

**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

**AGRADECIMIENTOS**

En primer lugar, quiero agradecer al doctor Queralt por guiarme en el desempeño de este trabajo, sobre todo por la formación crítica dentro de este campo de la ciencia de datos.

En segundo lugar, agradecer a todos los profesores y a todos los compañeros del Máster que me ayudaron y tuvieron su entera disposición para ayudarme a sacar adelante un máster tan completo como éste.

Por último, y no menos importante, deseo agradecer de corazón a mis padres, a mi prima, a mis amigos y a mi pareja por el apoyo recibido durante este periodo importante de mi vida.

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***

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

***ABSTRACT***

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

**ACRÓNIMOS Y ABREVIACIONES**

ACF Función de Autocorrelación

AI Inteligencia Artificial

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

TS Serie Temporal

**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:
* En la carpeta “data” se almacenan los ficheros CSV con los datos de las series temporales analizadas.
* La carpeta “docs” contiene el presente informe y el resumen ejecutivo.
* La carpeta “pics” guarda las gráficas obtenidas en el análisis de cada serie temporal.
* Por último, la carpeta “notebooks” reúne todos los notebooks de las series temporales con los gráficos, los análisis y los modelos utilizados para determinar la predicción de las mismas.

**Tabla de contenido**

[**AGRADECIMIENTOS** 3](#_Toc112983011)

[**1.** **RESUMEN** 5](#_Toc112983012)

[**1.1.** **CONTEXTO** 20](#_Toc112983013)

[**1.2.** **ESTADO DEL ARTE** 21](#_Toc112983014)

[**1.3.** **OBJETIVO Y DESARROLLO** 21](#_Toc112983015)

[**1.4.** **DELIMITACIONES** 22](#_Toc112983016)

[**2.** **MARCO TEÓRICO** 23](#_Toc112983017)

[**2.1.** **SERIES TEMPORALES** 23](#_Toc112983018)

[*2.1.1.* *Definición* 23](#_Toc112983019)

[*2.1.2.* *Componentes* 24](#_Toc112983020)

[*2.1.3.* *Estacionariedad* 25](#_Toc112983021)

[*2.1.4.* *Descomposición* 26](#_Toc112983022)

[**3.** **MODELOS DE PREDICCIÓN DE SERIES TEMPORALES** 26](#_Toc112983023)

[*3.1.1.* *ETS* 27](#_Toc112983024)

[*3.1.2.* *ARIMA* 27](#_Toc112983025)

[*3.1.3.* *PROPHET* 27](#_Toc112983026)

[**4.** **MARCO PRÁCTICO** 27](#_Toc112983027)

[**4.1.** **SERIE TEMPORAL 1: DESEMPLEO** 27](#_Toc112983028)

[*4.1.1.* *Desempleo sin intervención* 27](#_Toc112983029)

[*4.1.1.1.* *Modelo a batir: ETS* 29](#_Toc112983030)

[*4.1.1.2.* *Modelo a batir: ARIMA* 29](#_Toc112983031)

[*4.1.1.3.* *Modelo a batir: PROPHET* 30](#_Toc112983032)

[*4.1.1.4.* *Predicción con ETS* 31](#_Toc112983033)

[*4.1.2.* *Desempleo con intervención* 32](#_Toc112983034)

[*4.1.2.1.* *Modelo a batir: ETS* 32](#_Toc112983035)

[*4.1.2.2.* *Modelo a batir: ARIMA* 32](#_Toc112983036)

[*4.1.2.3.* *Modelo a batir: PROPHET* 33](#_Toc112983037)

[*4.1.2.4.* *Predicción con ETS* 34](#_Toc112983038)

[**4.2.** **SERIE TEMPORAL 2: TURISMO** 34](#_Toc112983039)

[*4.2.1.* *Turismo sin intervención* 34](#_Toc112983040)

[*4.2.2.* *Turismo con intervención* 34](#_Toc112983041)

[**4.3.** **SERIE TEMPORAL 3: MATRICULACIONES DE VEHÍCULOS** 34](#_Toc112983042)

[*4.3.1.* *Matriculaciones sin intervención* 34](#_Toc112983043)

[*4.3.2.* *Matriculaciones con intervención* 34](#_Toc112983044)

[**4.4.** **SERIE TEMPORAL 3: HIPOTECAS** 35](#_Toc112983045)

[*4.4.1.* *Hipotecas sin intervención* 35](#_Toc112983046)

[*4.4.2.* *Hipotecas con intervención* 35](#_Toc112983047)

[**5.** **CONCLUSIONES** 35](#_Toc112983048)

[**6.** **REFERENCIAS** 35](#_Toc112983049)

**Índice de ilustraciones**

[Ilustración 1. Series estacionarias. Fuente: Francisco Parra, 2019. 23](#_Toc112982329)

[Ilustración 2: Descomposición de una serie temporal. Fuente: Elaboración propia 23](#_Toc112982330)

[Ilustración 3: Evolución del desempleo en España. Fuente: Elaboración propia. 25](#_Toc112982331)

[Ilustración 4. Media y Varianza del desempleo sin intervenir. Fuente: Elaboración propia. 26](#_Toc112982332)

[Ilustración 5. Comparativa ETS-ARIMA del desempleo sin intervenir. Fuente: Elaboración propia. 27](#_Toc112982333)

[Ilustración 6. Modelo Prophet del desempleo sin intervenir. Fuente: Elaboración propia. 28](#_Toc112982334)

[Ilustración 7. Predicción del desempleo sin intervenir. Fuente: Elaboración propia. 28](#_Toc112982335)

[Ilustración 8. Gráfica desempleo intervenido. Fuente: elaboración propia. 29](#_Toc112982336)

[Ilustración 9. Comparación ETS-ARIMA del desempleo intervenido. Fuente: Elaboración propia. 30](#_Toc112982337)

