

Modelado y predicción clásico de series temporales aplicado a epidemiología en infecciones

Francisco Gallego Perona

10 de agosto de 2018

Índice

1. Introducción y Estado del arte	6
1.1. Contexto y objetivo general del trabajo	6
1.2. Tipos de procesos estudiados	8
1.2.1. Procesos Autorregresivos	8
1.2.2. Procesos de Media Móvil	8
1.2.3. Procesos Mixtos	9
1.2.4. Estacionalidad en ARIMA	9
1.3. Estado del arte	9
2. Análisis de objetivos y metodología	10
2.1. Análisis de objetivos	10
2.2. Metodología	12
2.2.1. Estudio preliminar de la serie	12
2.2.2. Estudio de Estacionariedad de la serie	12
2.2.3. Aplicación y estudio de otros modelos más simples	13
2.2.4. Filtro de Modelos	13
2.2.5. Elección de los mejores modelos para 3, 6, y 13 meses	15
2.3. Herramientas utilizadas	15
3. Diseño y resolución del trabajo realizado para la serie MARSA	16
3.1. Estudio preliminar de las series	16
3.1.1. Representación de la serie	16
3.1.2. Test Avanzado de Dickey-Fuller para estudio de la estacionariedad .	20
3.1.3. Descomposición de la serie	21
3.1.4. Análisis de ACF y PACF	22
3.2. Aplicación de métodos más simples	24
3.2.1. Aplicación del método Naïve	24
3.2.2. Aplicación del método de Media Simple	25

3.2.3. Aplicación de Suavizado exponencial con método Holt-Winters	26
3.3. Generación de modelos ARIMA	29
3.3.1. Filtrado por RMSE comparando con el mejor modelo Holt-Winters	30
3.3.2. Filtrado por Criterio de Ljung-Box	30
3.3.3. Filtrado mediante test de invertibilidad	31
3.3.4. Filtrado de modelos por Parsimonia	33
3.3.5. Filtrado de modelos por Complejidad	34
3.3.6. Último filtrado de modelos y comprobación de predicciones	35
4. Resolución para las series <i>Staphylococcus Aureus</i> y <i>Levofloxacino</i>	45
4.1. <i>Staphylococcus Aureus</i>	45
4.2. <i>Levofloxacino</i>	51
5. Conclusiones y vías futuras	56
6. Bibliografía	57

1. Introducción y Estado del arte

1.1. Contexto y objetivo general del trabajo

Según la Organización Mundial de la Salud (OMS), la resistencia a los antibióticos es una de las mayores amenazas para la salud a nivel mundial. Uno de los objetivos importantes en la lucha contra este tipo de amenaza es la optimización del uso de medicamentos.

Como respuesta a este problema, los organismos internacionales y los responsables de los servicios de salud definen protocolos de aplicación que permitan mantener la eficacia de los antibióticos, y evitar o limitar la aparición de resistencias bacterianas debidas a un consumo inadecuado. Estos protocolos se establecen en los programas de uso racional de los antibióticos, conocidos como Antimicrobial Stewardship Program (ASP). Un equipo ASP es un grupo interdisciplinar de clínicos del hospital (intensivistas, microbiólogos, farmacia hospitalaria y otros especialistas médicos) que definen acciones a llevar a cabo en las distintas dimensiones del problema. La aplicación efectiva de los ASP requiere de las utilidades que pueden proporcionar las TIC para facilitar el acceso a los datos relevantes en la toma de decisiones y el trabajo colaborativo.

Así, una de las utilidades que las TIC, y en especial la Inteligencia Artificial, pueden proporcionar a los clínicos es el desarrollo de modelos predictivos para la predicción de comportamientos futuros a partir los datos recolectados en los centros.

El grupo al que pertenecen los tutores de este trabajo desarrolla una línea de actividad en este ámbito. En los últimos años el grupo ha desarrollado e implantado en el Hospital Universitario de Getafe una plataforma inteligente, WASPSS, para ayudar en la gestión del ASP en el hospital, abordando los procesos clínicos implicados de forma integrada. En concreto, han abordado el desarrollo de técnicas y herramientas para: un modelo de inteligencia de negocio para ASP, el análisis de series temporales de datos, el apoyo a la decisión del tratamiento antibiótico y el soporte a la aplicación de guías clínicas.

El trabajo objeto de esta memoria se encuadra en esta línea de actividad, estando focalizado sobre el modelado y predicción de series temporales. En particular, se manejan series de datos de infecciones bacterianas detectadas a lo largo de un período de nueve años y series de administración de antibióticos en dicho período.

Como resultado del estudio se obtendrá un conjunto de modelos matemáticos útiles para planificar la aplicación de antibióticos en el centro hospitalario.

Este trabajo pertenece al área del Aprendizaje Computacional. Los modelos basados en análisis de series temporales reservan una parte de los datos pasados como conjunto de entrenamiento para construir el modelo. Los restantes datos hasta el final de la serie son usados como conjunto de test para evaluar la precisión de las predicciones. Una vez obtenido un modelo fiable puede ser aplicado a la predicción de valores futuros.

En particular, desarrollamos y comparamos modelos Naïve, modelos de Media Simple, Modelos de Suavizado Exponencial con el método Holt-Winters y modelos ARIMA.

Los resultados de este estudio serán útiles para compararlos con los de otros modelos basados en Redes Neuronales y Máquinas de Vector Soporte. Además, constituyen un paso previo necesario para el estudio de las relaciones causales entre series, que siempre parten del modelado ARIMA de las series componentes.

Todas las series estudiadas tienen la misma longitud; los datos abarcan desde enero de 2009 hasta enero de 2018, en total 109 meses. Cada punto de cada serie corresponde a un mes y contiene un recuento de los datos recolectados a lo largo de ese mes. Estas series son:

Incidencia de *Staphylococcus Aureus* El *Staphylococcus Aureus* es una bacteria que se aloja en el ser humano de manera habitual y que no tiene efectos negativos en personas sanas. Provoca infecciones nosocomiales o adquiridas en el recinto de atención a la salud por diversas causas. Esta cepa es resistente a los tratamientos utilizados para otras bacterias, por lo que es interesante realizar un estudio de cómo varía su incidencia en el tiempo y ver si podemos obtener un modelo que genere buenas predicciones en períodos de tiempo razonables.

La incidencia se puede expresar de la siguiente forma:

$$\text{Incidencia} = \frac{\text{Incidencias Nuevas (encontradas)}}{\text{Total Poblacional}}$$

Incidencia de MARSA MARSA es un tipo de estafilococo resistente a la meticilina. Es interesante estudiarlo por separado respecto a al *Staphylococcus Aureus*, ya que se trata de un tipo concreto que no mejora con el uso de antibióticos que normalmente curan las infecciones por estafilococos.

DOT de Levofloxacino El levofloxacino es un antibiótico efectivo para el tratamiento de un gran número de bacterias. Se considera un antibiótico de amplio espectro, por lo que se aplica en infecciones comunes antes de elegir un antibiótico de espectro más específico.

DOT, o días de terapia se refiere al número de días totales que el paciente recibe el tratamiento independientemente de las dosis suministradas. Como veremos en este estudio, se trata de una serie con estacionalidad bastante marcada en comparación a las otras dos series estudiadas, ya que interviene el personal médico en el proceso de generación de la serie.

1.2. Tipos de procesos estudiados

1.2.1. Procesos Autorregresivos

Un modelo ARIMA (autoregressive integrated moving average) permite expresar cualquier valor de la variable de una serie temporal como una combinación lineal de sus propios valores pasados. Llamamos proceso ARIMA al mecanismo de población que genera a la serie temporal. El modelo se construye a partir de una muestra de datos temporales y representa al verdadero proceso subyacente, que es inobservable. Si el modelo es una buena aproximación del proceso, tiende a imitar el comportamiento de éste.

Los procesos autorregresivos (a partir de este momento los nombraremos como AR) son los procesos ARIMA más simples, junto a los de media móvil.

Un proceso AR tiene la siguiente forma:

$$z_t = C + \phi_1 z_{t-1} + \phi_2 z_{t-2} + \dots + \phi_k z_{t-k} + a_t$$

Donde z_t es el valor actual de la serie, que es dependiente de valores anteriores z_{t-1}, z_{t-2} y sucesivos, multiplicados por el llamado término autorregresivo coeficiente AR ($\phi_1, \phi_2, \dots, \phi_k$). Por otra parte, C es una constante relacionada con la media de la serie y los coeficientes AR.

La variable a_t representa una perturbación aleatoria.

Cuanto mayor es el orden de un proceso AR, más lejanos en el tiempo son los valores anteriores de autorregresión que influyen en el valor actual.

Los procesos AR teóricos tienen las siguientes características:

- El ACF teórico decae exponencialmente o con un ligero decaimiento sinusoidal.
- El PACF teórico tiene picos en los primeros p lags y luego cae a 0.

1.2.2. Procesos de Media Móvil

Los procesos de media móvil (MA) siguen una estructura similar a los procesos autorregresivos. A continuación mostramos un ejemplo de proceso de media móvil de orden 2 o MA(2):

$$z_t = C - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_k a_{t-k} + a_t$$

Como vemos, ahora los valores de z_t dependen de $a_{t-1}, a_{t-2}, \dots, a_{t-k}$, las perturbaciones aleatorias ocurridas en instantes de tiempo anteriores.

En los procesos MA teóricos ACF y PACF tienen comportamientos invertidos respecto a los procesos AR:

- El ACF teórico tiene picos en los primeros q lags y luego cae a 0.
- El PACF teórico decae exponencialmente o con un ligero decaimiento sinusoidal.

1.2.3. Procesos Mixtos

Los procesos mixtos tienen la siguiente forma:

$$z_t = C + \phi_1 z_{t-1} - \theta_1 a_{t-1} + a_t$$

El ejemplo anterior es un proceso con componente AR(1) y componente MA(1) y se expresa como ARMA(1, 1).

Estos procesos son más complejos y explorar gráficas de autocorrelación no es tan sencillo. Por ello se utilizan otras técnicas como la aplicación de la EACF (Extended Autocorrelation function), aunque en este estudio no se va a hacer uso de ella.

En un proceso mixto, decaen tanto ACF como PACF.

En un proceso ARIMA o ARMA integrado, se modela la dinámica de la serie diferenciada. Es decir, se modela la diferencia de los dos últimos valores de la serie en función de las diferencias anteriores. El orden de un modelo ARIMA se expresa como una terna (p, d, q) , donde p y q son los órdenes de los procesos AR y MA, respectivamente, y d es el orden de diferenciación que definiremos más adelante. Información más detallada se puede encontrar en [1].

1.2.4. Estacionalidad en ARIMA

Los modelos ARIMA también pueden presentar estacionalidad. Un modelo con estacionalidad básico y muy común en la práctica es el $ARIMA(0, 1, 1)_s$, que tiene las mismas componentes no estacionales que el modelo no estacional $ARIMA(0, 1, 1)$, pero la componente MA(1) no se detecta en el lag 1 como sucedería al ser no estacional, sino que se presenta en el lag s . De igual manera ocurriría con un modelo que presente componente AR estacional.

Esto hace más complejo el estudio de las funciones de autocorrelación, al tener que disociar la parte estacional de la parte no estacional.

El ejemplo anterior quedaría de la siguiente manera:

$$z_t = z_{t-s} + C - \phi_s a_{t-s} + a_t$$

Donde z_{t-s} es el valor de la serie para el instante de tiempo $t-s$ y $\phi_s a_{t-s}$ es el coeficiente MA multiplicado por el valor de media móvil en el instante $t-s$.

1.3. Estado del arte

Uno de los modelos predictivos más extendido hoy en día son las redes neuronales. Son modelos matemáticos basados en las interconexiones del cerebro y permiten una relación no lineal entre la variable a predecir y los predictores.

Este tipo de modelo es bastante potente, sobre todo si nos referimos a su versión LSTM, suficientemente potente como para aprender comportamientos pasados de la serie son importantes en la predicción y cuáles no.

Para el médico experto, los modelos de aprendizaje modernos son como cajas negras, con una entrada y una salida, pero poco explicativos o nada en la práctica, sin embargo ARIMA es fácil de expresar y de entender y ha sido utilizado desde hace bastantes años, por lo que los médicos están más familiarizados con este tipo de modelos. Cada vez se va abriendo el camino a los modelos modernos, ya que presentan una muy buena capacidad de predicción.

Un detalle a tener en cuenta es que es un cambio total en la forma de ver el problema, transformándolo en un problema de aprendizaje supervisado separando una entrada (X) y una salida (Y).

Otro modelado bastante extendido es el modelado bayesiano. Merece la pena hacer mención de los modelos bayesianos aplicados a series temporales cortas. Para detectar la estacionalidad en los modelos ARIMA necesitaremos una cantidad de datos mínima, como veremos en el estudio que abarca este documento. Un tipo de modelo bayesiano es el BSTS (Bayesian structural time series) muy utilizado en econometría.

2. Análisis de objetivos y metodología

2.1. Análisis de objetivos

El objetivo del estudio es el de ofrecer al personal médico un conjunto de modelos matemáticos con capacidad predictiva, útiles para tomar decisiones en la aplicación de antibióticos para tratamiento de infecciones contraídas en el centro hospitalario.

En base a la predicción obtenida, el médico podrá estudiar de una forma visual y cuantitativa la evolución y tendencia de los datos de incidencia de las bacterias MARSA y Staphylococcus Aureus, en conjunción con los volúmenes mensuales de aplicación del antibiótico Levofloxacino.

El trabajo se encuadra dentro del área del Aprendizaje Computacional. Los modelos basados en análisis de series temporales reservan una parte de los datos pasados como conjunto de entrenamiento para construir el modelo. Los restantes datos hasta el final de la serie son usados como conjunto de test para evaluar la fiabilidad de las predicciones. Una vez obtenido un modelo fiable puede ser aplicado a la predicción de valores futuros.

En particular, un primer objetivo concreto es la construcción de modelos ARIMA para las tres series indicadas. Usar modelos ARIMA en este dominio es aconsejable puesto que este tipo de modelos son bien aceptados en Microbiología y son relativamente fáciles de

interpretar por el usuario final, a diferencia de otras técnicas de Inteligencia Artificial, como Redes Neuronales o Máquinas de Vector Soporte. Además, dentro de los modelos lineales, los ARIMA teóricamente proporcionan las mejores predicciones que se pueden obtener realizando un promedio ponderado de los valores pasados de una serie. En este contexto, mejor significa que minimiza el error cuadrático medio en las predicciones.

Un segundo objetivo es la construcción de modelos más simples que el ARIMA para comparar su fiabilidad predictiva con el primero. Obtendremos modelos Naïve, modelos de Media Simple y Modelos de Suavizado Exponencial con el método Holt-Winters. Estas comparaciones servirán para discutir la necesidad de aplicar modelos de predicción más complejos o usar estos modelos más simples. Así pretendemos comprobar experimentalmente en qué medida se cumplen las expectativas teóricas sobre el modelo ARIMA como el mejor predictor. En particular, queremos saber si en este dominio concreto basta usar métodos simples para obtener predicciones útiles para el clínico y razonablemente precisas. Ello es importante porque la metodología Box-Jenkins de construcción de modelos ARIMA es relativamente compleja, costosa en tiempo y difícil de automatizar, dependiendo en buena medida de juicios visuales subjetivos del analista.

Un tercer objetivo es contribuir a la automatización del proceso de construcción de modelos ARIMA, combinando técnicas de fuerza bruta con poda basada en los siguientes criterios:

- Criterio de capacidad de predicción usando el RMSE (Root Mean Square Error o Error Cuadrático Medio).
- Criterio de complejidad con los valores de AIC (Criterio de información de Akaike) de los modelos.
- Criterio basado en la autocorrelación del modelo usando las gráficas de ACF (Función de autocorrelación) y PACF (Función de autocorrelación parcial).
- Criterio de invertibilidad y estacionariedad.

Al comienzo del estudio, haciendo uso de un algoritmo de fuerza bruta, obtendremos una batería de modelos, que en posteriores pasos iremos filtrando siguiendo los criterios anteriormente comentados.

El estudio será totalmente univariable, no tendremos en cuenta la influencia de unas series en otras series para estudiar los resultados obtenidos en las pruebas.

El uso de covariables requiere técnicas que exceden en mucho el alcance de este trabajo, como la construcción de funciones de transferencia basadas en regresión dinámica o los modelos VAR.

2.2. Metodología

A continuación explico, de manera resumida, los conceptos básicos necesarios para entender las etapas del proceso de experimentación y el análisis llevado a cabo.

2.2.1. Estudio preliminar de la serie

En primer lugar, realizaremos un estudio gráfico de la serie, viendo si hay patrones que se repitan a lo largo del tiempo, posibles tendencias o presencia de estacionalidad.

También veremos si existen valores atípicos (outliers) en los datos. Estudiaremos si existen variaciones en la media y la varianza que hagan necesarias transformaciones matemáticas para estabilizar la serie de cara a la obtención de buenos modelos ARIMA.

Aplicaremos una descomposición en las componentes principales y se representarán los residuos de la serie. Por último, haremos un estudio de las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) para ver posibles correlaciones presentes en los datos. Estas correlaciones, como veremos en posteriores secciones, podrán ser no estacionales o estacionales.

2.2.2. Estudio de Estacionariedad de la serie

Para que una serie sea estacionaria (entendemos estacionariedad como estacionariedad débil) debe tener media, varianza y función de autocorrelación constante en el tiempo. Para ello haremos uso de la media móvil y de la desviación típica móvil, representando ambas para ver si es necesaria alguna transformación.

En caso de que la media no sea constante en el tiempo aplicaremos un tipo de transformación llamada *diferenciado regular*, que consiste en lo siguiente:

$$w_t = z_t - z_{t-1}$$

El nuevo valor en el instante de tiempo t se calcula como el valor original en dicho instante menos el valor anterior, por lo que si tenemos un aumento de la media a medida que avanzamos en el tiempo, esto hace que dicha subida se atenúe.

Podemos diferenciar tantas veces como necesitemos para obtener una media constante, pero no suele ser necesario diferenciar en más de dos niveles.

También podemos aplicar un diferenciado estacional. Como veremos en las pruebas realizadas, muchos de los modelos obtenidos requieren este tipo de operación, por lo que podremos saber, una vez acabado el proceso de selección, si necesitamos aplicar un diferenciado estacional para generar buenos modelos. El diferenciado estacional indica que algunos de los valores de la serie están variando de forma estacional o periódica.

$$w_t = z_t - z_{t-s}$$

Donde s es el periodo de la componente estacional.

Ambas diferenciaciones se pueden aplicar a la vez y el orden de aplicación es indiferente.

2.2.3. Aplicación y estudio de otros modelos más simples

En este punto ya habremos obtenido información gráfica y estadística de la serie estudiada, por lo que iremos aplicando modelos de más simple a más complejo antes de llegar al estudio y generación de modelos ARIMA. Algunos de estos modelos se explican en [3] El orden de estudio de los modelos más simples será el siguiente:

- Método Naïve
- Método de Media Simple
- Método Holt-Winters

2.2.4. Filtro de Modelos

Filtro por RMSE Una vez generados tanto los modelos más simples como los modelos Holt-Winters y ARIMA se ha elegido una primera forma de filtrado por el RMSE obtenido en las predicciones de cada modelo.

