



UNIVERSIDAD
AUSTRAL | INGENIERÍA

Análisis de Series Temporales

Contaminación del Aire

Equipo #1:
Maria Cecilia Arce
Maria Virginia Ainchil
Alejandra Cecco

Trabajo Práctico #1

10 de Julio, 2024.

Resumen Ejecutivo

El presente trabajo práctico se realizó con una base de datos que mide la calidad del aire, tomando en cuenta 13 variables; de las cuales 4 son variables continuas. Se tomaron en cuenta 3 de estas variables para seleccionar las series temporales en las que se realizó el análisis exploratorio, la descomposición de las series y la aplicación de modelos ARIMA; con el objetivo de poder predecir futuros valores para estas variables.

Se ejecutaron distintos modelos, y se midieron los performance de los mismos, para determinar cuál de ellos es el que mejor explica y predice cada una de las series temporales.

Contenido

Resumen Ejecutivo.....	1
Contenido.....	2
Introducción.....	3
Propiedades de las Series Temporales.....	5
Estacionariedad.....	7
Estacionariedad de la Serie Temporal Benceno (C6H6).....	8
Estacionariedad de la Serie Temporal Temperatura (T).....	11
Estacionariedad de la Serie Temporal Humedad Relativa (RH).....	14
Modelos SARIMA.....	16
Modelos para Serie C6H6.....	17
Modelos para Serie T.....	18
Modelos para Serie RH.....	19
Performance de los Modelos Seleccionados.....	20
Performance Modelo C6H6.....	21
Performance Modelo T.....	22
Performance Modelo RH.....	23
Análisis de Residuos de los Modelos.....	23
Residuos Modelo C6H6.....	24
Residuos Modelo T.....	26
Residuos Modelo RH.....	27
Pronóstico.....	28
Pronóstico de 24 horas para C6H6.....	29
Pronóstico de 24 horas para T.....	29
Pronóstico de 24 horas para RH.....	29
Modelo SARIMA-X.....	30
Performance Modelo SARIMA-X C6H6 con la variable exógena T.....	32
Residuos Modelo SARIMA-X C6H6 con la variable exógena T.....	32
Prueba de Hipótesis HEGY.....	33
Conclusiones.....	35
Bibliografía.....	36
Anexo 1 (cuadro comparativo métrica de los modelos).....	37
Anexo 2 (Transformación de Box-Cox en variable C6H6).....	38

Introducción

Desde hace ya algunos años, el tema del cuidado ambiental, ha tomado protagonismo como una prioridad para el correcto desarrollo de la humanidad. En este trabajo se pretende analizar particularmente la calidad del aire que respiramos.

Se considera que los contaminantes de la atmósfera en zonas urbanas, son responsables del aumento de enfermedades respiratorias en los seres humanos; y se ha descubierto, que algunos contaminantes en particular, como es el caso del benceno (C_6H_6), aumentan la probabilidad de contraer cáncer.¹ Es por eso, que una estimación precisa sobre la distribución de contaminantes en el aire que respiramos, puede ser relevante para tomar medidas de prevención en temas de salud.

Hoy en día, el seguimiento de la contaminación del aire en zonas urbanas, se lleva a cabo mediante redes de estaciones fijas distribuidas estratégicamente. Estos equipos se encargan de estimar de forma selectiva y precisa las concentraciones de muchos componentes presentes en la atmósfera.

En este contexto, el análisis de series temporales se ha establecido como una herramienta crucial para monitorear y comprender la dinámica de la calidad del aire a lo largo del tiempo. Esta metodología permite identificar patrones estacionales, tendencias a largo plazo, variaciones diarias y eventos extremos que afectan la concentración de contaminantes atmosféricos. Al analizar datos recopilados de estaciones de monitoreo distribuidas estratégicamente, es posible evaluar la eficacia de las políticas de control de la contaminación, así como anticipar y mitigar los impactos adversos en la salud pública y el medio ambiente.

El conjunto de datos con el que se realizó este trabajo práctico, contiene 9358 observaciones; que provienen del promedio por hora registrado en una serie de 5 sensores químicos integrados en un dispositivo con el objetivo de medir y controlar la calidad del aire. Dicho dispositivo, desarrollado por Pirelli Labs, se colocó en una zona significativamente contaminada de una ciudad italiana.

Los datos utilizados pertenecen a los valores registrados durante un año (de marzo del 2004 hasta febrero del 2005), todos los días, cada hora. Las variables

¹ S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia. (2008). "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario". ScienceDirect. <https://www.sciencedirect.com/science/article/abs/pii/S0925400507007691?via%3Dihub>

registradas por el sensor y presentes en el dataset, son 14. Como se mencionó anteriormente, el objetivo de este trabajo es analizar las series temporales de 3 variables, por lo que se seleccionaron las 3 variables continuas, más significativas para el análisis:

- **C6H6:** corresponde al promedio de benceno registrado en el aire. La unidad en la que se trabaja esta variable es microrg/m³.
- **T:** se refiere a la temperatura del aire. Está medido en °C. Como se mencionó anteriormente, el estudio se realizó en Italia, por lo que las temperaturas bajan en el invierno italiano (noviembre, diciembre, enero) y suben en el verano (junio, julio, agosto).
- **RH:** se refiere a la humedad relativa. Se mide en %.

Marco Teórico

Propiedades de las Series Temporales

Se conoce como serie temporal, a una secuencia de N observaciones (datos) ordenadas y equidistantes cronológicamente sobre una característica (serie univariante o escalar) o sobre varias características (serie multivariante o vectorial de una unidad observable en diferentes momentos).²

El primer objetivo del análisis de una serie temporal es elaborar un modelo estadístico que describa adecuadamente la procedencia de dicha serie, de manera que las implicaciones teóricas del modelo resulten compatibles con las pautas muestrales observadas en la serie temporal. Una vez generado este modelo, puede utilizarse para:

- Describir la evolución observada de la serie, así mismo como las relaciones contemporáneas y dinámicas entre sus componentes (en caso de tratarse de una serie multivariada).
- Prever la evolución futura de dicha serie y poder generar predicciones.

Para llevar a cabo el trabajo práctico, lo primero que se realizó fue graficar las 3 variables mencionadas, para entender los datos y analizar su comportamiento.

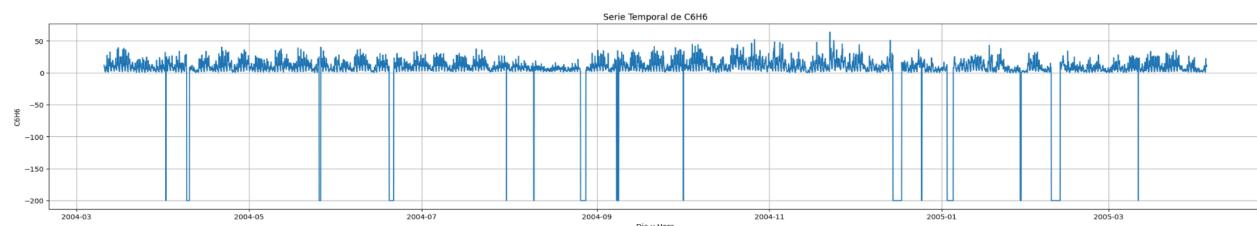


Figura 1. Serie Temporal Benceno C6H6 con datos originales.

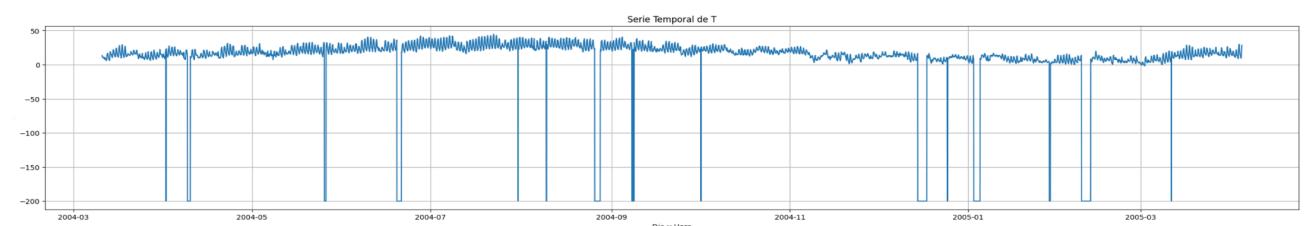


Figura 2. Serie Temporal Temperatura (T) con datos originales.

² Mauricio, José Alberto. (2007). "Introducción al Análisis de Series Temporales". Universidad Complutense de Madrid.

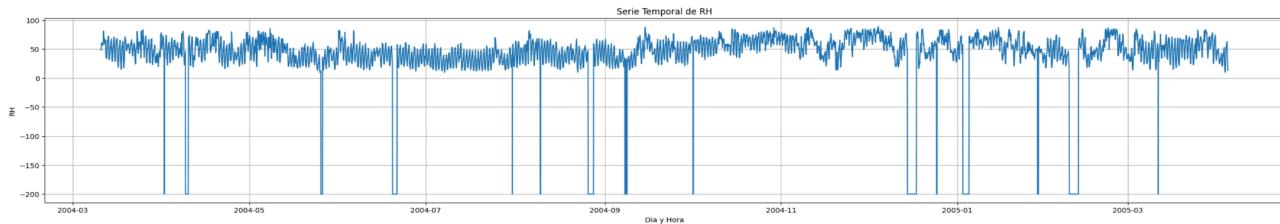


Figura 3. Serie Temporal Humedad Relativa (RH) con datos originales.

El dataset con el que se trabajó, no tenía valores nulos; pero al momento de analizar los datos, se descubrió que las 3 series temporales, tenían picos que iban hasta el valor -200; este valor de -200 se imputa automáticamente cuando el dispositivo tiene una falla de calibración. Se realizó una corrección a estos datos, para modificarlos, por valores más reales a través del método “Forward Fill”, el cual consiste en propagar el último valor válido registrado hacia adelante en el tiempo. De esta manera se logró sacar los picos que perturban tanto las series. El resto del trabajo se realizó con esta modificación sobre los datos originales.

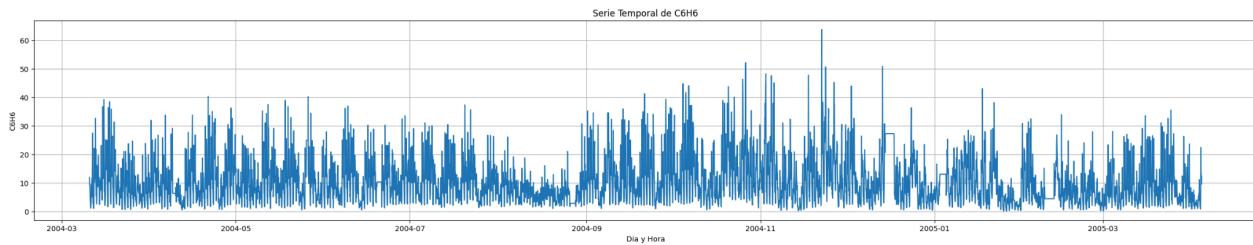


Figura 4. Serie Temporal Benceno (C6H6).

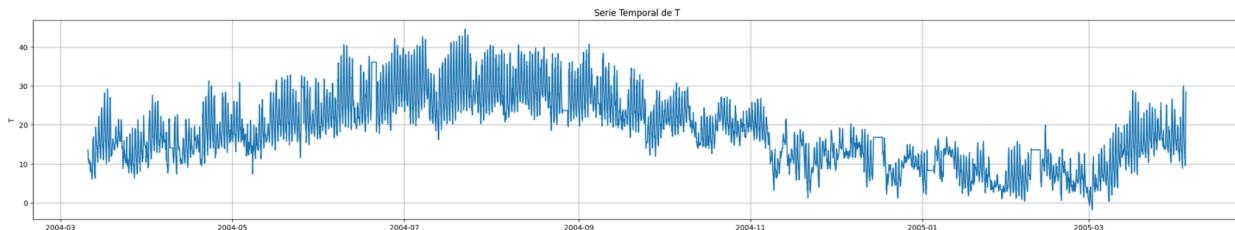


Figura 5. Serie Temporal Temperatura (T).

