Modelos de Predicción de Series Temporales: Gripe y COVID-19

Javier de Dios González

Resumen— En este artículo se trata la predicción de series temporales. Este tipo de datos prevalece en muchas disciplinas científicas y de ingeniería. La previsión de series temporales es una tarea crucial en ámbitos como el sanitario. En este trabajo, se emplea los modelos ARIMA y SARIMA para predecir valores con datos de la gripe y el COVID-19. Este proceso ha consistido en la obtención, análisis y preparación de los datos de forma detallada, para poder aplicar correctamente los modelos y obtener resultados. Los datos obtenidos se han analizado y esto ha permitido realizar la comparación entre los diferentes modelos aplicados. Finalmente, esta comparación se ha ampliado al resto de modelos más utilizados en el estado del arte de este ámbito.

Palabras clave—ARIMA, SARIMA, Serie Temporal, Transformer

Abstract— This article presents time series forecasting. This type of data is prevalent in many scientific and engineering disciplines. Time series forecasting is a crucial task in many ambits such as the healthcare one. In this project, the ARIMA & SARIMA models are used to predict values from Influenza and COVID-19 data. This process has consisted of obtaining, analyzing, and preparing the data in detail, in order to correctly apply the models and obtain results. The data obtained has been analyzed and this has allowed the comparison between the different applied models. Finally, this comparison has been extended to the rest of the most used models in the state of the art in this field.

Index Terms— ARIMA, SARIMA, Time Series, Transformer

1 Introducción - Contexto del trabajo

Una serie temporal es una sucesión de datos ordenados cronológicamente, normalmente espaciados en intervalos iguales. De este modo, el análisis de series temporales implica el trato con dos variables: la que se está estudiando y la del tiempo.

Las series temporales son especialmente utilizadas en el 'forecasting', este método consiste en usar un modelo para predecir el valor futuro de una serie temporal (extrapolación pronóstica), principalmente en función de su comportamiento pasado (autorregresión).

Esta es una de las técnicas más utilizadas en ámbitos como las finanzas, los negocios o en este caso o el pronóstico de la actividad de enfermedades como la gripe y similares. Las epidemias de gripe estacional crean enormes problemas económicos y de salud. Como referencia, los 'Centers for Disease Control and Prevention' (CDC) en 2018 anunciaron que, en promedio, cerca del 8% de la población estadounidense contrae la gripe cada temporada, con un rango de entre el 3% y el 11%, según la época del año.

A pesar de su importancia, en el seguimiento de esta enfermedad suele haber un retraso de al menos una semana debido a la recopilación y gestión de datos. Por este motivo, el pronóstico de estas afecciones es fundamental para el monitoreo en tiempo real y para que las agencias de salud pública asignen recursos para planificar y prepararse para posibles pandemias.

Se han desarrollado una variedad de métodos para pronosticar este tipo de series temporales y en este proyecto se han analizado y aplicado algunos de ellos. Este proceso se ha realizado tanto para datos de la gripe, ofrecidos en las publicaciones de los CDC [1], como para datos del COVID-19, ofrecidos por la WHO [2].

Adicionalmente, se ha dedicado una sección para hacer una comparativa entre los distintos modelos utilizados en el estado del arte del 'forecasting' de Influenza.

La idea de este proyecto surge tras un trabajo que realizo un grupo de Google en 2020 donde introducen los modelos transformer en el 'forecasting' de influenza además de tratar con todos los modelos del estado del arte de este ámbito [3]. El objetivo inicial del proyecto era reproducir de alguna manera los resultados obtenidos en este, intentando trabajar los modelos ARIMA y transformer, añadiendo un apartado para el COVID-19.

- *javierdediosg@gmail.com*
- *Computación*
- *Trabajo tutorizado por: Oriol Ramos Terrades (Ciències de la Computació)*
- *Curso 2021/22*

Finalmente, se realizó una adaptación acordando los objetivos siguientes:

- Obtener datos de la gripe y el COVID-19, analizarlos y prepararlos para su uso.
- Estudiar el modelo ARIMA y SARIMA aplicarlos a los datos obtenidos.
- Obtener resultados satisfactorios con alguno de los modelos.
- Realizar una comparativa entre los distintos modelos más utilizados en este ámbito.

2 Estado del Arte

La previsión de series temporales es una de las técnicas de ciencia de datos más aplicadas y es por esto por lo que se pueden encontrar diversos estudios que han utilizado datos de internet con el mismo objetivo que busca obtener este trabajo, realizar 'forecast' de ratios de ILI. Entre ellos se encuentran el de Twitter [4], Wikipedia [5] o Google [6].

Google tiene el GTF (Google Flu Trends) que utiliza un modelo linear que estima ratios de gripe actuales gracias a las búsquedas de ciertos términos. Otros estudios han aprovechado el GTF con nuevas técnicas para mejorarlo, como ARGO [7] en 2015 con un modelo autorregresivo.

Más recientemente, el modelo ARGONet [8] desarrollado encima de ARGO, obteniendo resultados de estado del arte parar el 'forecasting' de ILI.

Para el 'forcasting' de ILI también han sido utilizadas técnicas de Deep Learning. En 2018, se entrenó un modelo LSTM [9] para predecir la prevalencia de la gripe usando datos de Google, clima, contaminación del aire y datos de supervisión virológica. Y más tarde, en 2019, se adaptó un modelo Seq2Seq [10] con un mecanismo de atención similar para predecir la prevalencia de la gripe y mostró que su enfoque superó a ARIMA y LSTM.

Por otro lado, precisamente en [3] se introducen al 'forecasting' los modelos transformer, modelos que aprovechan mecanismos de autoatención para modelar datos de secuencia, y por lo tanto puede aprender dependencias complejas de varias longitudes de datos en series de temporales.

3 Datos

Para la gripe se han utilizado los datos ofrecidos por los CDC, estos disponen de informes semanales desde el 1997 hasta la actualidad, tanto a nivel estatal, como nacional de Estados Unidos.

Para el caso del COVID-19, los datos han sido obtenidos de la Organización Mundial de la Salud que ofrece una base de datos que almacena datos diarios de nuevos casos y muertes a nivel mundial desde el 1 de marzo de 2020.

