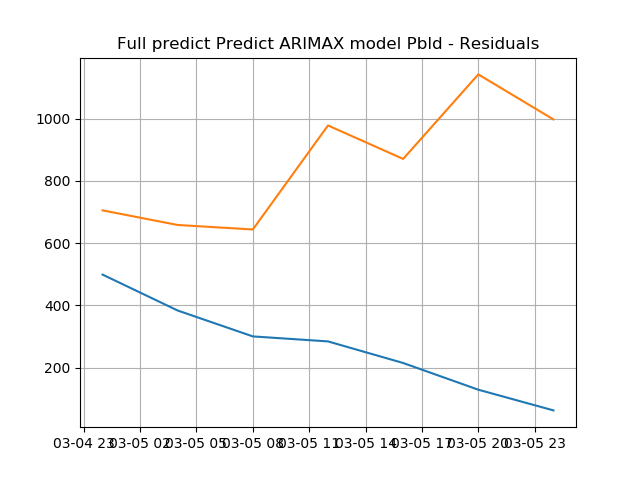
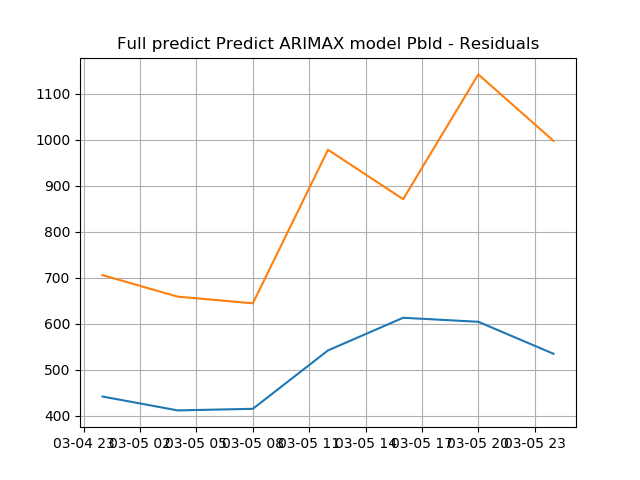


Figura : Predicción total tomando aproximado la tendencia mediante regresión lineal y un polinomio de grado 9

**Conclusiones**

**Bibliografía**

[] Directive 2012/27/EU of the European Parliament and of the Council of October 2012 on Energy Efficiency. 2030

[] Climate and Energy Policy Framework from the General Secretariat of the Council of the 23 and 24 October 2014.

[] Chia-Chien Hsu, Brian A. Sandford: *The Delphi Technique: Making Sense Of Consensus*. Practical Assessment,

Research & Evaluation, August 2007

[] Christopher Bennett, Rodney A. Stewart and Junwei Lu*: Autoregressive with Exogenous Variables and Neural Network Short-Term Load Forecast Models for Residential Low Voltage Distribution Networks*. Energies April 2014

[] Box, G.; Jenkins, G. Time Series Analysis: *Forecasting and Control; Holden-Day: San Francisco, CA, USA, 1970*

[] Engle, R.; Mustafa, C.; Rice, J. *Modelling peak electricity demand*. J. Forecast. 1992, 11, 241–251. 17. Darbellay, G.; [] Slama, M. *Forecasting short-term electricity demand: Do neural networks stand a better chance*? Int. J. Forecast. 2000, 16, 71–83.

[] Montero. R (2013): *Test de Causalidad* Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España

# Apéndices