

**Colegio Universitario de Estudios Financieros**

**Máster Universitario en Ciencia de Datos**

**MODELOS DE INTERVENCIÓN DE SERIES TEMPORALES: APLICACIÓN AL COVID-19**

**Alumno:** Gómez Rivas, Alejandro José

**Tutor:** Queralt Sánchez de las Matas, Ricardo A.

Madrid, 9 de septiembre de 2022

1. **Resumen**

Desde principios del 2020, cuando en China se estaba preparando para conmemorar la Fiesta de la Primavera, también conocida como el Año Nuevo Chino, el COVID-19, como nueva variante de enfermedad respiratoria aguda, azotó al país por completo y, en cuestión de pocos meses, el mundo entero. La pandemia ha causado el cierre de muchas empresas y una situación de incertidumbre para muchos datos socioeconómicos y bursátiles. En cambio, para otros negocios, especialmente los digitales, han conseguido aprovechar esta situación para impulsarse y obtener una relevancia y unos ingresos que nunca lo habían tenido antes.

Una predicción precisa de una posible futura propagación de esta pandemia u otra se vuelve crucial en tales situaciones. Por tanto, el principal objetivo de este Trabajo de Fin de Máster es realizar una investigación e implantación de distintos modelos predictivos del impacto del COVID-19 en distintas series temporales y determinar cuál de ellos tiene el mejor rendimiento de cara a predecir el futuro para que, de esta manera, las empresas y países de todo el mundo puedan estar preparados y conscientes para controlarlo.

***Palabras clave (corregir, no más de 5 terminaciones)***

Análisis de series temporales, serie estacionaria, ETS, ARIMA, Prophet

**Abstract (revisar si se entiende bien)**

Since the beginning of 2020, when China was preparing to celebrate the Spring Festival, also known as the Chinese New Year, COVID-19, as a new variant of acute respiratory disease, struck the entire country and, for a few months, the whole world. The pandemic caused the closure of many companies and a situation of uncertainty for many socioeconomic and for the stock market data. On the other hand, in many businesses, especially digital ones, they have been able to take advantage of this situation to boost themselves and obtain relevance and incomes that they had never had before.

An accurate prediction of a possible future spread of this or another pandemic becomes crucial in such situations. Therefore, the main objective of this Thesis is to conduct research and implement different predictive models of the impact of COVID-19 in different time series and determine which of them has the best performance in predicting the future and, in this way, companies and countries can be prepared and aware to control it.

***Keywords***

Time-series analysis, stationary time-series, ETS, ARIMA, Prophet

**Lista de acrónimos y abreviaciones (actualizarlo)**

ACF Función de Autocorrelación

ARIMA Modelo Autorregresivo Integrado de Media Móvil

EDA Análisis Exploratorio de Dato

ETS Modelo de Suavizado Exponencial

IDE Entorno de Desarrollo Integrado

MAPE Error Porcentual Absoluto Medio

MSE Media de los Errores al Cuadrado o Error Cuadrático Medio

RMSE Raíz del Error Cuadrático Medio

SARIMA Modelo Estacional Autorregresivo Integrado de Media Móvil

**Repositorio GitHub**

Con el fin de facilitar la información del presente informe, se incluye una carpeta comprimida con los notebooks de Python usados para el Trabajo de Fin de Máster.

El repositorio se ubica en este [hipervínculo](https://github.com/algoriv92/TFM), y el contenido del mismo se desglosa de la siguiente manera: (hay que actualizarlo cuando se actualice el reposit.)

* La carpeta “data” contiene los datos utilizados para la realización del informe.
* La carpeta “notebooks” contiene todos los notebooks de todas las series temporales que se han empleado para el presente informe.