3.1 Gripe

Los CDC, ofrecen varias opciones a la hora de descargar los datos: Fuente (Source), Área de supervisión (Surveillance área) y Estaciones (Seasons). Hay disponibles dos fuentes de datos:

La WHO/NREVSS (World Health Organization/National Respiratory and Enteric Virus Surveillance System), con un total de 380 laboratorios entre ambas organizaciones, ofrece una versión de supervisión virológica de los datos, es decir, una versión "más técnica", con los totales de todos los diferentes tipos de especímenes de gripe. Mientras que la ILINET (U.S. Outpatient Influenza-like Illness Surveillance Network), es una red que consta con aproximadamente 3.200 proveedores de atención médica en los 50 estados. Esta fuente ofrece unos datos menos específicos, ya que recopila datos de todos los pacientes con ILI (influenza-like illness), es decir, afecciones con unos síntomas parecidos a la gripe. Para este sistema, una ILI se define como una fiebre igual o por encima de los 37,8º además de tos o dolor de garganta.

Respecto al área de supervisión, los CDC ofrecen 4 agrupaciones distintas de los datos: la Nacional, a nivel nacional de Estados Unidos; por HHS Regions, una división del terreno estadounidense agrupando los estados en 10 regiones con una sede de salud y servicios humanos en común; por divisiones de censo, un reparto del terreno estadounidense agrupando los estados en 9 divisiones, ya sea por leyes o regulaciones, cultura o factores económicos; o Estatal, a nivel de cada uno de los estados que forman Estados Unidos.

Por último, las Estaciones, es una selección sobre que estación o estaciones de datos te interesan. Una estación es el conjunto de las últimas 12 semanas de un año junto a las primeras 12 del siguiente. Ej. la estación 2010-11, agruparía de la semana 40 a la 52 de 2010 y de la 1 a la 12 de 2011.

Acorde con [3], se usará la versión de ILINET de los años 2010 al 2018. En un principio en este proyecto estaba previsto utilizar la versión de los datos a nivel nacional ya que, en comparación con el estatal, se trata de un dataset más reducido. Pero finalmente se acabó usando la versión estatal ya que sería más fácil evaluar los resultados.
Este es el aspecto del dataset utilizado:

#	Column	Non-Null Count	Dtype
0	REGION TYPE	23012 non-null	object
1	REGION	23012 non-null	object
2	YEAR	23012 non-null	int64
3	WEEK	23012 non-null	int64
4	%WEIGHTED ILI	23012 non-null	object
5	%UNWEIGHTED ILI	23012 non-null	object
6	AGE 0-4	23012 non-null	object
7	AGE 25-49	23012 non-null	object
8	AGE 25-64	23012 non-null	object
9	AGE 5-24	23012 non-null	object
10	AGE 50-64	23012 non-null	object
11	AGE 65	23012 non-null	object
12	ILITOTAL	23012 non-null	object
13	NUM. OF PROVIDERS	23012 non-null	object
14	TOTAL PATIENTS	23012 non-null	object

Este dataset ofrecía muchas columnas de datos, de las cuales se han acabado usando:

- 'REGION': contiene el nombre del estado. Utilizada para tratar los datos de forma separada por estados. Ej. Alabama.
- 'YEAR': año al que pertenecen los datos. Utilizada para generar la columna 'DATE'. Contiene valores desde el 2010 al 2018.
- 'WEEK': numero de la semana del año al que pertenecen los datos. Segunda columna utilizada para generar 'DATE'. Contiene valores del 1 al 52.
- 'TOTAL_PATIENTS': número total de personas atendidas en un proveedor de atención médica perteneciente a la ILINET. Utilizada para generar la columna 'ILIRATIO'.
- 'ILITOTAL': número total de pacientes diagnosticados con una ILI (afección similar a la gripe). Segunda columna utilizada para generar 'ILIRATIO'.
- 'DATE: fecha válida formada gracias a la función to_datetime de pandas.
  Ej. 'WEEK': 40 + 'YEAR': 2010 = 'DATE": 2010-10-04 (el día 4 de octubre pertenece a la 40ª semana del año).
- 'ILIRATIO': ratio entre el total de pacientes y los que fueron diagnosticados con una ILI.

$$'ILIRATIO' = \frac{'ILITOTAL'}{'TOTAL\_PATIENTS' + \varepsilon}$$

Donde ε = 10^{-9}, para evitar posibles divisiones entre 0.

Una vez organizadas las columnas que se utilizaran y limpiados todos los valores null, este es el aspecto del dataset:

#	Column	Non-Null Count	Dtype
0	REGION	22529 non-null	object
1	YEAR	22529 non-null	int64
2	WEEK	22529 non-null	int64
3	ILITOTAL	22529 non-null	int32
4	TOTAL PATIENTS	22529 non-null	int32
5	DATE	22529 non-null	datetime64[ns]
6	ILIRATIO	22529 non-null	float64

Este dataset como tal, solo será utilizado para hacer unos gráficos de referencia iniciales, como la Figura 1, Figura 2 y Figura 3.

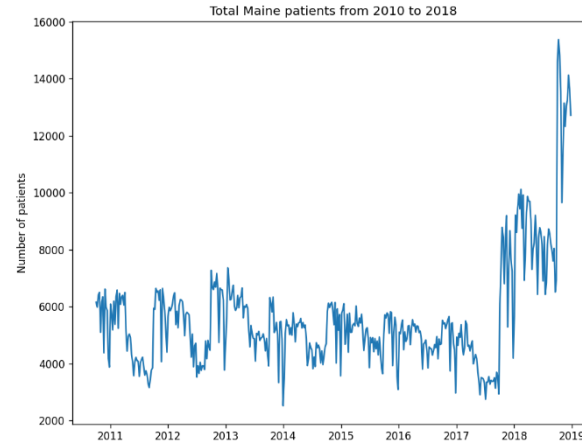

Figura 1. Pacientes totales en Maine del 2010 al 2018.

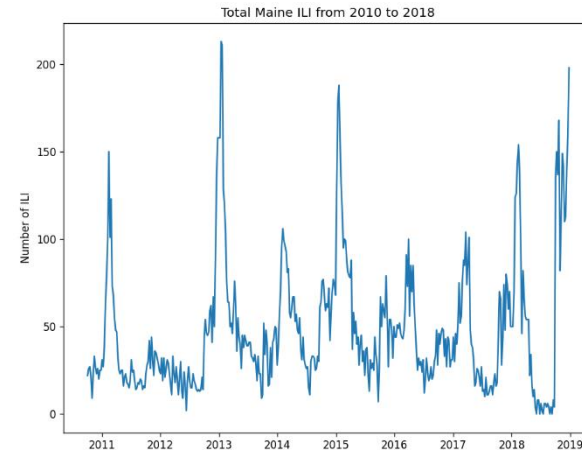

Figura 2. Numero de ILI en Maine del 2010 al 2018.