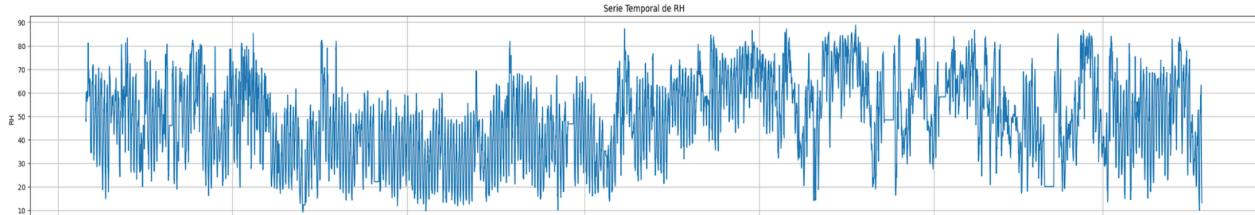


Figura 6. Serie Temporal Humedad Relativa (RH).

Las 3 series temporales muestran patrones distintos, por lo que es importante determinar si se trata de series estacionarias o no en cada caso, para determinar luego el modelo a implementar.

Estacionariedad

Se dice que una serie temporal es estacionaria cuando sus propiedades estadísticas, como la media, la varianza y la autocorrelación, son constantes a lo largo del tiempo. La estacionariedad en una serie puede ser:

- Estricta o Fuerte: cuando la distribución conjunta de cualquier conjunto de valores $Y_{t_1}, Y_{t_2}, \dots, Y_{t_k}$ es la misma que la distribución conjunta de $Y_{t_1}, Y_{t_2}, \dots, Y_{t_k} + h$ para cualquier h .
- Débil: cuando la media, la varianza y la autocovarianza son constantes en el tiempo.

Para poder implementar un correcto modelo univariante que explique y prediga el comportamiento de las variables analizadas en el tiempo, es necesario que las series cumplan con el principio de estacionariedad. Existen varios métodos para determinar si una serie es o no estacionaria, entre los que podemos mencionar:

- **Gráficos ACF Y PACF:** visualizar el comportamiento de una serie temporal ayuda a determinar si hay cambios evidentes en la media o en la varianza a lo largo del tiempo. La función de autocorrelación (ACF) de una serie estacionaria generalmente disminuye rápidamente; si no lo hiciera, estaríamos hablando de una serie no estacionaria. Estos gráficos son esenciales para identificar las características de la serie temporal y para seleccionar el orden adecuado de los

modelos ARIMA y SARIMA.

- **Tests de Raíces Unitarias:**

- **Test ADF (Dickey-Fuller Aumentado):** establece como H_0 que la serie temporal tiene una raíz unitaria; es decir que no estacionaria.
- **Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin):** es lo contrario a la prueba anterior. Establece como H_0 que la serie temporal es estacionaria, por lo que no presenta raíz unitaria.
- **Phillips-Perron:** es similar al ADF, pero maneja de manera distinta la autocorrelación. Establece como H_0 que la serie tiene una raíz unitaria, es decir no estacionaria.

Una vez que se realiza la exploración gráfica de la serie temporal, al igual que los test estadísticos, y los mismos sugieren que la serie no es estacionaria, el siguiente paso es diferenciar dicha serie o bien aplicar alguna transformación a los datos para hacerla estacionaria antes de modelar.

Muchas series temporales no pueden considerarse generadas por procesos estocásticos estacionarios (son no estacionarias) porque presentan ciertas tendencias claras en su evolución temporal (no presentan afinidad hacia algún valor constante en el tiempo), porque su dispersión no es constante, porque son estacionales, o por varias combinaciones de estos motivos.

A continuación se presenta el trabajo realizado a cada una de las series temporales para lograr la estacionariedad de las mismas.

Estacionariedad de la Serie Temporal Benceno (C₆H₆)

A la serie temporal que de la variable Benceno se le realizó una descomposición aditiva.

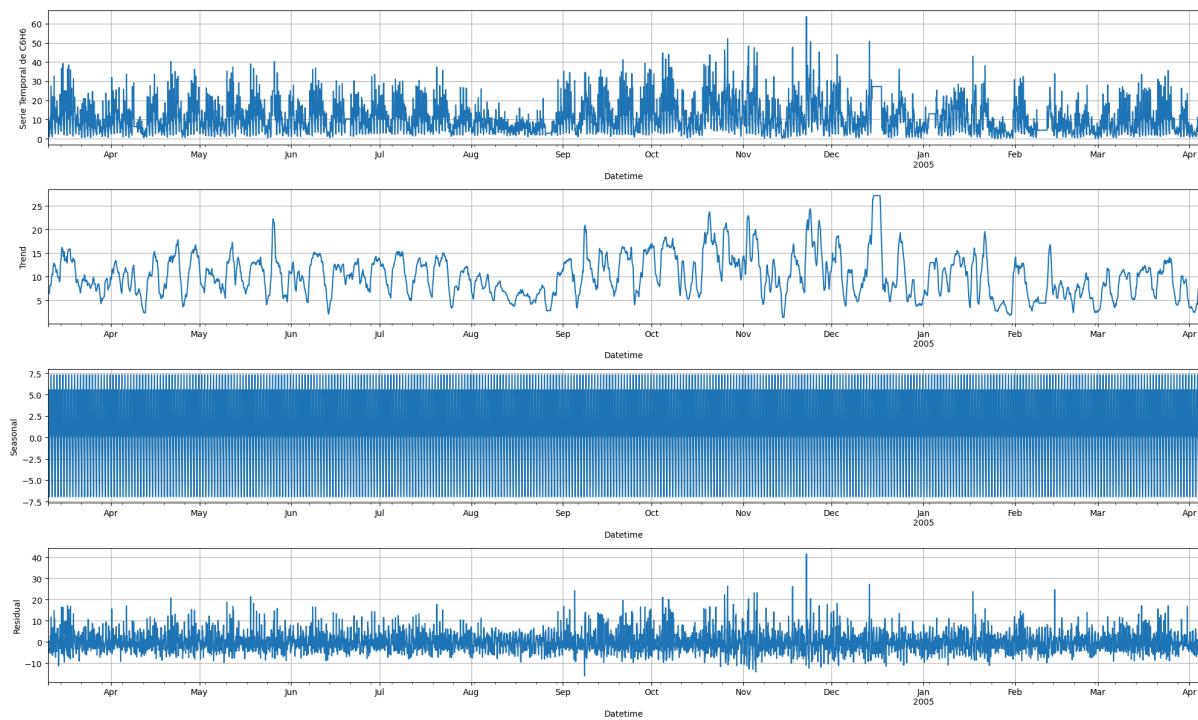


Figura 7. Descomposición aditiva de C6H6.

- La serie no parece tener tendencia, si en cambio una alta volatilidad.
- Se puede percibir una estacionalidad diaria en los datos: cada 24 horas la cantidad de benceno sigue un comportamiento similar, con dos picos uno por la mañana y otro por la tarde que probablemente corresponden al comportamiento del tráfico.
- Los residuos registran un random noise; a simple vista, no se puede definir como ruido balnco.

Gráficos ACF y PACF

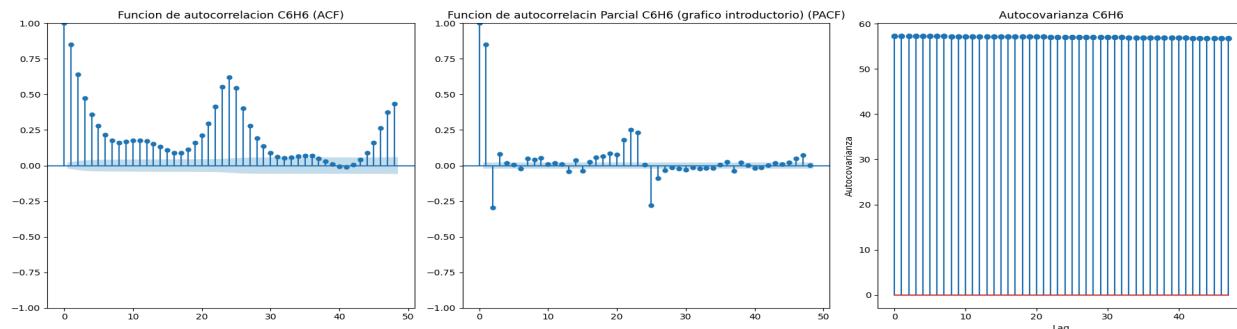


Figura 8. Gráficos ACF, PACF, Autocovarianza de C6H6.

El modo de caída exponencial (curva) del gráfico FAC que se repite cada 24 horas se puede interpretar como un patrón fuerte de estacionalidad diario. El hecho de que todos los valores estén fuera de la banda de confianza significa que las correlaciones son significativas, es decir, no son al azar. Los cortes abruptos en los gráficos de PFAC también indican estacionalidad. En el gráfico de autocovarianzas, los valores muy similares pueden interpretarse como una fuerte autocorrelación y una serie con memoria larga, lo que significa que los valores pasados de la serie tienen una influencia duradera sobre los valores futuros. También podría ser un indicio de que la serie no es estacionaria en la media.

Para validar lo observado visualmente, se realizaron distintos tests para probar la estacionariedad de la serie “C6H6”.

- **Test ADF (Dickey-Fuller Aumentado):** En esta prueba se obtuvo un p-valor de 0.00, que es menor que 0.05. Acorde a este resultado, se rechaza la H₀, por lo que se asume la H₁, indicando que la serie temporal es estacionaria. Previamente a aplicar el test se realizó una diferenciación correspondiente a la estacionalidad de 24 horas, por lo que no se necesitaría una diferenciación adicional para su modelado.
- **Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin):** El p-valor fue 0.1, que es mayor que 0.05, por lo que no se puede rechazar la H₀ y se asume que la serie es estacionaria.
- **Test PP (Phillips-Perron):** El p-valor obtenido fue 0,00 que es menor que 0.05, por lo que se rechaza la H₀. Se asume entonces que la serie es estacionaria.

Por último, se utilizó la función **ndiffs** de la librería pmdarima en Python, obteniendo como resultado 0, para confirmar que la serie no necesita más diferenciación.

En los siguientes gráficos de ACF (Autocorrelation Function) y PACF (Partial Autocorrelation Function) después de la diferenciación de 24 horas, aún se observa una marcada componente de estacionalidad de 24 horas.

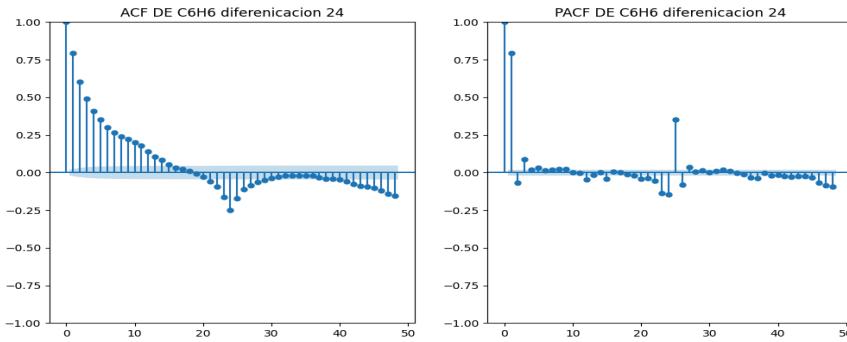


Figura 9. Gráficos ACF, PACF con serie C6H6 diferenciada.

Estacionariedad de la Serie Temporal Temperatura (T)

A la serie temporal de la variable temperatura se le realizó una descomposición aditiva.



Figura 10. Descomposición aditiva de T.

En la descomposición de esta serie podemos observar:

- La serie tiene estacionalidad, la cual ocurre por razones de estación: en los meses de verano la curva sube registrando mayores temperaturas y durante los meses de invierno la curva decrece. Esto genera una tendencia clara en los datos.
- Al mismo tiempo, se puede percibir una estacionariedad diaria de los datos: cada 24 horas la temperatura sigue un comportamiento similar registrando menores valores en las horas de la noche y alcanzando su valor máximo al medio día.
- Los residuos registran un random noise; se podría indicar a simple vista que no muestran correlación significativa y no contienen información adicional para agregar al modelo.

Gráficos ACF y PACF

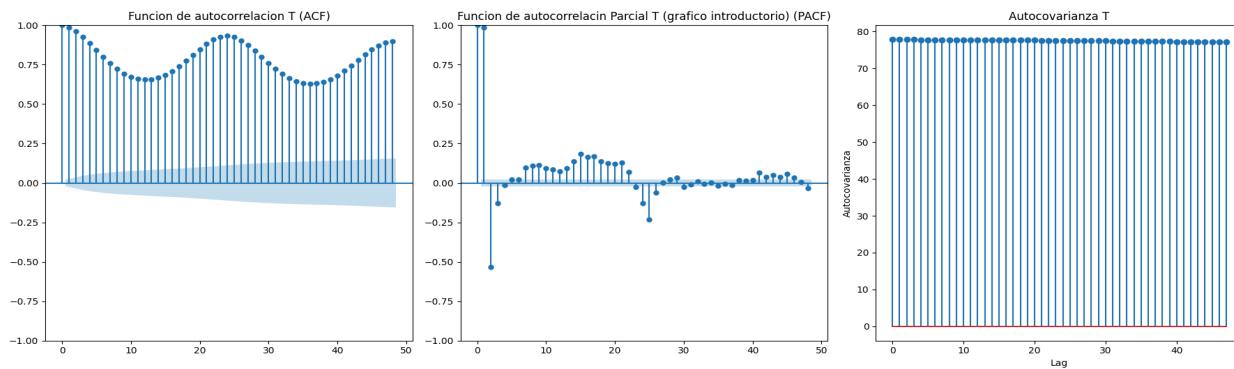


Figura 11. Gráficos ACF, PACF, Autocovarianza de T.