**Tabla de contenido (añadir también tabla de gráficos)**

[**1.** **Resumen** 3](#_Toc112777481)

[**2.** **Introducción** 13](#_Toc112777482)

[**2.1.** **Contexto** 13](#_Toc112777485)

[**2.2.** **Estado del arte** 14](#_Toc112777490)

[**2.3.** **Objetivos** 15](#_Toc112777493)

[**2.4.** **Estructura** 15](#_Toc112777496)

[**2.4.1.** **Obtención de los datos y manipulación** 16](#_Toc112777498)

[**2.4.2.** **EDA** 16](#_Toc112777500)

[**2.4.3.** **Definición** 16](#_Toc112777502)

[**2.4.4.** **Experimentación** 16](#_Toc112777504)

[**2.4.5.** **Evaluación** 16](#_Toc112777506)

[**2.5.** **Delimitaciones** 17](#_Toc112777508)

[**3.** **Marco teórico** 17](#_Toc112777513)

[**3.1.** **Series temporales** 17](#_Toc112777515)

[**3.2.** **ETS** 18](#_Toc112777519)

[**3.3.** **ARIMA** 18](#_Toc112777520)

[**3.4.** **PROPHET** 18](#_Toc112777521)

[**4.** **Marco práctico** 18](#_Toc112777522)

[**4.1.** **Obtención de los datos y preprocesado** 18](#_Toc112777523)

[**4.2.** **Modelos** 18](#_Toc112777524)

[**4.2.1.** **ETS** 18](#_Toc112777525)

[**4.2.2.** **ARIMA** 18](#_Toc112777526)

[**4.2.3.** **Prophet** 18](#_Toc112777527)

[**5.** **Resultados** 19](#_Toc112777528)

[**5.1.** **ETS** 19](#_Toc112777529)

[**5.2.** **ARIMA** 19](#_Toc112777530)

[**5.3.** **PROPHET** 19](#_Toc112777531)

[**6.** **Conclusiones** 19](#_Toc112777532)

[**7.** **Referencias** 19](#_Toc112777533)

1. **Introducción**

En este apartado se realizará una concepción general del proyecto con el fin de obtener una mejor comprensión del estudio para los lectores. Este capítulo está organizado de la siguiente forma. La primera sección pone en contexto los antecedentes de la enfermedad por COVID-19 y una descripción general de los métodos actuales de análisis y predicción de series temporales. El estado del arte de investigación del estudio se describe en la segunda sección.

Los propósitos y objetivos se describen en detalle en las siguientes dos secciones, seguidas de la metodología de investigación adoptada en el estudio. Luego se discuten brevemente las delimitaciones y la estructura de la tesis se presenta como una conclusión del capítulo.

* 1. **Contexto**

La enfermedad por coronavirus 2019 (COVID-19) fue identificada por primera vez en la ciudad de Wuhan, provincia de Hubei, China, en diciembre de 2019 y es causada por el coronavirus 2 del síndrome respiratorio agudo severo (SRAS) -CoV-2). Los pacientes mostraban un cuadro clínico de una neumonía grave, provocando insuficiencia respiratoria e incluso la muerte. Se contagia a través de aerosoles que se produce cuando el paciente tose o estornuda, por lo que las personas pueden infectarse con facilidad sin usar protección facial. Debido a su alto nivel de contagio, pronto se convirtió en una pandemia que ha afectado a la gran mayoría de países de todo el mundo y que, a día de hoy, sigue afectando, aunque a niveles más bajos que antes atrás gracias a los avances médicos y científicos. En el caso de España, el virus ha tenido un impacto bastante notorio.

El COVID-19 ha dado un gran impacto tanto en la salud de las personas como a la economía global. Los gobiernos de todos los países impusieron distintas políticas estrictas para hacer frente a la pandemia y “aplanar la curva” mediante el confinamiento domiciliario, restricciones de movilidad y el cierre temporal de muchas empresas y centros de distintos campos a excepción de aquellos que eran declarados como esenciales, como el sector sanitario o el de consumo, haciendo que muchas de estas empresas tuvieran que adaptarse a estas nuevas normas como buenamente pudieran.