El orden de comparación de modelos será de los más simples a los más complejos, filtrando por último los modelos ARIMA que superen la capacidad de predicción de todos los anteriores.

Como el RMSE es dependiente de la escala, no se deberá utilizar para realizar un estudio comparativo del error entre distintas series.

Filtro por Estacionariedad e invertibilidad de los modelos Vamos a aplicar varios test estadísticos para estudiar la estacionariedad e invertibilidad de los modelos. La invertibilidad asegura que los datos anteriores más próximos a la predicción tienen un peso mayor en dicha predicción que los datos más alejados. Esta propiedad hace que un proceso MA se pueda expresar como un proceso AR equivalente.

La invertibilidad se determina calculando las raíces de un polinomio construido a partir de los valores de los coeficientes MA.

Para estudiar la estacionariedad nos centraremos en los valores de los coeficientes AR, realizando el mismo estudio que para comprobar la invertibilidad.

Filtro por Parsimonia A continuación procedemos a filtrar por parsimonia. Con ello descartamos aquellos modelos ARIMA cuyos valores de p, q, d y P, Q, D sean mayores que un umbral establecido, dando modelos poco manejables y más difíciles de interpretar. A este principio de diseño se le denomina principio de parsimonia.

Como se comenta en los primeros capítulos del libro Forecasting with Dynamic Regression Models, de Alan Pankratz, se eligen los modelos de menor orden que expliquen correctamente el comportamiento de la serie.

El orden, no-estacional o estacional, es la suma de los valores p, q, d y P, Q, D, respectivamente. Los modelos ARIMA de menor orden tienden a dar las predicciones más precisas.

Filtro por AIC y BIC Despues de filtrar por RMSE, invertibilidad, estacionariedad y parsimonia, filtraremos por complejidad de los modelos, manteniendo aquellos cuyos valores de AIC y BIC sean menores, desechando los modelos cuya capacidad de generalización sea menor. En este caso se almacenan ambos a la hora de hacer las pruebas, pero realmente filtramos por AIC.

Así pues, cuanto mayor sea el valor de máxima verosimilitud, menor será el valor del criterio de información, y más se ajustará el modelo a los datos. Ésto depende de la forma del modelo y del grado de dispersión de los datos.

Dicho AIC viene dado por:

$$AIC = 2k - 2\ln(L)$$

$$AIC = N \log\left(\frac{SSE}{N}\right) + 2(k + 2)$$

Donde N es el número de observaciones, SSE es la suma de los errores al cuadrado y k es el número de predictores seleccionados para el modelo.

Por otra parte, el BIC (Bayesian Information Criterion) viene dado por:

$$BIC = k \ln(n) - 2\ln(L)$$

Donde, de nuevo, k es el número de parámetros, L es el valor de máxima verosimilitud y n es el número de observaciones. Igual que el AIC se basa en la máxima verosimilitud como medida de la bondad del ajuste.

No podemos comparar los valores de AIC o BIC obtenidos para los modelos Holt-Winters con los obtenidos en los modelos ARIMA, ya que en Holt-Winters siempre se computa la serie completa, mientras que en ARIMA, al realizar el proceso de diferenciado, se computan menos observaciones.

2.2.5. Elección de los mejores modelos para 3, 6, y 13 meses

Obtendremos tres grupos de modelos: para 3, 6 y 13 meses. De esta forma podremos comprobar si los modelos que tienen los mejores valores de predicción para los primeros meses siguen conservando esta cualidad para meses posteriores.

2.3. Herramientas utilizadas

Para la realización del trabajo de experimentación hemos utilizado las siguientes herramientas:

- El entorno de desarrollo ha sido Jupyter Notebook de Anaconda. Se trata de un entorno multiplataforma que facilita más de 1000 paquetes de código abierto. Destaca por ser usado en numerosos proyectos de aprendizaje computacional en Python, debido a su sencillez de uso y a la rapidez a la hora de generar documentación a la vez que se programa.
- Se ha usado la librería SARIMAX de statmodels para generar los modelos ARIMA, así como otras librerías utilizadas en análisis de datos como son pandas o numpy. De SARIMAX también se han usado los paquetes de adfuller para realizar el test de Dickey-Fuller o el seasonal_decompose para realizar una descomposición de las series.
- Para la representación de las series y de las posteriores predicciones se ha utilizado la librería plotly, haciendo necesario el uso de llamadas a su API, que facilita la generación de gráficas interactivas, bastante útil en el estudio de series temporales.
- Para almacenar los datos de los modelos obtenidos se ha utilizado la librería pickle. Debido al conjunto tan grande de modelos generados se vio necesaria una forma de almacenar los datos intermedios de cara a agilizar el filtrado. La librería pickle nos permite generar ficheros y cargarlos en un tiempo de ejecución bastante corto.
- En las sucesivas pruebas se ha usado un tipo de estructura llamada Dataframe, que organiza los datos en forma de tabla en la que se pueden hacer filtros y operaciones muy potentes de forma sencilla, haciendo que el manejo de los datos sea bastante más rápido y sencillo.

3. Diseño y resolución del trabajo realizado para la serie MARSA

3.1. Estudio preliminar de las series

3.1.1. Representación de la serie

El primer paso es representar la serie para tener una primera impresión de cómo varía en el tiempo e intentar detectar visualmente si necesita aplicarse alguna transformación. Este estudio es parte de la metodología Box-Jenkins. Las transformaciones se pueden ver en más detalle en [6]. A continuación, se muestra la serie MARSA en número de incidencias:

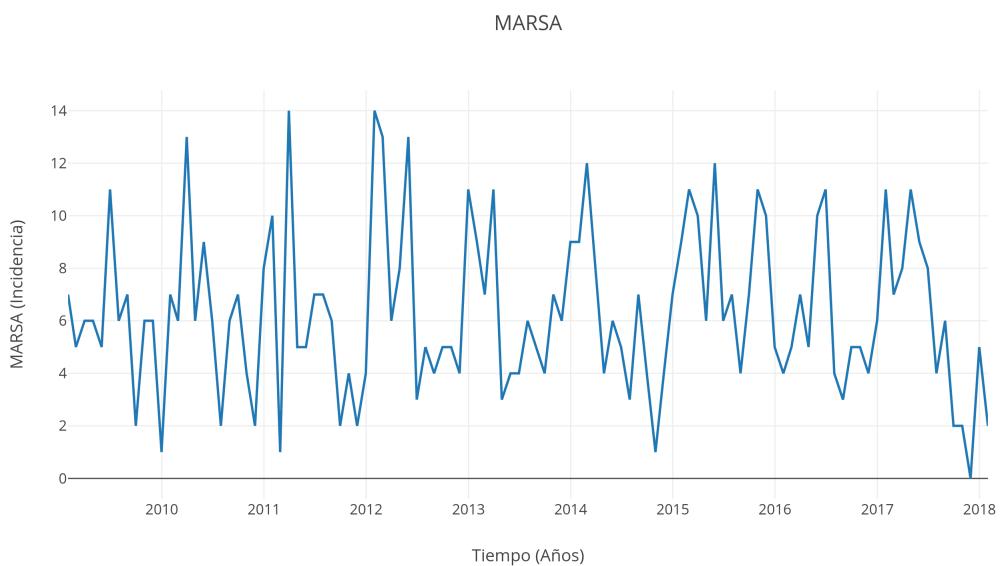


Figura 1. Serie MARSA

Como se puede observar en la Figura 1, es complicado distinguir un cambio brusco en la tendencia de la serie. Tampoco se intuye un gran cambio en la varianza, pero se puede observar una porción de la serie en la que hay un aumento de varianza que luego pasa a estabilizarse. Esta porción abarca desde comienzos de 2011 hasta mediados de 2012. A partir de esta fecha, ni la varianza ni la media parece presentar cambios abruptos.

Una vez representada la serie, un histograma puede ayudar a ver la distribución de los datos:

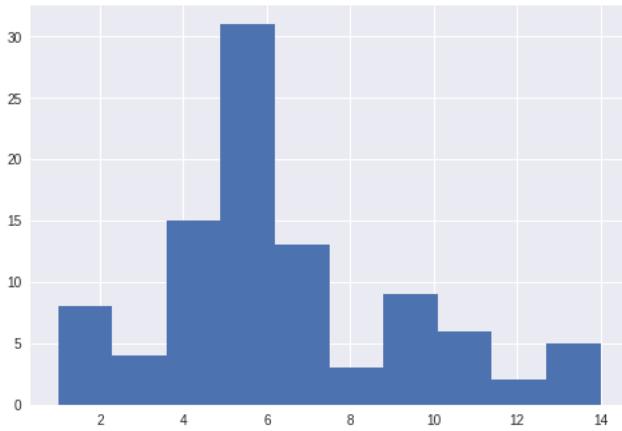


Figura 2. Histograma de serie MARSA

En este caso, podemos ver en el histograma representado en la Figura 2, que no estamos ante una distribución uniforme y no tiene forma de gaussiana simétrica. Esta forma indica un comportamiento exponencial en la que encontramos un pico en el valor 6. Aunque la distribución no parece seguir una forma normal, hay una mayor presencia de valores en el rango central (valores de 4 a 7). Esto sugiere la existencia de correlaciones en los datos, de forma que no se trata simplemente de ruido blanco.

Para ver si hay cambios de varianza y media con el nivel de la serie aplicamos un suavizado de tipo media móvil, pero la serie suavizada debemos tener en cuenta que no se debe usar para construir el modelo.

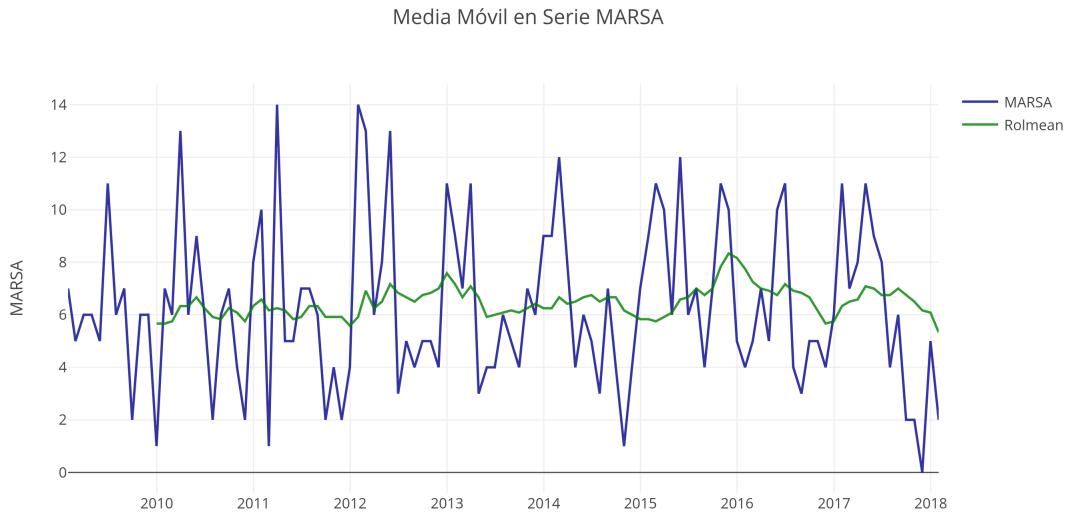


Figura 3. Representación de la media móvil de MARSA con ventana = 12

Como se puede ver en la Figura 3, el primer estudio acerca de la media parece dudoso; se mantiene prácticamente invariable en el tiempo, aunque vemos pequeñas subidas en los años 2012, 2013 y 2016, que nos hacen dudar de la estacionariedad. Será necesario aplicar un test estadístico objetivo para determinarlo.

En posteriores pruebas se mostrará que aplicando la diferenciación de la serie en un nivel se obtienen buenos modelos, por lo que, aunque podemos considerar la media estacionaria, se podría mejorar su estabilidad.

En la Figura 4 se muestra la media móvil aplicada en una ventana de seis meses en lugar de doce, se puede observar estacionalidad en los datos al presentar bajadas a principio de año, subidas hasta los meses de Mayo y Junio y de nuevo bajadas en los últimos meses del año. Esto hace que podamos afirmar que existe una estacionalidad ciertamente irregular entre 10 y 12 meses, lo que podría ocasionar problemas a la hora de detectar la estacionalidad de la serie.

Esto es coherente con lo que los clínicos observan en el dominio. Los brotes epidémicos de gripe son ciertamente estacionales, pero pueden retrasarse o adelantarse un poco en distintos años. Esta estacionalidad en la gripe, lleva al aumento del uso de antibióticos en ciertas épocas del año, y por tanto al posible aumento de resistencias bacterianas. Ello explica la estacionalidad débil e irregular de la serie de organismos multirresistentes (MARSA).

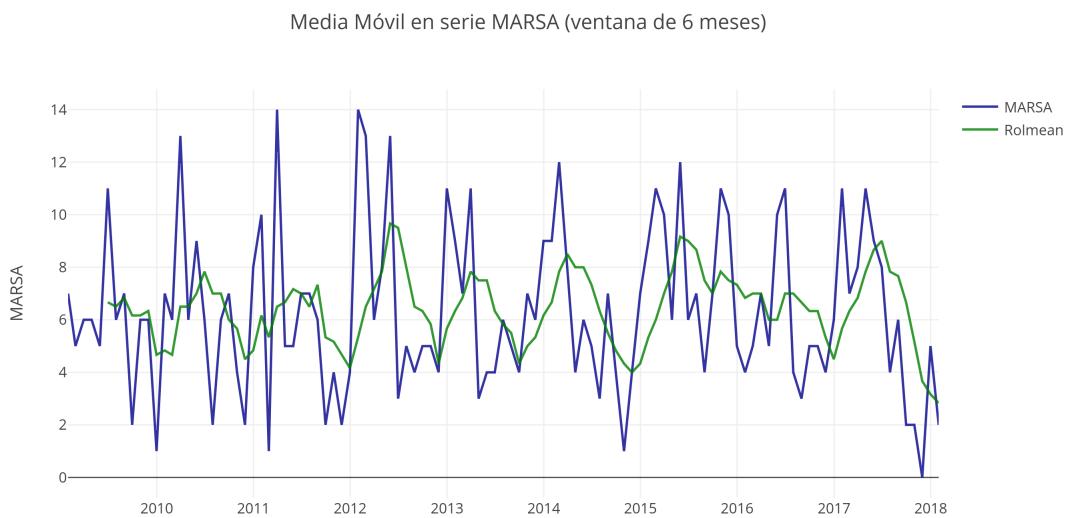


Figura 4. Representación de la media móvil de MARSA con ventana = 6

Antes de estudiar la varianza se va a aplicar la primera diferencia a la serie para tratar de estabilizar la media. La serie transformada se muestra en la Figura 5.

Differencing MARSA

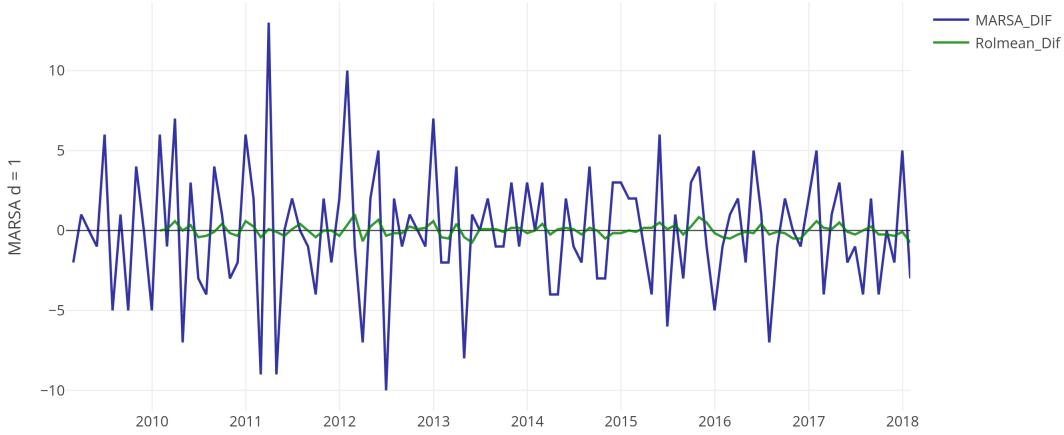


Figura 5. Media móvil de MARSA diferenciada ($d = 1$)

Se observa una media con mucha menos variabilidad. Este cambio hace que sea más sencillo estudiar la estabilidad de la varianza. En este caso se obtienen muy pocos valores que sobrepasen los rangos en los que se mueve la serie, por lo que el estudio gráfico indicaría que no sería necesaria una transformación Box-Cox para estabilizar la varianza.

Moving Standard Deviation

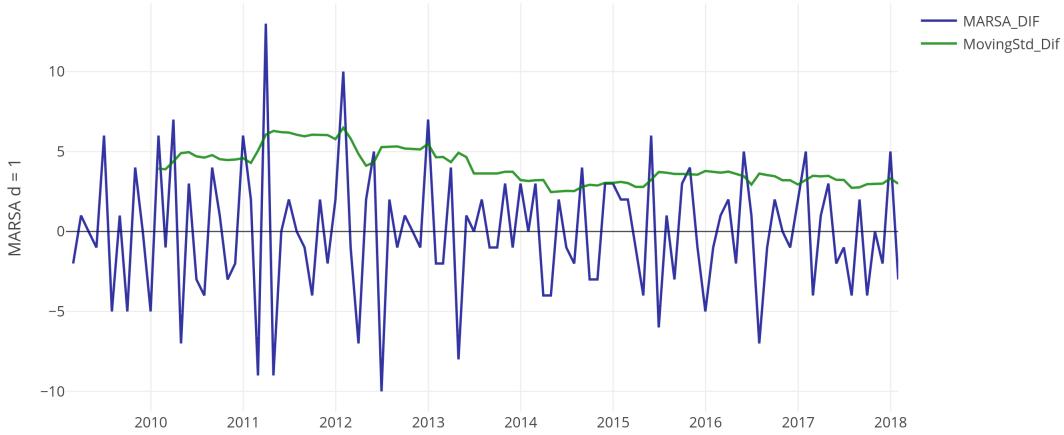


Figura 6. Desviación típica móvil de MARSA diferenciada ($d = 1$)

Ahora vemos cómo la varianza sufre una ligera bajada en la primera mitad de la serie, siendo estable en la segunda mitad. Para estabilizar la varianza utilizamos transformaciones Box-Cox, cuya forma general es:

$$z'_t = \frac{z_t^\lambda - 1}{\lambda}$$

En caso de tener una varianza que aumente con el nivel de la serie tendremos que aplicar un valor de $\lambda < 1$, mientras que si la varianza disminuye con el nivel se escogerá un valor de $\lambda > 1$. En nuestro caso la varianza no aumenta ni disminuye con respecto al nivel de la serie, pero sí vemos una subida en unos valores concretos de tiempo seguidos de una pequeña bajada. Haciendo una búsqueda del parámetro λ mediante el método boxcox de la librería scipy obtenemos un valor de 0.51, aproximadamente 0.5, lo que nos indica que la mejor transformación sería aplicar la raíz cuadrada.