Estos gráficos demuestran los ciclos de 24 horas claramente. En el gráfico ACF, se observa que no disminuye rápidamente a cero, estando todos los valores para los lags, muy por encima del área de significancia; lo que nos ayuda a descartar la estacionariedad de nuestra serie. La gráfica de autocovarianza, nos confirma esta idea; al tener una disminución tan lenta, significa que la serie tiene una fuerte dependencia temporal y confirmar así que se trata de una serie no estacionaria.

Para validar lo observado visualmente, se realizaron distintos tests para probar la estacionariedad de la serie “T”.

- Test ADF (Dickey-Fuller Aumentado): En esta prueba se obtuvo $0.019583 <$

p-valor de 0.05. Acorde a este resultado, se rechaza la H_0 , por lo que se establece que la serie temporal es estacionaria, y no necesitaría una diferenciación para hacer un modelado de la misma.

- **Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin):** En este test, se obtuvo como resultado $0.010 < 0.05$; lo que nos lleva a rechazar la H_0 . Con este test, se obtiene que la serie es no estacionaria; por lo que sería necesario aplicar una diferenciación a los datos.

Si el KPSS indica no estacionariedad y el ADF indica estacionariedad, entonces la serie es estacionaria por diferencia. Se debe utilizar la diferenciación para hacer que la serie sea estacionaria. Para comprobar esta idea, se utilizó la función `ndiffs` de la librería `pmdarima` en Python; obteniendo como resultado 1. Es por eso que se decidió hacer la diferenciación de la serie T, para lograr la estacionariedad de la misma.

Una vez implementada la primera diferenciación a la serie, se volvió a graficar, obteniendo los siguientes resultados:

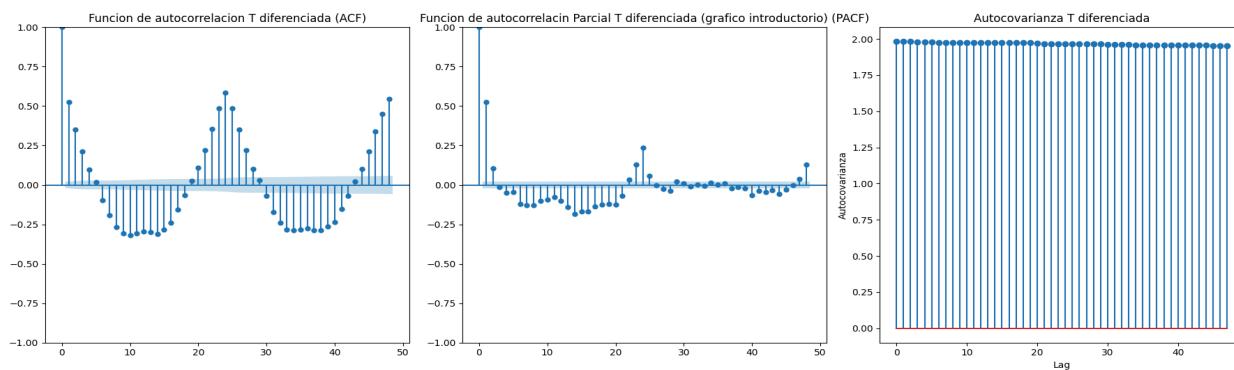


Figura 12. Gráficos ACF, PACF, Autocovarianza de T después de la primera diferenciación.

Ahora si se logra ver una disminución rápida en el gráfico ACF, lo que nos indica que la diferenciación realizada a nuestra serie original ha eliminado parte de la tendencia; así mismo, nos muestra un patrón oscilante, lo que sugiere la presencia de estacionalidad (cada 24 horas).

En el caso del gráfico PACF, muestra un corte brusco después de los primeros lags, lo que sugiere que hay términos autorregresivos en la serie.

En el caso del gráfico de autocovarianza, sigue habiendo una caída muy lenta de los datos; lo que nos indica que a pesar de que los datos fueron diferenciados, siguen

teniendo una fuerte dependencia temporal.

Es por esto, que se decidió realizar una segunda diferenciación a esta serie; con el objetivo de eliminar la estacionalidad de la serie que presenta cada 24 horas.

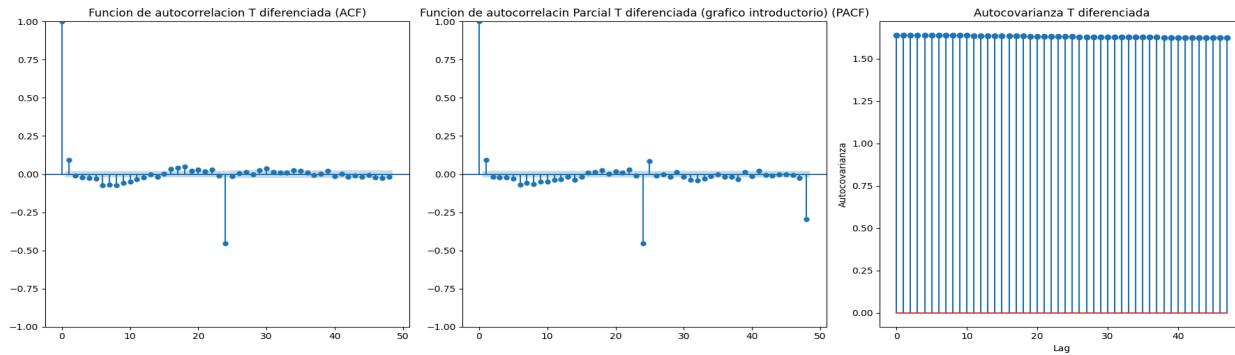


Figura 13. Gráficos ACF, PACF, Autocovarianza de T después de diferenciación de 24 horas.

Estacionariedad de la Serie Temporal Humedad Relativa (RH)

A la serie temporal de la variable temperatura se le realizó una descomposición aditiva.

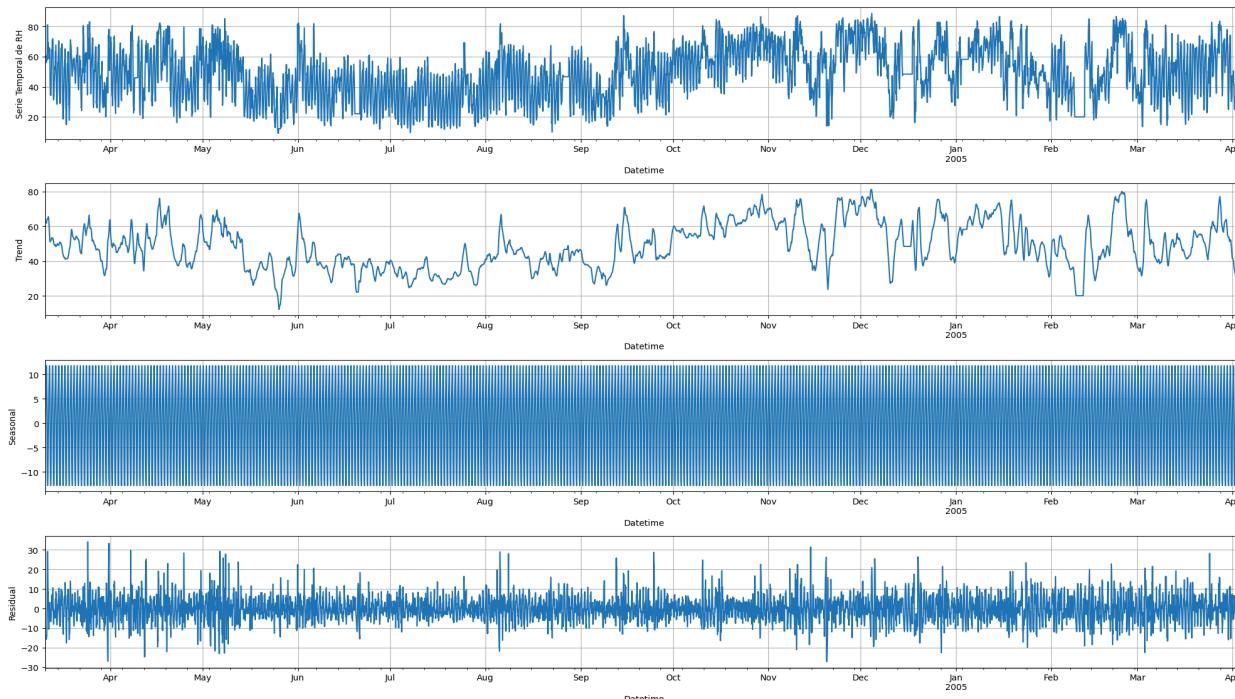


Figura 14. Descomposición aditiva de RH.

En la descomposición de la serie RH se puede observar:

- La serie no parece tener tendencia, si en cambio una alta volatilidad.
- Se puede percibir una estacionariedad diaria de los datos: cada 24 horas la humedad relativa sigue un comportamiento similar.
- El componente residual parece tener una varianza relativamente constante, lo que sugiere que la serie temporal original es estacionaria.

Gráficos ACF y PACF

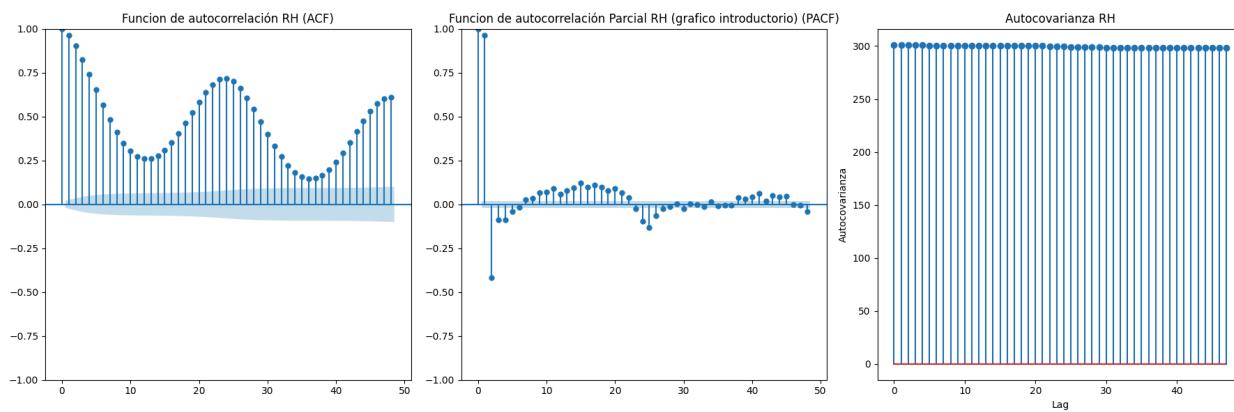


Figura 15. Gráficos ACF, PACF, Autocovarianza de RH.

- **Gráfico FAC:** Se observa un patrón importante de estacionalidad que se repite diariamente. Todos los valores están fuera de la banda de confianza, por lo tanto las correlaciones son significativas.
- **Gráfico PFAC:** Se observan cortes abruptos los cuales indican estacionalidad.
- **Gráficos Autocovarianzas:** Se observan valores muy similares lo que lleva a interpretar que es una serie de memoria larga.

Para validar lo observado visualmente, se realizaron distintos tests para probar la estacionariedad de la serie “RH”.

- **Test ADF (Dickey-Fuller Aumentado):** En esta prueba se obtuvo $0.0000 < p\text{-valor}$ de 0.05. Acorde a este resultado, se rechaza la H_0 , por lo que se establece que la serie temporal es estacionaria, y no necesitaría una diferenciación para hacer un modelado de la misma.

- **Test KPSS (Kwiatkowski-Phillips-Schmidt-Shin):** En este test, se obtuvo como resultado $0.010 < 0.05$; lo que nos lleva a rechazar la H_0 . Con este test, se obtiene que la serie es no estacionaria; por lo que sería necesario aplicar una diferenciación a los datos.
- **Test PP (Phillips-Perron):** Se utiliza para detectar la presencia de raíces unitarias en una serie temporal. Como el p-valor es $0.0000 < 0.05$, rechazamos la H_0 concluyendo que la serie es estacionaria.