[Ilustración 10. Modelo PROPHET para desempleo intervenido. Fuente: Elaboración propia. 30](#_Toc112982338)

[Ilustración 11.. Predicción del desempleo sin intervenir. Fuente: Elaboración propia. 31](#_Toc112982339)

**Índice de tablas**

[Tabla 1. MAPEs modelos para desempleo sin intervenir. Fuente: Elaboración propia. 31](#_Toc112982600)

[Tabla 2. Predicción del desempleo sin intervenir. Fuente: Elaboración propia. 31](#_Toc112982601)

[Tabla 3. MAPEs modelos para desempleo sin intervenir. Fuente: Elaboración propia. 34](#_Toc112982602)

[Tabla 4. Predicción del desempleo sin intervenir. Fuente: Elaboración propia. 34](#_Toc112982603)

1. **INTRODUCCIÓN**

En este apartado se dará 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 van a tender los datos temporales. Los métodos estadísticos, como el modelo ARIMA, los modelos de aprendizaje automático o los de redes neuronales, son métodos que han demostrado ser eficientes para los análisis de series temporales, y algunos de ellos ya han proporcionado información 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 la situación actual con el COVID-19 hace que sea una tarea ardua. Además, realizando una investigación por internet, como se ha comentado anteriormente, existen muchos estudios sobre el pronóstico que tienen los modelos de la enfermedad por COVID-19, pero hay escasas predicciones que puede tener el impacto de la misma en distintas series temporales de distintos campos.

Por tanto, resulta cautivador realizar un estudio empleando estos modelos predictivos en distintas series temporales y si éstos son válidos para el escenario actual y determinantes para detectar un posible evento no común, para poder evitarlo o hacer frente a ello mediante una correcta toma de decisiones. Por eso, se ha tomado la decisión de realizar este estudio bajo la tutela del director Queralt Sánchez de las Matas.

* 1. **OBJETIVO Y DESARROLLO**

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 pronóstico mediante el entrenamiento y el testeo de distintos modelos predictivos y determinar si son capaces de realizar detecciones de situaciones temporales atípicas y con cuánto margen de error.

*“Las técnicas matemáticas para evaluar la extensión y el impacto de una epidemia y ayudar a su control son muy variadas.”* (León, 2020). Para el desarrollo de este estudio, se ha seguido el mismo orden que tiene cada uno de los notebooks que se encuentran alojados en el repositorio. Una vez que se obtienen los ficheros donde se hallan los datos históricos de cada una de las series temporales, se realizan las siguientes técnicas:

* Preprocesamiento: se aplican distintos métodos de limpieza y tratamiento del dato, como el cambio del formato de la fecha, la detección de datos faltantes o nulos, el recorte del conjunto de datos para analizar una muestra, entre otros mecanismos.
* EDA: consiste en un análisis exploratorio de los datos. *“Es utilizado por los científicos de datos para analizar e investigar conjuntos de datos y resumir sus principales características, empleando a menudo métodos de visualización de datos”* (IBM, 2020)*.* Se les realizan una serie de cálculos estadísticos y de representaciones gráficas que ofrecen una comprensión los datos de un solo vistazo y que permite la identificación de distintos patrones comunes para, posteriormente, ver si es necesario realizar un tratamiento o modificación de los datos para adaptarlo a los modelos seleccionados para su testeo.
* Modelización: en esta fase se aplican los distintos modelos seleccionados y se les realiza un entrenamiento y test con la serie temporal incluyendo el momento COVID-19 para generar un resultado y un margen de error.
* Elección del modelo: de dichos resultados se selecciona el modelo que menos error haya generado y que más se haya aproximado a los datos reales.
* Previsión: con el modelo ganador, se le realiza un pronóstico para finales del año 2022, se guardan en un archivo CSV para contrastarlo cuando llegue la fecha y comprobar el margen de error que puede dar.

Además, a cada una de las series temporales se les realiza una intervención eliminando el momento del COVID-19 para observar el resultado final y compararlo con la serie original, con el objetivo de ver cuánta diferencia de error y de predicción dan cuando no existe un evento pandémico.

Por último, se da una conclusión final sobre los resultados con los modelos utilizados.

Para que fuera posible realizar toda esta metodología, se ha desplegado el proyecto mediante el uso del 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. **DELIMITACIONES**

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

Se han trabajado con varias series temporales actualizadas hasta estos últimos meses 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.

* 1. **SERIES TEMPORALES**
     1. *Definición*

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 de distintos efectos, con periodos diferentes, más cortos o más largos, como dos, cinco o diez años. Por ello, son difícilmente reconocibles y, a veces, no se separa de la tendencia. En este caso, al componente 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. *Estacionariedad*

En Matemáticas, un proceso estacionario es un proceso estocástico cuya distribución de probabilidad en un instante de tiempo fijo o una posición fija es la misma para todos los instantes de tiempo o posiciones, es decir, si no presenta cambios sistemáticos en su media (no hay tendencia) ni en su varianza, y sus variaciones periódicas han sido eliminadas.

Aplicando el concepto anterior a las series temporales, una serie se considera estacionaria cuando es estable a lo largo de un periodo de tiempo, esto es, cuando las media y la varianza son constantes en el tiempo.

La estacionariedad puede ser más o menos restrictiva en torno a su definición, por lo que se hablaría de dos tipos de estacionariedad:

* Serie estacionaria en sentido estricto o fuerte: se dice que una serie tiene esa definición si todos sus momentos, independientemente del orden, son invariantes en el tiempo.
* Serie estacionaria en sentido amplio o débil: se caracteriza así una serie estacionaria cuando todos sus momentos de primer y segundo orden son invariantes en el tiempo. Su ejemplo más simple es el de “ruido blanco”, es ahí donde la media y la varianza son siempre cero.