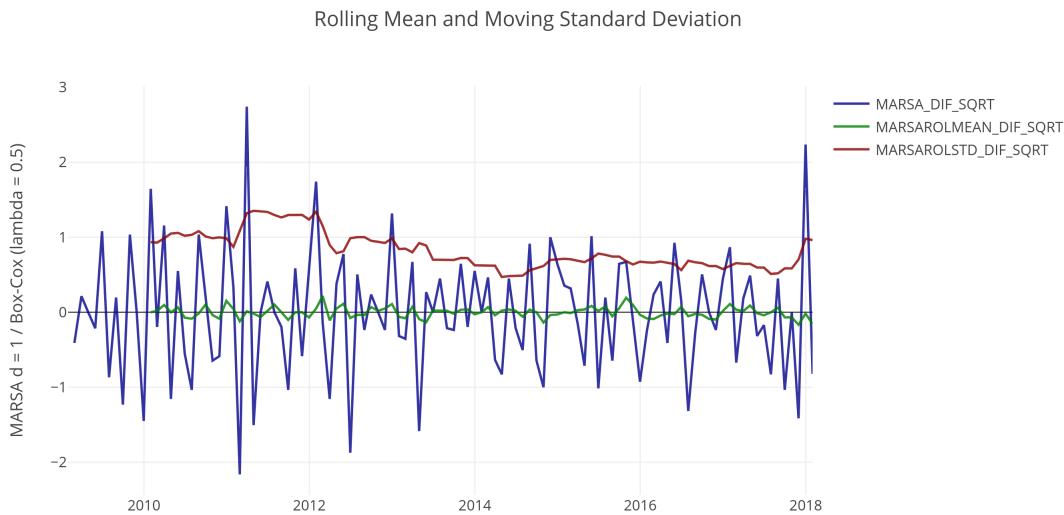


Figura 7. Serie MARSA transformada con $d = 1$ y $\lambda = 0.5$

Como observamos en la Figura 7, tanto media como varianza parecen estacionarias, pero la serie original sin transformar también parecía tener estas propiedades. Estas transformaciones requieren la aplicación de transformaciones inversas en los resultados de la predicción, por lo que antes de aplicarlos directamente vamos a realizar un test estadístico que nos indicará si la serie es o no es estacionaria.

3.1.2. Test Avanzado de Dickey-Fuller para estudio de la estacionariedad

Como última prueba antes de generar los modelos ARIMA, vamos a aplicar el Test de Dickey-Fuller, para seguir estudiando la estacionariedad de la serie. En este caso, las

transformaciones realizadas en la sección anterior indican que podríamos necesitar hacer la primera diferencia de la serie y estabilizar la varianza aplicando una transformación Box-Cox. Con el Test de DickeyFuller vamos a estudiar si existe la presencia de raíces unitarias, viendo si es estacionaria de orden 0. En este caso se considera que la hipótesis nula H_0 es la serie no estacionaria:

El valor del máximo lag usado en el test viene dado por la expresión:

$$12\left(\frac{n_{obs}}{100}\right)^{1/4}$$

El valor de p-value obtenido en adfuller es de $p = 3,30e^{-06}$; esto nos indica que la serie es estacionaria al ser mucho menor que el valor crítico de 0.05. De esta forma sabemos que puede no ser necesaria una transformación lambda y una diferenciación de la serie para obtener buenos modelos.

Viendo el resultado de este test podríamos descartar la opción de diferenciar la serie, pero como veremos en los tests posteriores, tendremos modelos de primera diferencia que se ajustan bastante bien en cuanto a predicción y explican corretamente el comportamiento de la serie, que supera por poco margen la longitud mínima que debe tener una serie de datos para el entrenamiento en ARIMA.

Como conclusión, la serie es estacionaria, aunque el test de Dickey-Fuller puede no detectar bien esta propiedad debido a la escasa longitud de la serie.

3.1.3. Descomposición de la serie

La descomposición de la serie nos va a ser útil para estudiar graficamente la tendencia y ver cómo se distribuyen los valores estacionalmente.

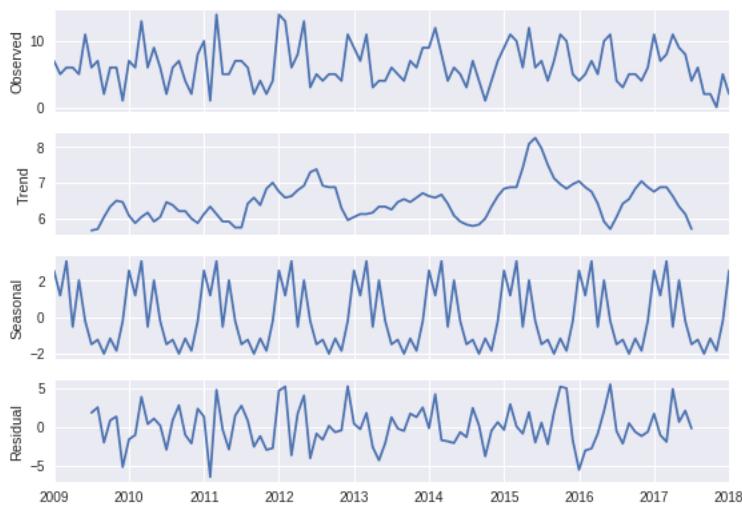


Figura 8. Descomposición de la serie MARSA

Podemos ver en la Figura 8 lo que ya hemos comentado anteriormente de la tendencia estacionaria con pequeñas subidas y bajadas a lo largo de la serie. Vemos una subida bastante

diferenciada del resto entre 2015 y 2016, pero en general, la serie se mantiene bastante estable.

Por otra parte, la estacionalidad del gráfico no nos da información acerca de si la serie sigue un patrón estacional, ya que se representa como la suma normalizada de todos los meses homólogos.

En cuanto a la representación de los residuos estamos ante un ejemplo de ruido blanco, ya que, podemos considerar que los picos que sobrepasan 2 veces el error estándar pueden ser debidos al factor de aleatoriedad presente en la serie.

3.1.4. Análisis de ACF y PACF

Un paso necesario en la metodología Box-Jenkins es la representación de la función de autocorrelación (ACF) y la función de autocorrelación parcial (PACF).

Para estudiar las ACF y PACF obtenidas para nuestra serie vamos a realizar una comparación con varias ACF y PACF teóricas. De esta forma sabremos si siguen algún patrón que nos haga elegir modelos simples muy usados en la práctica.

La función de auto-correlación (ACF) representa la correlación entre los valores de la serie para distintos espacios de tiempo. El cálculo de los coeficientes de la función se realiza de la siguiente forma:

$$p_k = \frac{\text{cov}(z_t, z_{t+k})}{\sigma_z^2}$$

Para saber si los valores obtenidos en la ACF son representativos usaremos el límite de dos errores estándar a partir del 0, de esta forma, los coeficientes que sobrepasen dicho valor se considerarán significativos y despreciaremos el resto.

Otra forma útil de medir la autocorrelación es la función de autocorrelación parcial (PACF), que mide la correlación entre z_t y z_{t+k} ignorando la dependencia de las componentes intermedias.

Explicado de otra forma, si tenemos el siguiente conjunto de ecuaciones de regresión:

$$\begin{aligned} z_t &= C_1 + \phi_{11}z_{t-1} + e_{1,t} \\ z_t &= C_2 + \phi_{21}z_{t-1} + \phi_{22}z_{t-2} + e_{2,t} \\ &\dots \\ z_t &= C_K + \phi_{K1}z_{t-1} + \phi_{K2}z_{t-2} + \dots + \phi_{KK}z_{t-K} + e_{K,t} \end{aligned}$$

Solamente tenemos en cuenta el valor del último coeficiente de cada ecuación, por lo que el peso de los valores intermedios se descarta, es decir, el valor del coeficiente viene dado por la correlación condicional de z_t y z_{t+k} .

En primer lugar vamos a estudiar ambas representaciones para la serie original sin aplicar ninguna transformación.

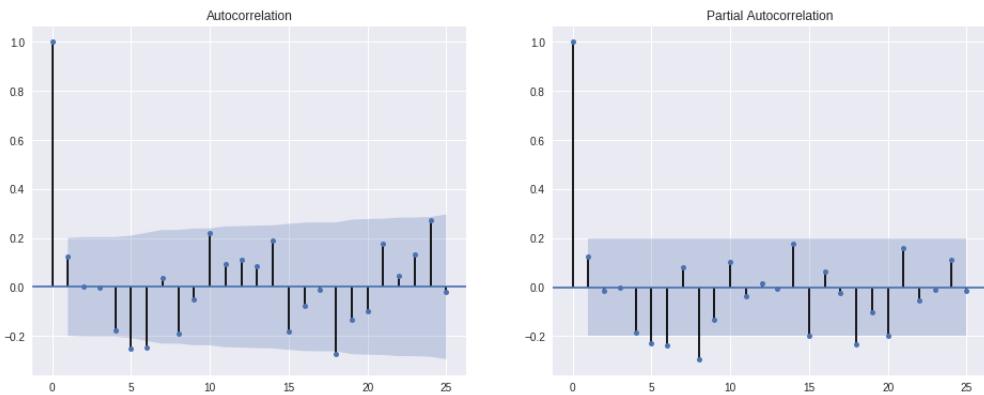


Figura 9. ACF y PACF de la serie MARSA original

ACF En la Figura 9 vemos unos picos en los lags 5 y 6, indicando una posible autocorrelación semestral. Por otra parte, viendo la ACF no observamos una caída lenta de los valores, por lo que, a simple vista no podemos indicar que se trata de una serie no estacionaria mirando los valores de las gráficas ACF y PACF.

Sabemos que en la práctica los patrones estacionales y no estacionales ocurren a menudo juntos, por lo que es útil tratar de estudiarlos por separado.

Si aislamos los lags múltiplos de 12 no encontramos estacionalidad anual, ya que no existen picos en dichos múltiplos que superen el valor de 2 veces la desviación estándar. Tampoco encontramos estacionalidad semestral, al encontrar el valor en el lag 12 muy por debajo del límite. Sin embargo, el valor obtenido en el lag 18 sí es significativo, aunque por muy poco, por lo que podemos descartarlo al tratarse de un lag muy lejano en tiempo y saber que la correlación entre valores tan lejanos no es útil al tratarse de series cortas. En caso de tratarse de series de una longitud considerablemente mayor sí podríamos tener en cuenta este valor si encontramos que la correlación es significativa y que el patrón tiene sentido en el ámbito en el que nos encontramos.

No vemos ni caída exponencial ni un corte a partir del cual los valores anteriores sean significativos y los posteriores no. Lo que sí vemos en el ACF es una tendencia a alternar el signo en los coeficientes, pero no hay una caída paulatina a 0. Los valores de correlación en la ACF son prácticamente despreciables a partir de orden 7 y esto, unido a lo que hemos comentado anteriormente indican estacionariedad en la serie.

PACF Si examinamos detenidamente la PACF encontramos que hay valores significativos a partir del lag 15, que podemos descartar al tratarse de valores muy cercanos al límite y ser lags muy avanzados.

Por otra parte tenemos que se repite la estructura vista en el ACF, presentando autocorrelación en los lags 5 y 6. Podemos descartar el valor del lag 8, que aunque sea el más significativo, daría lugar a modelos muy poco parsimoniosos. De hecho, los valores de los lags 5 y 6 sobrepasan por poco el umbral y darían lugar también a este tipo de modelos

poco parsimoniosos.

En la figura 10 se representan las funciones ACF y PACF para la serie diferenciada en un nivel.

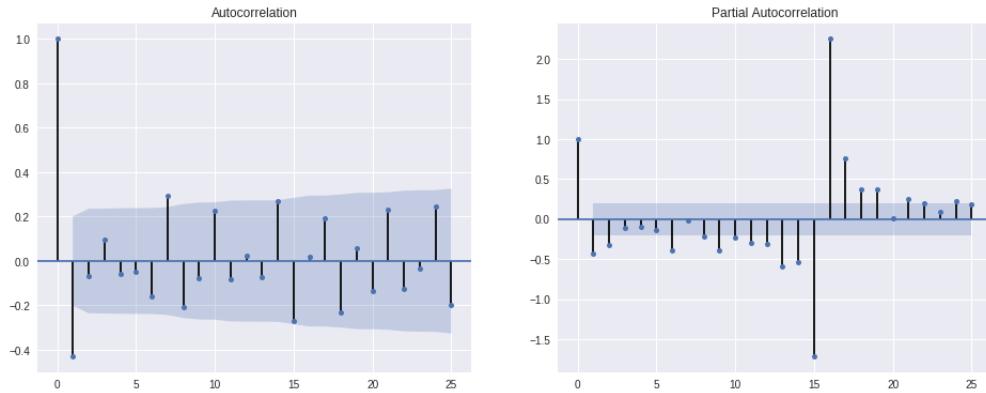


Figura 10. ACF y PACF de la serie MARSA diferenciada ($d = 1$)

ACF Como se puede ver en la gráfica ACF tenemos un valor en el lag 1 muy representativo, lo que quiere expresar que hay una fuerte correlación entre los instantes de tiempos t y los instantes inmediatamente posteriores $t + 1$.

Vemos también que aparece una correlación en el lag 7 y el resto son despreciables. Esta correlación espaciada en 7 a priori tiene una continuación en los lags 14 y 15, aunque al estar por debajo del valor mínimo no los vamos a considerar en el estudio.

PACF En el PACF, por otra parte, vemos que aparecen lags muy significativos en los lags 15 y 16, pero como se trata de lags muy distanciados, no tendremos en cuenta estos valores. Si nos ocupamos de estudiar los coeficientes de correlación, vemos que prácticamente todos sobrepasan el umbral mínimo establecido por dos veces desviación estándar, pero ni desciende gradualmente ni tiene un corte abrupto, por lo que mediante dicha gráfica no podemos encontrar un patrón claro.

3.2. Aplicación de métodos más simples

3.2.1. Aplicación del método Naïve

Este Método es el más sencillo y se basa en usar el último valor de los datos de entrenamiento como predicción. El método Naïve puede ser útil en series que no tengan una variación muy abrupta ni en media ni en varianza, por lo que los resultados de error suelen ser pobres en los casos estudiados en este documento, debido a su gran variabilidad. Aun así es un punto de partida muy simple, por lo que se ha elegido como primer filtro de modelos ARIMA generados por fuerza bruta.

$$y_{t+1} = y_t$$

El resultado obtenido al aplicar el método Naïve es el siguiente:

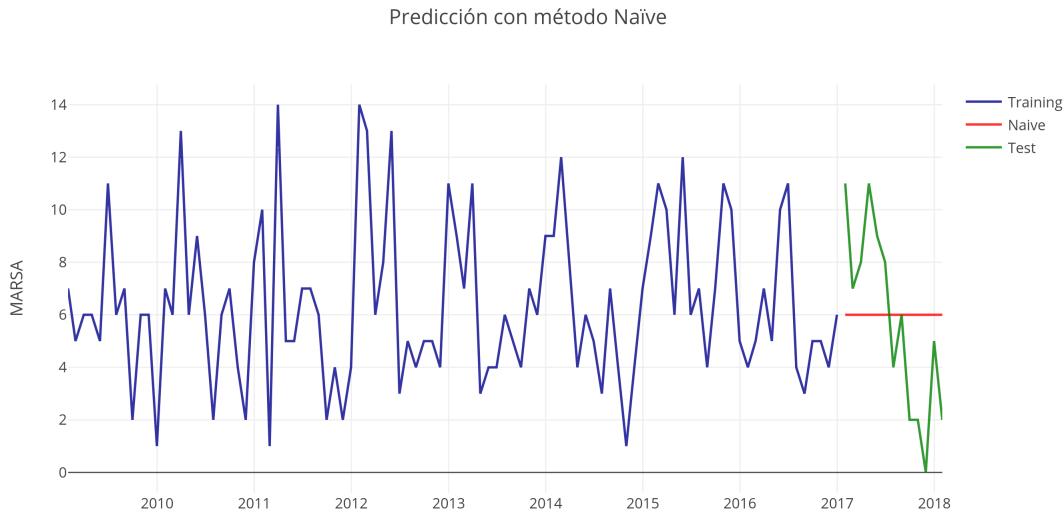


Figura 11. Predicción con método Naïve

Si observamos la figura 11, concretamente los datos usados para realizar el test (Datos en verde), estos presentan una bajada considerable bastante complicada de predecir, por lo que métodos de predicción simples como este no nos darán un buen valor de error.

$$RMSE_{Naive} = 3,47$$

Viendo que los valores de la serie oscilan entre 2 y 14, pero tienen una mayor distribución entre 4 y 7, el error obtenido, lejos de ser catastrófico, es bastante mejorable y, ya que este método no tiene en cuenta ni la posible estacionalidad presente en la serie, ni valores anteriores aparte del último, nos queda un gran margen de mejora.

3.2.2. Aplicación del método de Media Simple

El método de Media Simple es también una herramienta sencilla que utiliza la media de los valores de entrenamiento de la serie como predicción. Como se verá en el estudio realizado, no se obtienen mejores valores de error comparado con el método Naïve. En los casos de media estacionaria puede mejorar los resultados respecto a este último, pero en los datos de prueba contamos con una bajada bastante abrupta, por lo que serán bastante similares (e imprecisos) en la predicción.

$$\bar{x} = \sum_{n=1}^n x_i = \frac{x_1+x_2+x_3+\dots+x_n}{n}$$

La predicción con el método de Media Simple se representa en la figura 12.

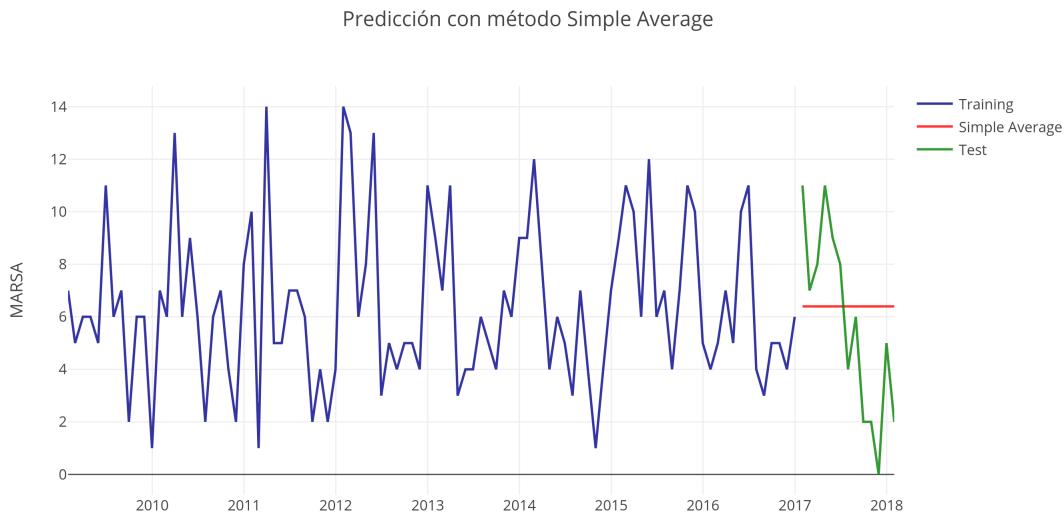


Figura 12. Predicción con método de media simple

El resultado de RMSE obtenido es:

$$RMSE_{SA} = 3,52$$

Como vemos, el valor de RMSE es ligeramente peor que en el caso del método Naïve, por lo que descartamos el valor obtenido y nos quedamos con el primer test realizado para filtrar los modelos.