Igualmente se utilizó la función ***ndiffs*** de la librería *pmdarima* en Python; obteniendo como resultado 0.

Se realizó una diferenciación de 24 hs. y se graficó:

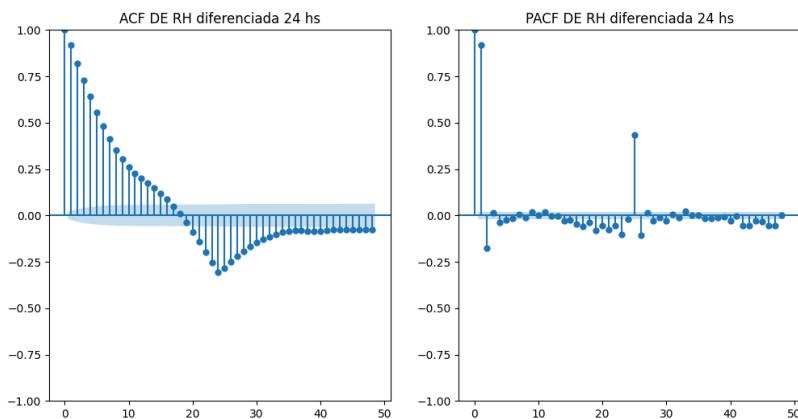


Figura 16. Gráficos ACF, PACF con serie RH diferenciada.

Modelos SARIMA

Un modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average) es una extensión del modelo ARIMA que incorpora componentes estacionales. Los modelos ARIMA son útiles para datos no estacionarios, pero a menudo, las series temporales también muestran patrones estacionales (repetitivos en intervalos de tiempo regulares). El modelo SARIMA permite capturar tanto la estructura no estacional como la estacional en los datos.³

³ PDF de clase.(Junio 2024). Del Rosso, Rodrigo. “Clase 3 - Modelos ARIMA”. Análisis de Series Temporales.

La combinación de los modelos ARMA(p,q) estacionales y no estacionales se expresa simplemente como: **$ARMA(p,q)(P,Q)s$** , donde:

- P corresponde al componente autorregresivo estacional.
- Q corresponde al componente de media móvil estacional.
- s representa el periodo estacional.

Como se mencionó anteriormente, las 3 series analizadas en este trabajo, tienen un comportamiento estacional (cada 24 horas los valores suben y bajan). A continuación se presentan los modelos realizados para cada una de las series.

Es importante mencionar que los parámetros tanto de la parte estacional como no estacional ($p,d,q(P,D,Q,s)$), fueron seleccionados a partir del análisis de las gráficas ACF y PACF anteriormente descritas. Después se fue variando, en base a los resultados que se fueron obteniendo de los Modelos y la significancia de los coeficientes. El parámetro s se fijó en 24, para las 3 series.

Modelos para Serie C6H6

Se analizaron 11 distintos modelos SARIMA, para la variable de Temperatura. Estos fueron los resultados obtenidos:

Modelos	Parámetros	AIC	BIC	LJUNG-BOX $P(Q) > 0.05$	JARQUE BERA $P(JB) > 0.05$
Modelo1	(2,0,2) (2,1,2,24)	47747	47811	0,88	0
Modelo 2	(2,0,2) (1,1,1,24)	47743	47793	0,88	0
Modelo 3	(2,0,2) (1,1,0,24)	50564	50578	0	0
Modelo 4	(2,0,1) (1,1,1,24)	47872	47915	0,48	0
Modelo 5	(2,0,0) (1,1,1,24)	47927	47962	0,39	0
Modelo 6	(1,0,1) (1,1,1,24)	47906	14942	0,41	0
Modelo 7	(2,0,2) (0,1,0,24)	52338	52359	0,56	0
Modelo 2_boxcox	(2,0,2) (1,1,1,24)	12263	12280	0,88	0,88
Modelo2_aut	(2,1,3) (1,1,1,24)	47846	47903	0,85	0

El modelo SARIMA seleccionado como “la mejor opción” dentro de las que se evaluaron es el **modelo2 SARIMA (2,0,2) (1,1,1,24)**, ya que el valor de AIC y BIC resultaron los menores. Estos fueron los resultados del mismo:

```
SARIMAX Results
=====
Dep. Variable: C6H6 No. Observations: 9357
Model: SARIMAX(2, 0, 2)x(1, 1, [1], 24) Log Likelihood -23864.895
Date: Sat, 06 Jul 2024 AIC 47743.790
Time: 20:03:49 BIC 47793.779
Sample: 03-10-2004 HQIC 47760.769
- 04-04-2005
Covariance Type: opg
=====
            coef    std err      z   P>|z|   [0.025   0.975]
-----
ar.L1      1.5693   0.023   68.350   0.000    1.524    1.614
ar.L2     -0.5848   0.020  -28.931   0.000   -0.624   -0.545
ma.L1     -0.6608   0.024  -28.094   0.000   -0.707   -0.615
ma.L2     -0.1934   0.009  -22.147   0.000   -0.211   -0.176
ar.S.L24   0.1421   0.008   17.638   0.000    0.126    0.158
ma.S.L24  -0.9393   0.003 -324.948   0.000   -0.945   -0.934
sigma2     9.6922   0.068  141.987   0.000    9.558    9.826
Ljung-Box (L1) (Q): 0.02 Jarque-Bera (JB): 22151.63
Prob(Q): 0.88 Prob(JB): 0.00
Heteroskedasticity (H): 0.86 Skew: 0.55
Prob(H) (two-sided): 0.00 Kurtosis: 10.47
=====
```

Modelos para Serie T

Se analizaron 11 distintos modelos SARIMA, para la variable de Temperatura. Estos fueron los resultados obtenidos:

Modelos	Parámetros	AIC	BIC	LJUNG-BOX P(Q) > 0.05	JARQUE BERA P(JB) > 0.05
Modelo 1	(1,1,1)(1,1,1,24)	25982	26018	1.00	0
Modelo 2	(1,1,2)(1,1,1,24)	25984	26027	0.99	0
Modelo 3	(1,1,3)(1,1,1,24)	25986	26036	0.99	0
Modelo 4	(1,1,1)(2,1,1,24)	25982	26025	0.99	0
Modelo 5	(1,1,2)(2,1,1,24)	25984	26034	0.99	0
Modelo 6	(1,1,3)(2,1,1,24)	25986	26043	0.99	0
Modelo 7	(1,1,0)(2,1,2,24)	25978	26021	0.95	0
Modelo 8	(1,1,0)(1,1,0,24)	28774	28795	0.89	0
Modelo 9	(2,1,0)(2,1,0,24)	27970	28006	1.00	0
Modelo 10	(1,1,0)(2,1,0,24)	27968	27997	0.95	0
Modelo 11	(1,1,0)(2,1,1,24)	25980	26016	0.95	0

El modelo SARIMA seleccionado como “la mejor opción” dentro de las que se evaluaron es el **Modelo 11: SARIMA(1,1,0)(2,1,1,24)**, ya que el valor de AIC y BIC resultaron los menores. Estos fueron los resultados del mismo:

SARIMAX Results						
Dep. Variable:		T	No. Observations:	9357		
Model:	SARIMAX(1, 1, 0)x(2, 1, [1], 24)		Log Likelihood	-12985.284		
Date:	Mon, 08 Jul 2024		AIC	25980.568		
Time:	14:07:42		BIC	26016.274		
Sample:	03-10-2004 - 04-04-2005		HQIC	25992.696		
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
ar.L1	0.1507	0.007	22.690	0.000	0.138	0.164
ar.S.L24	0.0871	0.009	9.985	0.000	0.070	0.104
ar.S.L48	0.0172	0.010	1.785	0.074	-0.002	0.036
ma.S.L24	-0.9392	0.004	-254.641	0.000	-0.946	-0.932
sigma2	0.9418	0.005	196.625	0.000	0.932	0.951
Ljung-Box (L1) (Q):		0.00	Jarque-Bera (JB):		100283.53	
Prob(Q):		0.95	Prob(JB):		0.00	
Heteroskedasticity (H):		0.52	Skew:		-0.80	
Prob(H) (two-sided):		0.00	Kurtosis:		18.98	

Modelos para Serie RH

Para la variable de Humedad relativa se analizaron sólo 7 distintos modelos SARIMA, ya que al probar con valores superiores a 1 aparecieron coeficientes no significativos. Estos fueron los resultados obtenidos:

Modelos	Parámetros	AIC	BIC	LJUNG-BOX P(Q) > 0.05	JARQUE BERA P(JB) > 0.05
Modelo1	(1,0,1)(1,1,1,24)	50917	50952	0,89	0,00
Modelo 2	(1,0,1)(2,1,1,24)	50918	50961	0,89	0,00
Modelo 3	(1,0,1)(0,1,1,24)	50928	50956	0,88	0,01
Modelo 4	(1,0,2)(1,1,1,24)	50918	50961	0,98	0,00
Modelo 5	(1,0,0)(1,1,1,24)	51221	51250	0,00	0,02
Modelo 6	(2,0,1)(1,1,1,24)	50918	50961	0,97	0,00
Modelo 7	(1,0,1)(1,1,2,24)	50918	50961	0,89	0,00

El modelo SARIMA seleccionado como “la mejor opción” dentro de las que se evaluaron es el **Modelo 1: SARIMA(1,0,0)(1,1,1,24)**, ya que el valor de AIC y BIC resultaron los menores. Estos fueron los resultados del mismo:

```
SARIMAX Results
=====
Dep. Variable: RH   No. Observations: 9357
Model: SARIMAX(1, 0, 1)x(1, 1, 1, 24) Log Likelihood: -25453.572
Date: Mon, 08 Jul 2024   AIC: 50917.145
Time: 01:24:29   BIC: 50952.851
Sample: 03-10-2004 - 04-04-2005   HQIC: 50929.273
Covariance Type: opg
=====
              coef    std err      z   P>|z|    [0.025    0.975]
-----
ar.L1      0.9415    0.003  280.425   0.000     0.935    0.948
ma.L1      0.1928    0.007   28.556   0.000     0.180    0.206
ar.S.L24    0.0397    0.009    4.193   0.000     0.021    0.058
ma.S.L24   -0.9620    0.003  -294.841   0.000    -0.968   -0.956
sigma2     13.5985   0.083   163.841   0.000    13.436   13.761
Ljung-Box (L1) (Q):          0.02   Jarque-Bera (JB): 42685.41
Prob(Q):                  0.89   Prob(JB):       0.00
Heteroskedasticity (H):    0.90   Skew:           0.76
Prob(H) (two-sided):       0.00   Kurtosis:      13.37
=====
```

Performance de los Modelos Seleccionados

Para determinar si un modelo es adecuado o no a las series temporales; es necesario partir nuestra serie en datos de entrenamiento (que entrenen al modelo seleccionado) y datos de Test (que prueben el desempeño del modelo al pronosticar). Se estableció para las 3 series analizadas, el 80% de los datos para entrenar y el 20% restante para testear cada modelo seleccionado.

Se tomaron en cuenta varias medidas, para determinar si los modelos son precisos o no a la hora de predecir:

- **MSE (Error Cuadrático Medio):** Calcula el promedio de los cuadrados de las diferencias entre los valores predichos y los valores observados.
- **RMSE (Raíz del Error Cuadrático Medio):** Toma la raíz cuadrada del MSE, para devolver los errores en las mismas unidades que los valores originales de la serie

temporal.

- **MAE (Error Absoluto Medio):** Calcula el promedio de las diferencias absolutas entre los valores predichos y los valores observados.
- **MAPE (Error Porcentual Absoluto Medio):** se expresa en porcentaje; es ideal para comparar diferentes series temporales que tienen distintas escalas o unidades de medición.

Para el caso de las 4 métricas mencionadas, es importante mencionar, que si da un valor igual a 0, se trata de una predicción perfecta. Entre más alto de el valor de estas métricas, significa mayores errores en las predicciones del modelo, por lo tanto peor desempeño del mismo.

