UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Predicción de energía eólica con modelos autorregresivos

Jaime Torrijos Moreno
Tutor: Ángela Fernández Pascual
Ponente: José Ramón Dorronsoro Ibero
Julio 2019

Predicción de energía eólica con modelos autorregresivos

AUTOR: Ángela Fernández Pascual
TUTOR: José Ramón Dorronsoro Ibero

Dpto. Ingeniería informática
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Julio de 2019

Resumen (castellano)

Este Trabajo Fin de Grado trata de definir la importancia del aprendizaje automático hoy en día. Además de mencionar distintos modelos de predicción más usados hoy en día, nos centraremos en el uso de modelos autorregresivos de cara a realizar predicción de energía eólica.

Así pues, empezaremos introduciendo y definiendo qué son series temporales, así como sus distintas propiedades, comportamientos y componentes, para posteriormente saber procesarlas, entender su descomposición y el cómo descomponerlas para, finalmente, aplicar modelos autorregresivos y realizar predicciones de producción de energía. Evidentemente, explicaremos los distintos modelos que emplearemos: autorregresivos, medias móviles y sus combinaciones y variantes que resultan en ARMA, ARIMA, SARIMA, ARMAX, ARIMAX y SARIMAX.

Abstract (English)

This bachelor tesis tries to emphasise the importance of the machine learning nowadays. We will name some of the predition models that are the most used. We will then focus on using autoregresive models to predict wind power.

With this purpose we will start introducing and defining a time series, its properties, behaviours and components. Later on we will explain how to process them, how to decomponse them and finally, we will use autoregresive models to make wind power prediction. Of course we will also explain what those autoregresive models are, their parameters and how they work. Those models are mainly the autorregresive model and moving average models, which in combination result in ARMA, ARIMA, SARIMA, ARMAX, ARIMAX and SARIMAX models.

Palabras clave (castellano)

Predicción, energía eólica, series temporales, autorregresión, aprendizaje automático.

Keywords (inglés)

Prediction, wind power, time series, autoregression, machine learning.

Agradecimientos

A Ángela, por aguantarme semana tras semana con esto; a Fran, por meterme un poco más en este mundillo y hacer que este trabajo de fin de grado tenga mucho más sentido del que ya tiene; a Jose por el apoyo en el sprint final.

INDICE DE CONTENIDOS

INDICE DE FIGURAS

INDICE DE TABLAS

1 Introducción

1.1 Motivación

Todos sabemos lo importante que es la informática hoy en día. Bueno, eso lo sabemos desde hace mucho, no hay novedad alguna en este aspecto. Lo que se sabe con menos frecuencia es la importancia y auge que está teniendo el aprendizaje automático, o, como se suele decir, el *machine learning*.

Sin darnos cuenta, empleamos cada día sistemas y aplicaciones que hacen uso de éste. Sin ir más lejos, nuestro dispositivo móvil aprende según lo que escribimos o nos recomienda anuncios en base a lo que buscamos o frecuentamos. Dejando a un lado el uso personal y particular y poniéndonos de lado de otros sectores, tenemos, por ejemplo, numerosos avances en diagnósticos médicos[1], contenidos censurables en una red social o reconocimiento de objetos instantáneamente, sin dejar a un lado, claro está, de dos temas muy de moda: coches autónomos o reconocimiento de voz[1].

Sin lugar a dudas, el aprendizaje automático está presente en muchos aspectos, es innegable. Incluso en el sector energético, el papel del aprendizaje automático juega un cada vez más importante papel[2]. De la mano del también de moda *big data*, España tiene muchos puntos a favor para ser una de las grandes fuentes de energía renovable del mundo. Pero evidentemente, el mercado eléctrico es, valga la redundancia, un mercado, y, por ello una predicción energética, cuanto mejor es, mayor beneficio obtiene para el parque que la produce, ya que la energía que se oferta debe ser la que se va a generar; en caso contrario el parque se ve penalizado. Además, es muy importante de cara a la imprescindible labor de mantenimiento de los parques. No cabe duda que realizar una labor de mantenimiento en hora punta de producción energética es impensable.

1.2 Objetivos

Dada la gran cantidad de modelos de predicción que hay actualmente en aprendizaje automático, parece imposible saber por dónde empezar. Suelen usarse los modelos que mejor resultado suelen dar en términos generales: SVM (máquina de soporte de vectores)[3], redes neuronales, algoritmos genéticos, *gradient boosting*, etc.

Sin embargo, en nuestro caso, a la hora de realizar predicción de energía eólica debemos contar con el factor de la temporalidad, punto débil de algunos modelos de clasificación-regresión. Por ello, vamos a analizar y procesar series temporales con el fin de usar modelos específicamente diseñados para estas: los modelos autorregresivos. Estos modelos tienen en cuenta numerosas propiedades de las series temporales y nos aportan un punto de enfoque único con este tipo de datos. Dichos modelos los desgranaremos aquí, así como el tratamiento de las series temporales, sus componentes y las distintas maneras de realizar predicción.

1.3 Organización de la memoria

La memoria está organizada en los siguientes capítulos:

- **Estado del arte:** en este apartado, explicaremos qué es una serie temporal, objeto central de este trabajo, así como sus propiedades, su tratamiento y los modelos que podemos aplicar a ésta.
- **Diseño:** aquí, hablaremos sobre cómo hacer la predicción de energía eólica, con todos los convenientes pasos en los datos usados, los modelos a usar, y la parametrización y el método de entrenamiento utilizado.
- **Integración, pruebas y resultados:** en esta sección mostraremos los resultados obtenidos de los diferentes modelos utilizados en la predicción. Incluiremos, además, gráficas que ayudarán a entender los diferentes modelos y resultados, además de, por supuesto, hacer más amena y sencilla la lectura.
- **Conclusiones y trabajo futuro:** siempre hay cosas que mejorar, continuar; más aún en el mundo del aprendizaje automático. Aquí planteamos qué camino seguir a partir de este trabajo, qué se podría mejorar y temas a los que se le podría darle una vuelta de cara a futuras consideraciones.

2 Estado del arte

2.1 Series temporales: introducción

Una serie temporal es un conjunto de observaciones x_t, cada uno de ellas en un momento de tiempo específico, t [4]. Dados estos datos, nos interesa saber los cambios de dicha variable en relación al tiempo y, una vez realizado dicho análisis, ser capaces de predecir sus valores futuros.

El uso de series temporales realmente se emplea mucho en otros aspectos no sólo relacionados con la meteorología, por ejemplo, también se usa en el medio ambiente (para predecir y observar las emisiones anuales de CO_2, por ejemplo), en el aspecto demográfico (nacimientos anuales), y, por poner un ejemplo más, en el aspecto económico (tasa de inflación o desempleo).

2.1.1 Representación

Para representar gráficamente y ver de un vistazo los datos, una serie temporal suele representarse en un gráfico temporal, representando en el eje de ordenadas (y) los valores de la serie y el tiempo en el eje de abscisas (x).

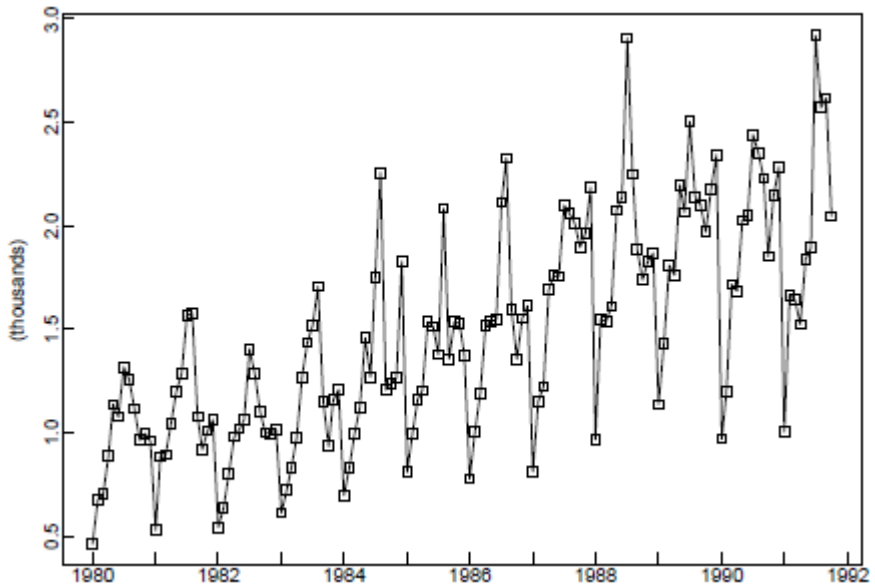


Figura 2.1: ejemplo de serie temporal de histórico de ventas de vino tinto australiano[5]

Notar que en el eje x el tiempo puede ir representado en cualquier unidad de éste: meses, días, años, horas, minutos, etc, todo en función al valor.