Una serie temporal, que es el tema del que se va a tratar en el presente trabajo, consiste en una secuencia de observaciones ordenadas de manera cronológica sobre una serie. Sus valores van ligados en el momento de tiempo en el que son medidos, de esta manera el análisis de la serie de tiempo implica el manejo en conjunto de dos variables: la variable objetivo (la que se va a estudiar) y la variable de tiempo. Los datos históricos de algunos activos empresariales o los datos macroeconómicos son claros datos típicos de series temporales.

El análisis de estas series temporales engloba un conjunto vasto de técnicas que permiten obtener la información y así observar el comportamiento que tuvieron dichas series temporales en el pasado y, de esta manera, poder tener opciones y mecanismos para poder predecir sus valores en un futuro estipulado. A día de hoy existen muchos métodos de análisis distintos para descubrir el patrón o predecir hacia dónde tienden los datos temporales. Los métodos estadísticos, como el modelo de media móvil integrada autorregresiva, también conocida como ARIMA, los modelos de aprendizaje automático o los de redes neuronales, como la memoria a largo o a corto plazo (LSTM), son métodos que han demostrado ser eficientes para los análisis de series temporales, y algunos de ellos ya han proporcionado información bastante relevante para saber hacer frente a situaciones atípicas.

* 1. **Estado del arte**

Los modelos de análisis de series temporales que se han aplicado para estudiar el seguimiento y la predicción de enfermedades infecciosas han obtenido, por lo general, unos resultados muy óptimos, pero todavía es un desafío encontrar una forma adecuada de analizar hacia dónde puede ir la tendencia del COVID-19, ya que se trata de una situación totalmente novedosa y de las que más ha impactado en la sociedad en lo que se lleva de siglo.

Como pronosticar la tendencia del virus sigue siendo todo un reto, la predicción de cualquier serie temporal teniendo en cuenta situaciones atípicas como la del COVID-19 hace que esto sea una tarea aún más ardua. Además, realizando una investigación en profundidad en Internet se ha comprobado que hay escasos estudios de cómo afectó el COVID-19 a distintas series temporales. Por tanto, habría que observar si estos métodos de análisis de series temporales son disponibles y válidos en el escenario actual y si podemos determinar el mejor método (si lo hay) para representar la tendencia de cada una de las series para que. de esta manera, se pueda predecir la tendencia de la misma y poder evitarlo mediante una correcta toma de decisiones.

* 1. **Objetivos**

*“Las técnicas matemáticas para evaluar la extensión y el impacto de una epidemia y ayudar a su control son muy variadas.”* (León, 2020). En este caso, la idea principal de este proyecto consistirá en analizar la situación de distintas series temporales tras la pandemia por COVID-19 y estudiar su predicción mediante una serie de modelos.

Para ello, se han obtenido los datos de distintas series temporales publicadas en las páginas web oficiales. se les ha realizado un tratamiento para poder graficarlos y realizar un análisis exploratorio para ver el impacto que ha tenido. A continuación, se ha realizado una elección y ajuste de distintos modelos como el Modelo Suavizado Exponencial o el ARIMA y, por último, se ha determinado cuál ha sido mejor y una conclusión final sobre los mejores modelos para predecir una propagación de una pandemia y para que, de esta manera, las empresas y los países puedan estar (continuar). Para que fuera posible realizar toda esa metodología, se ha desplegado todo el proyecto utilizando el lenguaje de programación **Python** (versión 3.9.7), y el IDE en el que se ha apoyado para su desarrollo ha sido **Jupyter Notebook** (versión 6.4.5).

* 1. **Estructura**

Como se ha mencionado en el apartado anterior, para poder cumplir con los objetivos establecidos en el presente trabajo, se va a usar la siguiente metodología de investigación estructurándolo en las siguientes fases:

* + 1. **Obtención de los datos y manipulación**

Esta fase detalla dónde se han obtenido los datos históricos de las distintas series temporales que se han empleado para el proyecto, su tratamiento y limpieza del dato. Se trata de un proceso importante, ya que ésta condiciona la calidad del resultado final.

* + 1. **EDA**

La visualización es un paso esencial para comprender los datos de un solo vistazo, ya que nos permite identificar patrones comunes y, posteriormente, especificar el modelo más apropiado para la serie temporal.