Performance Modelo C6H6

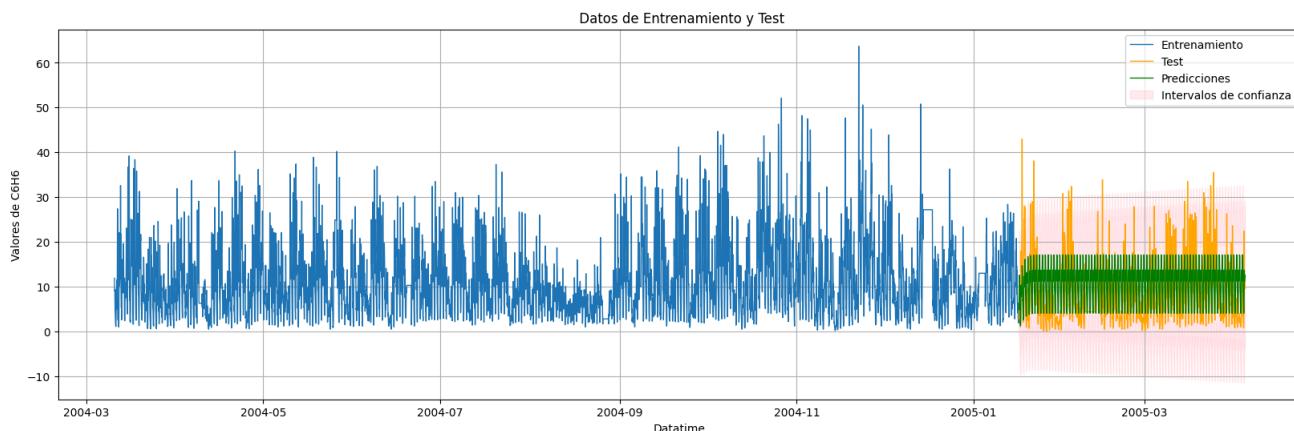


Figura 17. Performance Modelo SARIMA (2,0,2)(1,1,1,24) para C6H6 con train test y predicción.

Como se observa en el gráfico, las predicciones realizadas por el modelo SARIMA (2,0,2) (1,1,1,24) en esta serie temporal no son precisas. Aunque las predicciones siguen los valores observados, no logran capturar la volatilidad, es decir, los valores extremos de la serie.

Se calcularon las métricas correspondientes para poder comparar luego con otros modelos y evaluar su bondad. Los resultados obtenidos fueron los siguientes:

- MSE = 37.8980
- MAE = 6.1561
- RMSE = 5.0012

Performance Modelo T

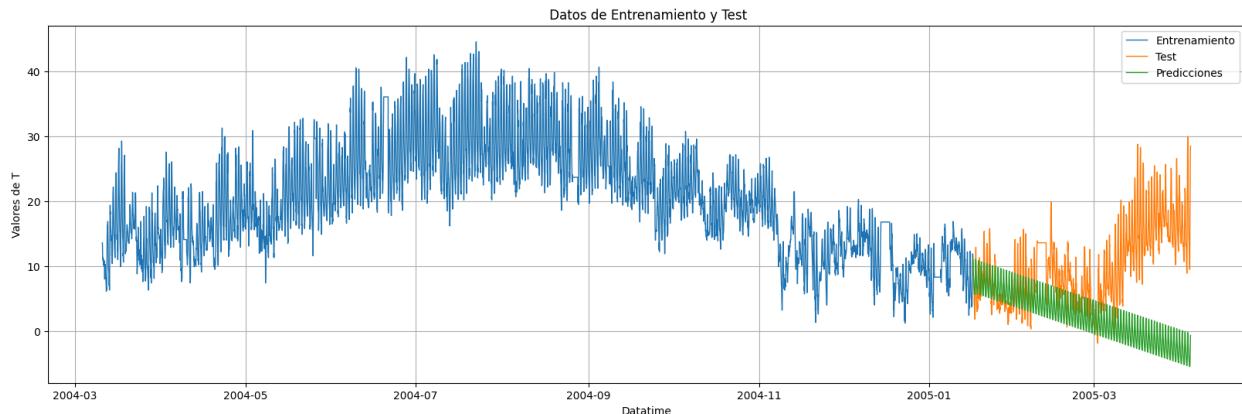


Figura 18. Performance Modelo SARIMA (1,1,0)(2,1,1,24) para T con train test y predicción.

Como se observa en el gráfico, las predicciones que realiza el Modelo **SARIMA(1,1,0)(2,1,1,24)** en esta serie temporal, son bastante malas. Una explicación puede ser, porque el modelo no es capaz de captar la tendencia que tiene nuestros datos. Al no contar con datos de años anteriores, no logra captar el comportamiento de que la temperatura empieza a aumentar en los meses de primavera, alcanzando los valores más altos durante los meses de verano. En este caso, esa información se pierde, o no es captada por el modelo; por lo que la predicción que nos arroja sigue la tendencia de la curva a disminuir. Esto se agrava conforme pasa el tiempo.

Se calcularon las métricas correspondientes para comprobar que el modelo seleccionado no es el adecuado para esta serie temporal. Estos fueron los resultados obtenidos:

- MSE = 123.3812
- MAE = 8.3726
- RMSE = 11.1077

Efectivamente, el modelo seleccionado SARIMA (1,1,0)(2,1,1,24), no resulta ser adecuado para explicar ni predecir los valores correspondientes a la serie temporal de la variable Temperatura.

Performance Modelo RH

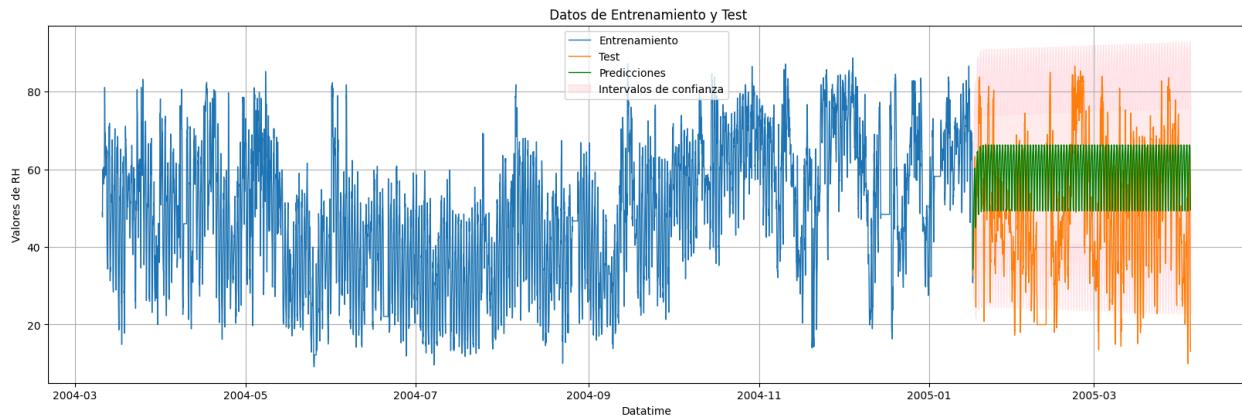


Figura 19. Performance Modelo SARIMA (1,0,1)(1,1,1,24) para RH con train test y predicción.

Como se observa en el gráfico, las predicciones que realiza el Modelo **SARIMA(1,0,1)(1,1,1,24)** en nuestra serie temporal, no son buenas.

Se calcularon las métricas correspondientes para comprobar que el modelo seleccionado no es el adecuado para esta serie temporal. Estos fueron los resultados obtenidos:

- MSE = 338.7695
- MAE = 15.20
- RMSE = 18.4056
- MAPE = 45.15

De hecho, el modelo SARIMA (1,0,1)(1,1,1,24) seleccionado, no es adecuado para explicar ni predecir los valores de la serie temporal de la variable Humedad Relativa.

Análisis de Residuos de los Modelos

El análisis de residuos para cualquier modelo de series temporales, es una etapa crucial en la evaluación y diagnóstico del mismo. Residuo es la diferencia entre el valor observado y el valor que el modelo logra predecir. Analizar los residuos, nos ayuda a verificar si el modelo ha logrado capturar adecuadamente los patrones en nuestra serie temporal o no. Y de esta manera, determinar si es un modelo adecuado para nuestros

datos. Para determinar que un modelo es el adecuado, los residuos del mismo deberían cumplir algunas condiciones:

- Comportarse como Ruido Blanco (no muestran patrones discernibles, tienen media cero y varianza constante).
- Deberían cumplir con una distribución normal. Se puede comprobar con el test de Shapiro-Wilk. Si $p\text{-valor}>0.05$, se comprueba que los residuos están normalmente distribuidos.
- Las autocorrelaciones de los residuos deben ser insignificantes. En caso de tener residuos altamente correlacionados, el modelo no es el indicado. Esto se puede determinar con el test de Ljung-Box. Si $p\text{-valor}>0.05$, nos indica que no hay correlaciones significativas en los residuos.

Residuos Modelo C6H6

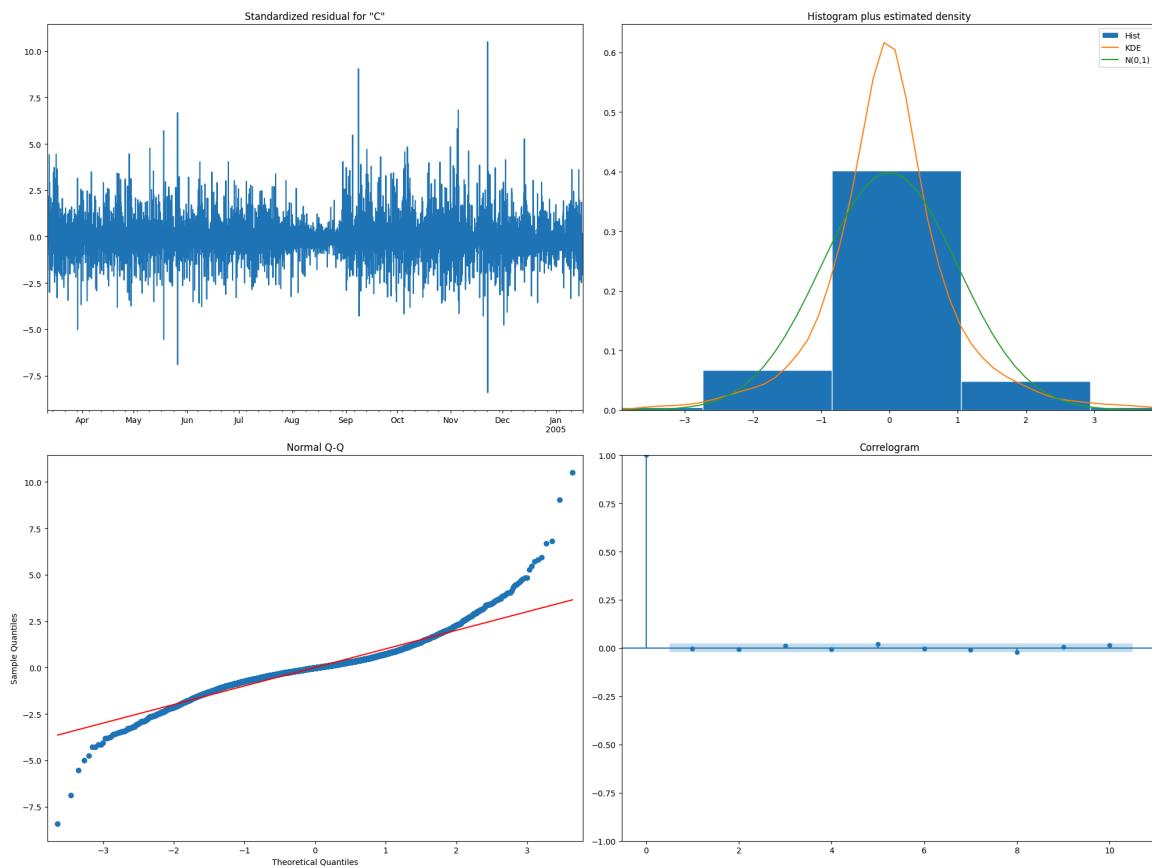


Figura 20. Análisis de Residuos del Modelo SARIMA (2,0,2)(1,1,1,24) para C6H6.

- **Residuales estandarizados:** Se observa una distribución aleatoria alrededor de cero. A simple vista, no se perciben patrones obvios.
- **Histograma con densidad estimada:** La curva teórica normal no coincide completamente con la curva real en varios puntos, indicando que los residuos no siguen una distribución normal.
- **Gráfico Q-Q normal:** Se compara los cuantiles de los residuos con los de una distribución normal. Los puntos deberían alinearse a lo largo de la línea roja que representa una distribución normal. Se observan desviaciones significativas en los extremos, lo que indica que los residuos no siguen una distribución normal.
- **Correlograma:** Muestra las autocorrelaciones de los residuos. Si los puntos están dentro de las bandas de confianza, podría indicar que el modelo ha capturado adecuadamente la estructura de autocorrelación de los datos.