Por otro lado, una serie no se va a considerar estacionaria cuando no es estable a lo largo del tiempo, es decir, que su media y su varianza no son constantes en el tiempo.

Es interesante que las series temporales sean estacionarias porque, de esta manera, se pueden obtener predicciones fácilmente por el patrón repetitivo que tienen.

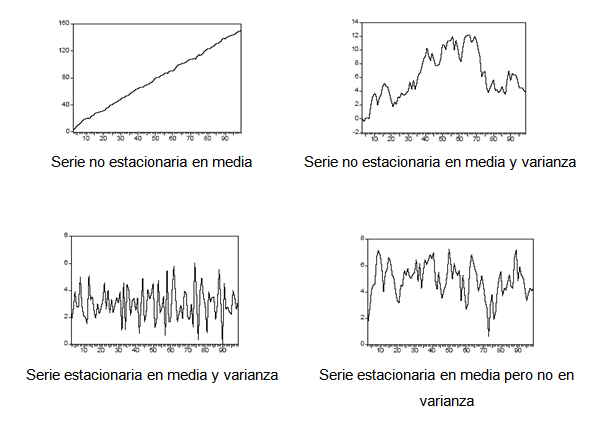


Ilustración 1. Series estacionarias. Fuente: Francisco Parra, 2019.

* + 1. *Descomposición*

Siguiendo el enfoque comentado en el apartado referente a la serie temporal y sus componentes, en el procedimiento de análisis exploratorio de las series temporales que se han empleado en el trabajo se le han incluido la descomposición de las mismas. *“En pocas palabras, es una técnica de análisis que divide una serie temporal en los tres componentes ya descritos: tendencia, estacionalidad y residuos”*. (Lewinson, s.f.)

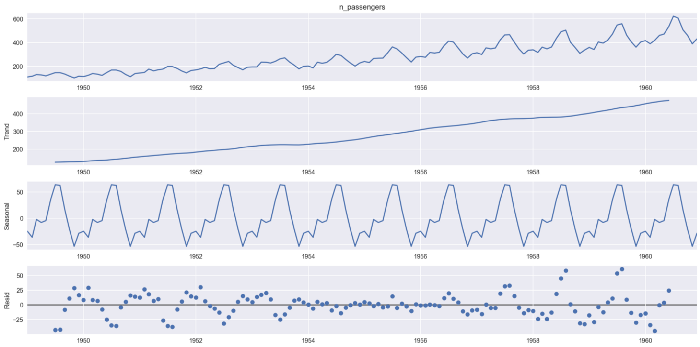


Ilustración 2: Descomposición de una serie temporal. Fuente: Elaboración propia

1. **MODELOS DE PREDICCIÓN DE SERIES TEMPORALES**

La predicción de las series de tiempo se basa de la siguiente forma: una vez que se conoce el comportamiento pasado de un sistema de datos, es posible realizar pronósticos sobre el comportamiento que ésta obtuvo a corto, medio o largo plazo.

Los métodos tradicionales estadísticos empleados para el pronóstico de las series temporales se basan en la construcción de funciones lineales a partir de las observaciones pasadas más recientes para proporcionar las predicciones futuras. Para este proyecto se han elegido tres modelos tradicionales, las cuales son ETS, ARIMA y PROPHET.

* + 1. *ETS*

El modelo ETS, por sus siglas en inglés significa *“Exponential Smoothing”*, modelo de Suavizado Exponencial, es uno de los métodos de predicción de series temporales más antiguos y estudiados. Es muy eficaz cuando los valores de las series temporales siguen una tendencia gradual y presentan un comportamiento estacional marcado en el cual los valores siguen un patrón cíclico repetido a lo largo de un número determinado de periodos de tiempo (ArcGIS, s.f.).

Este modelo se emplea a menudo en series temporales con las características descritas anteriormente, como para predecir la temperatura de un municipio durante una ola de frío para prepararse para las enfermedades relacionadas con esa estación, o bien aplicados en datos macroeconómicos-demográficos como es la tasa de desempleo.

El funcionamiento del modelo ETS se divide en dos partes: en una parte de previsión y otra parte que es la de validación. La primera se emplea para realizar la previsión de los valores de futuros periodos de tiempo, y la segunda se utiliza para determinar la precisión de la predicción de valores futuros que el modelo puede realizar para cada serie temporal. También se emplea para verificar la precisión del modelo dividiendo la serie temporal en entrenamiento y en test y, a través de la precisión y del error absoluto medio porcentual (MAPE), que es un indicador de desempeño del pronóstico que mide el tamaño del error, se comprueba la efectividad del modelo frente a los datos reales.

Cuanto más pequeño sea el MAPE, mucho más preciso va a ser el modelo aplicado a la serie temporal. Es muy útil este indicador cuando los modelos predictivos son muy similares entre sí y se quiere saber con exactitud la efectividad de uno u otro.

Existen distintos tipos de suavizado exponencial, pero todos se basan en separar las series temporales en varios componentes. Los valores de cada componente se estiman ponderando exponencialmente los componentes de los periodos de tiempo anteriores, de manera que la influencia de cada periodo de tiempo disminuya exponencialmente a medida que se avanza en el tiempo. Cada componente se define recursivamente a través de un planteamiento de modelo de estado-espacio, y cada componente depende de todos los demás componentes. La estimación de todos los parámetros se realiza mediante una estimación de máxima probabilidad.

En esta herramienta, todos los componentes son aditivos, de modo que el modelo de predicción es la suma de los componentes individuales. Si se utiliza un componente estacional, la herramienta utiliza el método estacional amortiguado de Holt-Winters. Si no se utiliza ningún componente estacional, la herramienta utiliza el método de tendencia amortiguada.