2.1.2 Tipos de series temporales

Siguiendo con las series temporales, procedemos ahora con su clasificación.

- Una serie es estacionaria cuando la media y/o la variabilidad de éstas se mantienen constantes durante el paso del tiempo.[6]

- Una serie es no estacionaria cuando la variabilidad y/o media varían a lo largo del tiempo. Una serie que no es estacionaria pueden mostrar cambios de varianza, mostrar una tendencia (esto es, que la media aumenta o disminuye a lo largo del tiempo), o mostrar efectos estacionales, en otras palabras, el comportamiento de la serie es similar cada cierto periodo de tiempo.

Cuando una serie es estacionaria, evidentemente realizar predicciones es más fácil, ya que la media es constante, y por ello podemos hacer una estimación de ésta con todos los datos. Un ejemplo muy simple de serie temporal estacionaria es ruido blanco (o *white noise* en inglés), donde la media y covarianza son cero.

2.1.3 Análisis

En la mayoría de casos, una serie temporal es la suma de varias componentes: la tendencia, la estacionalidad y la componente residual (o también llamada componente irregular), siguiendo la siguiente fórmula:

$$X_t = S_t + T_t + I_t$$

Siendo S_t la estacionalidad o componente estacionaria, T_t la tendencia e I_t la componente residual. La primera de ellas muestra las oscilaciones a lo largo de un período; la tendencia, un comportamiento ascendente o descendente de la media a largo plazo (no en un período únicamente); la componente residual muestra las variaciones aleatorias.

Es importante e interesante tratar de aislar estos componentes para su posterior análisis y tratamiento.

2.1.3.1 Análisis de la tendencia

Una serie temporal, en algunos casos simples, puede resultar seguir una tendencia lineal, es decir que sigue una función de la forma $T_t = a + bt$. Podemos estimar esta tendencia usando *least squares* o método de mínimos cuadrados.

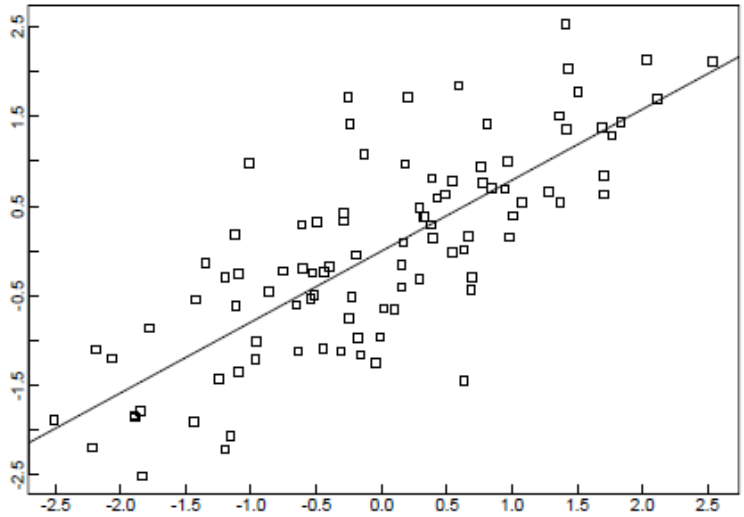


Figura 2.2: Ejemplo de estimación de tendencia por regresión por least squares

En otros casos, la tendencia de una serie temporal puede no seguir una recta sino que evoluciona a lo largo del tiempo. Por esta razón, es mejor estimar que una serie evoluciona a lo largo de un intervalo corto de tiempo. Para ello podemos usar la media móvil, que consiste en hacer la media móvil de los *n* datos anteriores. Evidentemente, cuanto mayor sea n, mayor influencia tendrán los datos más antiguos en el resultado de la estimación de la tendencia.

Otro método para estimar la tendencia de forma más general es no suponer ninguna hipótesis sobre la tendencia a corto plazo sino que suponemos simplemente que evoluciona en el tiempo. Así pues, decimos que la tendencia en el instante t es próxima a la tendencia en instante t-1, entonces tenemos que:

$$y_t = x_t - x_{t-1}$$

Diferenciar la serie equivale a asumir que la tendencia en t es el valor de la serie en el instante $t - 1$

2.1.3.2 Análisis de la estacionalidad

En una primera instancia, podemos considerar el efecto estacional de una serie temporal sobre cómo varía la media del período en relación a la media global.

Sin embargo, podemos, al igual que hemos hecho con la tendencia, no suponer ninguna hipótesis a corto plazo de la estacionalidad y sí suponer que evoluciona en el tiempo, pero lentamente.

Por ello, podemos, de nuevo, una serie diferenciada estacionalmente, siendo esta nueva serie:

$$y_t = x_t - x_{t-s}$$

Donde s es el período que marquemos a la serie temporal (por ejemplo, 365 en caso de ser anual, 12 en caso de ser mensual, ambos casos asumiendo que tenemos datos diarios; si éstos fuesen diarios, estos períodos se multiplicarían por 24), x_{t-s} es el valor de ese período en el pasado (lag).

2.1.3.3 Análisis de ruido o componente residual

No haremos especial hincapié en el análisis del ruido, ya que realmente debería seguir una cierta aleatoriedad en la serie, sin mostrar una tendencia ni estacionalidad, puesto que supone ya descompuesta y eliminada de la serie original.

2.1.3.4 Autocorrelación

La función de autocorrelación (también llamada a veces dependencia secuencial) se usa para encontrar patrones que se repiten dentro de una serie temporal, como por ejemplo, el ruido o la periodicidad.

La autocorrelación es la correlación de una serie X_t desplazada en el tiempo de esa misma serie X_t.

Esta función está comprendida entre -1 y 1, siendo 1 una autocorrelación perfecta y -1 lo contrario, siguiendo la formula

$$R(k) = \frac{E[(x_i - \mu) \cdot (x_{i-k} - \mu)]}{\sigma^2} \in [-1,1]$$

Siendo E el valor esperado y k el desplazamiento temporal que establezcamos.

Usualmente se suele definir, también, la autocovarianza, puesto que son muy similares: lo único en lo que difieren es en la constante que es la varianza[7].

2.2 Modelos a utilizar para predicción de series temporales

Habiendo explicado el análisis y proceso de descomposición de una serie temporal, vayamos ahora a utilizar distintos modelos para realizar predicción de series temporales. No cabe duda de que la tendencia y estacionalidad van a jugar un papel muy importante a la hora de la predicción y entrenamiento de los modelos; unos serán más sensibles y tratarán mejor la estacionalidad y/o tendencia, mientras que otros la ignorarán, o simplemente influirán negativamente en la predicción.

Dentro de la gran cantidad de modelos a emplear en aprendizaje automático, hemos de plantearnos que una serie temporal ha de tener un cierto orden (temporal) y que, por tanto, otros modelos no van a tener en cuenta este factor, como, por ejemplo, SVM (máquina de vectores de soporte), KNN (*k-nearest neighbours*) o una simple regresión logística. Por eso, podemos emplear modelos que sí tienen en cuenta dicha temporalidad, como son los modelos autorregresivos, modelos de media móvil y sus diferentes combinaciones, que resultan en ARMA, ARIMA y SARIMA.

2.2.1 Modelo autorregresivo (AR)

Un modelo autorregresivo establece que la variable de salida depende linealmente de sus valores previos. La notación de un modelo autorregresivo (lo denotaremos como AR a partir de ahora), viene definido como

$$AR(p) = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t$$

Donde $\phi_i \dots \phi_p$ son los parámetros del modelo, c es una constante y ε_t es ruido blanco (esto es, variables temporales aleatorias que no tienen correlación alguna, en otras palabras aleatoriedad). Así, un AR(1) es llamado como "modelo autorregresivo de primer orden", donde la variable de salida, X_t, está relacionada con únicamente un período anterior. Por tanto, siguiendo el ejemplo, en un modelo AR(2), la variable de salida estará relacionada con 2 períodos anteriores. Este modelo es un proceso estacionario, por lo que si la serie temporal no lo es, habría que hacerla estacionaria convirtiéndola a su serie de diferencias.