* + 1. **Definición**

En esta fase se van a estudiar algunos de los métodos y modelos de análisis de series temporales populares para obtener una visión general del concepto y la implementación de los mismos.

* + 1. **Experimentación**

Aquí tiene como objetivo realizar experimentos aplicando la definición explicada en la fase anterior. Los modelos analíticos se construirán con parámetros predefinidos (o se ajustarán con búsqueda de cuadrícula para algunos métodos de ML).

* + 1. **Evaluación**

En la última fase, se evaluarán los resultados de la última fase comparando criterios estadísticos. Primero se compararán con un modelo básico ingenuo para comprobar su validez y luego entre sí para determinar si existe un método con el mejor rendimiento.

* 1. **Delimitaciones**

Algunas de las delimitaciones que se han tenido durante la realización del proyecto han sido los siguientes:

* Se han trabajado con datasets actualizados hasta julio y agosto del 2022, pero hay algunos conjuntos de datos de series temporales con fecha de actualización más antigua, por lo que puede influir ligeramente en la precisión de los modelos.
* Para que cobre sentido este proyecto, se ha cerciorado y se ha descargado los datos históricos de las fuentes oficiales, por lo que ha habido algunas series temporales que se han descartado de este proyecto porque los datos no provenían de fuentes oficiales y no eran fiables o, porque para acceder a ellas había que dar una remuneración al autor de la web.
* Algunas series temporales presentes en el trabajo, aun estando alojados sus datos en fuentes oficiales, no cabía la posibilidad de descargarlo directamente en cualquier formato, por lo que se ha tenido que recopilar mecanográficamente los datos.

1. **Marco teórico**

En este apartado se va a proporcionar los conocimientos más importantes para que los lectores comprendan este trabajo. Primero se presenta una breve introducción de los métodos de análisis de series de tiempo utilizados en este proyecto, es decir, ETS, ARIMA y Prophet.

* 1. **Series temporales**

Por lo general, *“una****serie temporal****es una sucesión de observaciones de una variable (p. ej. ingresos, precios, etc.) tomadas en varios instantes de tiempo”* (Queralt Sánchez de las Matas, 2021). Son una secuencia que tiene intervalos constantes entre puntos de tiempo, como los datos que se toman diariamente y a la misma hora. Los datos de estas series se usan habitualmente en la estadística, en la economía, en las finanzas y en las predicciones y se aplican en gran medida en dominios con información temporal involucrada. El análisis de las series de tiempo se conoce como métodos analíticos que descubren dicho componente temporal en los datos, los cuales se pueden observar estadísticas y patrones significativos de los mismos.

Generalmente se logra aplicando modelos de series temporales para ajustar los datos para entrenarlos y testearlos, con el fin de obtener unos resultados predictivos previos y ver cuánto se aproximan a los datos reales. Y el pronóstico de las series temporales consiste en usar dichos modelos para predecir los valores futuros basados ​​en el conocimiento de las observaciones históricas.

En este proyecto, se va a considerar tres métodos de análisis de series temporales que se utilizarán para predecir la tendencia futura de varias series temporales que han pasado la pandemia por COVID-19 en el año 2020, las cuales se van a mostrar en el presente informe y se mostrarán los resultados obtenidos en los siguientes apartados para ver cuánto de efectivas son a la hora de predecir situaciones atípicas. Estos métodos han demostrado ser efectivos y populares durante un largo período de tiempo en sus respectivos campos, pero aún se desconoce si estos métodos aún pueden usarse y mantener su excelente rendimiento en una situación anómala como es esta pandemia.

* + 1. **Componentes**

Las series temporales presentan normalmente una o varias características que se denominan como componentes, que ayudan a explicar el comportamiento que tienen las series en el tiempo. Dichos componentes son los siguientes:

* **Tendencia.** Este componente hace referencia al comportamiento o el movimiento de la serie temporal a largo plazo. El movimiento que tiende a hacer es creciente o decreciente y no tiene por qué ser lineal. A veces incluye el ciclo, que es el componente que se va a definir a continuación.
* **Componente cíclico.** En este se refleja los comportamientos recurrentes, aunque no tienen por qué ser exactamente periódicos, con un periodo superior a un año. Muestran, habitualmente, cómo se suceden las etapas de, por ejemplo, la bonanza económica con las de las crisis, o al menos, desaceleración.