* + 1. *ARIMA*

El método ARIMA, por sus siglas en inglés significa “Autoregressive Integrated Moving Average”, modelo de Autorregresión Integrada de Media Móvil, es otro algoritmo de predicción clásico diseñado para ser empleado en series temporales estacionarias. Este método se basa en la combinación de tres parámetros (*p, d, q*) que son:

* AR: de Autorregresión (*p*), el cual quiere decir que el pasado predice el futuro de la serie temporal, por lo que un valor de la serie temporal en el tiempo es una función de los valores pasados.
* MA: de Media Móvil (*q*), significa que un punto de la serie temporal puede definirse en función de los errores de los valores pasados. Dichos errores son los obtenidos en el ajuste de la media ponderada respecto a la función original.
* Diferenciaciones (*d*), es el parámetro encargado de que logre que la serie temporal sea estacionaria.

*“La correlación es a la regresión, lo que la autocorrelación es a la autorregresión”* (Iglesias, 2021). Esto significa que, en una variable con autocorrelación, pueden utilizarse los valores pasados como predictores de valores futuros. De ahí la razón de ser del parámetro (𝑝) del modelo ARIMA, que refleja el número de valores de observaciones anteriores, o retardos (*lags*), que se utilizan para predecir una presente. Un ARIMA (1,0,0) utilizaría la observación inmediatamente anterior para explicar la presente.

El componente de Media Móvil (𝑞) indica el número de valores anteriores cuyos errores se utilizan para predecir un valor presente. Para comprender este componente hay que explicar la razón de ser de las diferenciaciones (𝑑). Los ARIMA se usan para explicar o modelizar procesos estacionarios (o estocásticos) los cuales sus propiedades no dependen del tiempo en el que se observan. Diferenciar la serie (restar a los valores presentes el inmediatamente anterior) puede servir para eliminar tendencias que alteren esas propiedades, indicando con la Diferenciación “𝑑” el número de veces que se repite ese proceso (o el número de diferenciaciones que se aplican).

En consecuencia, los procesos estocásticos están determinados por dos componentes o fuerzas principales: sus propios valores previos (explicados por su componente autorregresivo (AR) y las variaciones impredecibles producidas por multitud de variables que cambian e interaccionan a lo largo del tiempo de tal manera que son ostensiblemente aleatorias. Es el efecto de estas variaciones impredecibles lo que busca explicar el componente de medias móviles (Moving Averages, MA). Conceptualmente los modelos ARIMA (0, 0, 𝑞) o MA (𝑞) buscan explicar un valor presente a través de una función que lo relacione con las variaciones impredecibles de “𝑞” valores anteriores, o, en otras palabras, con el error de los términos de “𝑞” tiempos anteriores, de tal manera que un ARIMA (0,0,1) utilizaría el error de la observación inmediatamente anterior para explicar el valor presente y vendría definido por la ecuación.

* + 1. *PROPHET*

El modelo PROPHET es una herramienta de código abierto de Facebook utilizada para pronosticar datos de series de tiempo que ayuda a los científicos y estadísticos a comprender y predecir las series temporales.

Se basa en un modelo aditivo descomponible donde las tendencias no lineales se ajustan a la estacionalidad, donde se tiene en cuenta los efectos de las vacaciones. Este método de pronóstico va analizando y cambiando los parámetros en función de lo óptimo que haya sido el ajuste, permitiendo de este modo un ajuste automático de los parámetros del modelo. Así consigue una parametrización de las variables de forma automática gracias al aprendizaje realizado del comportamiento de la serie.

1. **MARCO PRÁCTICO**

En este apartado se lleva a cabo todo el marco teórico de una manera pragmática con datos reales. Para la realización de este proyecto, se han seleccionado cuatro series temporales, las cuales se han cogido los datos que se alojan en las páginas oficiales del gobierno español. Estas series temporales son el número de desempleados, el número de turistas, la matriculación de vehículos y el número de hipotecas.

* 1. **SERIE TEMPORAL 1: DESEMPLEO**

Para la obtención de los datos del desempleo español, se ha dirigido a la página web del Servicio Público de Empleo Estatal, concretamente en este [link](https://www.sepe.es/HomeSepe/que-es-el-sepe/estadisticas/datos-avance/paro.html). Una vez dentro, se obtienen los datos dentro del apartado “Datos de Series anuales”, seleccionando el que está registrado desde el año 2013 hasta el 2022, que resultaba una cantidad de tiempo óptima ya que hay suficientes datos para realizar el entrenamiento y el test del modelo, incluyendo el inicio de la pandemia, momento clave para determinar si los modelos empleados para esta serie son efectivos para predecir estas situaciones.

Cada serie se desglosa en dos partes: una parte que no se interviene en los datos en el momento por COVID-19 para ver cómo predicen los modelos, y otra parte que sí se interviene en los datos del desempleo, para ver la diferencia de pronósticos.

* + 1. *Desempleo sin intervención*

Comienza el análisis tomando los datos mensuales, en total hay 115 meses y, como se puede apreciar en la Ilustración 3, existe una tendencia decreciente hasta la raya roja que es marzo del 2020, momento en el que inicia la pandemia por COVID-19. También se puede comprobar un comportamiento estacional, pues los meses correspondientes al verano se generan empleo temporal y baja significativamente el paro de manera temporal. Pasa lo mismo con el mes de diciembre.

En el momento en el que comienzan las restricciones por COVID-19, muchos de los empleados con contratos fueron despedidos y otros pasados por ERTEs, lo que hace subir el número de parados. Este fenómeno se ha prolongado hasta el inicio del año 2021, momento en el que vuelve a haber una tendencia decreciente.



Ilustración 3: Evolución del desempleo en España. Fuente: Elaboración propia.