2.2.2 Modelo de medias móviles (MA)

También llamado moving average en inglés, es una técnica empleada para obtener una idea de las diferentes tendencias de una serie temporal. Un MA (abreviémoslo así a partir de ahora), resulta muy útil para predecir tendencias a largo plazo.

Partiendo de que la media muestral se define como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Un MA se basa en este concepto, pero la media va desplazándose por diferentes subconjuntos, y la realiza sobre el error medio. Podemos verlo como una ventana que se va desplazando y obteniendo la media sobre lo que está situado. Si en cambio tenemos un MA de 5, dicha ventana abarcaría cinco años.

Habitualmente un MA viene definido por el parámetro q, que es, justamente, el tamaño de la ventana sobre la que hacer la media. MA también es un proceso estacionario, al igual que AR.

$$MA(q) = \sum_{i=1}^{q} \phi_i \varepsilon_{t-i} + \varepsilon_t$$

Siendo ε_{t-i} son los términos de error y ε_t es un proceso de ruido blanco.

2.2.3 Modelo autorregresivo de medias móviles (ARMA)

Llamado en inglés a*utoregressive moving average model*, es un modelo que está formado por dos partes: una autorregresiva (AR), y una parte de media móvil (MA). Habitualmente el modelo se nombra de la forma:

$$ARMA\ (p, q) = \sum_{i=1}^{q} \phi_i \varepsilon_{t-i} + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i}$$

Donde p es el parámetro, u orden, que corresponde al AR y q, al MA.
Notar que el modelo ARMA es también un proceso estacionario. Es decir, que si el modelo no lo es, es necesario trabajar sobre una serie diferenciada.

2.2.4 Modelo autorregresivo integrado de medias móviles (ARIMA)

Al modelo ARMA vamos a añadirle una componente más, que es la parte integrada, que nos indica cuántas diferencias estacionarias son necesarias para lograr la estacionalidad. Si no es necesaria ninguna, simplemente es un ARMA.

La notación de un ARIMA viene definida por:

$$ARIMA(p, d, q)$$

Siendo p el orden de la autorregresión; d, el número de diferencias estacionarias requeridas; y q el número de períodos de predicción de error en la ecuación.

ARIMA funcionará bien si hay estacionariedad (media y varianza constante), de lo contrario, habrá que transformar los datos para poder usar un ARIMA.

Un modelo ARIMA(p,d,q) viene definido como:

$$ARIMA(p,d,q) = -(\Delta^d Y_t - Y_t) + \phi_0 + \sum_{i=1}^{q} \phi_i \varepsilon_{t-i} + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i}$$

Donde la primera parte, *d*, significa las "d" diferencias requeridas para hacer la serie en estacionaria, mientras que el resto de sumas corresponden a los ya vistos AR y MA. ϕ_0 *c*orresponde a una constante.

2.2.5 Modelo estacional autorregresivo integrado de medias móviles (SARIMA)

Por último, este modelo es usado cuando la serie temporal tiene estacionalidad. La notación es similar a la del ARIMA:

$$ARIMA(p,d,q)(P,D,Q)m$$

Los parámetros p,d,q corresponden a AR, a la parte integrada y a MA, respectivamente. Los parámetros P,D,Q simplemente representan a (p,d,q) pero para la parte estacional de la serie temporal. El parámetro m indica el número de períodos en cada estacionalidad.

2.2.6 Uso de variables exógenas: ARMAX, ARIMAX y SARIMAX

Una variable exógena es aquella variable que no está afectada por otras variables en un sistema. *Exo* significa "*fuera".* Por otro lado, una variable endógena es aquella que sí se ve influenciada por otros factores dentro de un sistema.[8]

Usando pues un modelo ARMA con variables exógenas (además de las endógenas), estaremos hablando de un ARMAX; si usamos un ARIMA, éste será entonces un ARIMAX, finalmente, si empleamos un SARIMA, con variables exógenas tendremos un SARIMAX.

Siguiendo algunos ejemplos anteriores: en la tasa de natalidad, una variable exógena de este ejemplo podría ser el salario percibido por los padres de un posible hijo; en el aspecto medioambiental, una variable exógena podría ser el número de volcanes entrados en erupción y la cantidad de gases emitida por estos en un período (estos aumentarán el nivel de CO2 en la atmósfera).

3 Diseño

En este trabajo vamos a utilizar distintos modelos enfocados a series temporales para realizar predicción de energía eólica. En particular, vamos a probar el modelo autorregresivo (AR), el modelo de medias móviles (MA), el modelo autorregresivo de media móvil (ARMA), el modelo autorregresivo integrado de media móvil (ARIMA), y finalmente dos modelos con que incluyen variables exógenas: ARMAX y ARIMAX.

Para ello, usaremos los datos del parque eólico de Sotavento[9], un parque eólico experimental que usa datos públicos. Contaremos con un conjunto de datos extraídos de la página web pertenecientes a los años 2013 al 2018, ambos inclusive, e incluyendo desde el día 1 de enero hasta el 31 de diciembre de cada año. El dato que se proporciona por día es la suma de los datos de las 24 horas de ese día completo.

Dado este rango de tiempo, los atributos que se facilitan son la dirección del viento (en grados), la velocidad (en km/h) y la energía producida (en KW). Realizaremos predicción eólica usando valores anteriores de producción siguiendo modelos de series temporales.

Usaremos para todo este trabajo el lenguaje de programación *Python*, un lenguaje muy completo en cuanto a paquetes y en el que el campo del aprendizaje automático tiene gran relevancia. Concretamente, nos ayudaremos de las librerías *Sklearn*[10], *Pandas*[11], *Statsmodels*[12] y *Mathplotlib*[13]

3.1 Preprocesado de los datos

Para más comodidad, cambiaremos las unidades en las que se presenta el *dataset* original por porcentaje en relación a la energía contratada por el parque. El parque tiene contratado 17.560KW/h, lo que supone en un día un máximo de 421440KW (es decir, lo multiplicamos por 24 horas) suponiendo el 100% de producción.

El *dataset* con el que contamos tiene *missing values*. Esto es importante, ya que en las series temporales dejar huecos sin valores hace que la estructura temporal se rompa. Aunque el tratamiento de valores faltantes tiene numerosas técnicas, usaremos uno de los más sencillos, ya que el número de *missing values* es muy bajo. Usaremos pues una interpolación lineal[14], que, usando esta fórmula, obtendremos los valores discretos dados otros dos valores (también discretos):

$$(y - y_1) = \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1)$$

De dicha fórmula, despejando, obtenemos que:

$$y = \frac{(x - x_1)}{(x_2 - x_1)}(y_2 - y_1) + y_1$$

3.2 Análisis de los datos

3.2.1 Descomposición

Como ya se ha visto, el análisis de una serie temporal consiste en el análisis de la descomposición de la serie en tendencia, estacionalidad y componente residual.