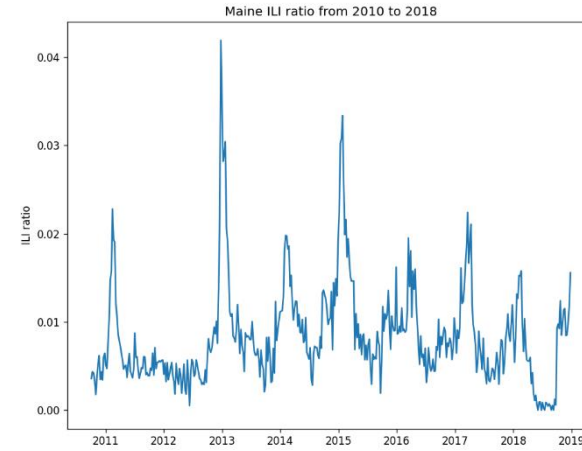

Figura 3. Ratio de ILI en Maine del 2010 al 2018.

En estas figuras se puede observar cómo curiosidad que a principios de 2018 hubo un gran pico de contagios. Esa temporada de gripe fue severa para todas las poblaciones de Estados Unidos y resultó en aproximadamente 41 millones de casos, 710.000 hospitalizaciones y 52.000 muertes. Este es el mayor número de casos desde la temporada de gripe de 2009 [11].

Una vez obtenido algo de contexto, hay que preparar los datos para los modelos, a partir de ahora ya no es necesario el dataset entero, así que solo se utilizaran las series temporales de cada estado. Estas están formadas por cada valor de 'ILIRATIO' y su fecha asociada. Ej. Maine:

DATE	ILIRATIO
2010-10-04	0.003577
2010-10-11	0.004349
2010-10-18	0.004217
2010-10-25	0.003384
2010-11-01	0.001765
…	…
2018-11-26	0.008434
2018-12-03	0.008536
2018-12-10	0.009985
2018-12-17	0.011723
2018-12-24	0.015572

[429 rows x 1 columns]

En el ejemplo del estado de Maine hay 429 filas y 1 columna. Cabe destacar que las fechas no cuentan como columna porque estas son el índice de cada fila de ILIRATIO.

3.2 COVID-19

El caso del COVID-19 se ha utilizado como otra oportunidad para aplicar lo aprendido con el caso de la gripe. Este es el aspecto del dataset antes de empezar a trabajar:

#	Column	Non-Null Count	Dtype
0	Date_reported	21550 non-null	object
1	Country_code	21550 non-null	object
1	Country	21550 non-null	object
2	WHO_region	21550 non-null	object
3	New_cases	21550 non-null	int64
4	Cumulative_cases	21550 non-null	int64
5	New_deaths	21550 non-null	int64
6	Cumulative_deaths	21550 non-null	int64

A diferencia de los datos ofrecidos por los CDC, estos son bastante auto explicativos y de las columnas que ofrece solo se utilizaran: 'Date_reported', 'Country' y 'New_cases'. Como son datos mundiales para este apartado, se ha decidido trabajar solo con los datos de España y Estados Unidos.

'Country' se utiliza para extraer los datos de España y Estados Unidos del resto de países, una vez creados dos grupos de datos para cada país, se ha creado una serie temporal como se hizo en el caso de los estados para la gripe.

Ej. Serie temporal de España:

Date_reported	New_cases
2020-01-11	1
2020-01-12	1
2020-01-14	1
2020-01-20	1
2020-01-23	1
…	…
2022-05-06	20828
2022-05-07	19355
2022-05-08	15993
2022-05-09	6054
2022-05-10	3638

[816 rows x 1 columns]

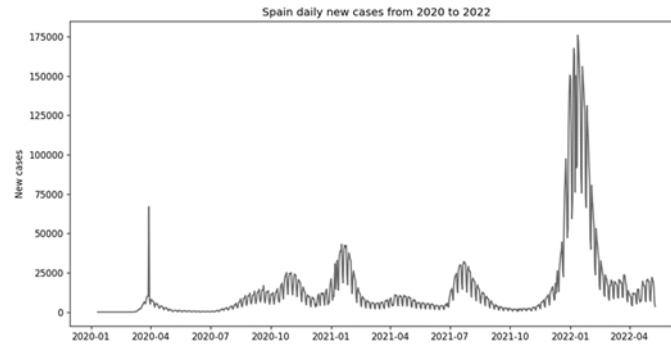

Figura 4. Nuevos casos de COVID-19 en España de 2020 a 2022.

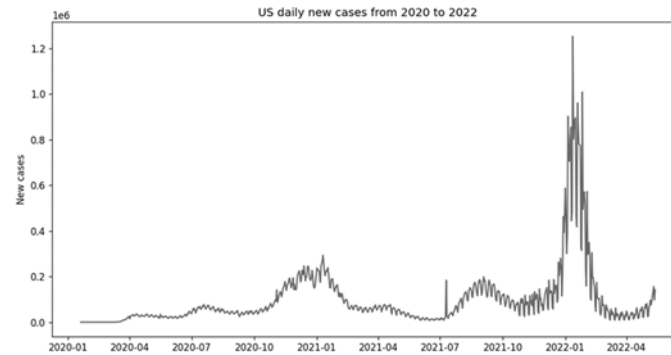

Figura 5. Nuevos casos de COVID-19 en Estados Unidos de 2020 a 2022.

4 MÉTODOS

4.1 ARIMA

El modelo ARIMA dispone de diferentes partes. Es por esto por lo que un modelo ARIMA necesita tres parámetros: 'p', 'd' y 'q'. Cada parámetro está relacionado con cada una de las partes que forman al modelo. El valor de 'p' es para el apartado AR y este quiere decir el número de valores autorregresivos. El 'd' es para el apartado I, que indica el número de veces que necesitaría el modelo diferenciar la serie para que fuese estacionaria. Y, por último, el 'q' del apartado MA, para saber cuántos términos MA se necesitan para eliminar toda autocorrelación en la serie.

Antes de aplicar un modelo autorregresivo como es el ARIMA, primero, hay que comprobar la estacionariedad de la serie. La estacionariedad, es un componente de una serie temporal contrario a la tendencia y la variación cíclica. Si una serie temporal es estacionaria y tiene un comportamiento particular durante un intervalo de tiempo determinado, se puede asumir que tendrá el mismo comportamiento en algún momento posterior. Esto es importante para este tipo de modelos ya que utilizan sus propios lags (una cantidad fija de tiempo pasado) como predictores.

Además, en una serie temporal estacionaria, la media, la variancia y la covariancia no dependen del tiempo y son constantes.

Hay distintos modos de comprobar la estacionariedad de una serie temporal, como con una gráfica de forma visual. Pero la manera más segura de comprobarlo es con el 'Augmented Dickey-Fuller Test'. Esta es una prueba de significancia estadística, lo que quiere decir que la prueba dará resultados en hipótesis, con hipótesis nula y alternativa. El ADF es una prueba de raíz unitaria, que prueba si una serie de tiempo no es estacionaria. La presencia de una raíz unitaria en la serie define la hipótesis nula, y la hipótesis alternativa define la serie temporal como estacionaria [12].