La prueba de Ljung-Box es una prueba de independencia de los residuos en series temporales, que verifica si los residuos están autocorrelacionados a través de diferentes rezagos. En este análisis, se realizó esta prueba con un máximo de 48 rezagos.

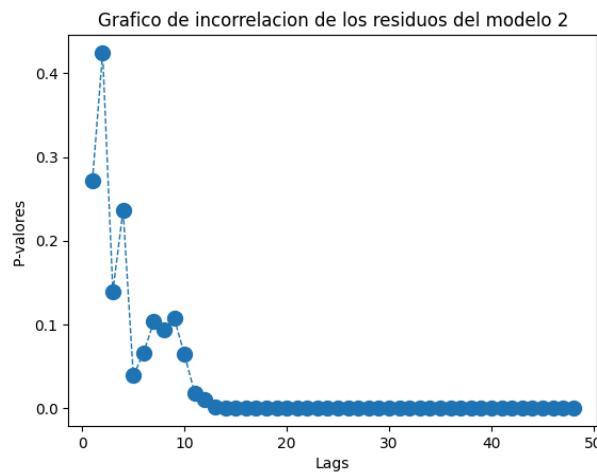


Figura 21. Correlación de

SARIMA (2,0,2)(1,1,1,24) para C6H6.

Residuos del Modelo

A partir del rezago 15, se observa una autocorrelación significativa en los residuos, lo cual indica que no cumplen con la condición de ruido blanco. Esto sugiere que el modelo no es adecuado para capturar completamente la estructura de los datos.

Al no cumplir la condición de ruido blanco se puede decir que no es un modelo confiable.

Residuos Modelo T

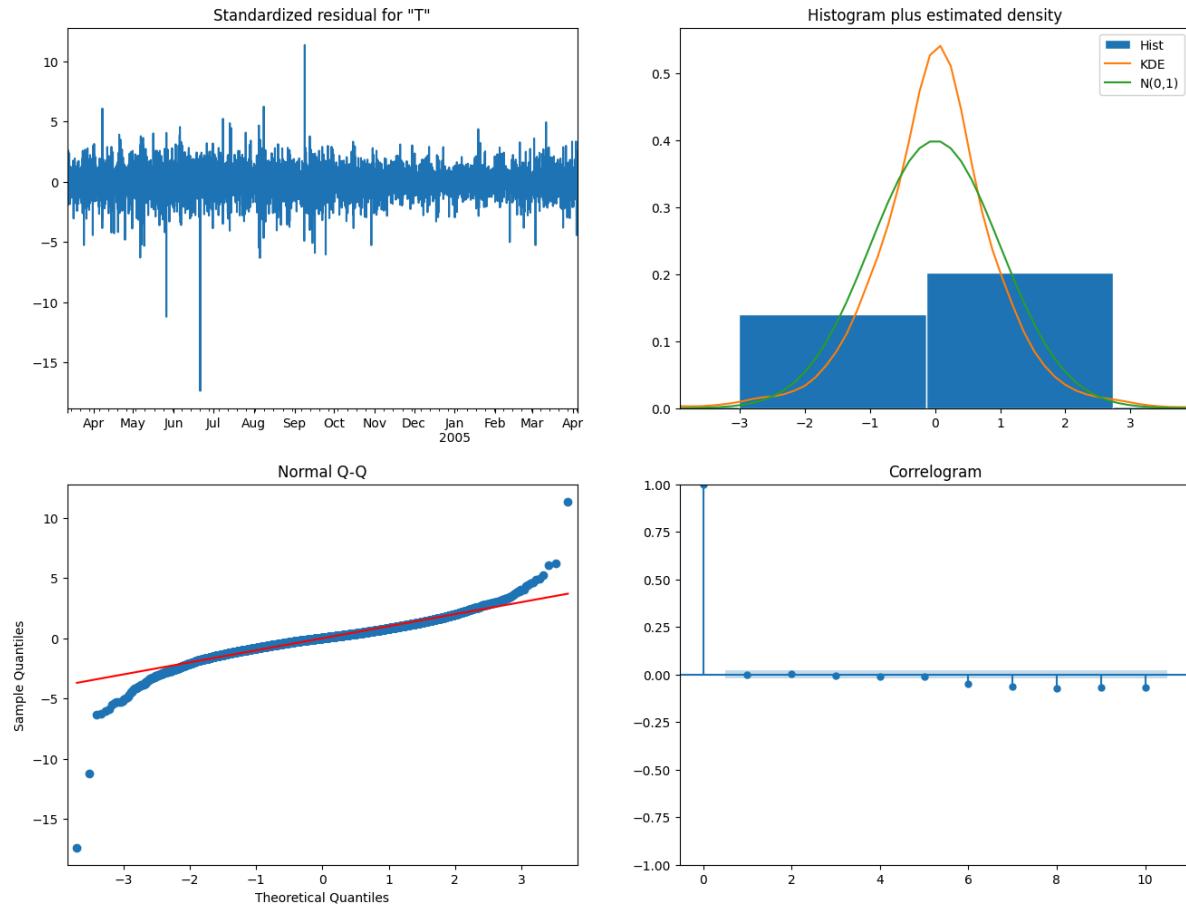


Figura 22. Análisis de Residuos del Modelo SARIMA (1,1,0)(2,1,1,24) para T.

En este análisis de residuos, se observan varias cosas:

- Los residuos parecen ser un ruido blanco, no se ve un comportamiento claro de los mismos; pero al mismo tiempo se ven dos picos importantes en el mes de Julio y Septiembre, los cuales deberían analizarse a detalle.
- Los residuos del modelo no tienen una distribución normal.
- Los residuos tienen autocorrelación.

El análisis de Residuos para el Modelo SARIMA (1,1,0)(2,1,1,24) para la serie temporal que corresponde a la Temperatura del aire, nos demuestra una vez más que no es el apropiado.

Residuos Modelo RH

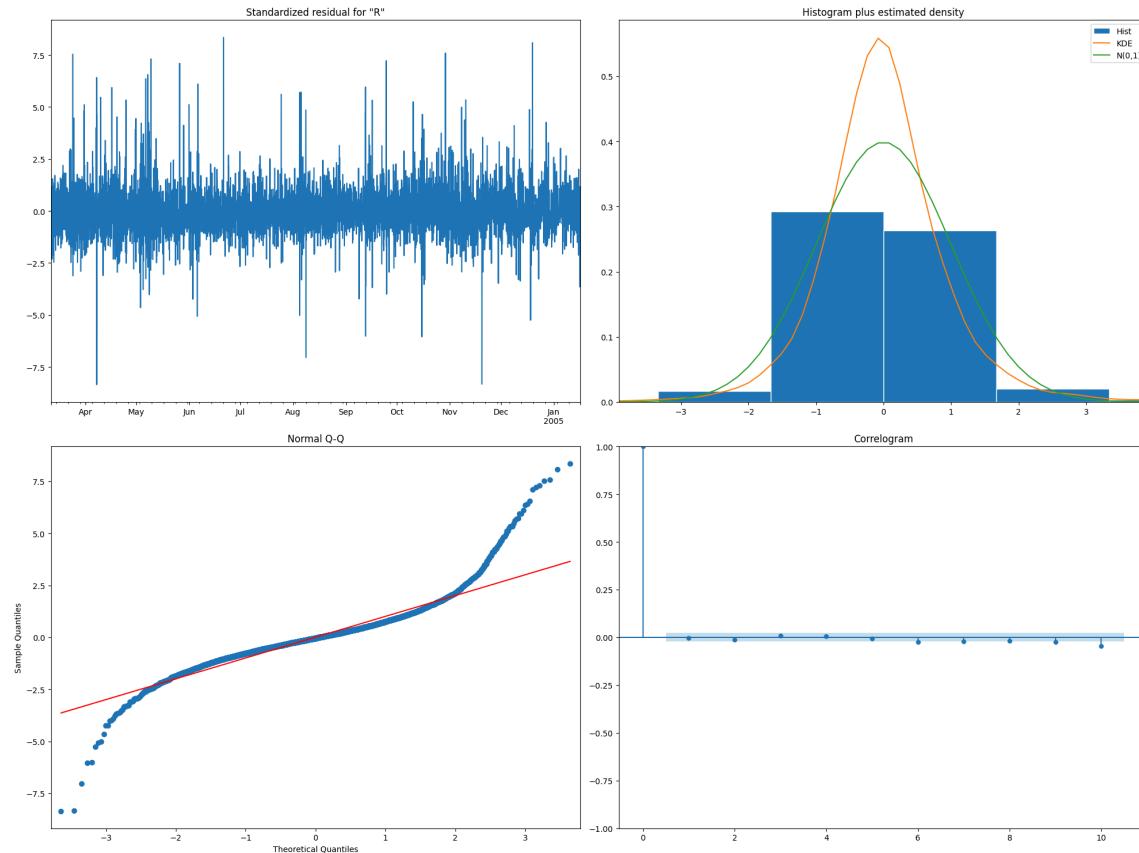


Figura 23. Análisis de Residuos del Modelo SARIMA (1,0,1)(1,1,1,24) para RH.

El gráfico de residuos estandarizados muestra que los residuos no tienen una tendencia clara en el tiempo. Esto sugiere que el modelo es capaz de capturar la variación en los datos. Sin embargo, hay algunas señales de que el modelo podría mejorarse. Por ejemplo, hay algunos valores atípicos en los residuos que podrían ser eliminados o tratados de otra manera. Además, el correlograma muestra una autocorrelación en los residuos. Esta autocorrelación podría ser eliminada mediante el uso de un modelo ARIMA o GARCH.

También se realizó la prueba de Ljung-Box en los residuos de un modelo. La prueba de Ljung-Box es una prueba de independencia de los residuos en series temporales para verificar si los residuos están autocorrelacionados a través de diferentes rezagos. Se realizó con un máximo de 48 rezagos.

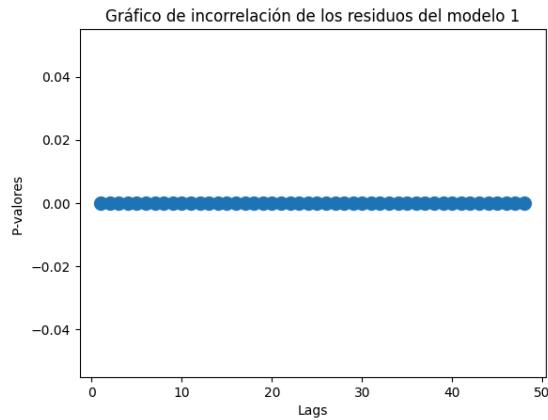


Figura 24. Correlación de Residuos del Modelo SARIMA (1,0,1)(1,1,1,24) para RH.

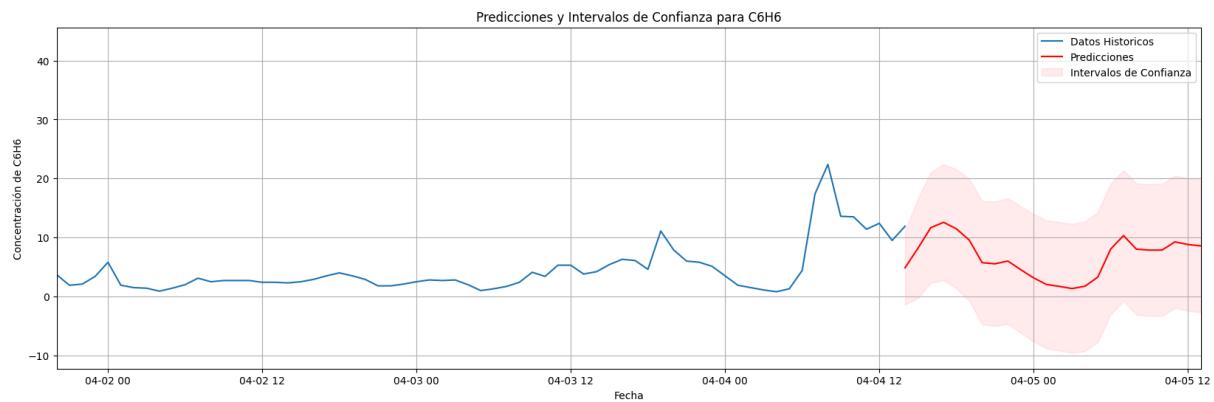
Como todos los p-values resultantes son iguales a 0.0, significa que hay evidencia estadística significativa de autocorrelación en los residuos del modelo. O sea, el modelo no está capturando completamente la estructura de autocorrelación en los datos. Esto sugiere que el modelo puede necesitar ser ajustado o mejorado para incorporar mejor la autocorrelación en los residuos.