Así pues, descomponiendo el dataset entero obtenemos:

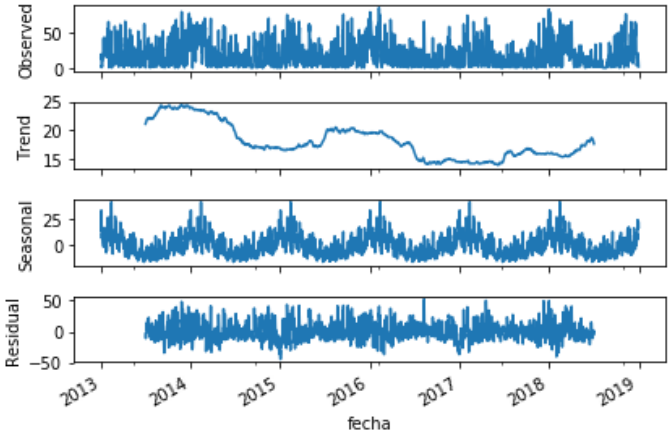


Figura 3.1: ejemplo de descomposición de una serie temporal

Y, en particular, representando gráficamente un año, pero utilizando para la descomposición el dataset entero, obtenemos:

3.2.1.1 Estacionalidad para los años 2014, 2015 y 2016

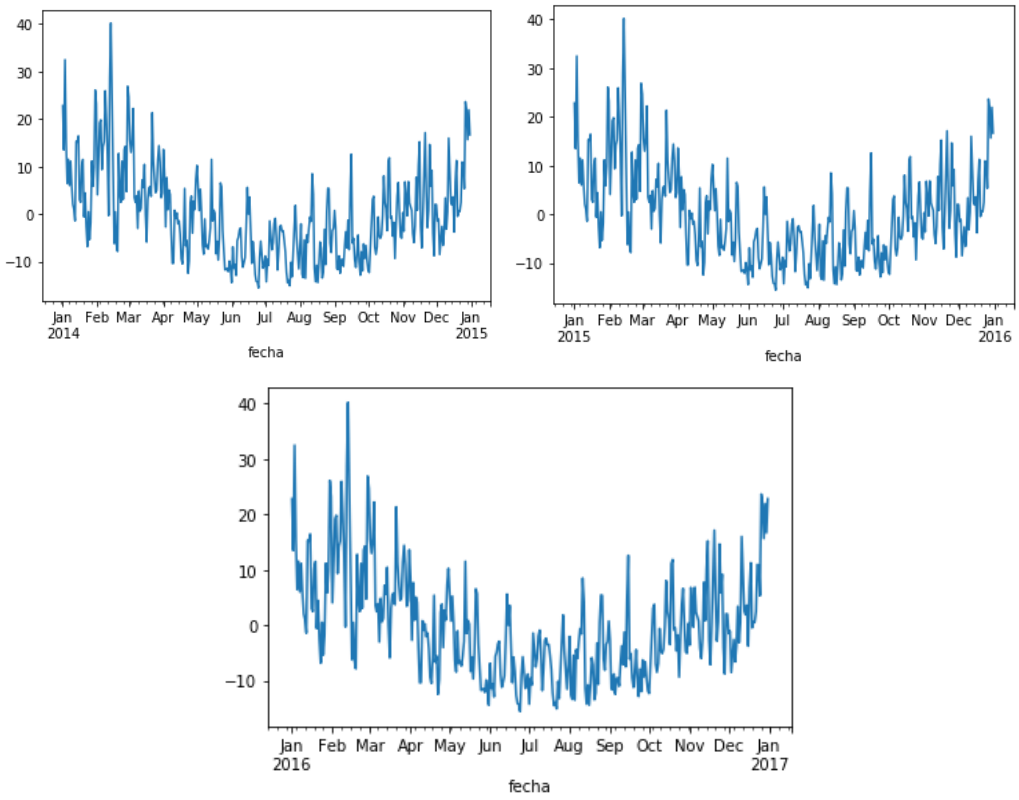


Figura 3.2 Ejemplo de estacionalidad de una serie temporal

Puede observarse que, en los primeros meses del año, la energía producida es mayor que en los meses de verano, repuntando en los últimos meses. Además, existe un máximo en el mes de febrero, patrón que se repite en todos los años. Se observa que el patrón en todos los años es el mismo, pudiendo decir que esta serie contiene una componente estacional anual.

3.2.1.2 Tendencia para los años 2014, 2015 y 2016

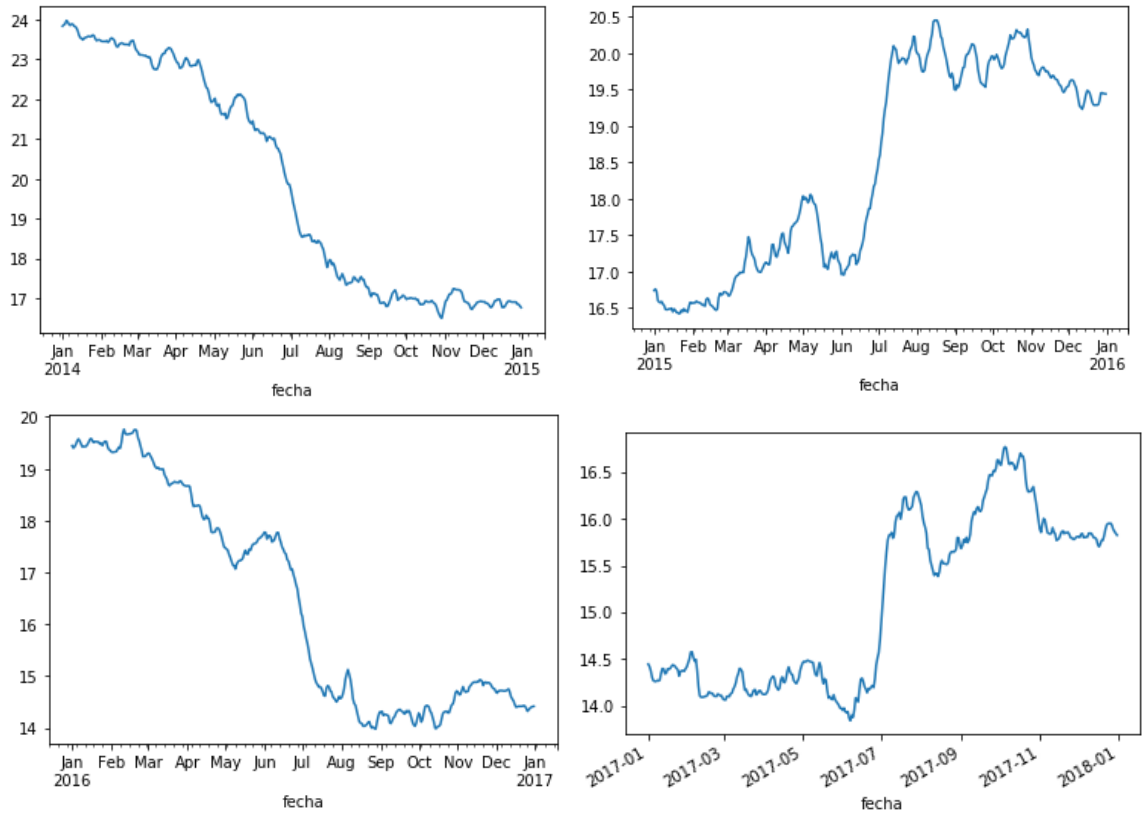


Figura 3.3 Ejemplo de tendencia de una serie temporal

Se observa que existe una tendencia que no es constante a lo largo del tiempo, puesto que, en diferentes años, la producción de energía es relativamente decreciente, en otros creciente, por lo que no se puede extraer una tendencia constante a lo largo de los años.