3.2.3. Aplicación de Suavizado exponencial con método Holt-Winters

El método Holt-Winters se basa en la división de la predicción en tres ecuaciones de suavizado:

- Ecuación para el nivel l_t
- Ecuación para la tendencia b_t
- Ecuación para el componente estacional s_t

Este método, a diferencia de los anteriormente comentados, incluye el factor de estacionalidad. Tiene dos variantes, la variante aditiva y la multiplicativa; el uso de la aditiva se prefiere si la estacionalidad se mantiene constante en la serie, mientras que el uso de la multiplicativa se prefiere si hay variaciones de estacionalidad proporcionales al nivel de la serie.

A continuación generamos modelos Holt-Winters usando ambas opciones, aditiva y multiplicativa, y el parámetro de estacionalidad desde 2 hasta 12 meses, por lo que se han tenido en cuenta todas las posibilidades.

Se han filtrado los modelos usando el valor de RMSE obtenido usando el método Naïve. Después de la operación tenemos los modelos de la tabla representada en la figura 13. En dicha tabla podemos observar en la primera columna los valores usados para generar los modelos, en la segunda y tercera columna tenemos los valores de BIC y AIC respectivamente y por último, en la cuarta columna los valores de RMSE de cada modelo calculado con la prueba de los 13 meses.

	Model	BIC	AIC	RMSE
1	[12, mul, add]	308.699649	267.670078	2.959413
2	[12, add, add]	296.445810	255.416239	3.204825
3	[3, add, mul]	271.873021	253.922584	3.246894
4	[4, add, add]	267.566521	247.051735	3.268332
5	[6, add, add]	288.128538	262.485056	3.301268
6	[5, mul, mul]	305.588073	282.508939	3.307633
7	[7, add, mul]	252.097482	223.889652	3.321229
8	[7, add, add]	250.713243	222.505413	3.376609
9	[10, add, add]	299.576826	263.675951	3.410162
10	[3, mul, mul]	289.510279	271.559842	3.413008

Figura 13. Mejores modelos Holt-Winters obtenidos

Los valores de RMSE para estos modelos son bastante mejores que en los dos casos anteriores y, al ser modelos más complejos ahora sí calculamos los valores de AIC y BIC para realizar una comparativa. En este caso nos vamos a quedar con el modelo con menor RMSE, pero vamos a estudiar graficamente el ajuste de los dos mejores modelos Holt-Winters así como su capacidad de predicción usando el último año como test. De esta forma tendremos una ligera idea de hasta dónde se puede llegar utilizando estos modelos en nuestro caso de estudio.

Método Holt-Winters - Ajuste del modelo

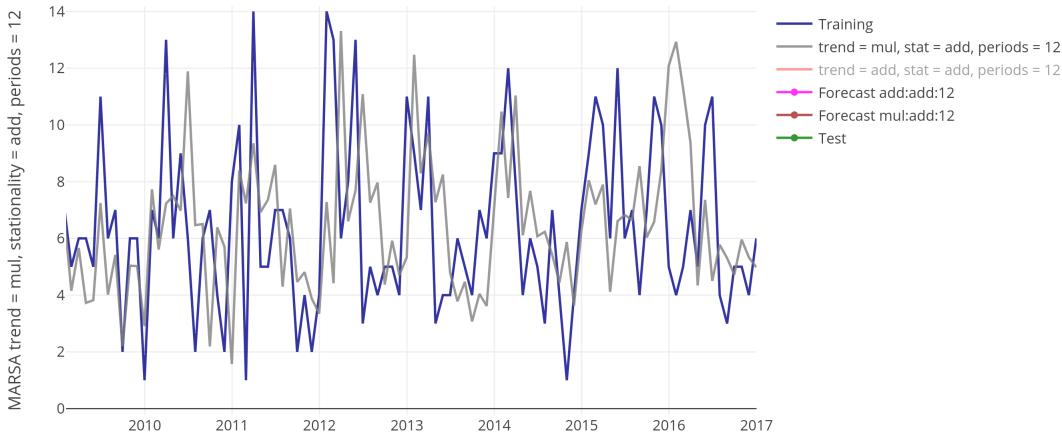


Figura 14. Holt-Winters con tendencia multiplicativa y estacionalidad aditiva

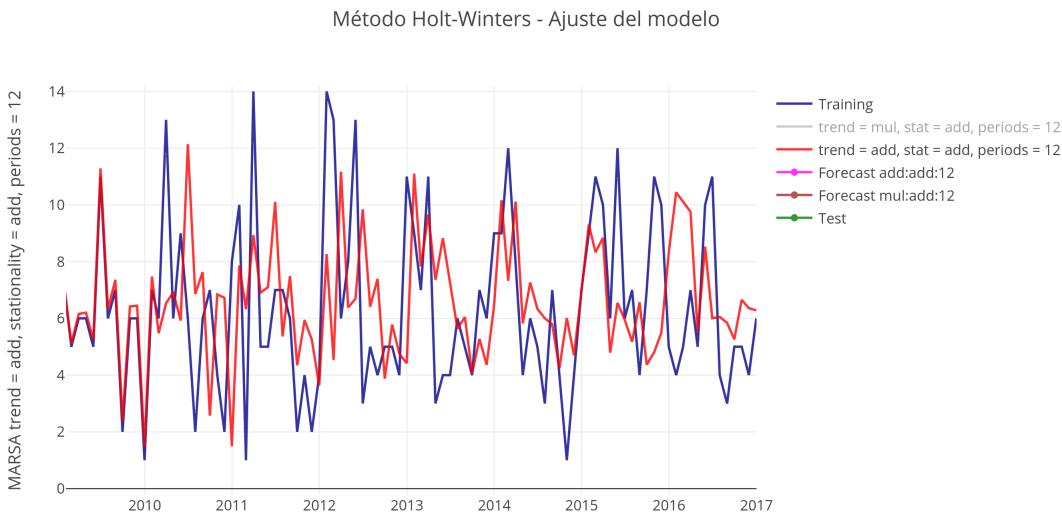


Figura 15. Holt-Winters con tendencia y estacionalidad aditivas

Como vemos en los ajustes de las figuras 14 y 15, el modelo con tendencia y estacionalidad aditivas parece ajustarse mejor al comportamiento de la serie, cosa que no tiene por qué ser mejor, como vemos en el cálculo de los dos RMSE.

Al final obtenemos que la serie con tendencia multiplicativa y estacionalidad aditiva predice bastante mejor y el valor de complejidad del modelo no es mucho mayor, por lo que podemos usar este para filtrar los modelos ARIMA generados. Tenemos que el RMSE elegido como último filtro de modelos más simples es:

$$RMSE_{holt-winters} = 2,96$$

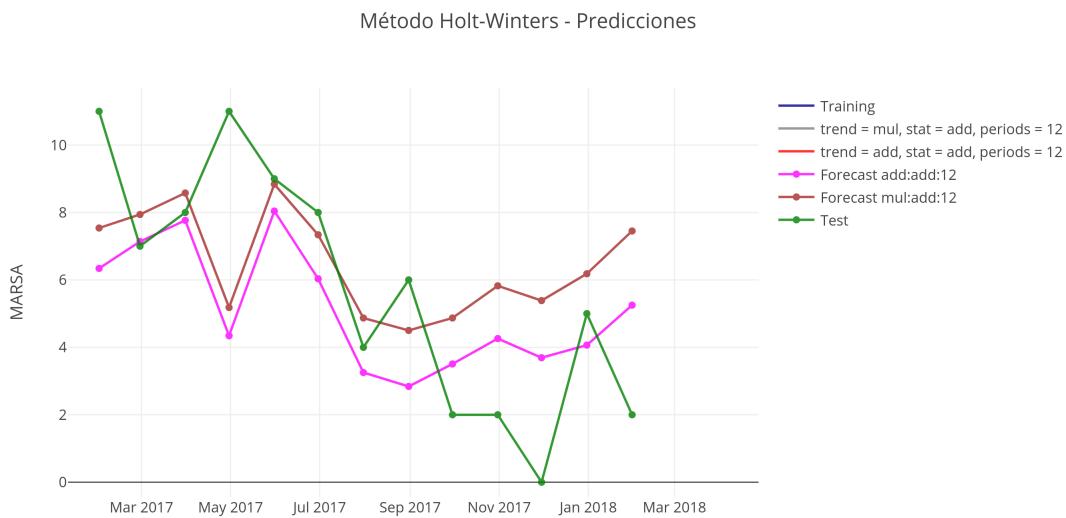


Figura 16. Predicción usando Holt-Winters

Como vemos en la figura 16, ambos resultados son muy similares y logran predecir, al menos, la tendencia intermedia de bajada de la serie. Al final vuelven ambas a un valor próximo a la media, que podría explicarse como un suavizado de la caída final de la serie.

Los resultados obtenidos no son del todo precisos, aunque no podemos decir que los modelos no expliquen de forma parcial el comportamiento de los datos de prueba.

Como conclusión, los datos obtenidos nos sirven como punto de referencia para establecer si los modelos ARIMA generados en el estudio se comportan de forma correcta o al menos consiguen mejores resultados.

3.3. Generación de modelos ARIMA

Para la generación de modelos se utiliza un algoritmo de fuerza bruta llamando a la librería SARIMAX y almacenando tanto los valores de los coeficientes de cada modelo, como su AIC, BIC, y error cuadrático para 3, 6 y 13 meses. En total se han generado más de 25000 modelos ARIMA en un total de aproximadamente 6 horas de cómputo.

	Model	BIC	AIC	RMSE	RMSE6	RMSE3
23786	[3, 1, 4, 3, 0, 2, 2]	440.190418	472.096933	2.462816e+00	1.690459e+00	1.769588e+00
526	[0, 0, 0, 4, 0, 3, 10]	307.039137	323.241951	2.467012e+00	1.375958e+00	1.371113e+00
2326	[0, 0, 3, 4, 0, 3, 10]	307.787138	330.066007	2.484566e+00	1.950533e+00	1.680879e+00
23654	[3, 1, 4, 2, 0, 1, 2]	454.646293	481.896998	2.493956e+00	1.807687e+00	1.853205e+00
9272	[1, 1, 0, 2, 0, 2, 8]	420.113856	434.254109	2.495692e+00	2.582087e+00	3.156834e+00
2926	[0, 0, 4, 4, 0, 3, 10]	309.685428	333.989649	2.497007e+00	2.131727e+00	1.808386e+00
17894	[2, 1, 4, 4, 0, 1, 2]	438.029551	467.341366	2.518169e+00	1.821086e+00	1.966427e+00
17642	[2, 1, 4, 2, 0, 0, 2]	453.873361	476.271088	2.525079e+00	1.782436e+00	1.936352e+00
23054	[3, 1, 3, 2, 0, 1, 2]	448.226704	473.000072	2.544298e+00	1.709910e+00	1.796285e+00
17654	[2, 1, 4, 2, 0, 1, 2]	449.908547	474.681915	2.551695e+00	1.825645e+00	2.009300e+00

Figura 17. Lista con algunos modelos ARIMA

3.3.1. Filtrado por RMSE comparando con el mejor modelo Holt-Winters

Usamos el mejor resultado de RMSE de los modelos Holt-Winters para realizar el primer filtrado de los modelos ARIMA. Así nos quedamos con un total de 279 modelos para realizar un estudio estadístico y ver si son modelos válidos estudiando su estacionariedad, su invertibilidad y sus residuos.

Todos los modelos obtenidos tienen parte estacional, detalle difícil de apreciar estudiando graficamente las funciones de autocorrelación. Tenemos una gran cantidad de modelos cuya periodicidad (valor del parámetro s) es 2, 8 y 10, indicando una posible componente de estacionalidad en dichos meses. Por otra parte, vemos por el filtro realizado, que los mejores modelos obtenidos no tienen diferenciación estacional (parámetro D = 0) y que hay un rango muy amplio de complejidad examinando el valor de AIC y BIC obtenido.

3.3.2. Filtrado por Criterio de Ljung-Box

El estadístico Q de Ljung-Box se utiliza para comprobar si una serie de observaciones en un período de tiempo específico son aleatorias e independientes. La autocorrelación puede disminuir la exactitud de un modelo predictivo basado en el tiempo, y llevar a la interpretación errónea de los datos.

Tiene la siguiente forma:

$$Q^* = T(T + 2) \sum_k^h (T - k)^{-1} r_k^2$$

Donde T es la longitud de la serie, r_k es el k-ésimo coeficiente de los residuos y h es el número de lags usados para realizar el test. Valores grandes de Q^* indican que hay correlación en los residuos de la serie.

Este estadístico demuestra que los valores de los datos son aleatorios e independientes hasta un cierto número de desfases. Si el estadístico es mayor que un valor crítico especificado, las autocorrelaciones para uno o más desfases podrían ser distintas de cero, lo que señalaría que los valores no son aleatorios ni independientes en el tiempo. Estos conceptos se pueden estudiar más profundamente en [2].

En nuestro caso, el criterio de Ljung-Box se usará para asegurar que no existe una dependencia entre los residuos.

Dep. Variable:	y	No. Observations:	96			
Model:	SARIMAX(2, 1, 0)x(2, 0, 1, 8)	Log Likelihood	-199.760			
Date:	Sat, 23 Jun 2018	AIC	411.520			
Time:	20:57:53	BIC	425.583			
Sample:	01-31-2009 - 12-31-2016	HQIC	417.145			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	-0.4465	0.106	-4.195	0.000	-0.655	-0.238
ar.L2	-0.2678	0.106	-2.518	0.012	-0.476	-0.059
ar.S.L8	-0.7973	0.163	-4.885	0.000	-1.117	-0.477
ar.S.L16	-0.3005	0.122	-2.472	0.013	-0.539	-0.062
ma.S.L8	0.8855	0.358	2.472	0.013	0.183	1.588
sigma2	9.0214	2.598	3.473	0.001	3.930	14.113
Ljung-Box (Q):	22.92	Jarque-Bera (JB):	1.76			
Prob(Q):	0.99	Prob(JB):	0.41			
Heteroskedasticity (H):	0.80	Skew:	0.20			
Prob(H) (two-sided):	0.58	Kurtosis:	3.62			

Figura 18. Ejemplo de modelo ARIMA

Como vemos en la tabla de la figura 18, tenemos el valor Q^* de Ljung-Box y el valor de probabilidad de Q. Nos fijaremos en este valor de probabilidad para hacer el filtrado de modelos, teniendo que cuando $p-value < 0,05$, queda demostrado que hay correlación entre los residuos (con una probabilidad del 95 %) por lo tanto se rechazarán dichos modelos. Por defecto estamos usando un nivel de significación para los intervalos de confianza de 0.05.

Si construimos el ACF de residuos de estos modelos, estos deben ser despreciables respecto a su error estándar, por lo que no existirá correlación en los residuos, dejando claro que el modelo captura correctamente los patrones existentes en los datos.

Representaremos en la última prueba las funciones de autocorrelación de residuos para comprobar que los modelos que filtramos mediante el criterio de Ljung-Box son correctos.

Veremos que los modelos tratados tienen unos residuos en forma de ruido blanco.

3.3.3. Filtrado mediante test de invertibilidad

Todo proceso estacionario $MA(q)$ puede ser expresado como un proceso $AR(\infty)$. Este es el llamado principio de invertibilidad. Principio en el que vamos a basarnos para estudiar la estacionariedad de los modelos.

Empezaremos estudiando las condiciones de invertibilidad para un proceso sencillo, en este caso un MA(2).

$$z_t = C + \theta_1 a_{t-1} - \theta_2 a_{t-2} + a_t$$

Las condiciones de invertibilidad del proceso anterior son:

$$\begin{aligned} |\theta_2| &< 1 \\ \theta_2 + \theta_1 &< 1 \\ \theta_2 - \theta_1 &< 1 \end{aligned}$$

Donde θ_1 y θ_2 son los coeficientes que multiplican a los términos de media móvil.

La invertibilidad requiere que todas las raíces de las ecuaciones características caigan fuera del círculo unidad. Tendríamos que expresar el proceso MA como un polinomio y calcular el valor absoluto de sus raíces. Si estos valores exceden en 1 quiere decir que estarán fuera del círculo unidad, por lo que el proceso será invertible [4].

El valor absoluto de las raíces se calcula como:

$$|z| = \sqrt[2]{r^2 + i^2}$$

Donde r es la parte real e i es la imaginaria.

De forma similar que con la invertibilidad de los procesos MA podemos comprobar la estacionariedad de los procesos AR, por lo que obtener los coeficientes AR y MA es necesario para seguir filtrando los modelos.

Después de aplicar el filtro por modelos invertibles y estacionarios tendremos 221 modelos ARIMA estadísticamente válidos, pero queremos tener un conjunto más pequeño, por lo que no nos podemos parar aquí. Representando un histograma con los modelos restantes podremos tener una idea de dónde poner los puntos de corte para seguir realizando el estudio.

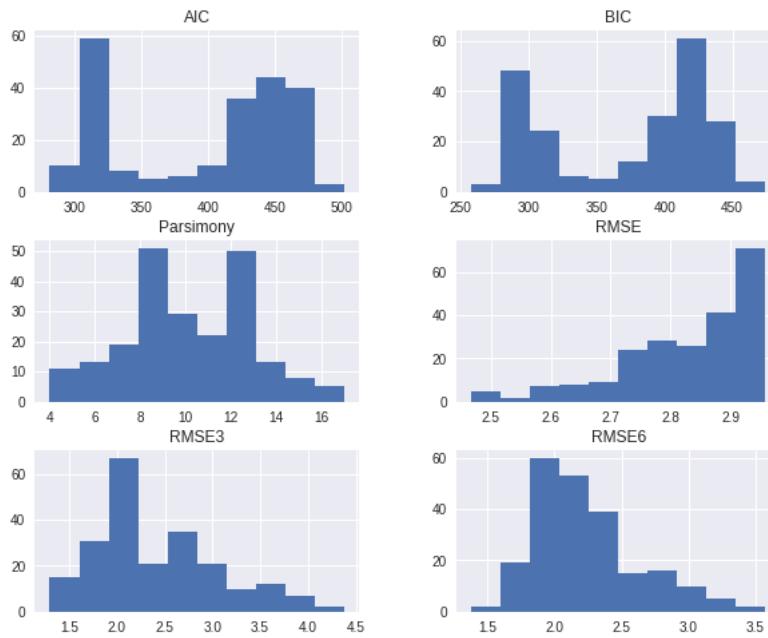


Figura 19. Distribución de los datos en modelos invertibles y estacionarios

Como se puede ver claramente en los histogramas de AIC y BIC de la figura 19, existen dos grandes grupos distribuidos de modelos. Uno de ellos se encuentra entre 300 y 350 y el otro es superior a 400. Queremos modelos lo más simples posibles, pero no podemos descartar todos los valores de este segundo grupo de modelos, ya que nos dan buenas predicciones. Por ello se ha elegido el límite de $AIC < 450$. De esta forma no vamos a descartar muchos modelos, pero sí aquellos que son excesivamente complejos comparados con la distribución de modelos que estamos manejando.