A continuación, un ejemplo del resultado obtenido de un ADF test en una serie temporal de uno de los estados:

ADF Statistic: -4.570520099360388
p-value: 0.00014633251823777013
Critical Values:
1%: -3.445721386098794
5%: -2.868316661451884
10%: -2.5703797268320376

El test de Dickey-Fuller aumentado devuelve un valor negativo y un valor p. Cuanto mayor sea la magnitud negativa del valor, mejor, ya que se busca que sea menor que los valores críticos. A parte, el valor de p debe estar por debajo del umbral de 0.05.

Con los datos obtenidos en el ejemplo, se puede rechazar la hipótesis nula y por ello afirmar que la serie temporal es estacionaria, ya que no existe una raíz unitaria.

Es importante destacar que el resultado de este test no es absoluto y por lo tanto solo nos informa de que es probable que la serie sea estacionaria.

Como se ha mencionado anteriormente, para ejecutar el modelo ARIMA es necesario encontrar los mejores valores de 'p', 'd' y 'q'. Para 'd' ya se ha comprobado que, para los datos de la gripe, el valor será 0, ya que, según los test realizados, las series no necesitan diferenciación.

Respecto a 'p', se puede averiguar su valor inspeccionando la Autocorrelación Parcial (PACF), como muestra la Figura 6. La PACF, se puede entender como la correlación entre la serie y un lag, después de excluir las contribuciones del resto de lags. De este modo, se sabrá si ese lag es necesario en el término AR o no [13].

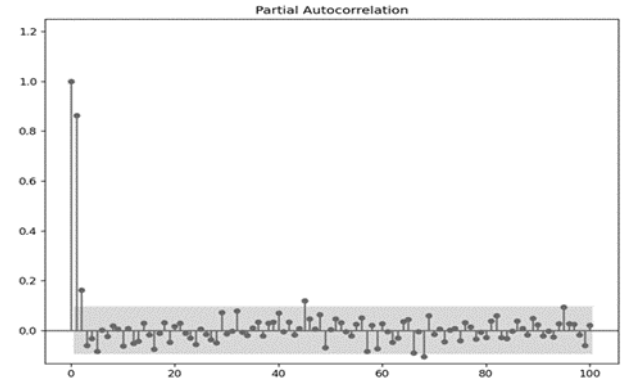

Figura 6. Autocorrelación parcial de la serie del estado de Maine.

En la Figura 6, se puede distinguir una zona sombreada, esta es la línea de significación, lo que quiere decir, que cualquier valor dentro de esta no se puede distinguir de 0. Sin tener en cuenta el lag 0, que siempre tendrá el valor máximo, solo tres lags están claramente por encima de la línea de significación. Por este motivo se considera que probablemente el mejor valor para 'p' será 3.

Por último, para el valor de 'q', el método es parecido al anterior, pero en este caso la gráfica que se inspecciona es la de la Autocorrelación (ACF), como muestra la Figura 7.

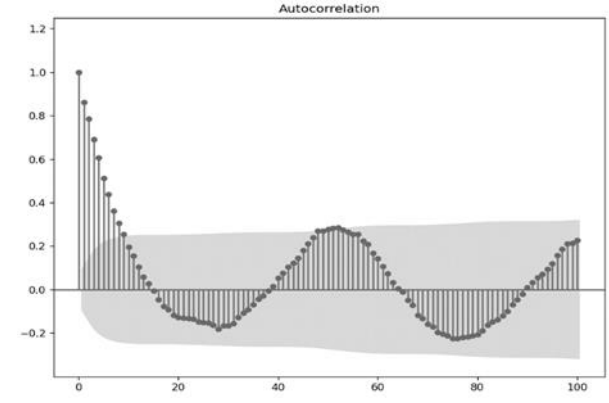

Figura 7. Autocorrelación de la serie del estado de Maine.

Se puede observar en la Figura 7 que en este caso son bastantes lo lags que sobrepasan la línea de significación. Pero lo más importante es que se puede distinguir un patrón muy claro. Esto hace sospechar de una posible estacionalidad, lo que tiene bastante sentido tratándose de datos de contagios de una afección como la gripe que suele tener los mismos picos en las mismas épocas cada año.

Por este motivo, se debería plantear el uso de un modelo SARIMA.

4.2 SARIMA

Como se comprobó anteriormente, los gráficos de ACF, mostraban un patrón que sugería estacionalidad, que para obtener unos mejores resultados es algo a tener en cuenta.

Pero un modelo ARIMA normal no es capaz de trabajar con esta característica, para ello se hará uso del SARIMA [14].

Si una serie temporal tiene una estacionalidad definida, SARIMA es mejor opción, ya que utiliza la diferenciación estacional. La diferenciación estacional es similar a la diferenciación regular, pero en lugar de restar términos consecutivos, se resta el valor de la temporada anterior.

Entonces, el modelo será representado como SARIMA(p,d,q)x(P,D,Q), donde, P, D y Q son SAR, orden de diferenciación estacional y términos SMA respectivamente y 'x' es la frecuencia de la serie temporal.

Como regla general, la configuración de los parámetros es muy parecida, pero hay que tener en cuenta que D nunca exceda uno. Y la diferenciación total 'd + D' nunca exceda 2.

5 Resultados

En el caso de la gripe se han utilizado ambos modelos, ARIMA y SARIMA para poder profundizar en la comparativa y buscar el modelo óptimo para esta serie. Mientras que para el COVID-19 se ha aplicación el modelo ARIMA para practicar con una fuente de datos distinta.

5.1 Gripe

ARIMA Como prueba, primero se ejecutó el modelo auto_arima de la librería pmdarima [15]. La característica de este modelo es que como su nombre indica, es automático, esto quiere decir que realiza varias iteraciones con diferentes valores de 'p', 'd' y 'q', buscando el mejor valor de AIC.

El criterio AIC (Akaike Information Criterion), es un método que proviene del campo de la probabilidad frecuentista y bayesiana. En resumen, el método con la puntuación más baja significa que pierde menos información y, por lo tanto, es mejor modelo [16].

Ej. de la ejecución ARIMA del auto_arima para la serie del estado de Maine:

Best model: ARIMA(3,0,0)(0,0,0)[0]
Total fit time: 0.955 seconds

Esto quiere decir, que ha determinado que el mejor modelo sería un ARIMA(3,0,0).