3.2.1.3 Componente residual para los años 2014, 2015 y 2016

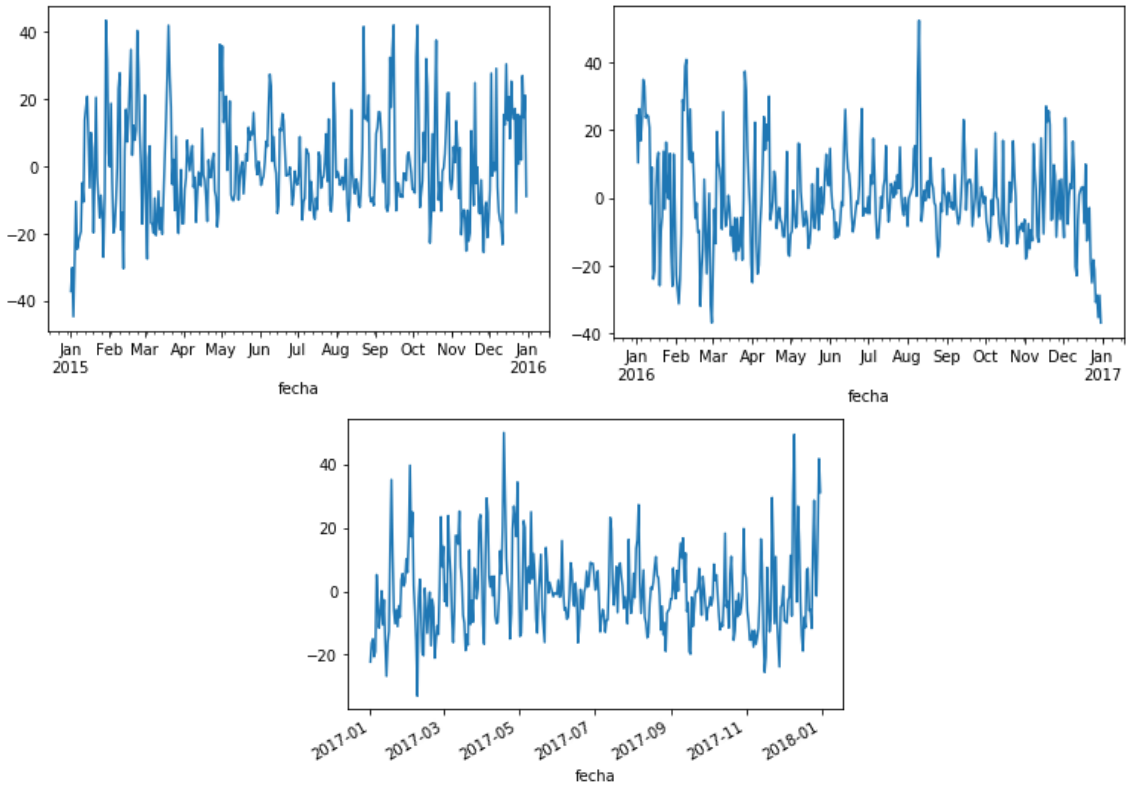


Figura 3.4 Ejemplo de componente residual de una serie temporal

Realmente, lo que se observa en la componente residual es ruido, una vez extraídas las demás componentes. Se espera que la componente residual sea irregular, ya que se ha aislado la tendencia y estacionalidad de ésta.

3.2.2 Autocorrelación

Gracias a la autocorrelación podemos encontrar patrones repetitivos en una serie temporal, como lo es la estacionalidad. Haciendo la autocorrelación de 365 retrasos (un año), vemos de nuevo que, anualmente, existen valores más altos en los primeros meses, bajando en los meses de verano y volviendo a subir en los últimos meses del año.

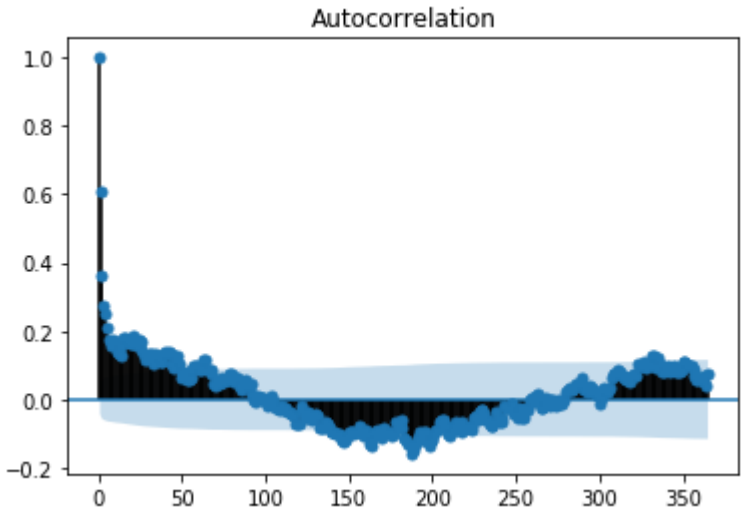


Figura 3.5 Autocorrelación de 365 períodos

Ya vista anteriormente, la función de autocorrelación está comprendida en el rango [-1, 1], siendo 1 una correlación perfecta. Como puede apreciarse, lógicamente, el primer dato del gráfico muestra una autocorrelación perfecta (de valor 1), ya que compara el valor consigo mismo.

3.3 Estimación de parámetros

Para estimar mejor los parámetros p, d y q de los modelos autoregresivos descritos anteriormente, realizaremos un proceso de train, test y validación a través de splits en el conjunto de train. Esto es, establecer un conjunto de train de longitud fija y partir el conjunto de test en varias particiones, cada una de ellas más grande que la anterior. En este caso, la longitud de test sigue la fórmula[15]

```
i * n_samples // (n_splits + 1) + n_samples % (n_splits + 1)
```

(nótese que el operador "//" indica división entera y "%", el módulo)

En otras palabras, a modo visual, se reduce en que dado un número n de splits, el tamaño de train de cada Split va a ir a aumentando mientras que el tamaño de test va a permanecer constante:

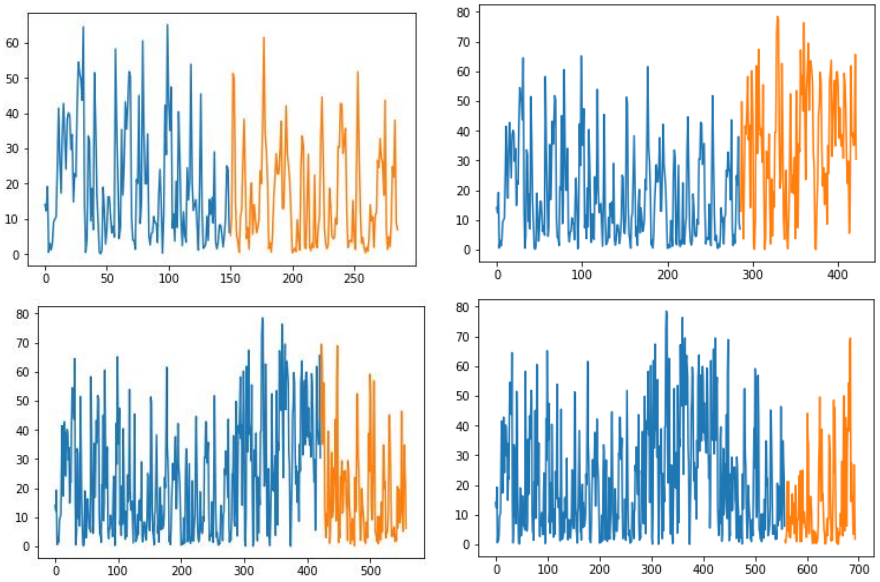


Figura 3.6 Ejemplo de conjuntos train-test de una serie temporal

Este método de validación es importante, ya que es distinto a los que nos solemos encontrar en otros modelos de aprendizaje automático, como KNN, donde el tamaño y patrones de test se escogen de manera aleatoria, sin importar el orden. En este caso, estamos trabajando con series temporales, donde el orden sí importa y no podemos permitir la aleatoriedad de los otros modelos. Además, este método de validación proporciona más robustez a la hora de la búsqueda de parámetros, ya que usa distintos conjuntos de validación.

Para hacer una estimación de los parámetros usados en los modelos de predicción autorregresivos (AR, MA, ARMA y ARIMA), realmente, en código, usaremos un modelo ARIMA, pero estableciendo los parámetros necesarios a 0 para hacer un modelo AR, MA.

En concreto, estableceremos los siguientes valores para los modelos:

$$ARIMA(p, d, q)$$
$$p = [0, 1, 3, 6, 10, 15]$$
$$d = [0, 1]$$
$$q = [0, 1, 2, 3, 5]$$

Así pues, si por ejemplo tenemos $p = 0$ y $d = 0$, realmente estaremos haciendo un modelo de medias móviles (MA), mientras que si tenemos un p = n y d, q = 0, estaremos haciendo un AR de parámetro n (siendo n un entero > 0).

Por otro lado, podemos especificar el tipo de tendencia ('nc' o 'c' siendo una tendencia no constante y constante, respectivamente) y el tipo de predicción ('linear', que realiza una predicción lineal diferenciando las variables endógenas — la producción —, mientras que 'levels', realiza una predicción sin diferenciar los datos) cuando d > 0[16].

3.4 Medida del error

En cuanto a la medida del error, usaremos el error cuadrático medio (MSE, o *mean square error* en inglés). La razón por la que usamos esta medida de error es porque es dependiente de la escala, a diferencia de otros que no lo son, como el MAPE. Además, con MAPE no

puede emplearse con porcentajes, medida que usamos en la producción al ir en proporción a la potencia instalada del parque.