Además, se llevó a cabo el test de Box-Pierce con el objetivo de determinar si los residuos de un modelo estadístico presentan autocorrelación significativa. El resultado de la prueba mostró un valor de p igual a cero. Rechazando la H₀ de independencia de los residuos. Se concluyó que los residuos están autocorrelacionados.

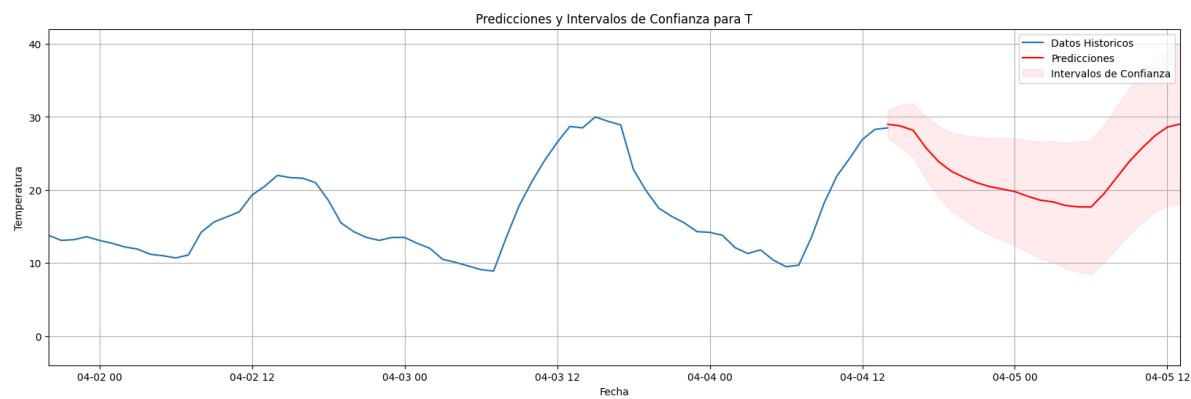
Pronóstico

En esta última parte del trabajo, se pretende probar cada uno de los Modelos seleccionados para cada serie temporal para predecir datos a futuro. Es importante establecer primero cuál es la ventana temporal más adecuada, para realizar dicho pronóstico. En nuestro caso, se estableció una ventana de 24 horas: ventana mínima. No sería apropiado ni seguro pronosticar una ventana mayor, ya que ninguno de los modelos cumplían las condiciones requeridas en los residuos para tener resultados confiables. Es decir, se constató que, conforme se avanza en el tiempo (lags) los residuos van teniendo mayor autocorrelación. Sólo se observan para los primeros 10-15 lags residuos sin correlación. Los modelos podrían predecir adecuadamente solamente la mitad de un día en el caso de las series temporales analizadas.

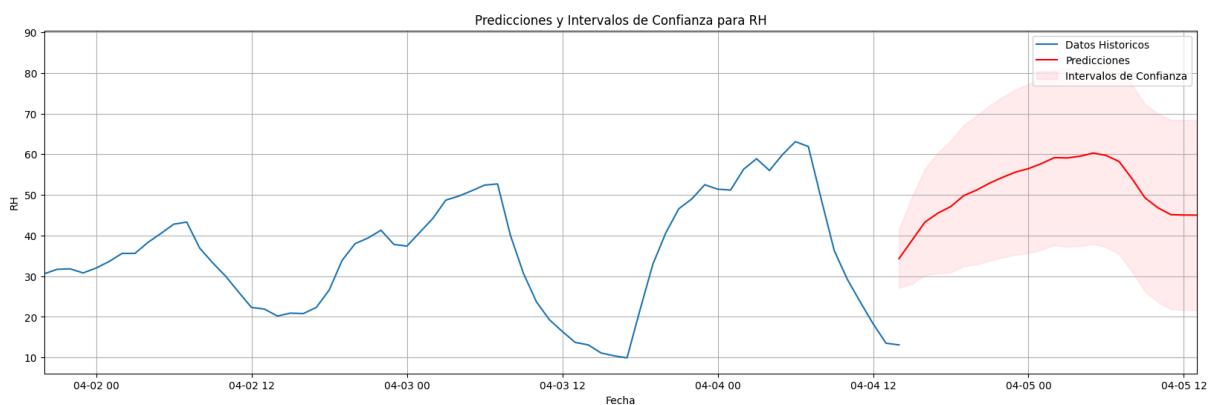
Pronóstico de 24 horas para C6H6



Pronóstico de 24 horas para T



Pronóstico de 24 horas para RH



Modelo SARIMA-X

El modelo SARIMA-X es una extensión del modelo SARIMA que incluye variables exógenas a la serie temporal analizada. Este modelo, no solamente considera los componentes estacionales y no estacionales de la serie temporal; sino que además, incorpora otras variables que podrían influir en el comportamiento de la serie temporal que se pretende modelar.

El modelo SARIMA-X se representa de la siguiente manera:

$$Y_t = \alpha + SARIMA(p, d, q)(P, D, Q)s + \beta X_t + \epsilon_t$$

Donde:

- α = Intercepto.
- β = coeficiente de la variable exógena.
- ϵ_t = término de error.

En este trabajo, se decidió aplicar el modelo SARIMA-X para aplicar a la serie temporal de benceno (C6H6) y agregar la variable exógena de Temperatura; con el objetivo de visualizar, si influye la temperatura en la concentración de benceno que se registra en el aire.

Se analiza el gráfico de autocorrelación cruzada de la serie C6H6 con la variable exógena T.

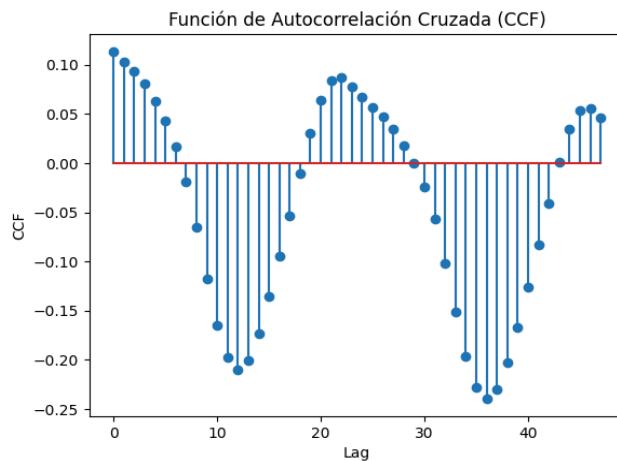


Figura 25. Gráfico CCF para Modelo SARIMA-X (2,0,2) (1,1,1,24) de C6H6 con variable exógena T.

En el gráfico de autocorrelación cruzada se observan autocorrelación positiva como negativa:

- Una correlación positiva en los lags cercanos a 0 sugiere una relación directa y casi instantánea, lo que significa que, en el mismo momento o con un pequeño desfase temporal, cuando la temperatura aumenta, la concentración de benceno también tiende a aumentar.
- Una correlación positiva con los lags positivos distintos de cero indica que, un aumento en la temperatura predice un aumento en la concentración de benceno después de cierto tiempo.
- Una correlación negativa en los lags negativos indica que un aumento en la temperatura en el pasado está asociado con una disminución en la concentración de benceno en el presente.

Se lleva a cabo el mismo procedimiento que se realizó con los modelos anteriores. Una vez que se determinan los parámetros para el modelo SARIMA-X (2,0,2) (1,1,1,24) exog=df_T_train. Se lleva a cabo el entrenamiento del mismo sobre la base de training (80% de los datos) para después testearlo en la base de testing (20% de los datos).

SARIMAX Results						
Dep. Variable:	C6H6	No. Observations:	7486			
Model:	SARIMAX(2, 0, 2)x(1, 1, [1], 24)	Log Likelihood	-19247.046			
Date:	Sun, 07 Jul 2024	AIC	38510.092			
Time:	20:36:53	BIC	38565.433			
Sample:	03-10-2004 - 01-16-2005	HQIC	38529.101			
Covariance Type:	opg					
	coef	std err	z	P> z	[0.025	0.975]
T	0.0198	0.025	0.802	0.423	-0.029	0.068
ar.L1	1.5797	0.025	63.648	0.000	1.531	1.628
ar.L2	-0.5944	0.022	-27.224	0.000	-0.637	-0.552
ma.L1	-0.6715	0.026	-26.318	0.000	-0.721	-0.621
ma.L2	-0.1905	0.010	-19.123	0.000	-0.210	-0.171
ar.S.L24	0.1408	0.009	15.517	0.000	0.123	0.159
ma.S.L24	-0.9366	0.003	-279.646	0.000	-0.943	-0.930
sigma2	10.1231	0.084	120.989	0.000	9.959	10.287
Ljung-Box (L1) (Q):	0.02	Jarque-Bera (JB):	15605.87			
Prob(Q):	0.88	Prob(JB):	0.00			
Heteroskedasticity (H):	1.20	Skew:	0.52			
Prob(H) (two-sided):	0.00	Kurtosis:	10.01			

Performance Modelo SARIMA-X C6H6 con la variable exógena T

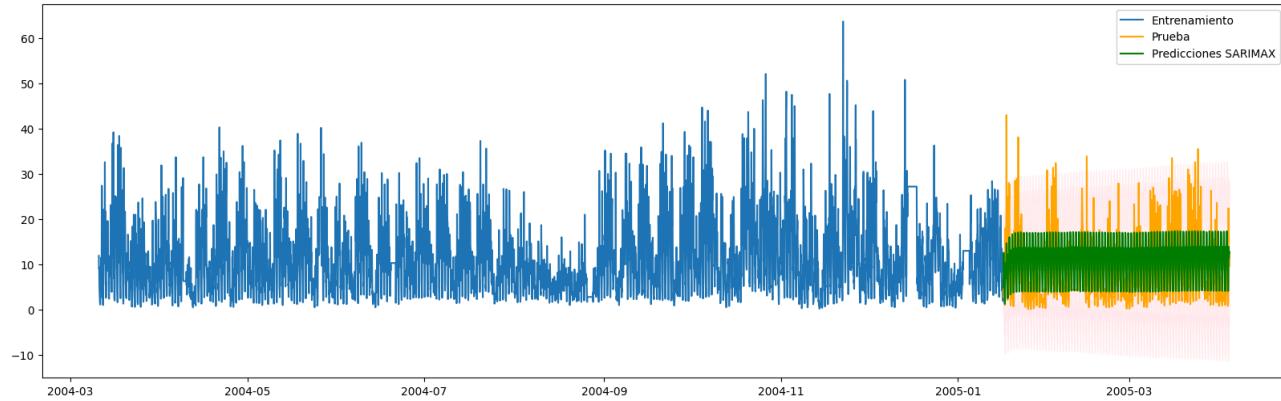


Figura 26. Performance del Modelo SARIMA-X (2,0,2) (1,1,1,24) de C6H6 con variable exógena T.

Se calcularon las mismas métricas para determinar si el modelo tiene un buen performance sobre los datos analizados; y se obtuvieron los siguientes resultados:

- MSE = 37.7549
- RMSE = 6.1445
- MAE = 4.9934

Las métricas muestran una ligera mejora respecto a las obtenidas con el modelo SARIMA, es decir, sin considerar la variable exógena de la temperatura.

Residuos Modelo SARIMA-X C6H6 con la variable exógena T

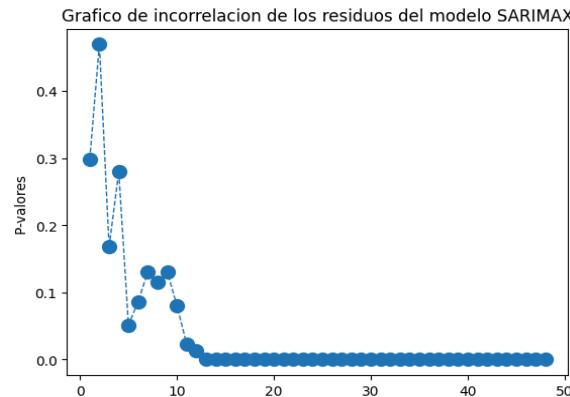


Figura 27. Correlación de Residuos del Modelo SARIMA-X (2,0,2) (1,1,1,24) de C6H6 con variable exógena T.

Al igual que los otros modelos anteriormente analizados, se observa una importante autocorrelación a partir del lag 13.

Se realizó el test de Ljung-Box con 24 rezagos y muestra un p-valor de 0, que es menor que 0.05, indicando que los residuos no cumplen con la normalidad. También el test de Box-Pierce y se rechaza la hipótesis de no autocorrelación de los residuos.

Se concluye que aunque el modelo SARIMA-X es ligeramente mejor que el modelo SARIMA, ninguno es confiable debido a que sus residuos no cumplen con la condición de ruido blanco.

