COVID-19: ¿realmente debemos alarmarnos?

Jorge Huertas Martín del Olmo

Álvaro Villadangos del Río

1. Introducción

Cuando tuvimos que hacer la propuesta para este trabajo, a mediados de febrero, el COVID-19 era un tema de actualidad, pero que veíamos como lejano y que en ningún caso imaginamos que fuese a tener la trascendencia que, desgraciadamente, ha tenido a nivel mundial. Una de las claves de la propagación del virus se debe a la gran cantidad de enfermos asintomáticos[1] (aquellos que tienen el virus y pueden contagiarlo a otras personas, pero no muestran síntomas), así como haber desoído las advertencias y recomendaciones de la Organización Mundial de la Salud (OMS), que ya el 30 de enero había declarado el COVID-19 como una emergencia sanitaria de preocupación internacional[2], y que, finalmente, el 13 de marzo lo declaró como pandemia global.

Actualmente, cada país ha planteado la lucha contra el virus de formas similares pero distintas, abogando unos por los test masivos, otros por el distanciamiento social, y otros por el aislamiento total de la población. Sin embargo, Tedros Adhanom Ghebreyesus, Director General de la OMS, ha afirmado que aplicar o centrarse solo en una de estas medidas no es suficiente, sino que es necesaria una combinación de todas ellas.[3]

Los efectos más evidentes de la pandemia son el trágico número de víctimas que ya se ha cobrado, así como de los riesgos para la salud y seguridad pública que ha originad. Pero por otro lado, es muy importante el impacto que ha tenido el coronavirus COVID-19 en la sociedad y en la economía. Las medidas que muchos países están tomando han forzado por un lado a despedir (y en España a realizar ERTEs) a una gran cantidad de empleados, y los que han podido mantener su puesto se ven obligados a teletrabajar, con los problemas y riesgos de ciberseguridad que ello conlleva.[4]

El escenario futuro que se nos plantea es extremadamente incierto, con los gobiernos de todos los países tomando medidas diariamente y reaccionando continuamente ante las últimas informaciones, tratando de erradicar el virus lo antes posible. Por ello, un análisis de la serie temporal de los contagiado por COVID-19 a nivel mundial y el intento de hacer una predicción de cómo va a evolucionar en el futuro nos parece muy interesante.

2. Material y Métodos

La base de datos utilizada

Fuente de información empleada (de dónde se han obtenido los datos).

Análisis inicial de los datos: periodo considerado, periodicidad de los datos, presencia de estacionalidad, tendencia, etc.

Problemas identificados en los datos (si es que los hubo) y cómo se han solucionado. Por ejemplo, si hay atípicos, valores faltantes, etc.

Herramienta de análisis: en nuestro caso RStudio (hay que indicarlo). Pero es necesario indicar también (y referenciar correctamente) los paquetes que se han usado y para qué se han empleado.

Modelos que se han probado y por qué, así como la metodología seguida para compararlos.

Los gráficos/figuras se numerarán consecutivamente con números romanos, en negrita, tamaño 9 y título debajo de la figura: Ejemplo:

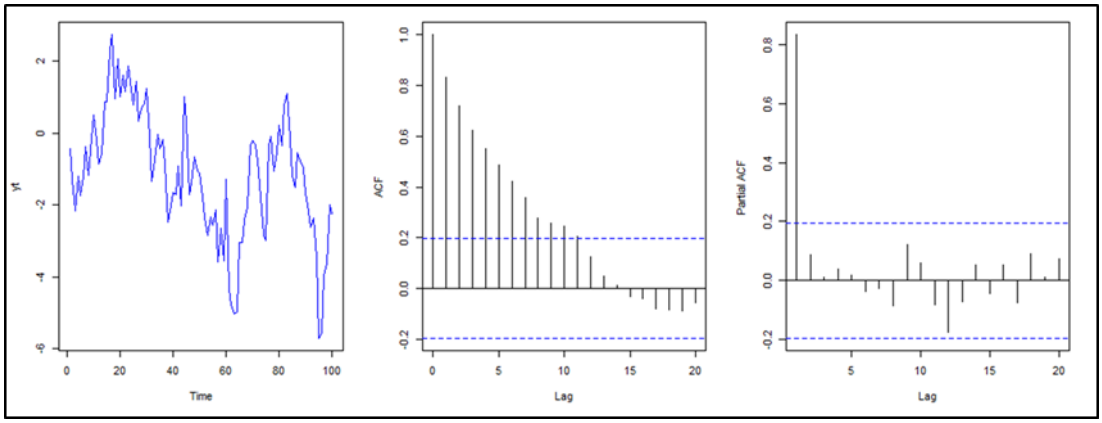


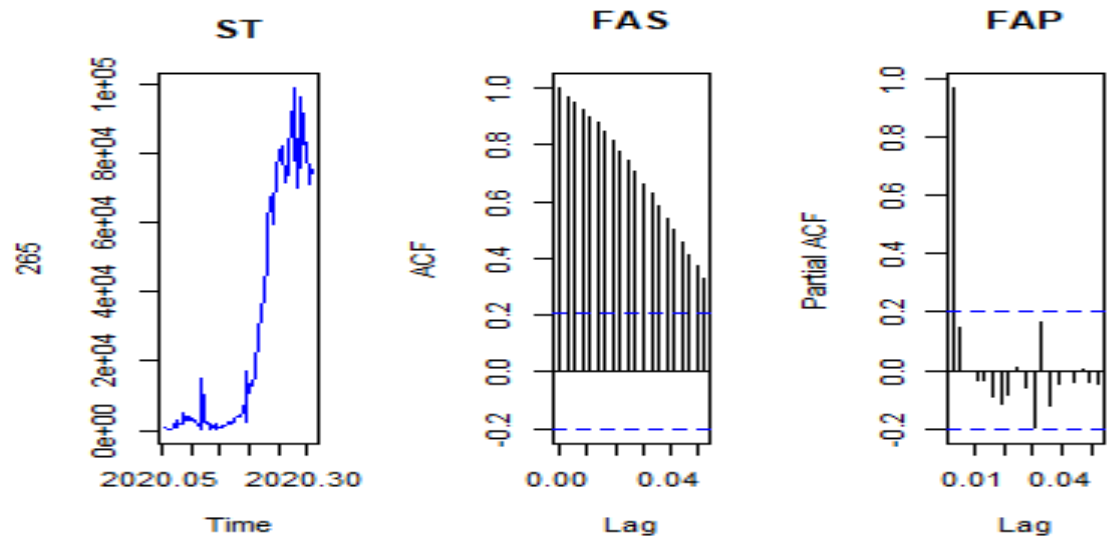
Figura 1: Serie temporal con su FAS y FAP

Extensión máxima del apartado: 800 palabras y 3 gráficos.

Resultados y Discusión

Nuestro dataset es básicamente el registro diario de nuevos casos de COVID-19 en el mundo entero. Al ser registros diarios nos los encontramos en formato serie temporal desde el 22 de enero de 2020. Para series temporales lo más óptimo es utilizar un modelo ARIMA ajustado a los datos.

Para ello primero comprobaremos si hay NAs y haremos un plot de los datos para ver la serie.



Como vemos la serie temporal es bastante corta de apenas 3 meses o 90 días. En ella podemos

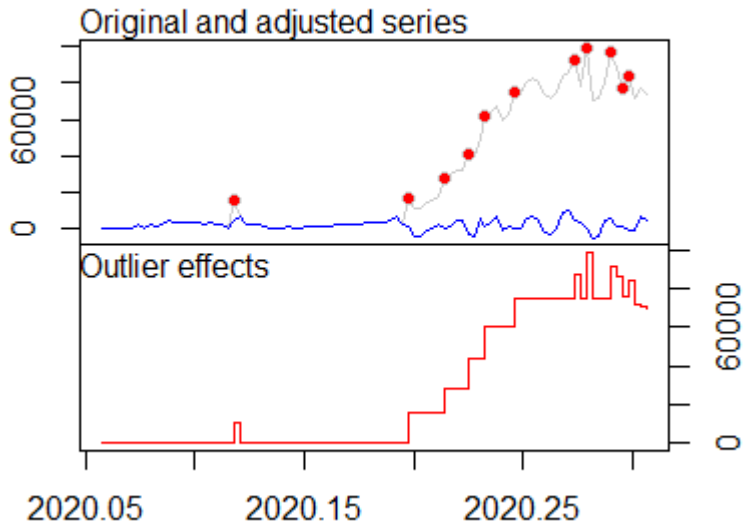
observar la evolución del virus donde se aprecia una tendencia bastante plana durante la primera mitad y luego da un salto enorme para volver a estabilizarse. Hay que tener en cuenta que en cuanto una persona se contagia no es añadida al data set automáticamente, sino que los gobiernos deben realizar pruebas para conocer estos datos. Por este motivo, el salto se debe a que en esas fechas es cuando saltaron las alarmas y los gobiernos empezaron a hacer pruebas y obtener positivos en COVID-19.

Continuamos analizando las FAS y FAP.

Respecto a las FAS vemos un decaimiento muy lento y progresivo de las componentes significativas. Este decaimiento se debe principalmente al componente AR de la serie. Así, será preciso hacer una diferenciación para posteriormente poder revisar las FAS.

Respecto a las FAP, podemos ver una única componente significativa con valor igual a 1. En R se programó para que la primera componente se relacionara consigo misma por lo que no aporta información. Por esto, no la tendremos en cuenta por tanto no hay componentes significativas en este gráfico. De esta manera, probablemente el modelo tendrá un componente AR(0).