En cuanto a los valores de RMSE para 13 meses vemos que los más numerosos se encuentran en torno a 2.9, pero seguimos teniendo un gran número de modelos que superan al método Holt-Winters. Para los valores de RMSE en 6 y 3 meses sucede algo similar, tenemos una concentración en los menores valores de RMSE.

3.3.4. Filtrado de modelos por Parsimonia

Un criterio importante a la hora de elegir buenos modelos es que siga el principio de parsimonia. Queremos elegir modelos que expliquen bien los datos, es decir, que tengan un buen ajuste de los datos de entrenamiento y a su vez que sean generales. Preferimos estos modelos más generales ya que funcionan mejor en un rango más amplio de situaciones.

Imaginemos que la serie mantiene una tendencia constante a lo largo de los datos de entrenamiento y los datos que están fuera de este rango, usados para predecir, tienen una variación mínima en la tendencia. Un modelo menos sobreentrenado no será, a priori, más

preciso cuando se sigue la misma tendencia y no está sujeta a cambios, pero al ser menos generales funcionan peor para este tipo de pruebas con un comportamiento volátil.

Vamos a calcular el valor de parsimonia de la siguiente manera:

$$\text{Parsimonia} = p + d + q + P + D + Q$$

Para el caso de modelos sin parte estacional, obtendremos un valor máximo de 9 ($p=4$, $q=4$ y $d=1$) y contando con la parte estacional tendremos un valor máximo de 18 (el doble, al contar con parámetros P , D y Q). Podríamos elegir quedarnos con aquellos modelos cuyo valor de parsimonia esté por debajo de la mitad de este valor máximo, es decir, un valor de 9, pero elegimos 8 para seguir acotando lo máximo posible el número de modelos.

Después de este filtro obtenemos 69 modelos, eliminando por completo aquellos modelos con periodicidad 2.

	Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
526	[0, 0, 0, 4, 0, 3, 10]	307.039137	323.241951	2.467012	1.375958	1.371113	7
9272	[1, 1, 0, 2, 0, 2, 8]	420.113856	434.254109	2.495692	2.582087	3.156834	6
3272	[0, 1, 0, 2, 0, 2, 8]	430.543200	442.326744	2.560180	2.764787	3.517149	5
1126	[0, 0, 1, 4, 0, 3, 10]	304.758405	322.986570	2.573522	1.861413	1.665719	8
9512	[1, 1, 0, 4, 0, 2, 8]	330.149827	347.166902	2.599224	2.709429	3.342966	8
3512	[0, 1, 0, 4, 0, 2, 8]	333.510973	348.512916	2.620146	2.795743	3.515611	7
490	[0, 0, 0, 4, 0, 0, 10]	303.404616	313.531375	2.698109	2.380254	2.527901	4
9260	[1, 1, 0, 2, 0, 1, 8]	421.770278	433.553823	2.701078	2.914256	3.730380	5
21260	[3, 1, 0, 2, 0, 1, 8]	408.354095	424.669228	2.705398	2.601490	3.279927	7
15260	[2, 1, 0, 2, 0, 1, 8]	411.520030	425.582862	2.709676	2.742511	3.493909	6
442	[0, 0, 0, 3, 1, 1, 10]	302.466210	312.592968	2.722297	1.855262	2.082300	5

Figura 20. Ejemplo de modelos después de filtrar por parsimonia

Este filtro por parsimonia nos ha eliminado una gran cantidad de modelos cuyo valor de AIC y BIC era alto, pero todavía nos queda fijar un umbral para filtrar por complejidad.

3.3.5. Filtrado de modelos por Complejidad

Como hemos visto anteriormente en el resultado del histograma del valor AIC de los modelos (figura 19), existen dos grandes grupos: uno cercano a 300 (poco complejos) y otro entre 400 y 500 (más complejos). Como no queremos deshacernos de posibles modelos con buena predicción vamos a fijar un valor de filtrado de $AIC < 450$, de esta forma desecharímos los modelos demasiado complejos pero conservaremos algunos de los que se encuentran en el segundo grupo.

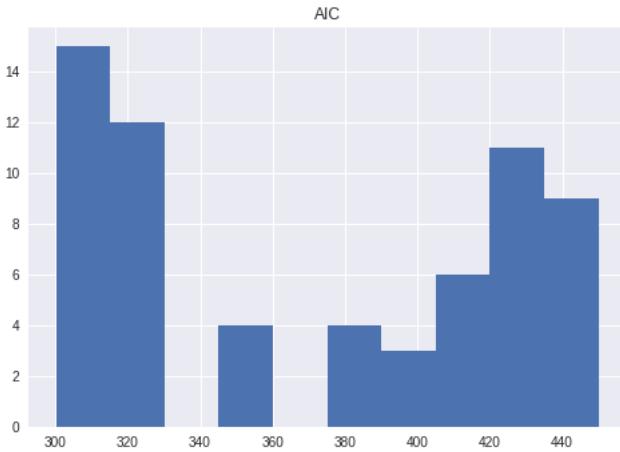


Figura 21. Histograma de AIC despu s de filtrar por complejidad

3.3.6. \'Ultimo filtrado de modelos y comprobaci n de predicciones

Por \'ltimo vamos a realizar 3 filtros por RMSE: uno teniendo en cuenta los mejores valores de RMSE para 13 meses, otro teniendo en cuenta los mejores para 6 meses y el \'ltimo para 3 meses. Veremos algunos de los modelos y sus predicciones, viendo el comportamiento de predicci n a corto, medio y largo plazo.

Mejores modelos para 13 meses El primer paso es el m s general; nos fijamos en aquellos modelos con menor valor de RMSE para 13 meses. En la tabla de la figura 22 tenemos los 5 mejores en cuanto a RMSE.

Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
526 [0, 0, 0, 4, 0, 3, 10]	307.039137	323.241951	2.467012	1.375958	1.371113	7
9272 [1, 1, 0, 2, 0, 2, 8]	420.113856	434.254109	2.495692	2.582087	3.156834	6
3272 [0, 1, 0, 2, 0, 2, 8]	430.543200	442.326744	2.560180	2.764787	3.517149	5
1126 [0, 0, 1, 4, 0, 3, 10]	304.758405	322.986570	2.573522	1.861413	1.665719	8
9512 [1, 1, 0, 4, 0, 2, 8]	330.149827	347.166902	2.599224	2.709429	3.342966	8

Figura 22. Modelos con mejor RMSE para 13 meses

Estos modelos, como se puede observar en la primera columna de la tabla, son todos estacionales, y las componentes de estacionalidad son mayores que las no estacionales.

La primera columna (Model) est  formada por los par metros del modelo, es decir, los valores p, d, q, P, D, Q y S. Los 3 primeros valores son los correspondientes a la parte no estacional y los 4 \'ltimos a la parte estacional.

Como vemos, S (periodicidad) toma valores 8 y 10; este \'ltimo confirma el estudio previo

realizado, al encontrar estacionalidad en un rango entre 10 y 12. confirma también que los mejores modelos Holt-Winters no iban desencaminados

Los modelos con periodicidad 10 tienen valores de AIC y BIC más bajos, por lo que serán más generales y preferibles para realizar predicciones.

El primer modelo destaca del resto, al tener mejores valores de RMSE, RMSE6 y RMSE3.

Si observamos los modelos podemos distinguir dos grupos muy similares, por lo que vamos a estudiar el modelo con menor error de cada uno de ellos.

Modelo $ARIMA(4, 0, 3)_{10}$ El modelo con el índice 526 parece ser el mejor; vamos a comprobar si realiza una buena predicción y a su vez podemos confirmar gráficamente que se trata de un buen modelo examinando sus residuos.

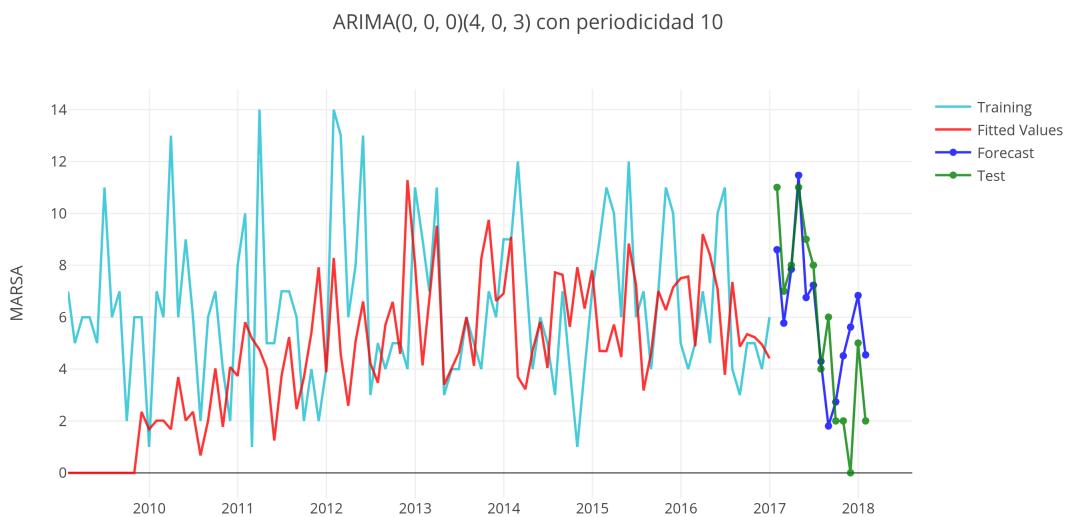


Figura 23. $ARIMA(0, 0, 0)(4, 0, 3)_{10}$

ARIMA(0, 0, 0)(4, 0, 3) con periodicidad 10

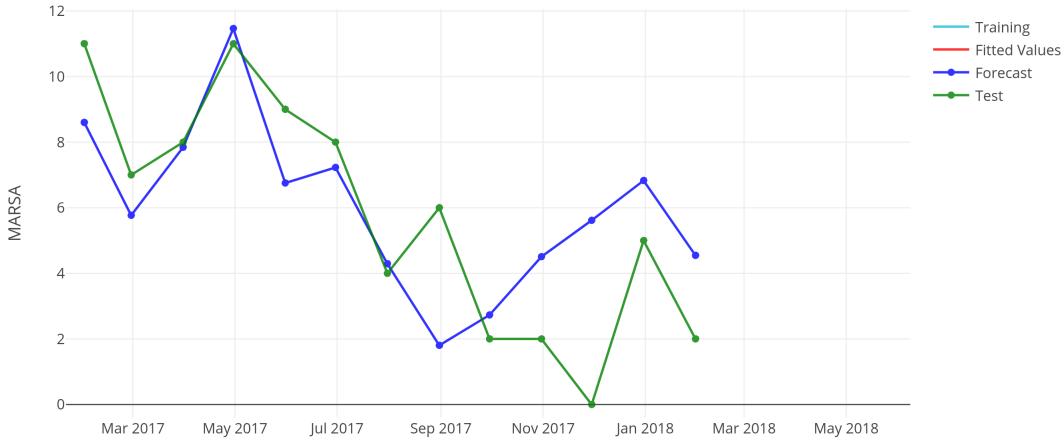


Figura 24. Predicción en $ARIMA(0, 0, 0)(4, 0, 3)_{10}$

En la figura 23 tenemos marcado en rojo el ajuste del modelo, en azul claro el fragmento de la serie utilizado para entrenamiento, en verde la parte de test y en azul oscuro la predicción. Como vemos en las figuras 23 y 24, el resultado de la predicción recoge bastante bien la tendencia de la serie, sobre todo en los primeros valores. Observamos que la bajada abrupta difícil de predecir usando modelos Holt-Winters ha sido correctamente intuida por el modelo.

En cuanto al ajuste, nos indica lo que ya sabíamos; es un modelo con bajo AIC, lo que quiere decir el modelo no tiene un gran ajuste a los datos de entrenamiento, dando lugar a una mayor generalización. Aun así, la predicción es muy buena.

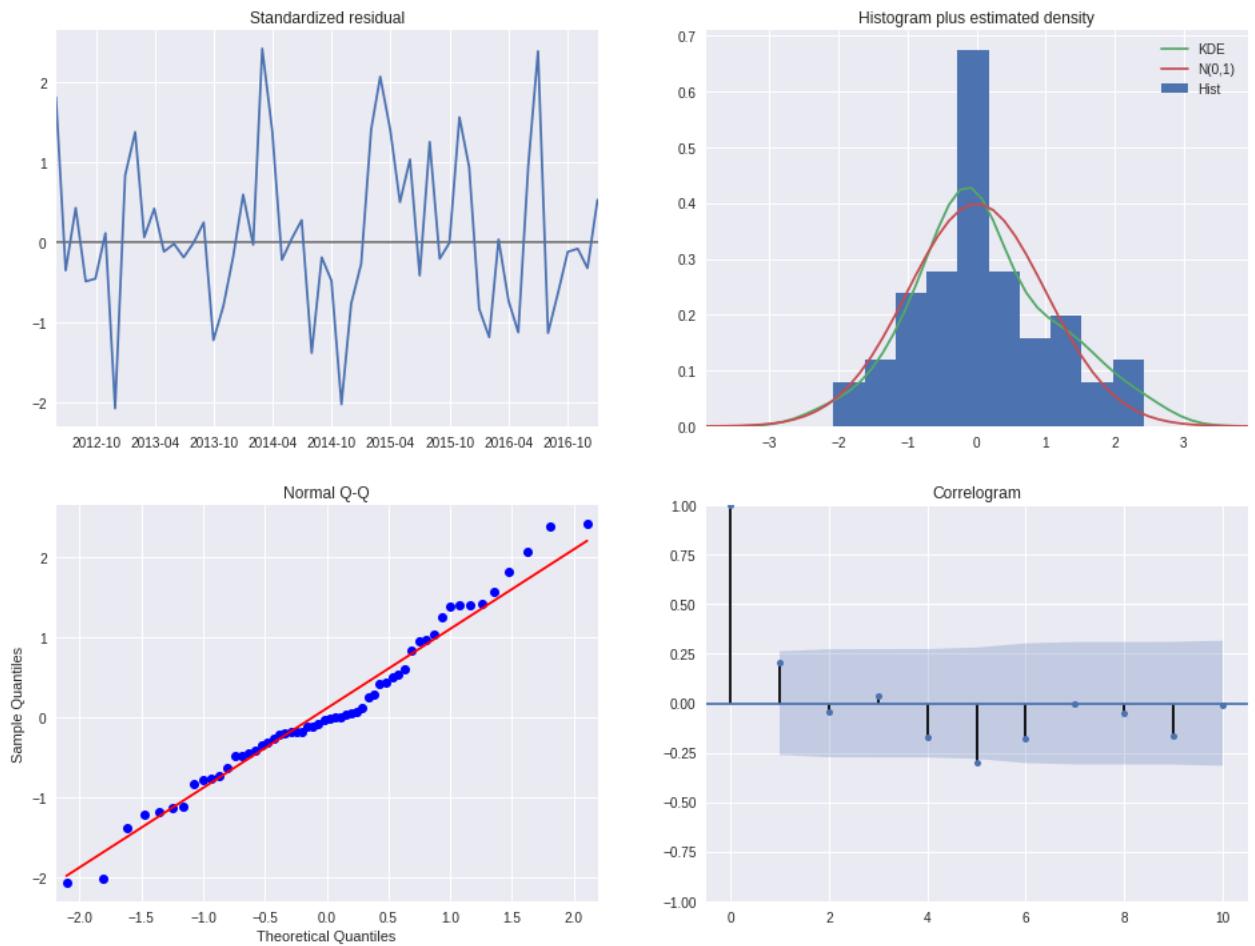


Figura 25. Residuos de $ARIMA(0,0,0)(4,0,3)_{10}$

En la figura 25 tenemos una representación gráfica de los residuos. En la primera imagen podemos ver los residuos estandarizados en el tiempo. Los residuos no deben sobrepasar el límite de 2 desviaciones típicas pero observamos que hay unos cuantos picos que sobrepasan este valor. Aceptaremos estos picos debido a la aleatoriedad de la serie.

Esto, apoyado en que el histograma resultado de los residuos del modelo tiene una forma laplaciana que podríamos considerar cercana a normal y que el Q-Q Plot no se desvía demasiado de la normalidad pero presenta fluctuaciones en los cuantiles centrales, nos hace elegir no descartar el modelo.

El correlograma obtenido también es correcto. Tenemos un valor que sobrepasa el límite de confianza pero lo podemos descartar, ya que no es un valor realmente significativo.

Modelo $ARIMA(1,1,0)(2,0,2)_8$ Este modelo presenta tanto parte no estacional como parte estacional.

ARIMA(1, 1, 0)(2, 0, 2) con periodicidad 8

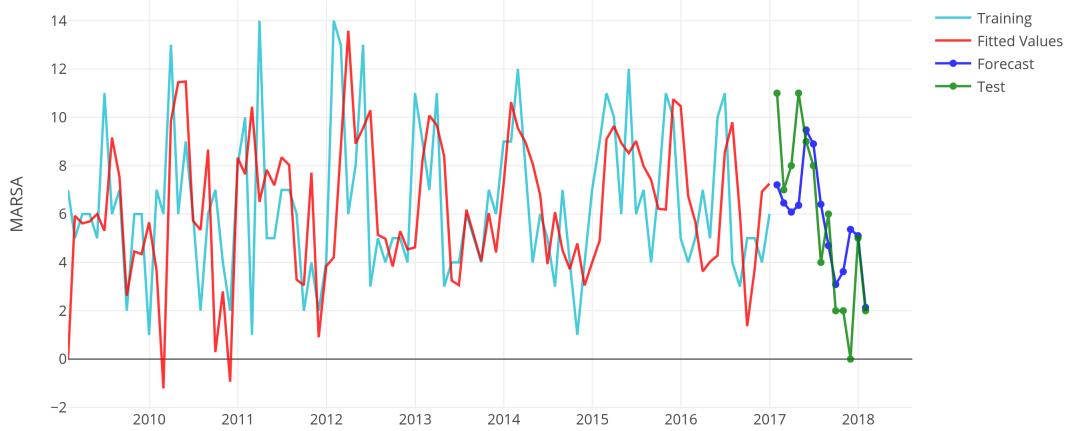


Figura 26. ARIMA(1, 1, 0)(2, 0, 2)₈

ARIMA(1, 1, 0)(2, 0, 2) con periodicidad 8



Figura 27. Predicción en ARIMA(1, 1, 0)(2, 0, 2)₈

Como podemos observar en la figura 26, este modelo se ajusta mucho mejor al subconjunto de datos de entrenamiento de la serie, pero su valor de predicción es peor. Aun así vamos a representar los valores de los residuos en el modelo para ver si tiene mejor o peor comportamiento que el caso exclusivamente estacional con periodicidad 10.

