



Series de Tiempo: Proyecto Final

Alejandra Lelo de Larrea Ibarra 00124433

Victor Quintero Mármol González 000175897

15 de mayo del 2018

Index

1. Introducción

2. Datos

3. Modelos Clásicos

4. Modelo Dinámico Lineal

5. Comparación de Modelos

6. Conclusiones

Introducción

- Datos de la cantidad de Monóxido de Carbono (CO) promedio en partículas por millón (ppm) para la Ciudad de México desde 1995.
- Norma vigente: el límite de CO en el ambiente es de 11 ppm para un promedio de 8 horas.
- Relevante en temas de salud:
 - ✓ *Puede causar afectaciones al corazón o al cerebro*
 - ✓ *Uno de los tipos más comunes de envenenamiento*
- Relevante en temas de políticas públicas:
 - ✓ *Indicador para definir la calidad del aire y declarar contingencias ambientales.*
 - ✓ *Parámetros de la SEDENA:*
 - Buena: 0.00-5.50
 - Regular: 5.51-11.00
 - Mala: 11.01-16.50
 - Muy mala: 16.51-22.00
 - Extremadamente mala: >22.00

Objetivo

- El objetivo de este trabajo es ajustar distintos modelos Estacionales Autorregresivos Integrados de Promedios Móviles (SARIMA), así como un Modelo Dinámico Lineal (DLM) para posteriormente comparar el desempeño de ambas metodologías.

Index

1. Introducción

2. Datos

3. Modelos Clásicos

4. Modelo Dinámico Lineal

5. Comparación de Modelos

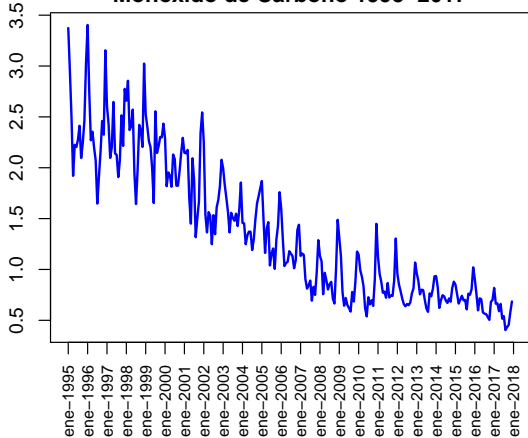
6. Conclusiones

Datos

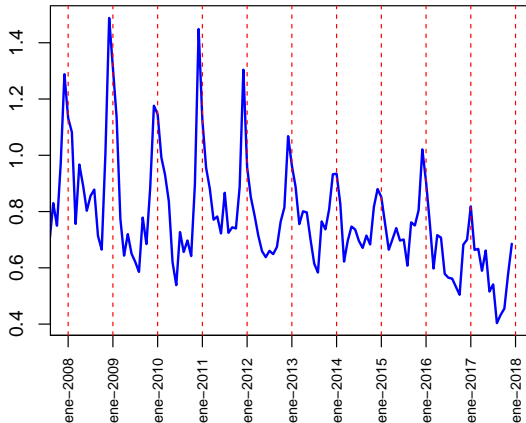
- Datos del nivel de CO por hora en la Ciudad de México obtenidos de la SEDEMA.
- Se agregan los datos mensualmente: 276 observaciones del promedio mensual de CO.
- Periodo muestral: enero de 1995 a diciembre del 2017.

Datos

Monóxido de Carbono 1995–2017



Monóxido de Carbono 2008–2017

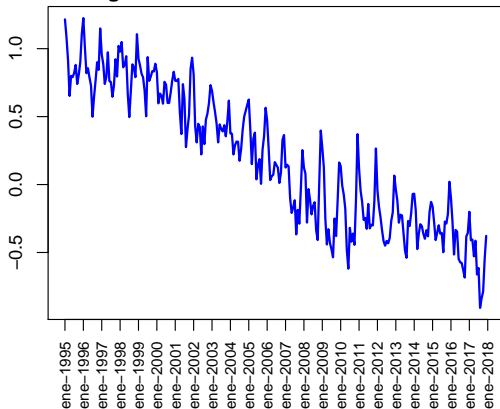


■ La serie tiene un componente estacional, tendencia decreciente y cambios en la varianza.

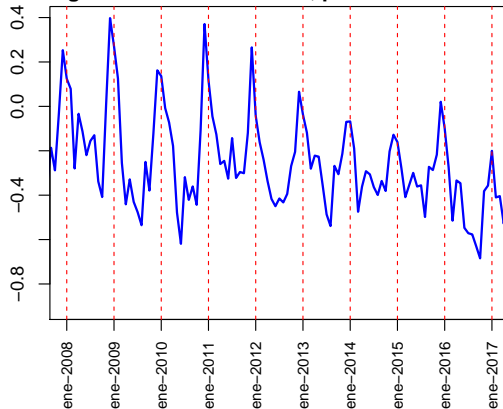
Datos

- Para tratar de eliminar los cambios en la varianza se obtiene el logaritmo del CO.