Para analizar la estacionariedad de la serie temporal, se realiza un cálculo de la media y de la varianza de todos los números de desempleados del conjunto de datos y se representa gráficamente. Como se indicó en el marco teórico, es una manera rápida de determinar si una serie temporal tiene alguna estructura dependiente del tiempo o no.

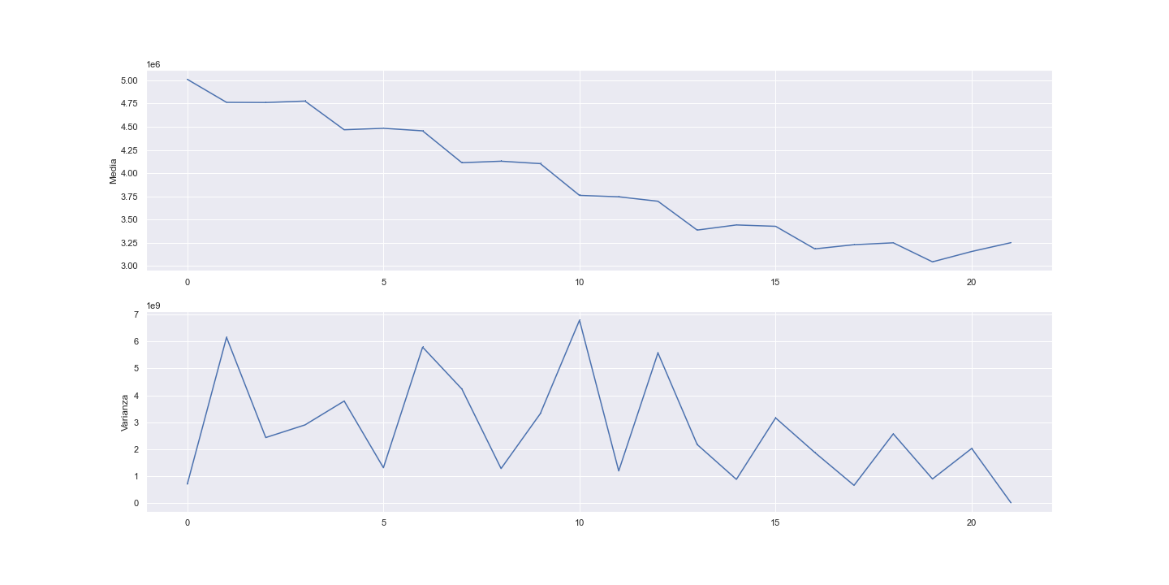


Ilustración 4. Media y Varianza del desempleo sin intervenir. Fuente: Elaboración propia.

A simple vista, la media posee unos valores constantes decrecientes en el tiempo, lo que nos indica que la serie es estacionaria, pero en varianza no se aprecia bien. Para asegurar que una serie es estacionaria o no, existe una prueba de raíz unitaria que detecta estadísticamente la presencia de conducta tendencial estocástica en una serie de tiempo mediante un contraste de hipótesis. Esta prueba nos proporciona una verificación rápida y evidente de si la serie es estacionaria o no (Ortiz, 2020).

* **La hipótesis nula (H0)** de la prueba determina que la serie temporal se puede representar mediante una raíz unitaria. En otras palabras, indica que la serie temporal no es estacionaria, lo que significa que tiene alguna estructura dependiente del tiempo.
* **La hipótesis alternativa (H1)**, que rechaza la hipótesis nula, indica que la serie temporal sí es estacionaria.

La interpretación del resultado del test se basa en el valor que dé el valor ‘p-value’. Un valor ‘p’ por debajo del umbral de significancia (0,05) significa que la hipótesis nula se rechaza. En este caso, el valor ‘p’ está por debajo del umbral, y se concluye que la serie es estacionaria.

* + - 1. *Modelo a batir: ETS*

Como los datos son estacionales, comienza la batida de modelos con el modelo ETS, modelo de Suavizado Exponencial. Tras realizar un autoETS se obtiene un ETS(MAM), es decir, un modelo multiplicativo de Hol-Winters con errores multiplicativos.

* + - 1. *Modelo a batir: ARIMA*

La serie, al ser estacionaria, no se aplica la transformación de los datos a logaritmos. Tras realizar el autoARIMA para encontrar el mejor modelo, se obtiene SARIMAX (Seasonal ARIMA) (1, 1, 0)x(1, 1, 0, 12), es decir, con auto-regresión de orden 1, la serie depende del último retardo y media móvil de orden 1, es decir, la serie depende del error ocurrido en el retardo anterior. El 12 es la periodicidad, que son 12 meses.

Realizando una comparativa gráfica se observa que ambos modelos no predicen en absoluto el momento COVID-19, pero ambos predicen con bastante exactitud el número de desempleados a partir de diciembre del 2021, lo que quiere decirse que, de una manera estadística, el efecto COVID-19 termina en esa fecha.