Ahora estudiaremos los outliers. Para ello usaremos la librería tsoutliers. Su función tso() nos devolverá 11 outliers. En consola podemos ver que la propia función nos indica de que tipo es cada outlier. Aun así, haremos un plot para visualizarlo.



En este plot encontramos dos escenarios. Por un lado, la gráfica original y las series ajustadas sin el efecto de los outlier, por el otro lado el efecto de los outliers aislado. Si lo observamos atentamente la mayoría de los outliers son de cambio de nivel y luego encontramos unos cuantos aditivos. Los aditivos no son significativos, en cambio, un análisis de la sucesión de 5 oultiers de cambio de nivel durante ese periodo de tiempo podría ayudar a entender el contagio.

A raíz de este análisis no procederemos a eliminar los outliers entendemos que aportar mucha información a la serie temporal y eliminarlos podría suponer destrozar la información.

Ahora procederemos a construir el modelo.

Para empezar, debemos marcar cual es el objetivo de nuestro modelo que será predecir con éxito los próximos 20 días. Para ello, extraeremos y apartaremos los 20 últimos días que disponemos en un test set y el resto de datos históricos los mantendremos en un train set.

Para aplicar un modelo ARIMA se deben cumplir que la serie sea estacionaria, es decir, tiene que ser constante en media y varianza. Lógicamente como nuestra serie de primera como tiene tendencia no es estacionaria porque la media cambia a lo largo de la serie. Por lo tanto, deberemos hacer x número de diferenciaciones hasta que sea estacionaria. Manualmente hemos hecho 2 porque luego más tarde calculando el modelo deberemos indicar hasta que orden querremos indicar las diferenciaciones. Como ya sabemos porque lo acabamos de calcular indicaremos que sean 2 diferenciaciones.

A continuación, comprobaremos si la serie temporal tiene heterocedastidad. Con la librería forecast y la función BoxCox.lambda() veremos si es necesaria realizar la transformación. Lambda nos devuelve un valor de 1 por lo que no se presenta heterocedasticidad. Esto ya se podía anticipar a simple vista, pero nunca está de más asegurarse.

Además, la serie tampoco presenta estacionalidad. Hay que tener en cuenta que es una serie muy cortar y según transcurra el tiempo y se vaya normalizando la situación veremos si realmente tiene un componente estacional.

Ahora automatizaremos el proceso de elección del modelo con auto.arima(). Como auto.arima utiliza atajos tunearemos los parámetro para que no los haga y cree un modelo mejor. Para ello indicaremos stepwise=False, approx=False, seasonal = False, trace=True.

A continuación, intentaremos mejorar a autoarima. Para ello, partiremos de un modelo naive ARIMA(0,1,0) con dos diferenciaciones porque son las que hemos visto anteriormente que eran necesarias. Desde este punto de partida haremos un bucle for que irá iterando con todos los modelos posibles dentro de lo razonable y escogerá en de mayor AIC. El resultado y mejor modelo sería un ARIMA(1,2,2) coincidiendo en este caso con auto.arima.