Prueba de Hipótesis HEGY

La prueba HEGY (Hylleberg, Engle, Granger y Yoo) se utiliza para detectar raíces unitarias estacionales en series temporales. Es una extensión de las pruebas de raíces unitarias convencionales (ADF, KPSS, Phillips-Perron) y está diseñada para series con datos trimestrales o mensuales, donde se pueden detectar raíces unitarias en diferentes frecuencias estacionales.

La idea principal de la prueba HEGY es descomponer la serie temporal en componentes que correspondan a diferentes frecuencias estacionales y luego realizar pruebas de raíces unitarias en cada uno de estos componentes.⁴

Pasos principales para realizar la prueba HEGY:

- **Transformación de la serie temporal:** La serie original se transforma en componentes estacionales utilizando polinomios estacionales.
- **Regresión auxiliar:** Se estima una regresión que incluye las variables transformadas y se realiza la prueba de raíces unitarias en cada componente.
- **Interpretación de resultados:** Se analizan los resultados de las pruebas para determinar si hay raíces unitarias estacionales en la serie temporal.

Aunque las series de este trabajo no son trimestrales ni mensuales, aplicamos de todas maneras la prueba de Hegy para verificar la presencia de raíces unitarias estacionales que no pudieron ser captadas por los otros tests, pensamos que por la variabilidad de los datos. Se verificó específicamente en la serie de C6H6. La prueba de Hegy es conocida por ser un método más robusto en estos casos.

⁴ Alonso, Julio C., Seeman, Paul. (2010) Prueba de HEGY en R: Una guía.

Para la metodología:

- Se separa la serie C6H6 en componentes estacionales, de tendencia y residuales.
- Se construye un modelo de regresión lineal utilizando desfases de la serie.
- Se aplica el test de Dickey-Fuller a los residuos del modelo para determinar si la serie residual es estacionaria.

El resultado da y p-valor: 1.1379926472091072e-29 < 0.05 por lo que se puede rechazar la H₀ de raíz unitaria. Concluimos que la serie es estacional y los residuos del modelo tampoco muestran evidencia de raíz unitaria.

Conclusiones

En este trabajo, se realizó un exhaustivo análisis de series temporales sobre 3 variables continuas presentes en un estudio de Contaminación de Aire: benceno, temperatura y humedad relativa. Como se demostró a lo largo del trabajo, resulta bastante complejo modelar comportamientos que tienen que ver con cuestiones climáticas; y se llegó a la conclusión que los modelos SARIMA no serían los adecuados para realizar pronósticos en el caso de las series temporales con las que se trabajó.

Se evaluaron distintos modelos SARIMA, con el objetivo de identificar alguno que ajustara de manera adecuada a cada serie temporal. Se tomaron en cuenta diversos criterios para medir la performance de los mismos: AIC, BIC, análisis de residuos, error cuadrático medio, entre otros. A pesar de los esfuerzos realizados a lo largo del trabajo, como la diferenciación de las series, la aplicación de transformación Box-Cox a los datos originales; no se alcanzó un nivel de precisión satisfactorio para las predicciones que se realizaron.

Uno de los factores que pudo haber influido en este resultado es la limitada cantidad de datos disponibles, ya que solo se disponía de registros correspondientes a un año. La falta de datos históricos impidió capturar adecuadamente patrones a largo plazo y variaciones que podrían mejorar la precisión de los modelos. Los datos adicionales de múltiples años podrían proporcionar una base más robusta para la modelización y permitir una captura más precisa de la estacionalidad y tendencias subyacentes. Otro factor, es la volatilidad de los datos con los que se trabajó; en el caso de las 3 series temporales, vemos mucha fluctuación entre los datos más bajos y más altos de cada día (periodos de 24 horas). Un modelo SARIMA, no logra captar esta volatilidad de los datos.

La variabilidad y los patrones no capturados por los modelos SARIMA sugieren que futuras investigaciones podrían beneficiarse de la exploración de técnicas de modelado más avanzadas. Entre estas, se podrían considerar métodos basados en aprendizaje automático.

Bibliografía

- Fuente del dataset utilizado: <https://archive.ics.uci.edu/dataset/360/air+quality>
- S. De Vito, E. Massera, M. Piga, L. Martinotto, G. Di Francia. (2008). "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario". ScienceDirect.
<https://www.sciencedirect.com/science/article/abs/pii/S0925400507007691?via%3Dihub>
- Peña, D. (2010). Análisis de Series Temporales. Madrid: Alianza.
- Mauricio, José Alberto. (2007). “Introducción al Análisis de Series Temporales”. Universidad Complutense de Madrid.
- Material de Clase (2024). Del Rosso, Rodrigo. Análisis de Series Temporales. Universidad Austral.
- Enders, W. Applied Econometric Time Series. Fourth Edition. Wiley Series.
- Alonso, Julio C., Seeman, Paul. (2010) Prueba de HEGY en R: Una guía.
https://www.researchgate.net/publication/254399805_Prueba_de_HEGY_en_R_Una_guia
- Chat GPT.

Anexo 1 (cuadro comparativo métrica de los modelos)

En este anexo se presenta un cuadro comparativo con las métricas analizadas para algunos de los Modelos desarrollados para cada una de las series temporales. Se seleccionó el “mejor modelo obtenido”, es decir, el de menor AIC. Y se comparó con el “peor modelo obtenido” el de mayor AIC. A continuación los resultados obtenidos:

Métricas de Modelos C6H6

Modelos	MSE	RMSE	MAE
Modelo2 (2,0,2) (1,1,1,24)	37,8980	6,1561	5,0012
Modelo2_aut (2,1,3) (1,1,1,24)	239,1702	15,4651	13,1275
SARIMAX (2,0,2) (1,1,1,24) exog = df_T	37,7549	6,1445	4,9934
Modelo boxcox destranformado	32,9408	5,7394	4,4625

Métricas de Modelos T

Modelos	MSE	RMSE	MAE
Modelo 11 (menor AIC, modelo seleccionado)	123,3812	8,3727	11,1077
Modelo 1 (segundo mejor modelo)	124,0532	8,3973	11,1379
Modelo 8 (mayor AIC)	1.615,6570	34,0386	40,1952

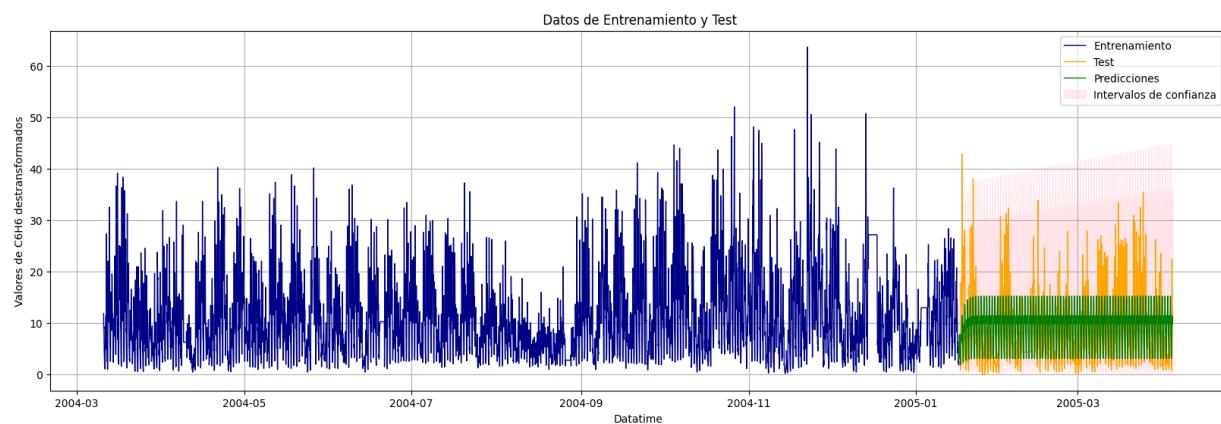
Métricas de Modelos RH

Modelos	MSE	RMSE	MAE
Modelo 1	338,76	15,2	18,4
Modelo 3	342,3	15,29	18,5
Modelo 5 (mayor AIC)	335,07	15,09	18,5

Anexo 2 (Transformación de Box-Cox en variable C6H6)

Se intentó mejorar la adecuación de los datos mediante la transformación Box-Cox para estabilizar la varianza y lograr la normalidad de los errores, requisitos del modelo SARIMA.

Se realizó la transformación de Box-Cox con un lambda = 0.26. Luego, se dividieron los datos en conjuntos de train y test (80%, 20%). Se entrenó el mejor modelo SARIMA (2,0,2)(1,1,1,24) seleccionado previamente por el criterio AIC y se calcularon las predicciones. Posteriormente, se transformaron los datos para obtener las predicciones reales y calcular las métricas.

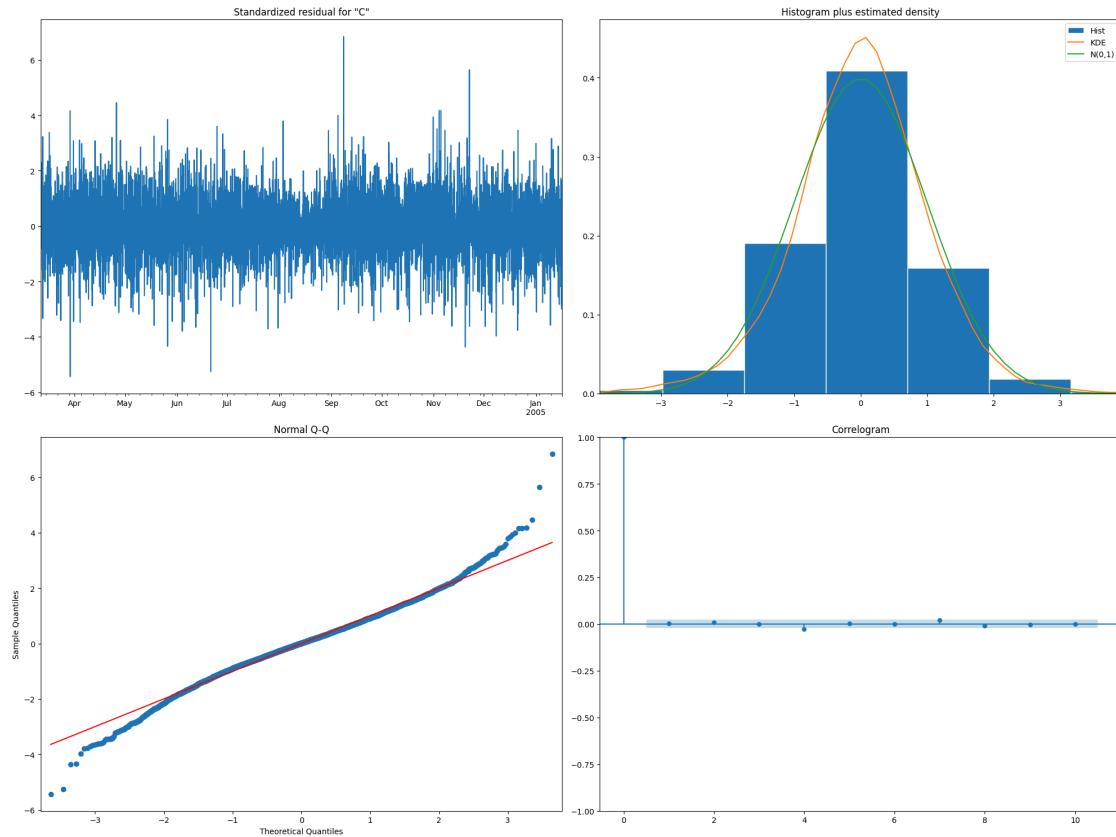


Los resultados fueron los siguientes:

- MSE = 32.9408
- RMSE = 5.7394
- MAE = 4.4625

Las métricas obtenidas fueron las mejores entre todos los modelos probados para la serie C6H6.

Se analizaron los residuos de modo gráfico y a través de Test Shapiro-Wilk (testea normalidad) y Test Box-Pierce (autocorrelación).



Test de Shapiro-Wilk: p-valor: $0.0 < 0.05$ se rechaza la H_0 de que los residuos siguen una distribución normal.

Test Box-Pierce: p-valor: $0.0 < 0.05$ se rechaza la H_0 de que no hay autocorrelación significativa en los residuos.

Concluimos que, aunque el modelo entrenado con datos transformados mediante Box-Cox tiene una mejor performance, no cumple con las condiciones de los residuos exigidas por los modelos SARIMAX, por lo cual sigue siendo un modelo poco confiable.