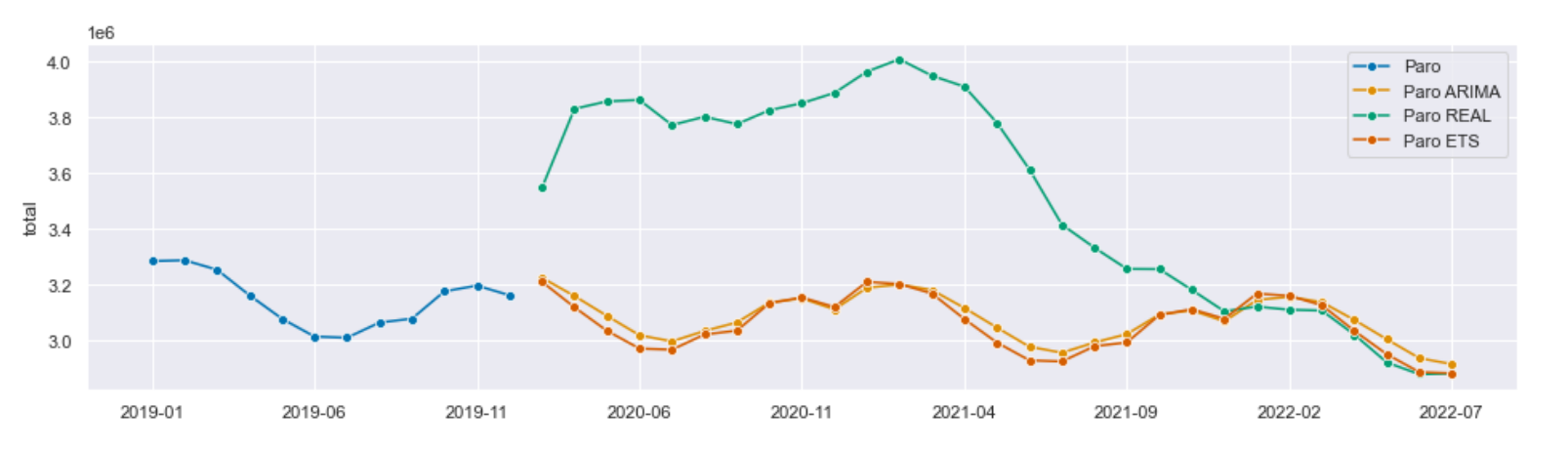


Ilustración 5. Comparativa ETS-ARIMA del desempleo sin intervenir. Fuente: Elaboración propia.

Realizando el cálculo del MAPE, se comprueba que el modelo ETS da un error ligeramente inferior (12,08% frente el 12,27% de error que predice ARIMA).

* + - 1. *Modelo a batir: PROPHET*

Modelando la serie temporal con este método se observa que hace el mismo efecto que con ETS y ARIMA, no predice el momento por COVID-19, pero luego consigue una mejor predicción más adelante. Su MAPE da un error del 13,98%, lo que indica que **el modelo ETS es el que menos error da**.

|  |  |
| --- | --- |
| Modelo | MAPE |
| ETS | 12,08% |
| ARIMA | 12,27% |
| PROPHET | 13,98% |

Tabla 1. MAPEs modelos para desempleo sin intervenir. Fuente: Elaboración propia.



Ilustración 6. Modelo Prophet del desempleo sin intervenir. Fuente: Elaboración propia.

* + - 1. *Predicción con ETS*

La predicción que toma para finales del año 2022 tiene una tendencia creciente y coincidente con el del resto de los finales de año, a excepción de la etapa de COVID-19. Tiene sentido, pues el modelo ETS al final es útil para conjunto de datos con estacionalidad como es este caso y hace un cálculo promedio móvil simple.

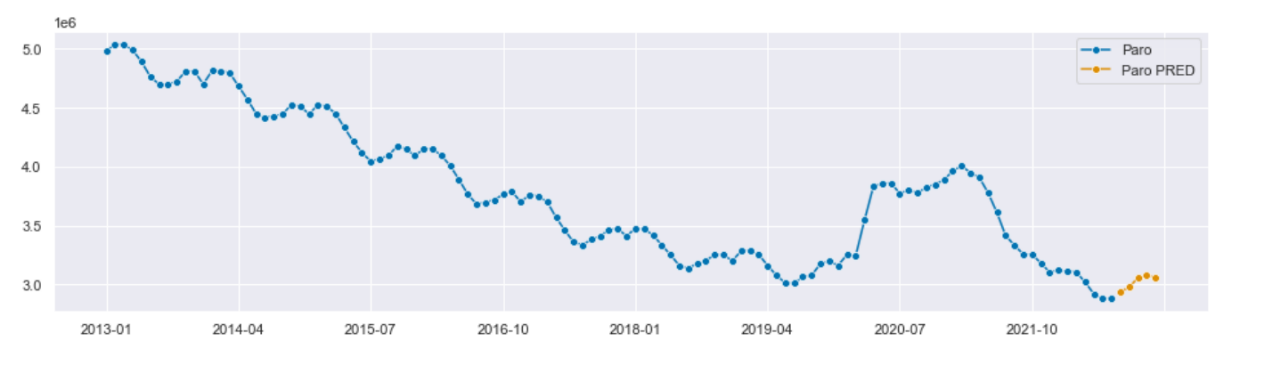


Ilustración 7. Predicción del desempleo sin intervenir. Fuente: Elaboración propia.

A continuación, los datos obtenidos en números:

|  |  |
| --- | --- |
| Periodo | Total desempleados |
| 2022-08 | 2.938.509 |
| 2022-09 | 2.978.051 |
| 2022-10 | 3.057.594 |
| 2022-11 | 3.080.354 |
| 2022-12 | 3.058.159 |

Tabla 2. Predicción del desempleo sin intervenir. Fuente: Elaboración propia.

* + 1. *Desempleo con intervención*

Para realizar la intervención, se ha realizado un entrenamiento-test con el modelo que menos error dio en el conjunto de datos del desempleo original hasta febrero del 2020, el mes antes del comienzo de la pandemia. Dicho modelo es el ETS. Después se realiza una predicción con los datos hasta el mes de diciembre del 2021, que se mencionó anteriormente que era el punto donde la predicción y los datos reales coincidían, concluyendo el momento por COVID-19. Por último, los datos obtenidos se añaden al dataset y, al representarlo, se obtiene esta serie temporal de desempleo intervenido. Las rayas rojas dividen el momento que se ha intervenido con el resto del conjunto de datos.