Con la librería lmtest analizaremos los coeficientes. Al analizar los coeficientes, vemos que el AR(1) no es significativo, ya que, tiene un P-valor muy elevado 7,12%. Con un valor tan alto estudiamos la posibilidad de eliminarlo manualmente. Viendo los AIC la diferencia entre un ARIMA(1,2,2) y un ARIMA(0,2,2) es muy pequeña y nos quedaremos con el ARIMA(0,2,2).

```
> coeftest(arima.fit.entrenamiento) #vemos si los coeficientes con significativos

z test of coefficients:

    Estimate Std. Error  z value  Pr(>|z|)
ar1  0.30116    0.16696   1.8037   0.07128 .
ma1 -1.73576    0.10991 -15.7929 < 2.2e-16 ***
ma2  0.85257    0.12130   7.0286 2.086e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

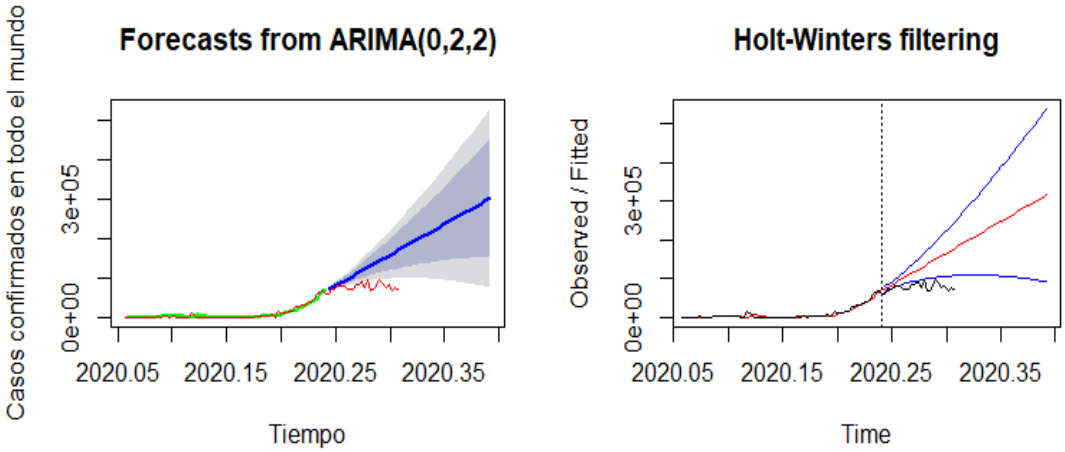
Continuamos con el test de Ljung-Box. Analizando los residuos con la función tsdiag() vemos que parece que los residuos se encuentran bien y que el modelo pasa el test.

Nuestro modelo ya estaría construido y pasaremos a la última parte del código que serio la predicción con un ARIMA(0,2,2). Nuestra intención es predecir los 20 últimos días del test set y luego los 30 días siguientes.

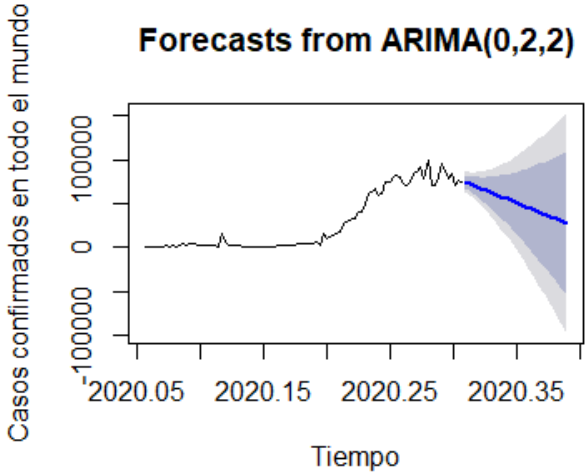
Pintamos la predicción y obtenemos el siguiente resultado. La línea negra representa la serie real hasta el día de hoy, la línea azul representa la predicción que hemos hecho a futuro hasta dentro de 30 días y el embudo azul representa lo que puede variar la predicción con un 95% de confianza. Partiendo de que la predicción (línea azul) debería ajustarse a la línea negra vemos

que se dispara hacia arriba mientras los datos de validación (línea negra) se estabilizan dejando de crecer. La conclusión que sacamos es que la predicción es muy mala.

Vistos los malos resultado probamos con el modelo Holt en las mismas condiciones: 20 días de validación y 30 días para predecir que pasara en el próximo mes con una confianza del 95%. Hacemos el mismo plot que hemos utilizado con el ARIMA, pero el resultado es igual de malo.



Antes de tirar la toalla, probaremos a predecir con toda la serie prescindiendo de la validación. Esto quiere decir que la predicción que haga únicamente diremos si es buena si nuestra intuición y lógica dice que es así.



En este caso vemos que la predicción es ha cambiado muchísimo. Ahora el modelo ha interpretado que hemos llegado al pico de la pandemia y por lo tanto esta prediciendo la desescalada. Esta desescalada parece que será progresiva durará al menos un mes más. Este cambio se debe a que en las dos últimas semanas se ha producido una estabilización en el número de contagios y le hemos añadido esta información que es muy explicativa para el modelo sobretodo en una serie tan corta. Pero como hemos dicho al no haber hecho ninguna validación no tenemos forma de comprobar si el modelo se ajusta a la serie. Dejando la intuición de lado, sabremos si la predicción es buena y que el modelo se ajusta a la serie cuando dispongamos datos de esos días, es decir, dentro de 2 semanas o un mes.

3. Limitaciones y Conclusiones

Principales limitaciones/problemas/debilidades del análisis realizado.

Conclusiones del análisis.

Extensión máxima: 500 palabras. Sin gráficos.

4. Referencias

[1]Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D. Y., Chen, L., & Wang, M. (2020). Presumed asymptomatic carrier transmission of COVID-19. *Jama*.

[2]Sohrabi, C., Alsafi, Z., O'Neill, N., Khan, M., Kerwan, A., Al-Jabir, A., ... & Agha, R. (2020). World Health Organization declares global emergency: A review of the 2019 novel coronavirus (COVID-19). International Journal of Surgery.

[3]Cohen, J., & Kupferschmidt, K. (2020). Countries test tactics in 'war'against COVID-19.

[4]Ahmad, T. (2020). Corona Virus (COVID-19) Pandemic and Work from Home: Challenges of Cybercrimes and Cybersecurity. Available at SSRN 3568830.