La fórmula[17] que sigue el error cuadrático medio es:

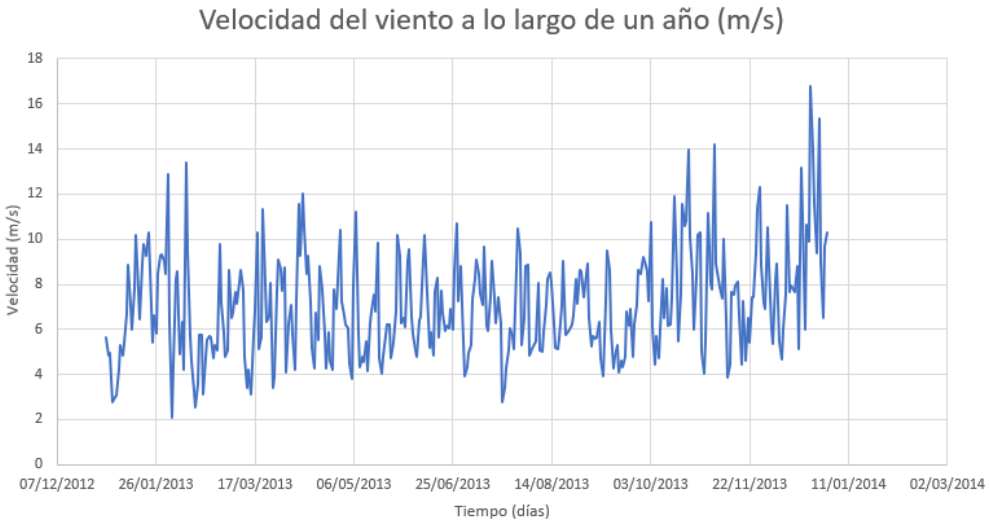
$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{Y}_i - Y_i\right)^2$$

Donde $\hat{Y}_i$ es un array o vector de valores predichos y Y_i es un vector de los valores originales. A menor valor del MSE, menor error y por tanto mejores resultados obtenidos en la predicción.

Una vez realizadas las etapas de entrenamiento, test y validación con AR, MA, ARMA y ARIMA, procederemos a quitar la estacionalidad de las particiones de datos para quitar la componente estacionaria de la serie temporal. Con diferenciación nos referimos a que dado un valor en un instante t, le restamos el valor del instante t-365, que, en nuestro caso, es un año anterior. Por esta razón, el conjunto de datos va a verse reducido ya que el año 2013 va a quedar inutilizable, puesto que no hay t-365 con la que aplicar la resta respecto de este año. Así, realmente estamos haciendo pruebas para los modelos autorregresivos mencionados con la serie temporal diferenciada y sin diferenciar para evaluar el comportamiento e influencia de la estacionalidad.

3.5 Uso de variables exógenas

Posteriormente, y para finalizar, realizaremos las pruebas con dos modelos más: ARMAX y ARIMAX, que son los modelos ARMA y ARIMA (respectivamente) con variables exógenas. Del conjunto de datos, disponemos, aparte de la producción de energía, que utilizamos siempre como variables endógenas (esto es, variables dentro del modelo en sí), de la dirección del viento, en grados, y la velocidad de éste, en metros/segundo.



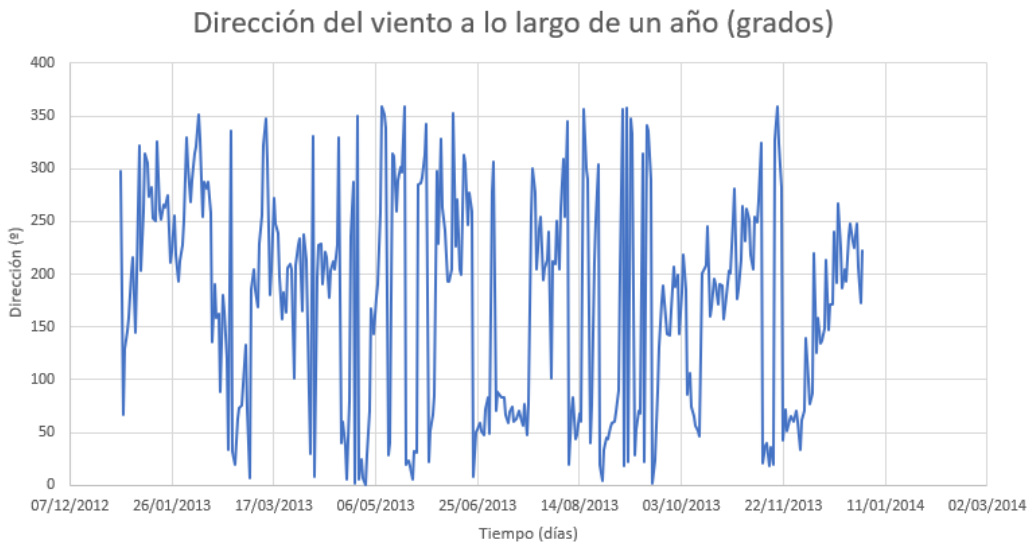


Figura 3.7 Representación gráfica de la velocidad y viento de nuestro conjunto de datos

Se observa que la variable dirección no tiene una correlación con la producción, por lo que nos decantaremos por usar como variable exógena la variable de la velocidad del viento. Para tener una relación más coherente en cuanto a la proporcionalidad con la variable endógena, la normalizaremos, como también hicimos con la producción, solo que esta vez cogemos el máximo del conjunto de datos que obtenemos y hacemos una proporción.

Sea pues el máximo de la velocidad del conjunto de datos: max($velocidad$)

$$Velocidad_{normalizada}\ (t) = \frac{velocidad(t) * 100}{máx\ (velocidad)}$$

En nuestro conjunto de datos, la máxima velocidad alcanzada durante los años 2013 y 2018 es 220 m/s, por lo que este valor será considerado el 100%. Si, por ejemplo, tenemos una velocidad de 50 m/s, normalizada quedará que

$$Velocidad_{normalizada(t)} = \frac{50\ * 100}{220} = 22{,}727\%$$

Representando gráficamente la variable exógena y endógena, podemos ver que están muy correlacionadas:

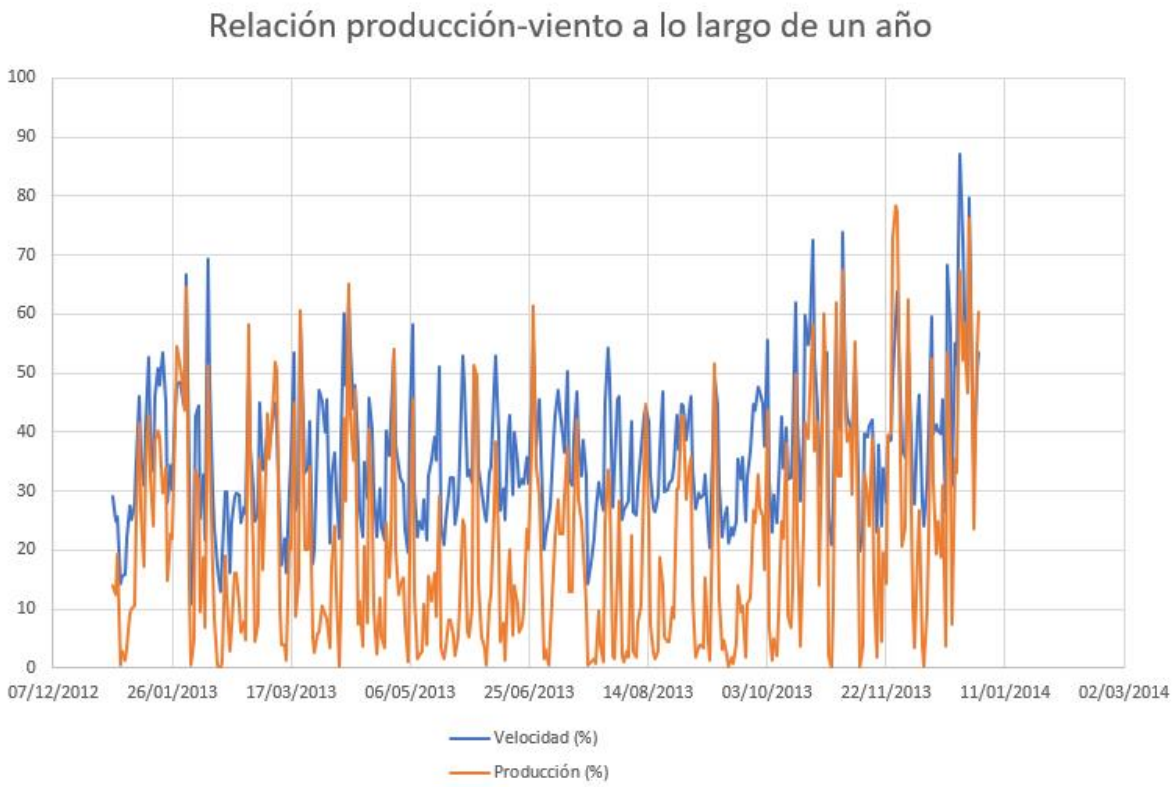


Figura 3.8 Relación gráfica de la producción y viento

Cabe destacar, además, de que los datos de velocidad son datos reales, a diferencia de los datos de la producción que, en el Parque de Sotavento, son simulados. Por esta razón, como aliciente, se espera que el error en las predicciones disminuya bastante. Esto no se da en predicción de energía eólica, puesto que los datos del viento también son predicciones, no disminuyendo en tanta medida el MSE como sí ocurre con datos reales de viento.