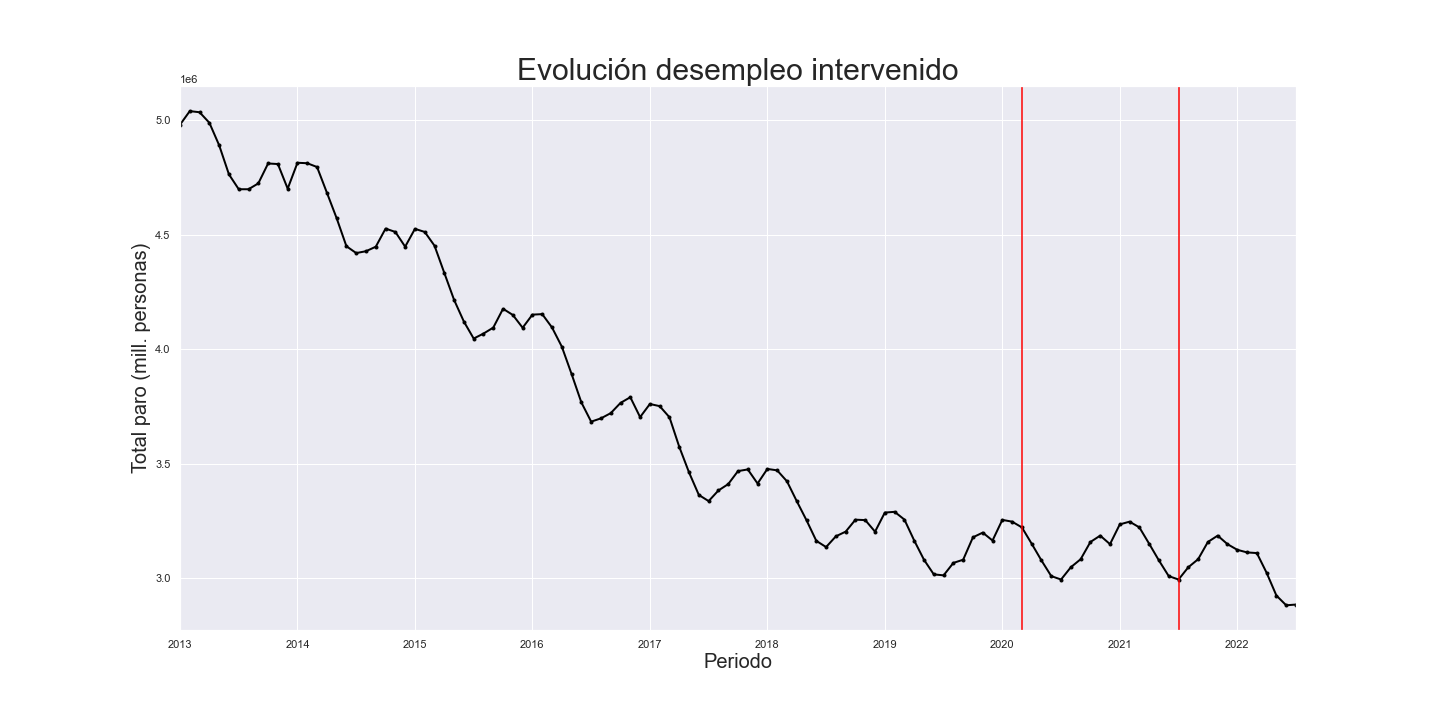
**

Ilustración 8. Gráfica desempleo intervenido. Fuente: elaboración propia.

La serie cumple con una tendencia decreciente que se aplana al final del periodo, cumple con la condición de serie estacionaria y su estacionalidad está marcada por el descenso del paro en la etapa de verano y el mes de diciembre, tal y como se explicó en el modelo de desempleo sin intervención.

* + - 1. *Modelo a batir: ETS*

La predicción obtiene muy buena precisión y un error MAPE de un 1,19%, lo que indica que este modelo es muy bueno para este tipo de series temporales en las que no hay saltos temporales irregulares y con una estacionalidad marcada. El modelo que se obtiene es el ETS(MAM), el mismo que se obtuvo en el desempleo sin intervención.

* + - 1. *Modelo a batir: ARIMA*

Con ARIMA pasa exactamente lo mismo: obtiene una muy buena predicción con un error del 1,36% y un modelo SARIMAX (1, 1, 0)x(1, 1, 0, 12)

Haciendo un contraste gráfico, no se ve apenas diferencias. Se podría predecir perfectamente con cualquiera de los dos, ya que apenas fallan para este tipo de series temporales.

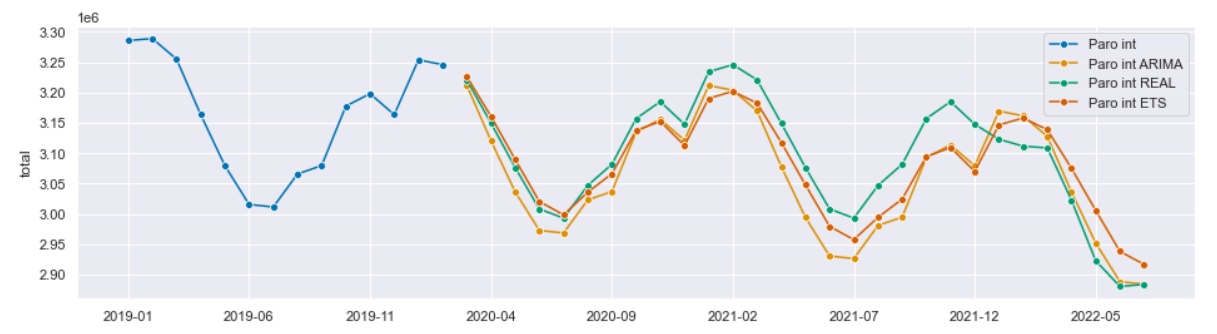


Ilustración 9. Comparación ETS-ARIMA del desempleo intervenido. Fuente: Elaboración propia.