Por último realizaremos el estudio de los residuos en este modelo.

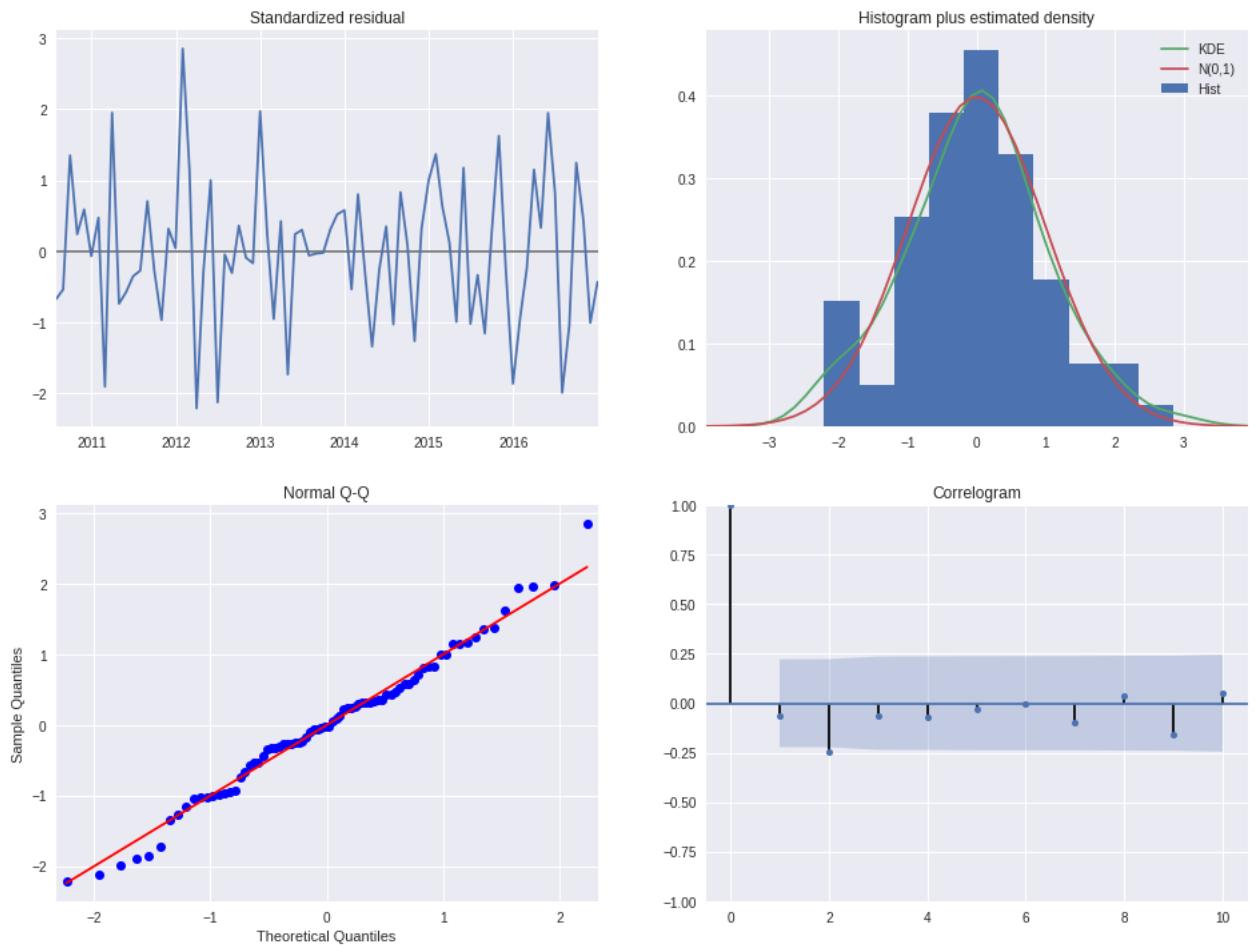


Figura 28. Residuos de $ARIMA(1,1,0)(2,0,2)_8$

En la figura 28 representamos los residuos del modelo. Si comparamos los residuos estandarizados en el tiempo con el caso anterior, esta vez obtenemos menos picos que sobrepasan los valores de 2 veces la desviación típica.

El histograma presenta una normalidad clara y el Q-Q Plot lo confirma. Por último, el correlograma cuenta con un valor que sobrepasa el límite de confianza pero lo podemos descartar de nuevo al encontrarse casi en el límite.

Por último mostraremos las predicciones de los 5 modelos seleccionados. Muchos de ellos tienen un comportamiento parecido al resultado obtenido en Holt-Winters, pero algunos como los que hemos seleccionado para su estudio en detalle tienen un mejor comportamiento y se ajustan más a los valores reales.

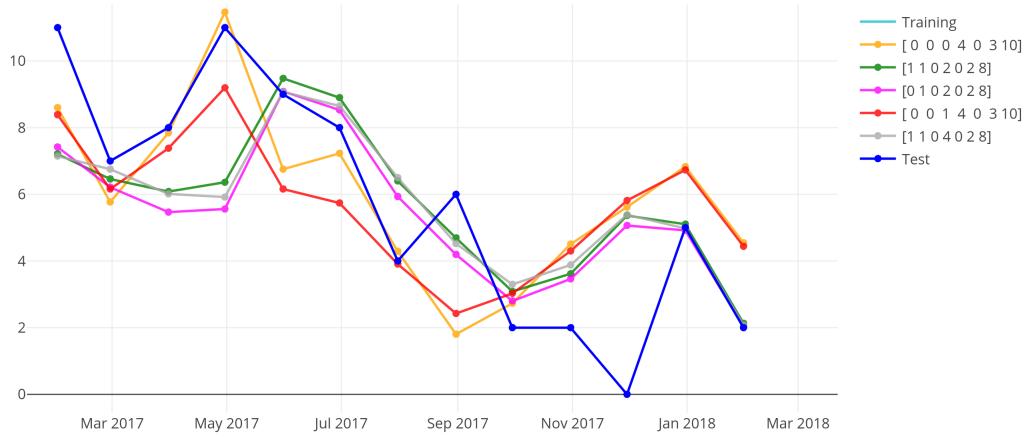


Figura 29. Predicción de modelos ARIMA seleccionados para 13 meses

Mejores modelos para 6 meses Esta vez hemos ordenado por aquellos modelos que predicen mejor en un rango más corto, concretamente en 6 meses.

Los 5 mejores modelos se muestran en la figura 30. Como vemos, el modelo que mejor valor de predicción tiene en 6 meses coincide con el mejor en 13, por lo que para abarcar un rango más amplio de modelos, vamos a pasar a estudiar el único modelo que difiere en estacionalidad de la lista.

	Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
	526 [0, 0, 0, 4, 0, 3, 10]	307.039137	323.241951	2.467012	1.375958	1.371113	7
	14061 [2, 0, 3, 2, 0, 1, 9]	418.592387	439.568987	2.915681	1.823100	2.039456	8
	18442 [3, 0, 0, 3, 1, 1, 10]	284.482098	300.244433	2.854906	1.837598	1.989042	8
	442 [0, 0, 0, 3, 1, 1, 10]	302.466210	312.592968	2.722297	1.855262	2.082300	5
	1126 [0, 0, 1, 4, 0, 3, 10]	304.758405	322.986570	2.573522	1.861413	1.665719	8

Figura 30. Modelos con mejor RMSE para 6 meses

Modelo ARIMA(2,0,3)(2,0,1)₉ Este modelo tiene unos valores de complejidad peores que el resto con periodicidad 10, por lo que tiene un mayor ajuste a los datos de entrenamiento, siendo menos general. Vemos en la figura 32 que en los primeros meses tiene una capacidad de predicción decente y que se intuye la bajada presente en el año 2017, para acabar convergiendo a la media.

ARIMA(2, 0, 3)(2, 0, 1) con periodicidad 9

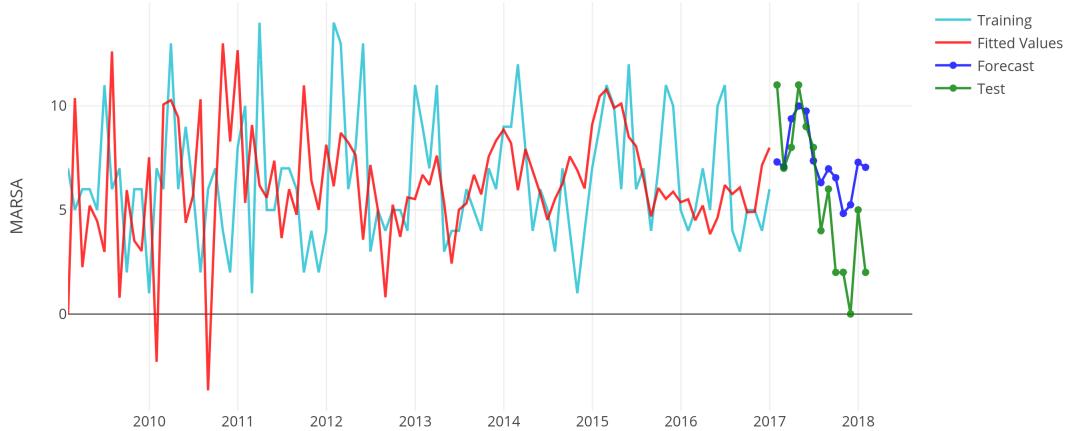


Figura 31. ARIMA(2, 0, 3)(2, 0, 1)₉

ARIMA(2, 0, 3)(2, 0, 1) con periodicidad 9

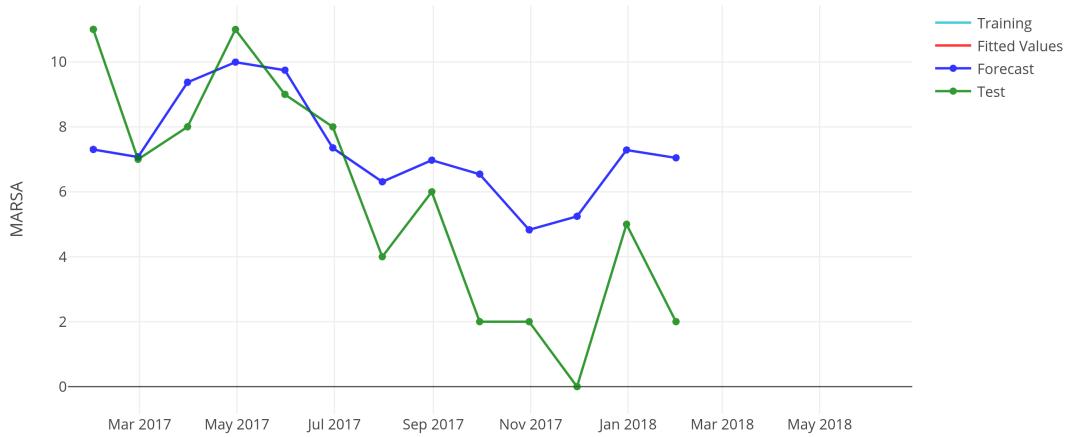


Figura 32. Predicción de ARIMA(2, 0, 3)(2, 0, 1)₉

Por otra parte, revisando los valores de residuos de la figura 33 podemos afirmar que se trata de un modelo que cumple los criterios de normalidad y de independencia en los residuos. Como vemos en el correlograma, no hay ningún lag que sobresalga del límite de confianza, el histograma es normal desplazado a la izquierda y el Q-Q Plot se puede considerar decente salvo los extremos, cosa que se repite en la mayoría de modelos.

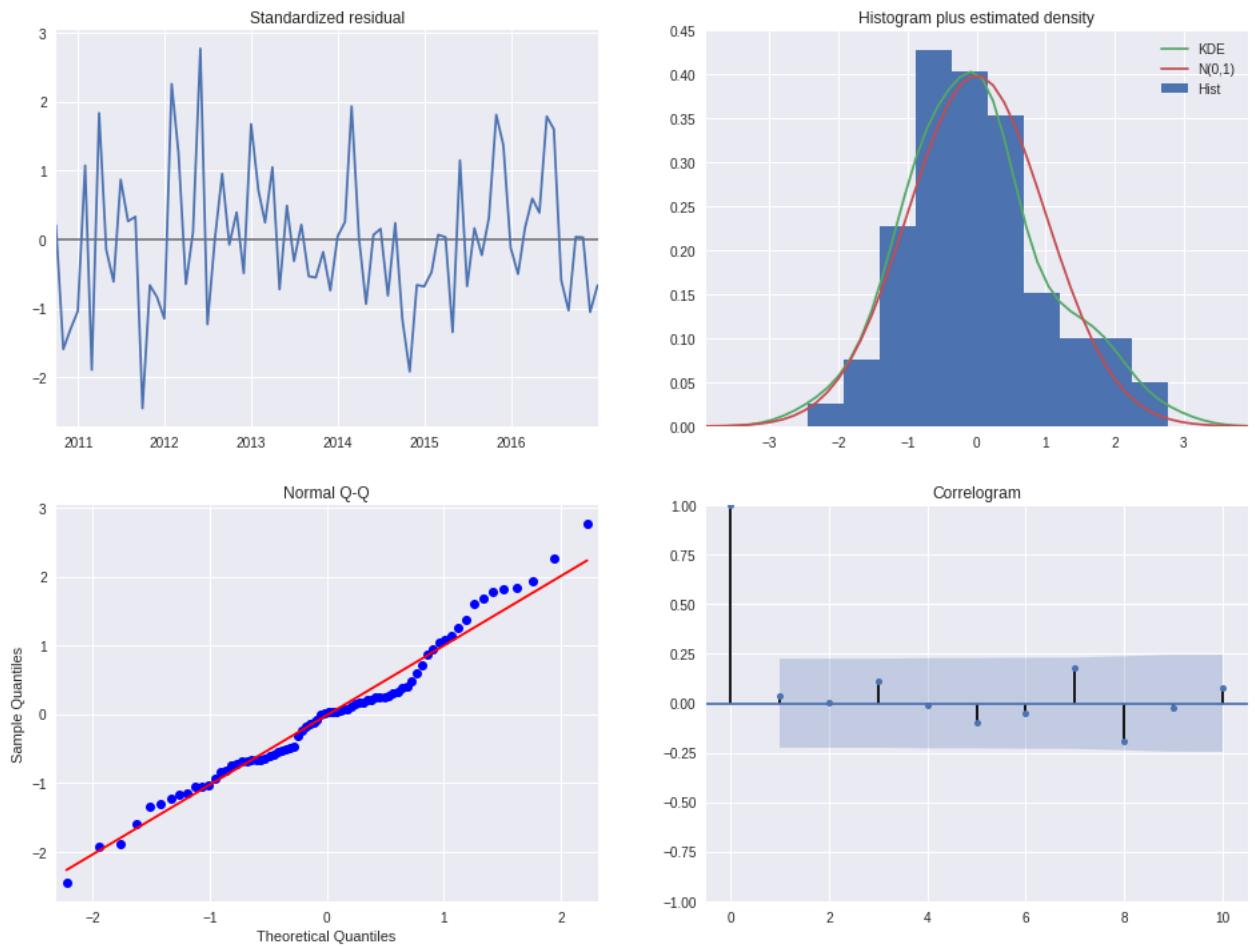


Figura 33. Residuos de $ARIMA(2,0,3)(2,0,1)_9$

Si vemos la predicción de todos los modelos ARIMA de la figura 34 tenemos que la mayoría recogen la subida desde Marzo hasta Mayo y la bajada desde Mayo hasta Septiembre, volviendo siempre a valores próximos a la media en los últimos meses, desde Septiembre hasta Enero de 2018.

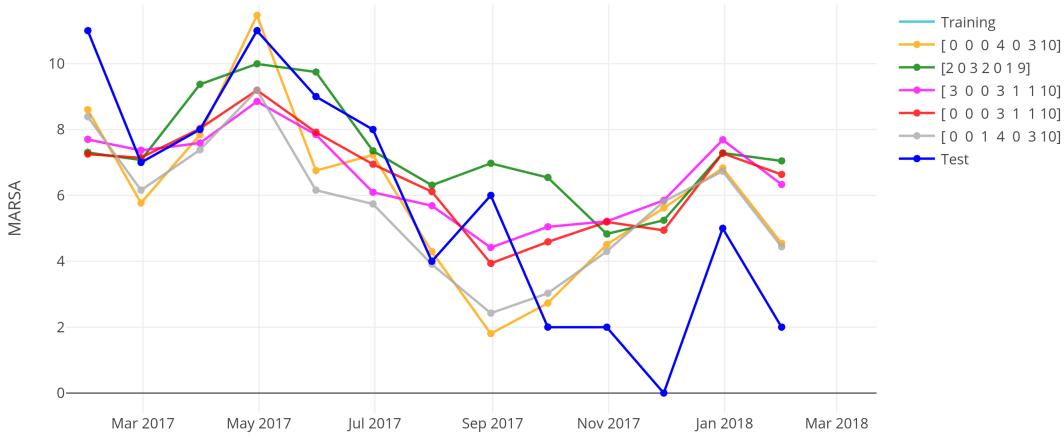


Figura 34. Predicción de modelos ARIMA seleccionados para 6 meses

Mejores modelos para 3 meses Por último, obtenemos los modelos que dan mejor resultado para 3 meses. Como vemos en la tabla de la figura, siguen apareciendo como mejores valores los modelos estacionales con periodicidad 10.

	Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
	526 [0, 0, 0, 4, 0, 3, 10]	307.039137	323.241951	2.467012	1.375958	1.371113	7
	24751 [4, 0, 1, 1, 0, 2, 7]	419.985975	441.424215	2.930903	2.188597	1.527891	8
	1126 [0, 0, 1, 4, 0, 3, 10]	304.758405	322.986570	2.573522	1.861413	1.665719	8
	18442 [3, 0, 0, 3, 1, 1, 10]	284.482098	300.244433	2.854906	1.837598	1.989042	8
	10630 [1, 1, 2, 3, 1, 0, 10]	296.678933	310.601822	2.895332	2.070299	2.004693	8

Figura 35. Modelos con mejor RMSE para 3 meses

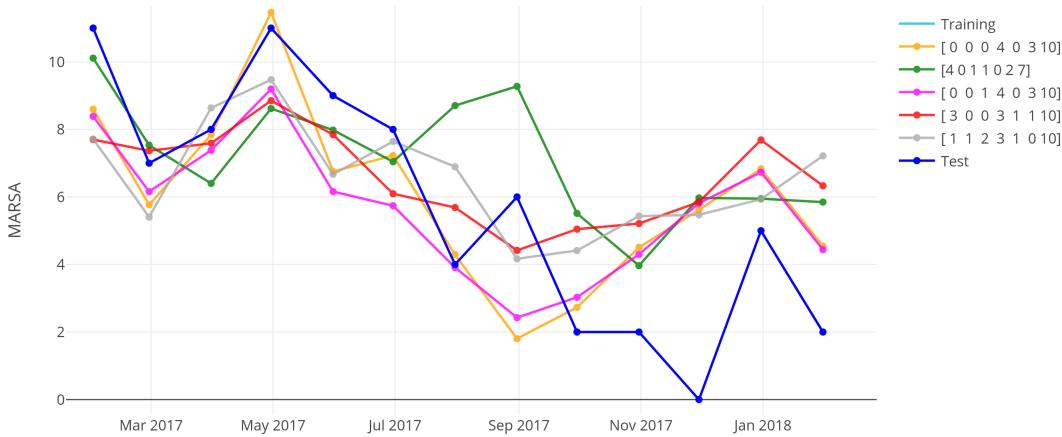


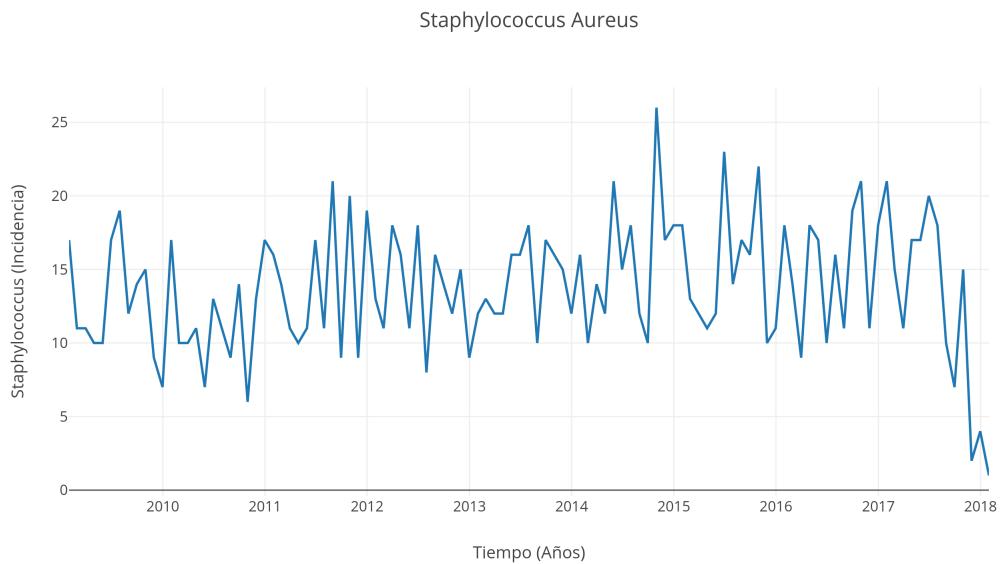
Figura 36. Predicción modelos ARIMA seleccionados para 3 meses

4. Resolución para las series *Staphylococcus Aureus* y *Levofloxacino*

A continuación, se mostrarán los resultados para las series *Staphylococcus Aureus* y *Levofloxacino*. No se repetirá el proceso completo, solo se indicarán los cambios necesarios en los filtros así como el comportamiento de la serie de forma gráfica y la predicción de los mejores modelos obtenidos.

4.1. *Staphylococcus Aureus*

La serie de Incidencia de *Staphylococcus Aureus* es la siguiente:



Como vemos en la figura 38, la serie no presenta una forma del todo normal, teniendo un ligero desplazamiento a la izquierda. El histograma no nos aporta más información que esta.

Descomposición de la serie

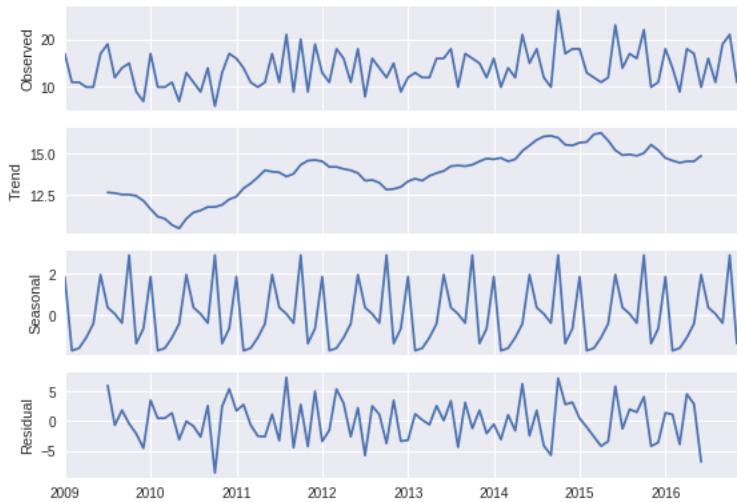


Figura 39. Descomposición de la serie

Al descomponer la serie vemos claramente una tendencia ascendente con el nivel de la serie, por lo que sería conveniente aplicar una operación de diferenciación para estabilizar la media.

ACF y PACF Representando ACF y PACF para la serie original vemos cómo en ambas tenemos un valor que pasa de los límites en el lag 8. También vemos que el resto de lags no son representativos hasta un orden muy alto, por lo que, al generar modelos muy poco parsimoniosos, descartamos la opción de tenerlos en cuenta.

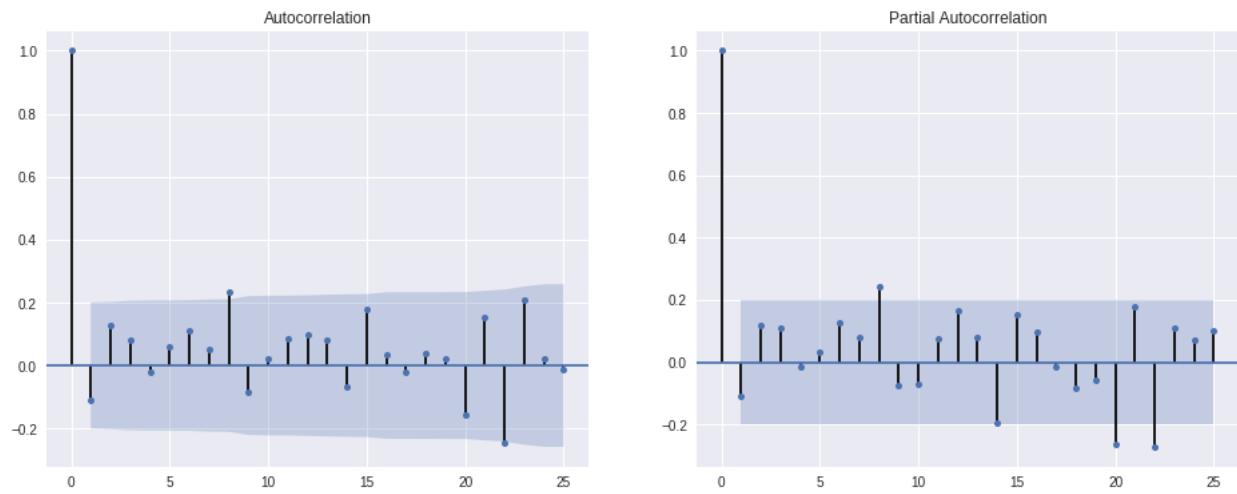


Figura 40. ACF y PACF de la serie original

Como sabemos que la serie puede no ser estacionaria debido a la subida en la tendencia que hemos detectado al hacer la descomposición, vamos a aplicar diferenciación a la serie para ver cómo varían ACF y PACF.

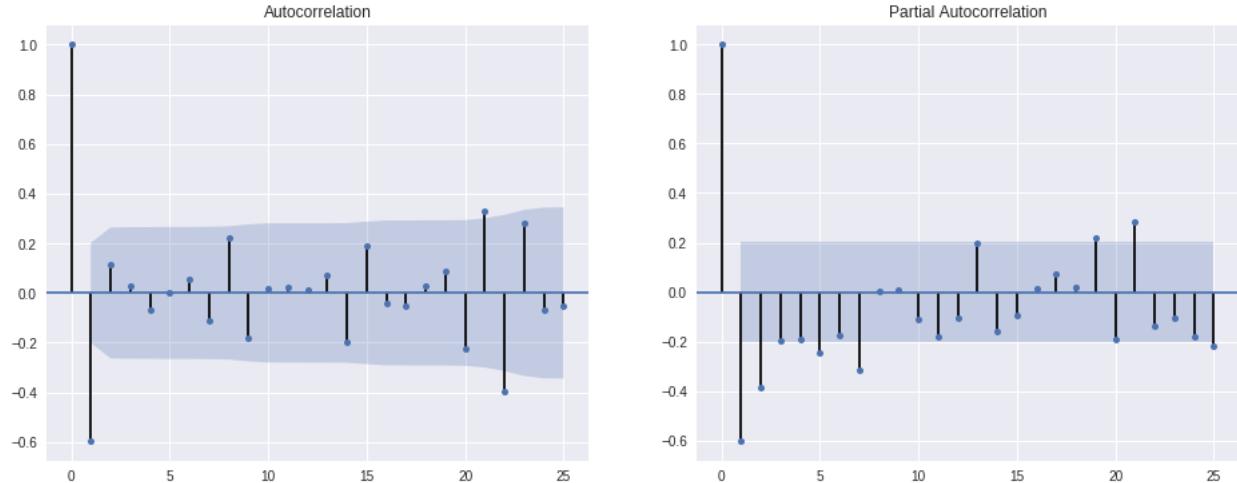


Figura 41. ACF y PACF de la serie diferenciada $d = 1$

Como vemos en la figura 41, al diferenciar la serie aparece en el ACF un valor representativo en el lag 1, lo que quiere decir que vamos a tener una fuerte dependencia entre un valor y su valor anterior. Por otra parte, la PACF muestra dos picos en los lags 1 y 2 con valores poco representativos a partir del lag 5. Podemos descartar estos lags al presentar un valor por encima del límite de forma muy ajustada. Por último vemos que el lag 7 vuelve a ser representativo, y el lag 14 y cercanos están próximos a dos desviaciones típicas; pero el

valor que de verdad nos llama la atención es en el lag 21. Parece que tenemos un caso de estacionalidad con periodicidad próxima a 7.

La parte no estacional del modelo podría ser una (1, 1, 2), al presentar diferenciación y ser un modelo mixto bastante común en la práctica, pero vamos a ver si esto se corresponde con la experimentación que llevamos a cabo en este trabajo.

Filtrado de modelos Realizando el mismo proceso de filtrado por modelos más simples tenemos el siguiente conjunto de modelos:

Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
12394 [2, 0, 0, 3, 0, 2, 10]	389.486710	406.757774	6.381706	5.921539	5.426506	7
12382 [2, 0, 0, 3, 0, 1, 10]	387.662205	402.774387	6.436770	5.529223	5.003679	6
15332 [2, 1, 0, 2, 1, 2, 8]	412.737578	428.376324	6.739195	3.982441	4.395704	8
24214 [4, 0, 0, 1, 1, 2, 10]	397.690334	415.085432	6.759484	4.386351	4.371058	8
9392 [1, 1, 0, 3, 0, 2, 8]	421.529833	437.269300	6.760113	4.279195	4.587283	7
Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
15332 [2, 1, 0, 2, 1, 2, 8]	412.737578	428.376324	6.739195	3.982441	4.395704	8
2012 [0, 0, 3, 1, 1, 2, 8]	390.758566	406.295120	6.783457	4.058222	4.472575	7
8012 [1, 0, 3, 1, 1, 2, 8]	392.746634	410.502696	6.783359	4.058661	4.477289	8
18212 [3, 0, 0, 1, 1, 2, 8]	406.409148	422.247908	6.789247	4.073863	4.531030	7
24212 [4, 0, 0, 1, 1, 2, 8]	408.355402	426.456841	6.819259	4.213886	4.725857	8
Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
24214 [4, 0, 0, 1, 1, 2, 10]	397.690334	415.085432	6.759484	4.386351	4.371058	8
15332 [2, 1, 0, 2, 1, 2, 8]	412.737578	428.376324	6.739195	3.982441	4.395704	8
18214 [3, 0, 0, 1, 1, 2, 10]	395.698352	410.919063	6.765416	4.407346	4.397549	7
2012 [0, 0, 3, 1, 1, 2, 8]	390.758566	406.295120	6.783457	4.058222	4.472575	7
8012 [1, 0, 3, 1, 1, 2, 8]	392.746634	410.502696	6.783359	4.058661	4.477289	8

Figura 41. Mejores modelos ARIMA (RMSE) para 13, 6 y 3 meses

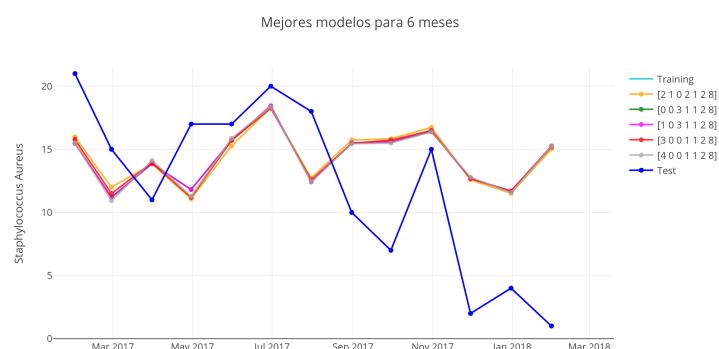
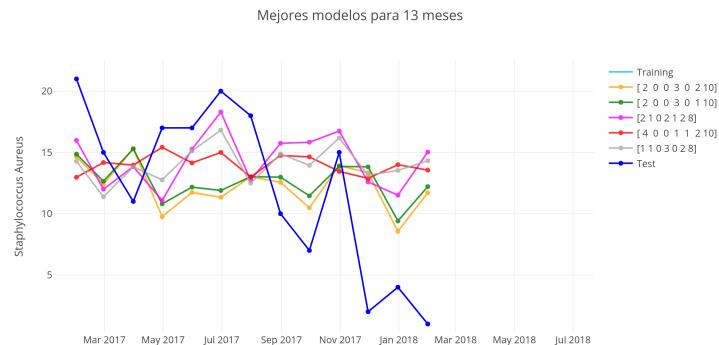
Como vemos en la figura 41, los mejores modelos en cuanto a RMSE son estacionales de periodicidad 8 y 10. Esto coincide con el estudio realizado con la serie MARSA, por lo que, aunque no estudiemos la relación o dependencia entre series podemos afirmar que se trata de dos series con un comportamiento parecido, aunque no idéntico.

Los resultados de predicción son bastante más pobres que en MARSA y, aunque conseguimos un error bastante menor que con el método Holt-Winters, no acaba de representar correctamente los valores de test. En la Figura 42 tenemos los mejores modelos Holt-Winters, en los que el valor mínimo de RMSE es superior a 7.

	Model	BIC	AIC	RMSE
0	[8, add, mul]	319.239778	288.467600	7.312901
1	[10, add, add]	357.299278	321.398404	7.361431
2	[12, add, add]	342.883061	301.853490	7.387579
3	[12, add, mul]	342.939691	301.910120	7.391547
4	[3, add, add]	293.175140	275.224703	7.411613
5	[5, add, add]	323.589560	300.510426	7.432302
6	[6, add, add]	313.153745	287.510264	7.454852
7	[6, add, mul]	314.703540	289.060058	7.484653
8	[8, mul, add]	335.406617	304.634438	7.502223
9	[8, add, add]	302.429170	271.656992	7.585848
10	[11, add, mul]	346.371576	307.906353	7.655852

Figura 42. Mejores Modelos Holt-Winters

Podemos ver en las predicciones en la figura 43 que no se ajustan muy bien a los datos de prueba. En todos los casos son modelos en los que hay una ligera fluctuación en la media, por lo que cuadran bien desde Mayo hasta Agosto de 2017, pero esto no intuye bien la bajada meses posteriores. Puede que esto sea debido a que se utilizan pocos niveles de autorregresión, pero si estamos buscando los modelos siguiendo el principio de parsimonia, estos niveles más complejos se deben descartar para dejar paso a modelos más generales.



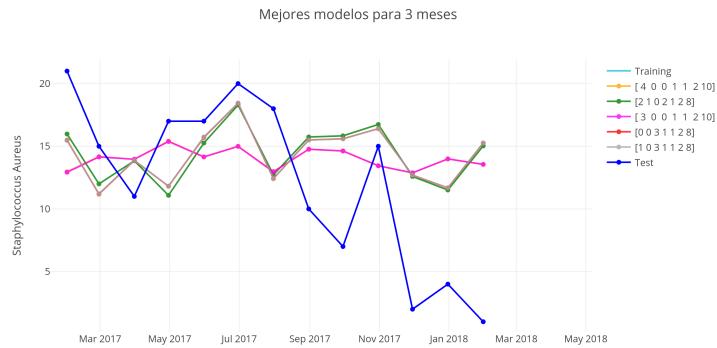


Figura 43. Mejores predicciones según RMSE

La limitada capacidad predictiva de los modelos ARIMA para esta serie está, muy probablemente, asociada a la caída abrupta en la incidencia de nuevos casos de infección por *Staphylococcus Aureus* que aparece hacia el final de la serie, en 2017. Se trata de un comportamiento distinto al observado en el resto de la serie, y por tanto, difícil de predecir con un modelo lineal del tipo ARIMA.

Estos casos suelen deberse a una intervención exógena que altera la dinámica de la serie. En el dominio de Microbiología, tales cambios pueden ocurrir en respuesta a una modificación en la política de administración de antibióticos seguida por la Comisión de Infecciones del hospital. Este cambio, a su vez, afecta a las poblaciones de bacterias, produciendo el cambio observado en la serie.

Es posible modelar las intervenciones y superponerlas al modelo ARIMA, lo que seguramente permitiría mejores predicciones. También se pueden aplicar técnicas de detección de cambios que confirmen la existencia de un cambio significativo en la dinámica de la serie. Tanto el modelado de intervenciones, como las técnicas de detección de cambios, escapan al alcance de este TFG, pero es recomendable estudiarlas en profundidad como trabajo futuro.

4.2. Levofloxacino

Al estudiar la serie de Levofloxacino estamos ante un proceso controlado, por lo que es menos atractivo obtener información útil de dichos datos. Aun así, tendremos un ejemplo claro de serie estacional y podremos debatir si es necesario el uso de modelos ARIMA o podría bastar con alternativas más sencillas.