A menudo, los ciclos económicos resultan de la superposición o yuxtaposición de distintos efectos, con periodos diferentes, más cortos o más largos (dos años, diez o veinte años, etc.). Por ello, son difícilmente reconocibles. Por ello, a veces no se separa de la tendencia. A la componente, en este caso, se le denomina como "*ciclo-tendencia*".

* **Componente estacional.** Consiste en movimientos de oscilación de un periodo de tiempo. La estacionalidad no se presenta únicamente cuando el periodo amplio es el año, a veces hay estacionalidades mensuales o semanales en series diarias, o estacionalidades diarias en series horarias, como son las series de cotizaciones bursátiles, por ejemplo. La estacionalidad siempre es de un periodo fijo y conocido. Es importante que las estacionalidades tengan un periodo no superior al anual para que no se confundan con las componentes cíclicas.
* **Componente irregular.** También llamado como “ruido”, recoge las alteraciones de la serie, pequeñas en su incidencia, y sin una pauta periódica ni tendencial reconocible. Se considera que está ocasionada por múltiples factores, de pequeña entidad y diferentes ritmos temporales, que no se pueden estudiar individualmente. Esto en la teoría, porque en la práctica lo que ocurre es que la consideración de una serie como compuesta por componentes tendenciales, cíclicas y estacionales no deja de ser un modelo y, como tal, una representación aproximada e imperfecta, aunque valiosa, del mundo real. La componente irregular recogería, en consecuencia, la incapacidad del modelo para explicar a la perfección el comportamiento de la serie temporal.
  1. **ETS**
  2. **ARIMA**
  3. **PROPHET**

1. **Marco práctico**
   1. **Obtención de los datos y preprocesado**
   2. **Modelos** 
      1. **ETS**
      2. **ARIMA**
      3. **Prophet**
2. **Resultados**
   1. **ETS**
   2. **ARIMA**
   3. **PROPHET**
3. **Conclusiones**
4. **Referencias**

<https://towardsdatascience.com/time-series-diy-seasonal-decomposition-f0b469afed44>

<https://www.madrimasd.org/blogs/matematicas/2020/06/01/148025>

<https://estrategiastrading.com/series-estacionarias/#:~:text=Qu%C3%A9%20es%20una%20serie%20estacionaria,y%20tampoco%20siguen%20una%20tendencia>.

<https://bookdown.org/content/2274/portada.html>

**Algunos proyectos y tesis**

<https://repositorio.upct.es/bitstream/handle/10317/9108/tfm-yuc-ana.pdf?sequence=1&isAllowed=y>

COVID-19: La importancia de los datos

<https://www.upm.es/?id=c56d5d3135531710VgnVCM10000009c7648a____&prefmt=articulo&fmt=detail>

TS Analysis Methods for Predicting COVID-19 Case Trend

<https://www.diva-portal.org/smash/get/diva2:1617426/FULLTEXT01.pdf>

Series temporales para predecir contagios y muertos por COVID-19

<https://www.colibri.udelar.edu.uy/jspui/bitstream/20.500.12008/25413/1/Aportes_6_Rodriguez_Modelos.pdf>

Predicción del nº de parados debido a la influencia del COVID-19 en Canarias

<file:///C:/Users/alexg/Downloads/Influencia%20del%20COVID-19%20en%20las%20predicciones%20del%20numero%20de%20parados%20en%20Canarias%20usando%20Google%20Trends.pdf>

Forecasting the COVID-19 economic recovery

<https://alex-oktay.com/wp-content/uploads/2021/08/2021b.pdf>

Predicción ventas

<https://upcommons.upc.edu/bitstream/handle/2117/187451/tfm-memoria-sofiacamara.pdf?sequence=1&isAllowed=y>