* + - 1. *Modelo a batir: PROPHET*

Con este modelo se obtienen unas predicciones ligeramente peores, pues este modelo es más preciso para modelos que tengan puntos de datos atípicos grandes, eventos irregulares o con tendencias de crecimiento no lineal que se aproximen a un límite. El MAPE ha dado un resultado del 5,73%, lo que indica que el modelo ETS vuelve a triunfar para este conjunto de datos del desempleo intervenido en el momento COVID-19.

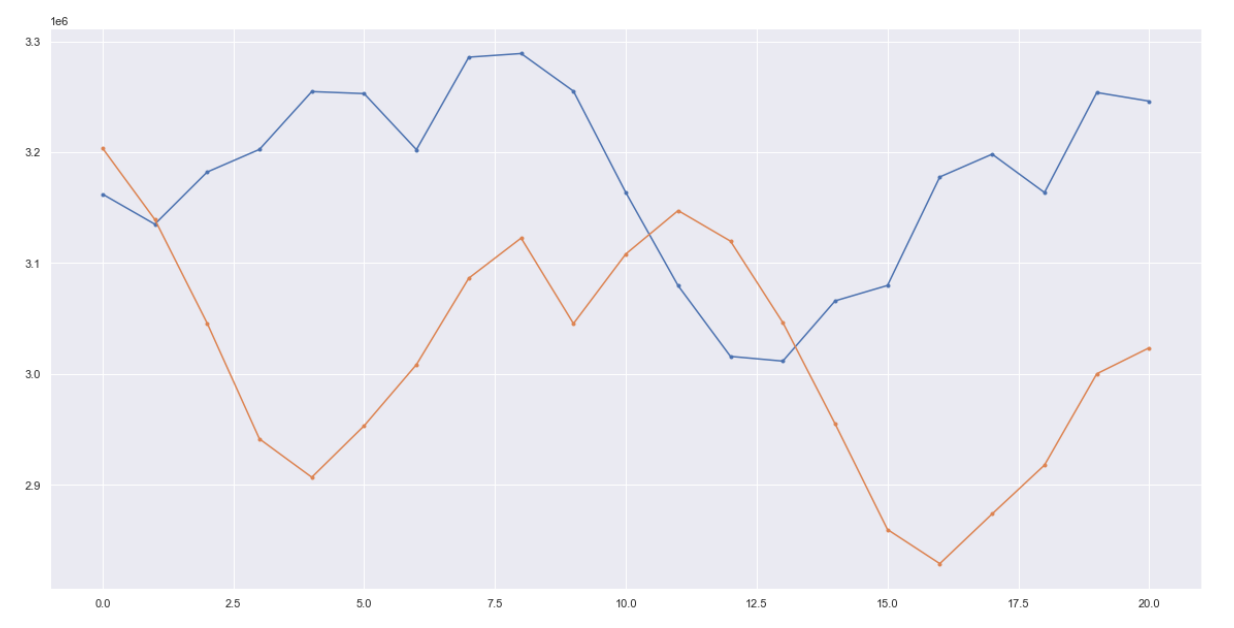


Ilustración 10. Modelo PROPHET para desempleo intervenido. Fuente: Elaboración propia.

|  |  |
| --- | --- |
| Modelo | MAPE |
| ETS | 12,08% |
| ARIMA | 12,27% |
| PROPHET | 13,98% |

Tabla 3. MAPEs modelos para desempleo sin intervenir. Fuente: Elaboración propia.

* + - 1. *Predicción con ETS*

*.* La predicción que toma para finales del año 2022 tiene una tendencia creciente y coincidente con el del resto de los finales de año, a excepción de la etapa de COVID-19. Tiene sentido, pues el modelo ETS al final es útil para conjunto de datos con estacionalidad como es este caso y hace un cálculo promedio móvil simple.

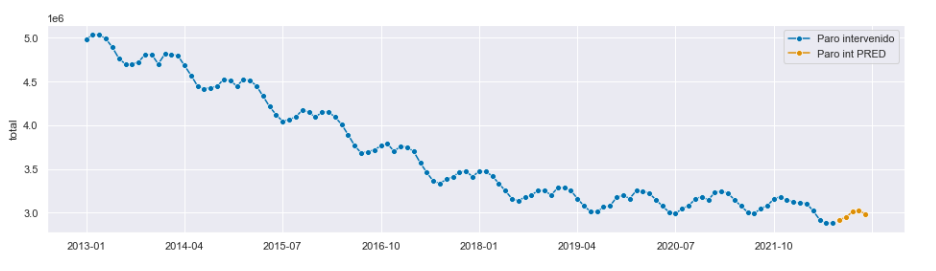


Ilustración 11.. Predicción del desempleo sin intervenir. Fuente: Elaboración propia.

A continuación, los datos obtenidos en números:

|  |  |
| --- | --- |
| Periodo | Total desempleados |
| 2022-08 | 2.938.509 |
| 2022-09 | 2.978.051 |
| 2022-10 | 3.057.594 |
| 2022-11 | 3.080.354 |
| 2022-12 | 3.058.159 |

Tabla 4. Predicción del desempleo sin intervenir. Fuente: Elaboración propia.

* 1. **SERIE TEMPORAL 2: TURISMO**
     1. *Turismo sin intervención*
     2. *Turismo con intervención*
  2. **SERIE TEMPORAL 3: MATRICULACIONES DE VEHÍCULOS**
     1. *Matriculaciones sin intervención*
     2. *Matriculaciones con intervención*
  3. **SERIE TEMPORAL 3: HIPOTECAS** 
     1. *Hipotecas sin intervención*
     2. *Hipotecas con intervención*

1. **CONCLUSIONES**
2. **REFERENCIAS**

<https://otexts.com/fpp2/ts-objects.html>

<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>

<https://analisisdedatos.net/analisis/ST/estacionariedad.php>

**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>