4 Integración, pruebas y resultados

4.1 Validación sin serie estacional diferenciada

Los resultados obtenidos para los distintos modelos autorregresivos son los siguientes. Nótese que el número o los números entre paréntesis corresponden al valor o valores que mejor resultado han obtenido en la validación.

Sin aplicar diferenciación para la componente estacionaria		
	Horizonte 0	Horizonte 10
AR (10)	1,0183	7,8916
MA (5)	1,51005	7,8529
ARMA (10,2)	0,2869	1,9393
ARIMA (10,1,1)	0,1561	0,5210
ARMAX (10,2)	0,1867	0,5187
ARIMAX (6,1,1)	0,1012	0,4971

Tabla 4.1 Resultados de la validación para modelos autorregresivos sin aplicar diferenciación

Podemos observar que el modelo de medias móviles es el que menos aporta al resto de modelos en general, ya que es el mayor error obtiene en relación al resto, pero en combinación con la componente autorregresiva, el resultado mejora hasta caer a un 0,2869 de error MSE. Finalmente, considerando la componente de tendencia, el resultado es algo mejor, con un 0,1561 de error, debido a que hemos establecido en el modelo un comportamiento de la tendencia específico, siendo no constante y no lineal (por 'levels' o intervalos). Así, el modelo realiza predicciones en términos del nivel original y no en términos de la serie diferenciada (como hace por defecto 'linear'). Notar, además, que d > 1 no está completamente soportado en la API de statsmodels[18], por lo que hemos acotado d a 0 y 1.

4.2 Validación con serie estacional diferenciada

Aplicando diferenciación para la componente estacionaria de los datos, obtenemos una mejora en AR y MA, pero empeora ligeramente en ARMA y ARIMA.

Aplicando diferenciación para la componente estacionaria		
	Horizonte 0	Horizonte 10
AR (15)	0,7394	2,1304
MA (2)	1,2076	3,1050
ARMA (10,1)	0,7010	2,1241
ARIMA (6,1,1)	0,3195	0,4002
ARMAX(10,2)	0,2538	0,6221
ARIMAX(6,1,1)	0,2398	0,5987

Tabla 4.2 Resultados de la validación para modelos autorregresivos aplicando diferenciación

Notar que el MSE es el error realizado entre la serie no diferenciada, es decir, revertida, puesto que las medidas no son las mismas (y se ven influenciadas por periodos pasados) y que, además, muchas veces son negativas influyendo en el MSE.

4.3 Validación: horizontes y resultados generales

Veamos ahora la evolución del MSE conforme a los horizontes predichos (en nuestro caso, diez). Es importante destacar que es esperable notar una peor predicción cuanto más alejado estemos del valor de origen (cero). El modelo, finalmente, tiende a hacer una media de los datos y por esta razón, se estanca, teniendo en consecuencia un MSE constante en comparación con los horizontes anteriores para ese modelo.

Este comportamiento puede observarse fácilmente en las gráficas de más abajo, donde en cualquier caso el error permanece constante cuanto mayor es el horizonte.

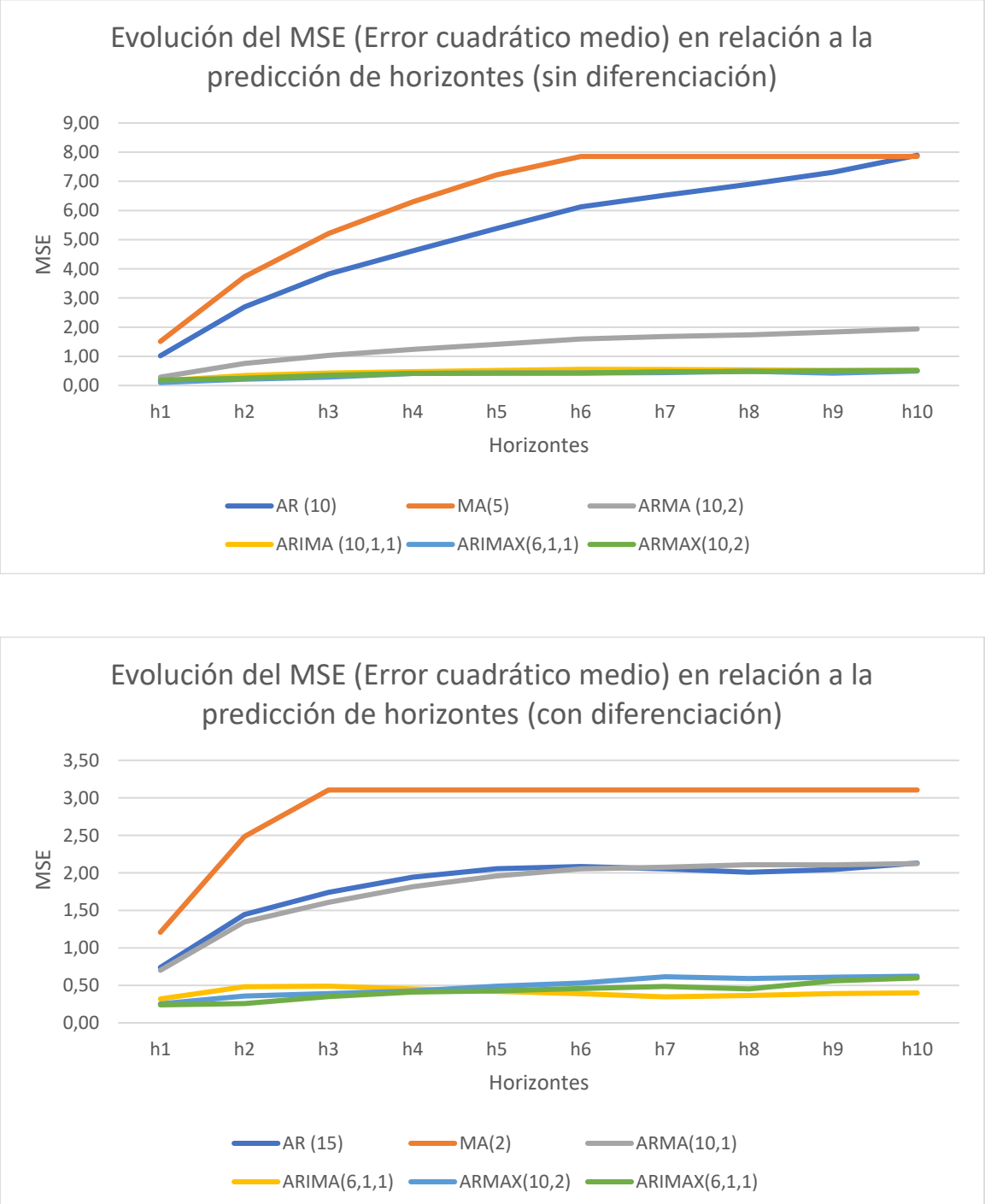


Figura 4.1 Evolución del MSE en relación a la predicción de horizontes

Vemos ahora los distintos errores para los modelos probados para los distintos parámetros (p, d, q, en caso de que existan en cada uno de los modelos).