Esto coincide con el análisis que se ha hecho anteriormente, donde se ha determinado que los mejores valores serian: p = 3 y d = 0. El problema se encuentra en el apartado MA, en este caso ha comprobado que sin tener en cuenta la estacionalidad, el mejor parámetro para 'q' es 0.

Estos son los resultados del modelo ARIMA(3,0,0) para la serie del estado de Maine:

SARIMAX Results

Dep. Variable:	y	No. Observations:	285
Model:	SARIMAX(3, 0, 0)	Log Likelihood	1261.775
Date:	Sat, 21 May 2022	AIC	-2513.549
Time:	12:39:39	BIC	-2495.287
Sample:	0	HQIC	-2506.228
	- 285		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0011	0.000	2.947	0.003	0.000	0.002
ar.L1	0.7551	0.039	19.120	0.000	0.678	0.833
ar.L2	0.2156	0.076	2.826	0.005	0.066	0.365
ar.L3	-0.0945	0.054	-1.740	0.082	-0.201	0.012
sigma2	8.287e-06	3.68e-07	22.515	0.000	7.57e-06	9.01e-06

Los resultados ofrecen mucha información así que se analizará por partes:

SARIMAX Results

Dep. Variable:	y	No. Observations:	285
Model:	SARIMAX(3, 0, 0)	Log Likelihood	1261.775
Date:	Sat, 21 May 2022	AIC	-2513.549
Time:	12:39:39	BIC	-2495.287
Sample:	0	HQIC	-2506.228
	- 285		
Covariance Type:	opg		

En este apartado, a parte de la información básica, como el nombre del modelo, con sus parámetros y el número de observaciones, también ofrece cuatro valores que ayudan a la hora de evaluar el resultado [17]:

- Log Likelihood: este valor no es muy usado, pero por lo general, cuanto mayor sea, mejor.
- AIC: descrito anteriormente, cuanto menor el valor, mejor.
- BIC: (Bayesian Information Criterion) muy parecido al AIC, ambos valores penalizan la adición de más parámetros para compensar posible overfitting, aunque el BIC castiga más. Mismo objetivo, cuanto más pequeño, mejor.
- HQIC: (Hannan–Quinn Information Criterion) similar a los anteriores, pero usado menos frecuentemente.

Es importante mencionar que ninguno de los valores anteriores significa mucho por sí mismo. Estos valores son útiles a la hora de comparar modelos entre ellos.

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0011	0.000	2.947	0.003	0.000	0.002
ar.L1	0.7551	0.039	19.120	0.000	0.678	0.833
ar.L2	0.2156	0.076	2.826	0.005	0.066	0.365
ar.L3	-0.0945	0.054	-1.740	0.082	-0.201	0.012
sigma2	8.287e-06	3.68e-07	22.515	0.000	7.57e-06	9.01e-06

En este segundo apartado, aparece la tabla de coeficientes y en las filas se encuentran: ar.L1, ar.L2 y ar.L3, que representan los 3 lags AR añadidos en la configuración del modelo.

- La columna 'coef', representa la importancia de cada elemento.
- La columna 'std err' es una estimación del error del valor predicho. Dice como de fuerte es el efecto del error residual en sus parámetros estimados.

- La columna 'z' es igual a los valores de 'coef' divididos por 'std err'. Por lo tanto, es el coeficiente estandarizado.
- La columna 'P>|z|' es el valor p del coeficiente. Este valor es el mismo valor p que se mencionó en el test ADF.
- Las dos últimas columnas representan los intervalos de confianza.

Por el momento no hay otro modelo con el que compararlo así que la mejor manera de cuantificar como de buenos han sido los resultados es a través de una representación gráfica como la de la Figura 8.

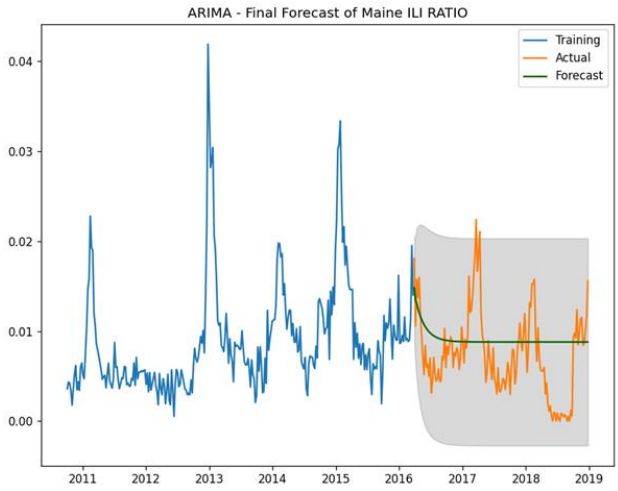

Figura 8. Resultado de las predicciones del modelo ARIMA(3,0,0) para la serie del estado de Maine.

En la Figura 8, el modelo parece dar un pronóstico direccionalmente correcto. Y los valores reales se encuentran mayoritariamente dentro de la banda de confianza del 95%. Aun así, es evidente que los valores predichos no se ajustan a los valores reales.

SARIMA Después de obtener los resultados del modelo ARIMA, como se ha mencionado anteriormente en el análisis de la autocorrelación, se podía sospechar una estacionalidad, por este motivo a continuación se ejecutará el modelo SARIMA. Para la ejecución de este modelo, se ha utilizado el mismo auto_arima de pmdarima, pero esta vez, se ha activado la estacionalidad y se ha añadido una frecuencia de 52, ya que las series están formadas por valores semanales.
Ej. de la ejecución SARIMA del auto_arima para la serie del estado de Maine:

Best model: ARIMA(1,0,1)(2,1,2)[52]
Total fit time: 1907.721 seconds

En este caso, ha determinado que el mejor modelo sería un SARIMA(1,0,1)(2,1,2)[52].

Cabe destacar la diferencia en los tiempos de ejecución, de 0.955 segundos para el modelo ARIMA, a 1907.721 segundos (≈32 min) para el SARIMA.

A continuación, los resultados del modelo SARIMA(1,0,1)(2,1,2)[52] para la serie del estado de Maine:

SARIMAX Results

Dep. Variable:	y	No. Observations:	285
Model:	SARIMAX(1, 0, 1)x(2, 1, [1, 2], 52)	Log Likelihood	998.893
Date:	Sat, 21 May 2022	AIC	-1981.786
Time:	15:23:04	BIC	-1954.178
Sample:	0	HQIC	-1970.653
	- 285		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	0.0003	0.001	0.426	0.670	-0.001	0.002
ar.L1	0.8627	0.035	24.592	0.000	0.794	0.931
ma.L1	-0.1562	0.075	-2.087	0.037	-0.303	-0.010
ar.S.L52	-0.4533	3.261	-0.139	0.889	-6.845	5.938
ar.S.L104	-0.1604	0.453	-0.354	0.723	-1.048	0.727
ma.S.L52	-0.4345	3.154	-0.138	0.890	-6.616	5.747
ma.S.L104	-0.0870	2.637	-0.033	0.974	-5.256	5.082
sigma2	8.972e-06	1.48e-06	6.053	0.000	6.07e-06	1.19e-05

Se puede ver que la disposición de los resultados es la misma que para el modelo ARIMA, la diferencia más notable es que encontramos más elementos en la tabla de coeficientes, estos son los parámetros que añade el modelo SARIMA.