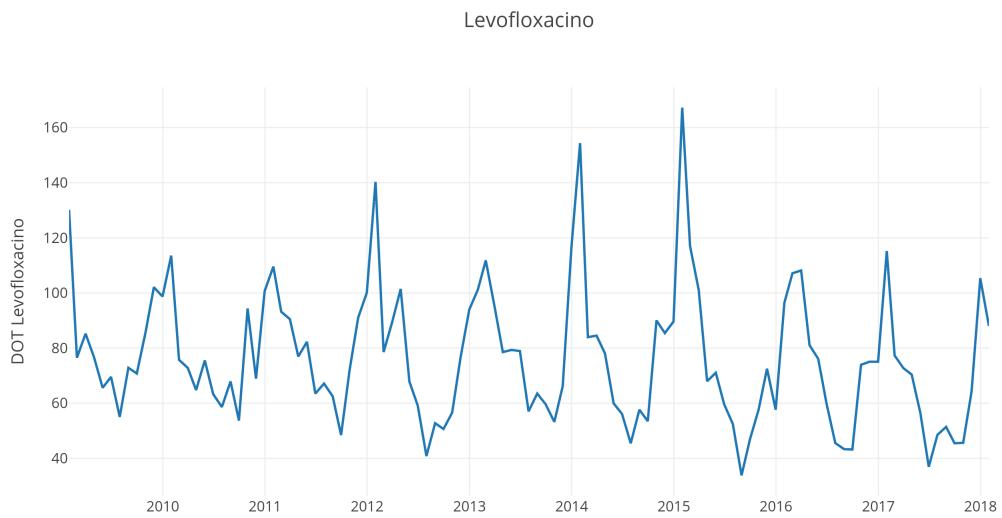


Figura 44. Serie DOT de Levofloxacino

A simple vista, la serie del Levofloxacino es claramente estacional.

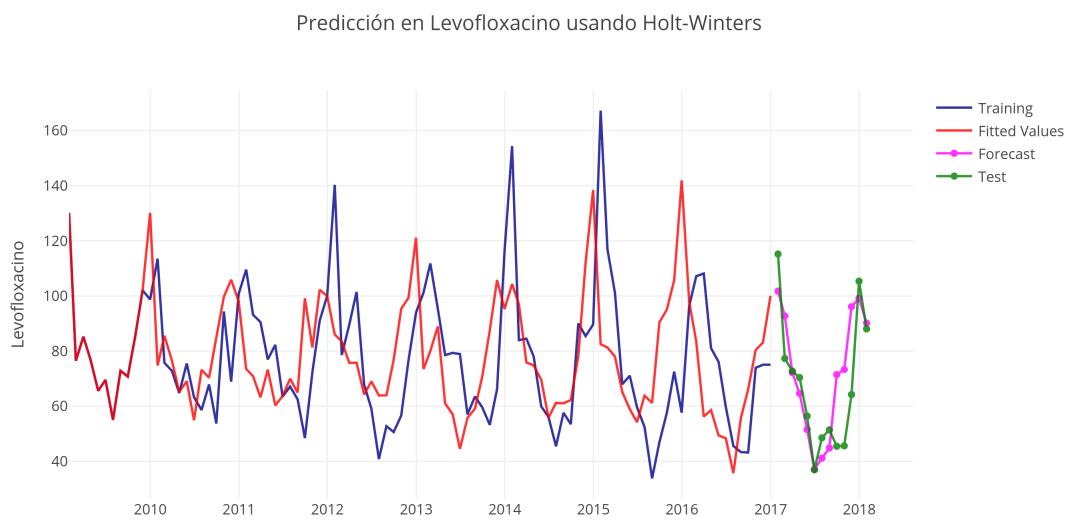


Figura 45. Predicción de Levofloxacino con Holt-Winters

En la figura 45 vemos la predicción con el método Holt-Winters usando un modelo aditivo tanto en tendencia como en estacionalidad con un periodo de estacionalidad de 11. Con un

método muy simple como el suavizado exponencial se obtiene un resultado realmente bueno.

Podemos ver en la figura 46 que, al usar modelos ARIMA en esta serie obtenemos resultados muy parecidos a los obtenidos con Holt-Winters.

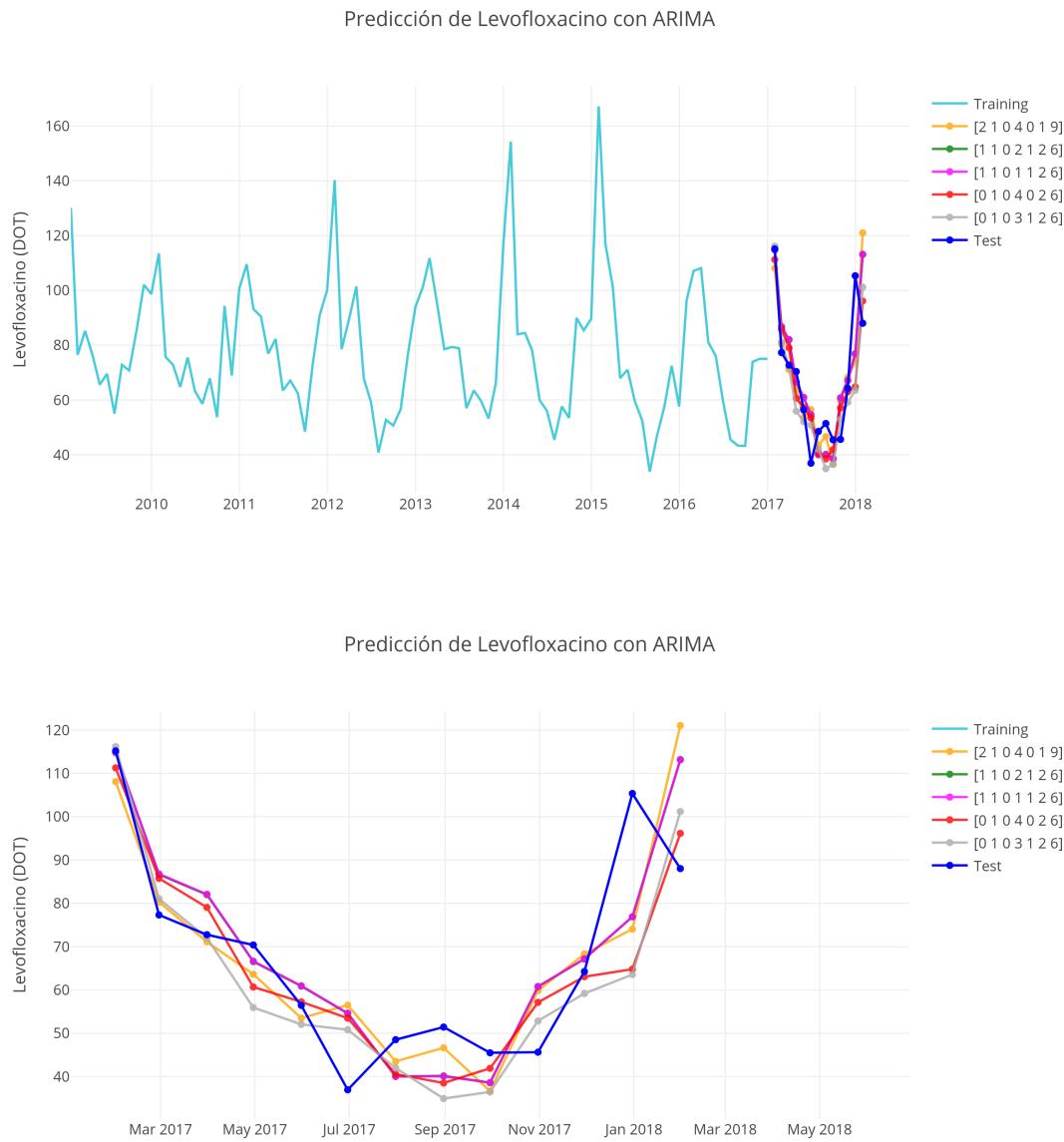


Figura 46. Predicción de Levofloxacino con ARIMA

Por último vamos a hacer una comparación de los resultados obtenidos usando el método Holt-Winters y los diferentes modelos ARIMA.

Model		BIC	AIC	RMSE
0	[11, add, add]	687.136983	648.671760	15.372036
1	[12, add, mul]	584.716463	543.686892	15.930038
2	[11, add, mul]	686.498243	648.033020	16.539003
3	[12, add, add]	593.476766	552.447195	16.815266
4	[6, mul, add]	625.656136	600.012654	22.182445
5	[6, add, add]	627.431517	601.788035	22.671206
6	[6, add, mul]	615.256135	589.612654	22.690526
7	[2, mul, add]	634.891795	619.505706	23.010694
8	[7, add, mul]	667.256328	639.048498	23.398406
9	[9, mul, mul]	635.633148	602.296622	23.421806
10	[8, mul, mul]	630.121769	599.349591	23.457869
11	[9, add, mul]	637.451783	604.115257	23.498515
12	[9, add, add]	644.945834	611.609308	23.609994
13	[8, add, mul]	632.043634	601.271455	24.032106
14	[8, add, add]	641.759977	610.987799	24.214855
15	[10, mul, mul]	638.837646	602.936772	24.252655

Figura 47. Mejores modelos Holt-Winters por RMSE

Como vemos en la tabla de la figura 47, los mejores modelos en cuanto a RMSE tienen un periodo de estacionalidad próximo a 11. También podemos observar algunos modelos con estacionalidad 6 cuyo valor de predicción es bastante bueno en comparación con otros métodos más simples.

Para comparar estos valores con los obtenidos en los resultados de los modelos ARIMA generados, representaremos de nuevo las tablas de mejores modelos por RMSE para 13, 6 y 3 meses.

Model		BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
3222	[0, 1, 0, 1, 1, 3, 6]	623.153689	634.396165	11.947705	7.821195	7.940220	6
3342	[0, 1, 0, 2, 1, 3, 6]	625.061518	638.552490	11.993259	7.903937	8.120524	7
3282	[0, 1, 0, 2, 0, 3, 6]	680.333924	694.318324	12.203095	9.928678	9.807987	6
15272	[2, 1, 0, 2, 0, 2, 8]	680.494363	696.901001	12.636949	13.319224	13.415139	7
15499	[2, 1, 0, 4, 0, 1, 7]	594.183170	611.578268	12.781120	15.650176	17.497160	8

Figura 48. Mejores modelos ARIMA por RMSE para 13 meses

Model		BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
3222	[0, 1, 0, 1, 1, 3, 6]	623.153689	634.396165	11.947705	7.821195	7.940220	6
3342	[0, 1, 0, 2, 1, 3, 6]	625.061518	638.552490	11.993259	7.903937	8.120524	7
3450	[0, 1, 0, 3, 1, 2, 6]	635.488491	649.064570	14.757830	8.272284	7.488418	7
15501	[2, 1, 0, 4, 0, 1, 9]	512.035837	528.380247	14.967230	8.647215	5.195218	8
3570	[0, 1, 0, 4, 1, 2, 6]	587.408070	602.628781	15.350230	8.793656	7.956008	8

Figura 49. Mejores modelos ARIMA por RMSE para 6 meses

Model	BIC	AIC	RMSE	RMSE6	RMSE3	Parsimony
15501 [2, 1, 0, 4, 0, 1, 9]	512.035837	528.380247	14.967230	8.647215	5.195218	8
9330 [1, 1, 0, 2, 1, 2, 6]	670.505831	684.490231	13.707524	9.179832	6.845643	7
9210 [1, 1, 0, 1, 1, 2, 6]	668.537419	680.191086	13.746214	9.235677	6.910467	6
3510 [0, 1, 0, 4, 0, 2, 6]	635.864275	651.703034	14.095694	8.928946	7.424730	7
3450 [0, 1, 0, 3, 1, 2, 6]	635.488491	649.064570	14.757830	8.272284	7.488418	7

Figura 50. Mejores modelos ARIMA por RMSE para 3 meses

Los mejores valores coinciden en estacionalidad con los modelos Holt-Winters y se mueven en un rango muy próximo de complejidad (viendo el valor obtenido en el AIC de los modelos). Esto hace que se refuerce la idea de estacionalidad en 6 meses aportada por los modelos Holt-Winters. Los modelos Holt-Winters están a la par en predicción con los ARIMA en este caso.

Podemos comprobar, representando ACF y PACF, que los lags con picos más significativos se encuentran cada 6 meses, manteniendo valores significativos en lags cercanos a estos. Comprobamos que la estacionalidad se repite en para el lag 12 y 18, lo que confirma que los modelos que se están obteniendo describen correctamente el comportamiento del proceso que genera la serie.

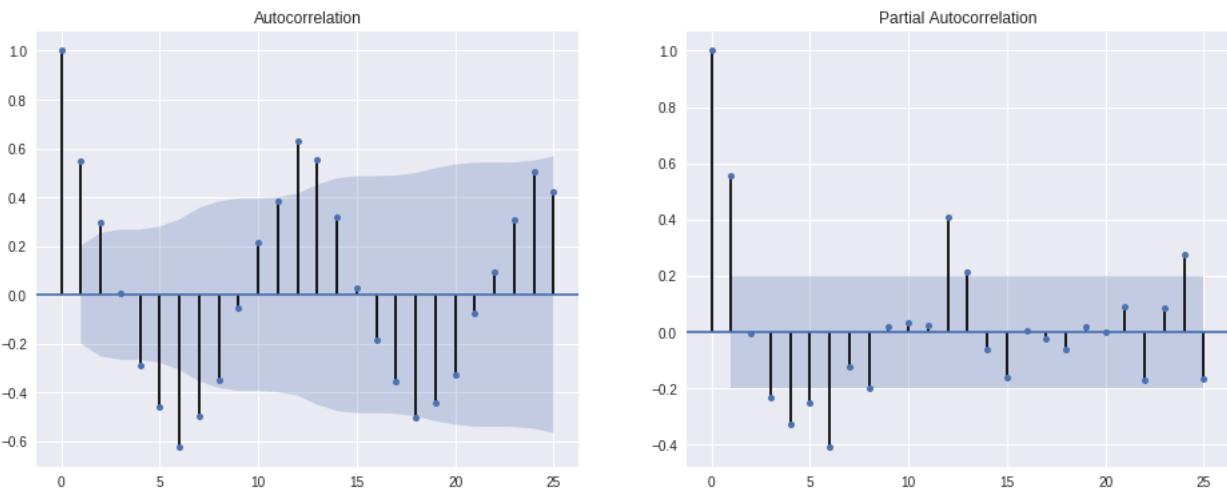


Figura 51. ACF y PACF para el Levofloxacino

¿Por qué es útil tener un modelo ARIMA de una serie temporal de administración de antibióticos, cuando es algo que viene predeterminado por el protocolo de aplicación seguido por los clínicos?

No olvidemos que el objetivo final es obtener las mejores predicciones posibles, y que para ello es necesario tener en cuenta las covariables que afectan a la incidencia de infecciones bacterianas. Un modelo de función de transferencia basado en regresión dinámica permitiría tener en cuenta las relaciones causales entre series y, previsiblemente, obtener predicciones más precisas en las series de incidencia bacteriana. Ese tipo de análisis más avanzado se está implementando en un TFM complementario. La metodología que se sigue en ese trabajo exige la obtención previa de un modelo ARIMA para cada una de las series relacionadas, que es lo que proporcionamos en este trabajo.

5. Conclusiones y vías futuras

A lo largo de este estudio experimental se han cubierto los objetivos concretos planteados en la sección 3.1.

En primer lugar, se han construído y validado modelos ARIMA para las series de incidencia de bacterias y de aplicación de antibióticos. Además, se han construído modelos más simples que el ARIMA para comparar su habilidad predictiva con el primero. Así se han obtenido modelos Naïve, modelos de Media Simple y Modelos de Suavizado Exponencial con el método Holt-Winters. Estas comparaciones servirán para discutir la necesidad de aplicar modelos de predicción más complejos o usar estos modelos más simples. Así pretendemos comprobar experimentalmente en qué medida se cumplen las expectativas teóricas sobre el modelo ARIMA como el mejor predictor.

En particular, queremos saber si en este dominio concreto basta usar métodos simples para obtener predicciones útiles para el clínico y razonablemente precisas. Ello es importante, porque la metodología Box-Jenkins de construcción de modelos ARIMA es relativamente compleja, costosa en tiempo y difícil de automatizar, dependiendo en buena medida de juicios visuales subjetivos del analista.

Además, se ha intentado automatizar el proceso de selección de modelos ARIMA para series temporales siguiendo la metodología Box-Jenkins, comparando los resultados obtenidos con otros modelos más simples y combinando técnicas de fuerza bruta con poda basada en una serie de criterios estadísticos.

Uno de los objetivos marcados era el de ofrecer un conjunto reducido de modelos ARIMA al personal médico para tener una visión del comportamiento de las series en el tiempo. Este objetivo se ha tratado, tanto desde un punto de vista clásico, analizando funciones de autocorrelación como desde un punto de vista orientado a las pruebas.

Por otra parte, el estudio nos ha aportado la capacidad de saber en qué caso sería bueno aplicar modelado ARIMA u optar por modelos Holt-Winters (muy extendidos en microbiología) más simples para realizar predicciones. Sabemos, después de la experimentación, que series con muy marcada estacionalidad se pueden modelar usando Holt-Winters, obteniendo resultados realmente buenos, pero no para procesos que generan series más próximas

a ruido blanco, más complejas y que requieren de transformaciones para obtener buenas predicciones. En estas últimas entraría el uso de ARIMA.

En cuanto a trabajos futuros, este estudio proporciona un primer paso para aplicar técnicas de detección de cambios en la dinámica de las series y para modelar las intervenciones que dan lugar a estos cambios.

También aportamos los primeros pasos para realizar un estudio más complejo utilizando covariables, construyendo funciones de transferencia basadas en regresión dinámica.

Finalmente, otra de las vías futuras separadas de este trabajo es la de estudiar el mínimo número de datos necesarios para generar modelos ARIMA y comprobar cuál es el horizonte temporal de validez de los modelos. Esencialmente, se trataría de ver durante cuánto tiempo, las dinámicas modeladas a partir de un conjunto limitado de datos, siguen proporcionando predicciones fiables sin necesidad de construir un nuevo modelo ARIMA mensualmente, cada vez que la serie es extendida en un nuevo punto. Se podría diseñar un experimento seleccionando conjuntos de datos de predicción de forma incremental.

6. Bibliografía

Referencias

- [1] PANKRATZ, A., *Forecasting with Dynamic Regression Models* (1991) John Wiley & Sons, Inc.
- [2] HYNDMAN, ROB J., *Thoughts on the Ljung-Box test* 24 de febrero de 2014. Recuperado de <https://robjhyndman.com/hyndtsight/ljung-box-test/>
- [3] ANALYTICS VIDHYA, *7 methods to perform Time Series forecasting (with Python codes)*. 8 de febrero de 2018. Recuperado de <https://www.analyticsvidhya.com/blog/2018/02/time-series-forecasting-methods/>
- [4] CHARLES ZAIONTZ, *Invertibility of MA(q) Processes*. Recuperado de <http://www.real-statistics.com/time-series-analysis/moving-average-processes/invertibility-ma-processes/>
- [5] KAGGLE, *SARIMAX on mean visits*. Diciembre de 2017. Recuperado de <http://www.real-statistics.com/time-series-analysis/moving-average-processes/invertibility-ma-processes/>
- [6] HYNDMAN, ROB J., *Forecasting: principles and practice*. 17 de abril de 2016. Recuperado de <https://www.otexts.org/fpp>

- [7] HATALIS, KOSTAS, *Tutorial: Multistep Forecasting with Seasonal ARIMA in Python*. 12 de abril de 2018. Recuperado de <https://www.datasciencecentral.com/profiles/blogs/tutorial-forecasting-with-seasonal-arima>