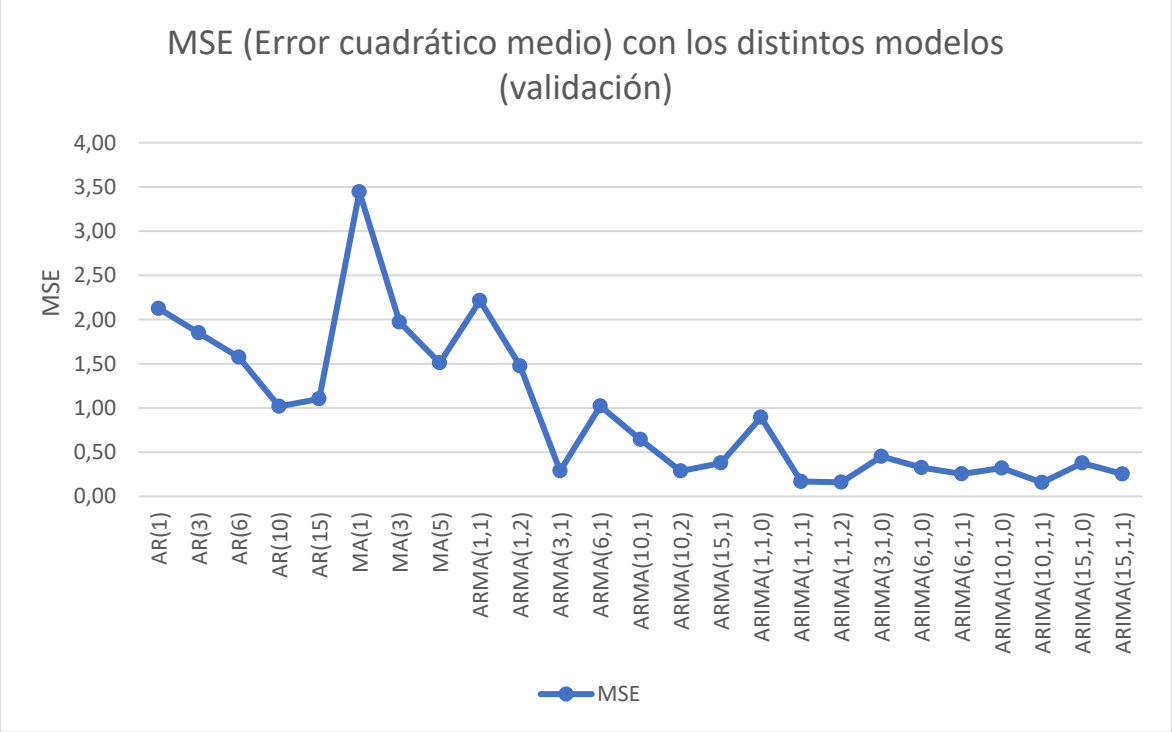


Figura 4.2 Representación gráfica del MSE acorde a los distintos modelos

Nótese que en esta tabla no se incluyen los modelos para todas las combinaciones de los valores de los parámetros indicados anteriormente (p, d, q). En algunos casos, la API de statsmodels ha lanzado una excepción con un mensaje de error, del tipo "*HessianInversionWarning: Inverting hessian failed, no bse or cov_params available"* o "*ConvergenceWarning: Maximum Likelihood optimization failed to converge. Check mle_retvals*", lo que viene siendo errores en el cálculo de la máxima verosimilitud de los datos. En cualquier caso, no es algo preocupante por la cantidad de valores y combinaciones para los que hemos probado p, d y q.

4.4 Mejora significativa con el uso de variables exógenas

Por último, los resultados de los modelos ARMAX y ARIMAX (esto es, ARMA y ARIMA con variables exógenas), muestran una clara mejora en las predicciones, ya que, como se dijo anteriormente, hay una muy clara correlación entre la velocidad del viento y la producción de energía, aportando mucho valor a la predicción.

Podemos concluir pues que el mejor modelo seleccionado es un ARIMAX(6,1,1), seguido de un ARMAX(10,2). Sin utilizar variables exógenas, el mejor modelo escogido es un ARIMA(10,1,1), sin emplear la serie diferenciada estacional en ninguno de los tres modelos mencionados.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

Es muy frecuente usar modelos de predicción como son la máquina de soporte de vectores o redes neuronales. Sin embargo, a veces no se aprovecha el factor temporal que tienen las series temporales, que es, justamente, algo importante que puede dar muchas pistas a la hora de realizar una mejor predicción. Por eso, es importante adecuarse a esta particularidad y en este trabajo se ha estudiado la familia de modelos más sencillos: los modelos autorregresivos.

Por el contrario, los modelos autorregresivos no son tan usados como por ejemplo, las redes neuronales, por lo que la implementación de éstos todavía no está totalmente completada, tiene casos en los que para algunos parámetros el modelo no está soportado totalmente, o bien la documentación es escasa. Este ha sido uno de los problemas principales de este trabajo, pues el comportamiento del modelo no era el esperado usando la sintaxis y código que son habituales en *Python* para otros modelos, como KNN, regresión logística y otros modelos incluidos en la librería *sklearn*.

5.2 Trabajo futuro

De cara al futuro, sería interesante continuar avanzando en más modelos autorregresivos. El siguiente que vendría inmediatamente después a los que hemos usado sería el modelo SARIMA o SARIMAX (éste último si utilizamos las variable exógena del viento), que trata la estacionalidad mejor que ARIMA, no teniendo que hacer una serie diferenciada. Sin embargo, nos hemos encontrado con el problema descrito anteriormente: la API de *statsmodels* no está lo suficientemente bien documentada para hacer funcionar el modelo correctamente. Hay que considerar, además, que el período que nosotros usamos es alto (365) y que el conjunto de datos es relativamente grande, aumentando así el coste computacional.

Justamente, otro enfoque que se puede dar de cara a la implementación de modelos autorregresivos es utilizar otro lenguaje de programación. Python es un lenguaje interpretado y es lento, pero la gran cantidad de paquetes, su legibilidad y su amplio uso hace que tenga una gran importancia. Es probable que, por la lentitud, *Python* pueda tener problemas si estos modelos se aplican en predicción de energía eólica en un sistema de producción, donde los valores de la energía predicha deben tardar un tiempo muy bajo en predecirse, de lo contrario no se consideran válidos para el cliente. Por ello, podría considerarse implementar los modelos en C, bien desde cero, o bien usando un paquete de utilidades para la regresión, diferenciación y demás, haciendo más llevadera la parte de cálculos matemáticos que lleva a cabo el modelo.

Referencias

[1]: Niharika G. Maity: Machine learning for improved diagnosis and prognosis in healthcare:
https://ieeexplore.ieee.org/document/7943950

[2]: Machine learning can boost the value of wind energy. Carl Elkin: https://deepmind.com/blog/machine-learning-can-boost-value-wind-energy/

[3]: Support vector machine-based short-term wind power forecasting, Jianwu Zeng: https://ieeexplore.ieee.org/document/5772573

[4]: Peter J. Brockwell, Richard A. Davis: Introduction to time series and forecasting, second Edition, capítulo 1.

[5]: Peter J. Brockwell, Richard A. Davis: Introduction to time series and forecasting, second Edition, capítulo 1.

[6]: Peter J. Brockwell, Richard A. Davis: Introduction to time series and forecasting, second Edition, sección 1.4.

[7]: De la Horra, Julián.: Estadística aplicada. Díaz de Santos, 2003

[8]: Jason Brownlee: Taxonomy of Time Series Forecasting Problems: https://machinelearningmastery.com/taxonomy-of-time-series-forecasting-problems/

[9] Página web del parque eólico Sotavento: http://www.sotaventogalicia.com/es/datos-tiempo-real/historicos

[10]: Scikit-learn: machine Learning in Python:
https://scikit-learn.org/

[11]: Pandas: Python Data Analysis Library:
https://pandas.pydata.org

[12]: Statsmodels: statistics in Python
https://www.statsmodels.org/

[13]: Matplotlib: Python plotting
https://matplotlib.org/

[14]: Interpolation methods, Paul Bourke:
http://paulbourke.net/miscellaneous/interpolation/

[15]: Sklearn API: TimeSeriesSplit: https://scikitlearn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

[16]: StatsModels ARIMA API: https://www.statsmodels.org/devel/generated/statsmodels.tsa.arima_model.ARIMA.predict.html

[17]: Mean square error: definition and examples, *Bob Bruner*: https://study.com/academy/lesson/estimation-of-r-squared-variance-of-epsilon-definition-examples.html

[18]: ARIMA model issue: https://github.com/statsmodels/statsmodels/issues/4047