Llama la atención que a pesar de que en teoría este es un modelo que se ajusta mejor a nuestros datos, la comparativa de los valores de AIC y BIC no es favorable. Esto es porque son modelos distintos y no se pueden comparar. Igual que el ARIMA(3,0,0) tenía un mejor AIC que, por ejemplo, un ARIMA(2,0,0). El SARIMA(1,0,1)(2,1,2)[52] tiene un mejor AIC que un SARIMA(0,0,1)(2,1,2)[52].

Una vez más para comprobar lo correctos que son los valores predichos, se utiliza la representación gráfica en la Figura 9.

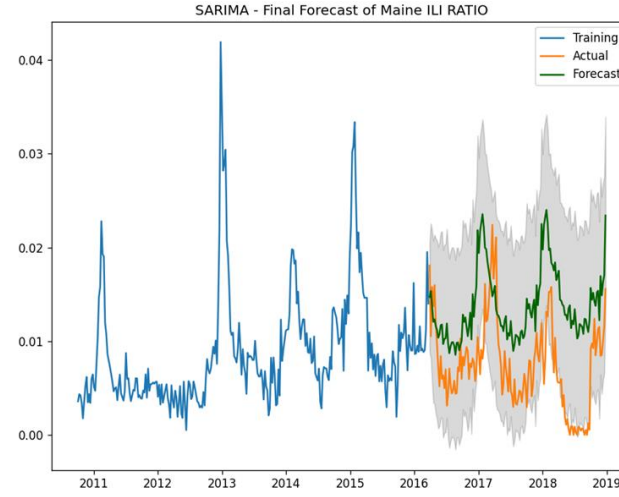

Figura 9. Resultado de las predicciones del modelo SARIMA(1,0,1)(2,1,2)[52] para la serie del estado de Maine.

Estos son unos resultados mucho más satisfactorios, no solo los valores reales se encuentran mayoritariamente dentro de la banda de confianza del 95%, pero, además, la predicción tiene una forma muy parecida a la serie actual [18].

5.2 COVID-19

A la hora de comprobar la estacionariedad de la serie de COVID-19 con test como el ADF mencionado anteriormente, no se llega a un resultado concluyente. Para este caso se puede descartar la estacionalidad, pero para ayudar a reducir el aumento de la media, se aplica el logaritmo a la serie como muestra la Figura 10.

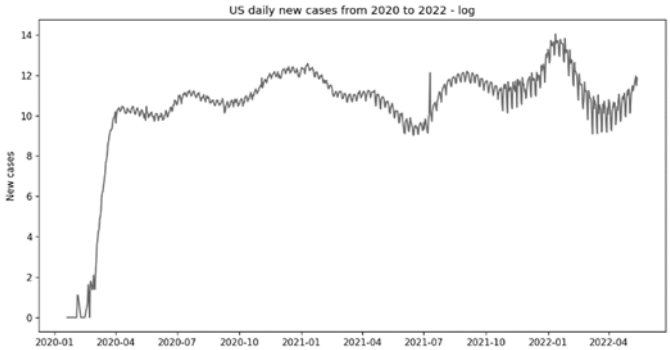

Figura 10. Logaritmo de los nuevos casos de COVID-19 en Estados Unidos de 2020 a 2022.

Tras esta transformación, en los test de estacionariedad se puede rechazar la hipótesis nula. Por lo tanto, el siguiente paso ya puede ser ejecutar el modelo ARIMA.

Se ejecuta el mismo modelo auto_arima pero esta vez para la serie del COVID-19 de Estados Unidos y este es el resultado:

Best model: ARIMA(1,0,3)(0,0,0)[0]
Total fit time: 2.492 seconds

Tras determinar que el mejor modelo para estos datos es el ARIMA(1,0,3), se pueden comprobar los resultados de este modelo:

SARIMAX Results

Dep. Variable:	y	No. Observations:	547
Model:	SARIMAX(1, 0, 3)	Log Likelihood	-23.026
Date:	Sun, 22 May 2022	AIC	56.053
Time:	01:41:04	BIC	77.575
Sample:	0	HQIC	64.465
	- 547		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.9997	0.000	2014.873	0.000	0.999	1.001
ma.L1	-0.0399	0.018	-2.236	0.025	-0.075	-0.005
ma.L2	-0.1009	0.026	-3.890	0.000	-0.152	-0.050
ma.L3	-0.0645	0.024	-2.650	0.008	-0.112	-0.017
sigma2	0.0629	0.001	43.873	0.000	0.060	0.066

Después del análisis de resultados del apartado anterior, se puede observar que los resultados no parecen muy prometedores. Pero esta es la mejor configuración que se ha encontrado para estos datos, así que, para comprobar los resultados se vuelve a generar una gráfica como la Figura 11.

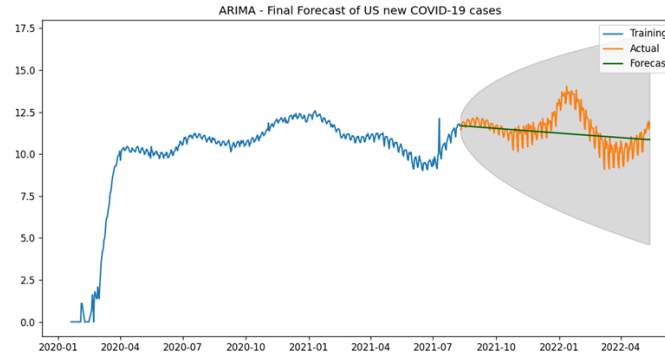

Figura 11. Resultado de las predicciones del modelo ARIMA(1,0,3) para la serie de Estados Unidos.

La situación de estos resultados es muy parecida a la obtenida en la ejecución del modelo ARIMA para los datos de la gripe. La dirección del pronóstico parece correcta y la totalidad de los datos reales se encuentra dentro de la banda de confianza del 95%. Por lo que estos resultados son bastante satisfactorios tras los problemas que han causado.

6 DISCUSIÓN DE LOS RESULTADOS

Como el foco principal del trabajo estaba en los datos de la gripe, a continuación, se analizarán con más profundidad.

Además de gráficamente, para evaluar los resultados obtenidos se han calculado la correlación y el error cuadrático medio (RMSE). Por un lado, la correlación de Pearson se puede definir como un índice para medir el grado de relación entre dos variables. Y por otro, el RMSE mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. La división de los datos es de 66% y 33% para train y test respectivamente para todos los modelos.

Se ha realizado una ejecución de cada modelo para cada uno del total de 37 estados que se han analizado y los resultados se encuentran en las Figuras 12 y 13.

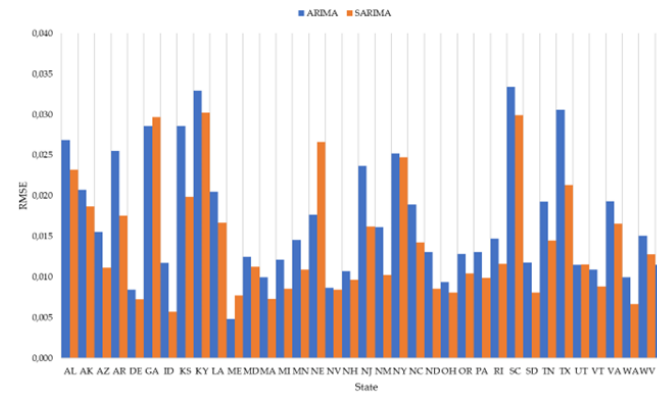

Figura 12. RMSE del modelo ARIMA y SARIMA.

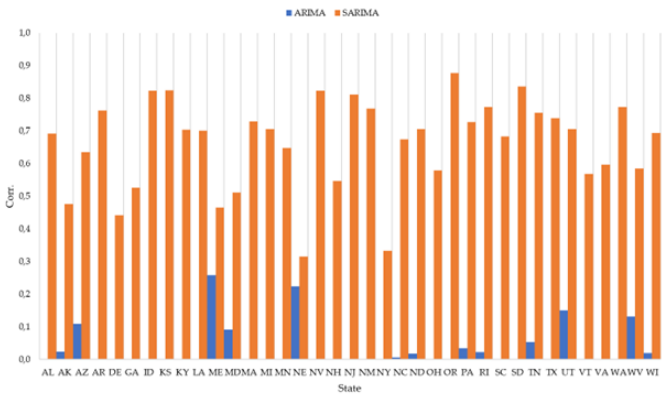

Figura 13. Correlación de Pearson del modelo ARIMA y SARIMA.

Estos valores concuerdan con los resultados gráficos obtenidos en los ejemplos de las Figuras 8 y 9. Para una mejor visualización, también se han calculado la media de ambos coeficientes entre todos los estados:

- RMSE medio para el modelo ARIMA: 0,01705.
- RMSE medio para el modelo SARIMA: 0,0142.
- Correlación de Pearson media para el modelo ARIMA: -0,0515.
- Correlación de Pearson media para el modelo SARIMA: 0,6624.

Teniendo esto en cuenta, ha habido estados con mejores resultados que otros, a lo largo de este documento se ha usado el estado de Maine debido a que ha dado uno de los mejores resultados para todos los modelos. Pero Idaho ha dado también muy buenos resultados para el modelo SARIMA en concreto, como muestra la Figura 14.

En su defecto también ha habido estados como Kentucky que han dado resultados poco satisfactorios, como muestran las Figuras 15 y 16.

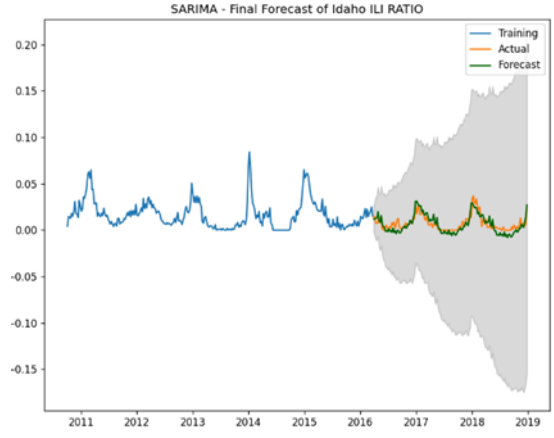

Figura 14. Resultado de las predicciones del modelo SARIMA para la serie del estado de Idaho.

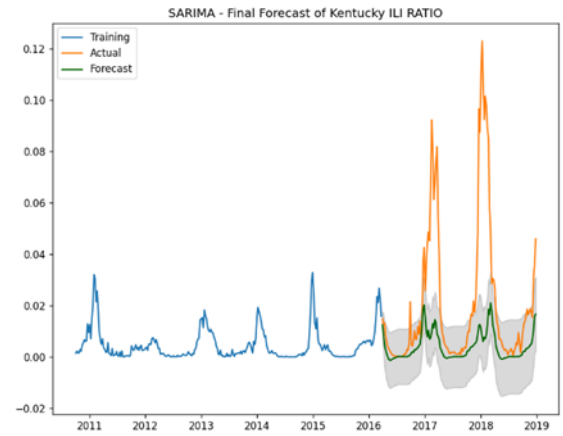

Figura 15. Resultado de las predicciones del modelo SARIMA para la serie del estado de Kentucky.

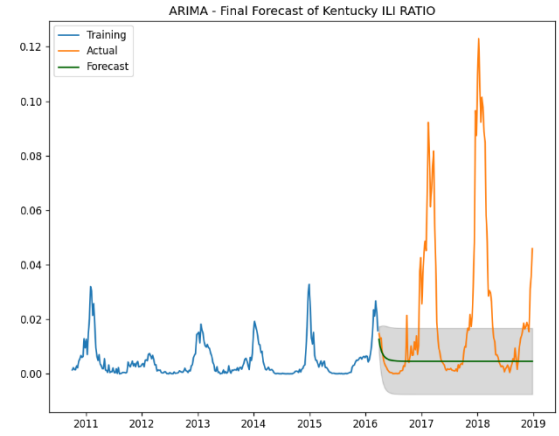

Figura 16. Resultado de las predicciones del modelo ARIMA para la serie del estado de Kentucky.

En [3] el objetivo principal es generar y aplicar un modelo transformer en el ámbito del 'forecasting' para series temporales y compararlo con el estado del arte. Para ello, deciden utilizar los datos de la gripe en distintos modelos como: el ARIMA, el LSTM y el Seq2Seq. Establecen al ARIMA como el modelo base y comparando los resultados de todos los modelos se obtiene una tabla como esta:

Modelo	Pearson Correlation	RMSE
ARIMA	0.769 (+0 %)	1.020 (-0 %)
LTSM	0.924 (+19.9 %)	0.807 (-20.9 %)
Seq2Seq	0.920 (+19.5 %)	0.642 (-37.1 %)
Transformer	0.928 (+20.7 %)	0.588 (-42.4 %)
ARIMA (TFG)	-0.051	0.017
SARIMA (TFG)	0.662	0.014

Los resultados demuestran que el resto de los modelos suponen un cierto nivel de mejora sobre el modelo ARIMA. El Transformer es el que ofrece los mejores resultados sobre todo en el apartado del RMSE con más de un 40% de diferencia,

mientras que en la correlación prácticamente todos ofrecen una mejora del 20% aproximadamente. Estos resultados concuerdan con los estudios de estado del arte.

Los resultados de este proyecto, aunque satisfactorios respecto al trabajo realizado, quedan un poco por debajo de estos valores. Aun así, esta desemejanza tiene sentido ya que ha habido diferencias en el trato de datos entre ambos proyectos.

La primera intención era reproducir el proceso de [3] pero estos no son muy explícitos en cuanto a su trato de los datos. La mayor diferencia es que en su caso se predecía una semana tomando 10 como referencia, mientras que en este proyecto se tomaban como referencia más o menos 6 años, para predecir los 3 siguientes. Esto resulta en menos precisión y, por tanto, en resultados ligeramente inferiores.

La decisión de utilizar los datos de este modo fue tomada tras investigar un como utilizaba la mayoría de los usuarios el modelo ARIMA en diferentes ejemplos y análisis encontrados en Internet, donde prácticamente nadie utilizaba el método de 'sliding window' que usan ellos en su proyecto.

Una razón posible por la que ellos decidieron hacer uso de esta técnica es porque su trabajo se centraba en los transformers, además del resto de modelos y no solo en el ARIMA, por tanto, quizá este enfoque era más adecuado.

7 Conclusiones

Finalmente, tras haber acabado el trabajo, se puede hacer un balance con perspectiva.

Si que es cierto que la idea inicial de lo que sería este proyecto fue sutilmente adaptada, ya que en un principio se tenía pensado llegar a trabajar con el modelo transformer. Pero esta idea se descartó en las primeras fases del trabajo por su complejidad y consumo de tiempo. En su lugar se decidió dedicar todo el tiempo disponible en entender bien los datos y los modelos ARIMA y SARIMA.

Por este motivo se han cumplido todos los objetivos que se acabaron estableciendo.

Se ha realizado un análisis y estudio tanto de los datos de la gripe como de los modelos, obteniendo resultados satisfactorios con el modelo SARIMA. Con esto se ha ganado una experiencia que se ha utilizado para obtener también unos resultados decentes con el modelo ARIMA para los datos del COVID-19.

La parte más problemática resultó ser la planificación y las primeras fases del trabajo, por la falta de experiencia en este ámbito. Pero una vez superada esta, el resto del trabajo se pudo completar sin grandes imprevistos. Por lo tanto, este proyecto ha acabado también siendo un buen problema de gestión de tiempo y de adaptación a imprevistos.

Bibliografia

[1] Centros para el Control y la Prevención de Enfermedades, "Datos clave sobre la influenza", https://espanol.cdc.gov/flu/about/. 2021.

[2] World Health Organization, "WHO Coronavirus (COVID-19) Dashboard", https://covid19.who.int/data. 2022.

[3] Neo Wu, Bradley Green, Xue Ben, Shawn O'Banion, "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case", https://arxiv.org/abs/2001.08317. 2020.

[4] Michael J Paul, Mark Dredze, David Broniatowski, "Twitter improves influenza forecasting", https://pubmed.ncbi.nlm.nih.gov/25642377/. 2014.

[5] David J. McIver, John S. Brownstein, "Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time", https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3990502/. 2014.

[6] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, Larry Brilliant, "Detecting influenza epidemics using search engine query data", https://pubmed.ncbi.nlm.nih.gov/19020500/. 2009.

[7] Shihao Yang, Mauricio Santillana, S C Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO", https://pubmed.ncbi.nlm.nih.gov/26553980/ . 2015.

[8] Fred S Lu, Mohammad W Hattab, Cesar Leonardo Clemente, Matthew Biggerstaff, Mauricio Santillana, "Improved state-level influenza nowcasting in the United States leveraging Internet-based data and network approaches", https://pubmed.ncbi.nlm.nih.gov/30635558/. 2019.

[9] Liyuan Liu, Meng Han, Yiyun Zhou, Yan Wang, "Bioinformatics Research and Applications", https://link.springer.com/chapter/10.1007/978-3-319-94968-0_25. 2018.

[10] Kenjiro Kondo, Akihiko Ishikawa, Masashi Kimura, "Sequence to Sequence with Attention for Influenza Prevalence Prediction using Google Trends", https://dl.acm.org/doi/abs/10.1145/3365966.3365967?casa_token=ZCSh7orZH6QAAAAA:pgAg4x8YvjN1de_gAQbFH5LfpwcrSzt1X-CpOv1UtURMiiilSR94qIGrw2SQ5D9waumVfcH9-ngC. 2019.

[11] CDC, "Summary of the 2017-2018 Influenza Season", https://www.cdc.gov/flu/about/season/flu-season-2017-2018.htm. 2019.

[12] Yugesh Verma, "Complete Guide To Dickey-Fuller Test In Time-Series Analysis", https://analyticsindiamag.com/complete-guide-to-dickey-fuller-test-in-time-series-analysis/. 2021.

[13] Xiaoyu Sun, "Understand ARIMA and tune P, D, Q", https://www.kaggle.com/code/sumi25/understand-arima-and-tune-p-d-q/notebook. 2018.

[14] Adhistya Erna Permanasari, Indriana Hidayah, Isna Alfi Bustoni, "SARIMA (Seasonal ARIMA) implementation on time series to forecast the number of Malaria incidence", https://ieeexplore.ieee.org/document/6676239. 20130.

[15] pmdarima, "Project description", https://pypi.org/project/pmdarima/. 2022.

[16] Shachi Kaul, "Probabilistic Model Selection with AIC/BIC in Python", https://medium.com/analytics-vidhya/probabilistic-model-selection-with-aic-bic-in-python-f8471d6add32. 2020.

[17] Nikol Holicka, "Interpreting ARMA model results in Statsmodels for absolute beginners", https://medium.com/analytics-vidhya/interpreting-arma-model-results-in-statsmodels-for-absolute-beginners-a4d22253ad1c#:~:text=The%20log%2Dlikelihood%20value%20is,log%2Dlikelihood%2C%20the%20better. 2019.

[18] Selva Prabhakaran, "ARIMA Model – Complete Guide to Time Series Forecasting in Python", https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/. 2021.