Log-Monóxido de Carbono 1995-2017



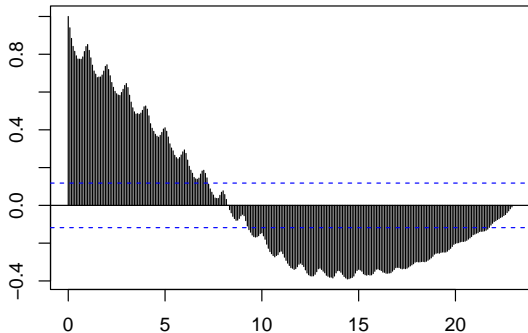
Log Monóxido de Carbono, periodo 2008-2017



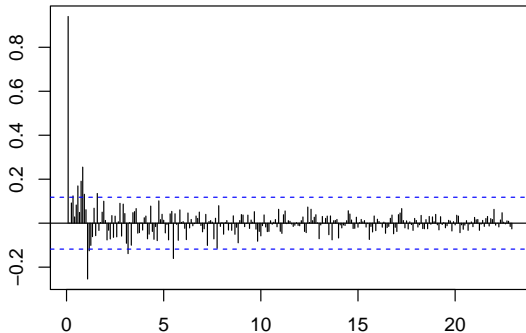
- Varianza es más homogénea, persiste la tendencia y la estacionalidad.

ACF y PACF Log-CO

ACF para la serie mensual del Log-CO



PACF para la serie mensual del Log-CO

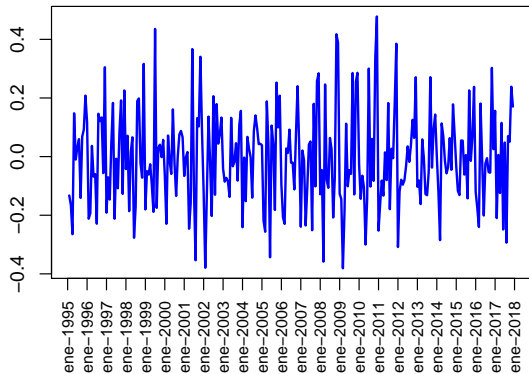


- ACF: varios rezagos significativos, comportamiento de una serie con componente estacional.
- PACF: algunos rezagos significativos, parece que se corta rápidamente.

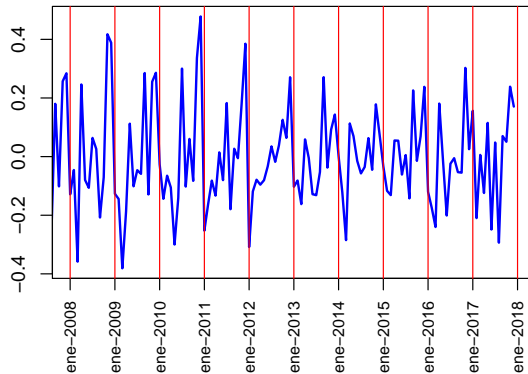
Diferencia Orden 1

- Para eliminar la estacionalidad diferenciamos de orden uno.
- Persiste la estacionalidad.

$\Delta \text{Log-Monóxido de Carbono}$

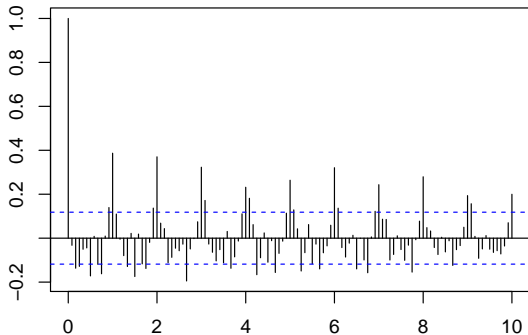


$\Delta \text{Log-Monóxido de Carbono, periodo 2008-2017}$

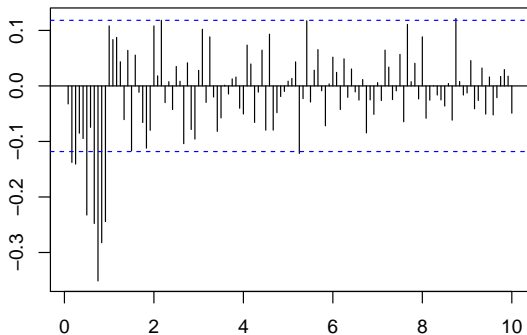


ACF y PACF Diferencia Orden 1

ACF para la serie mensual $\Delta\text{Log-CO}$



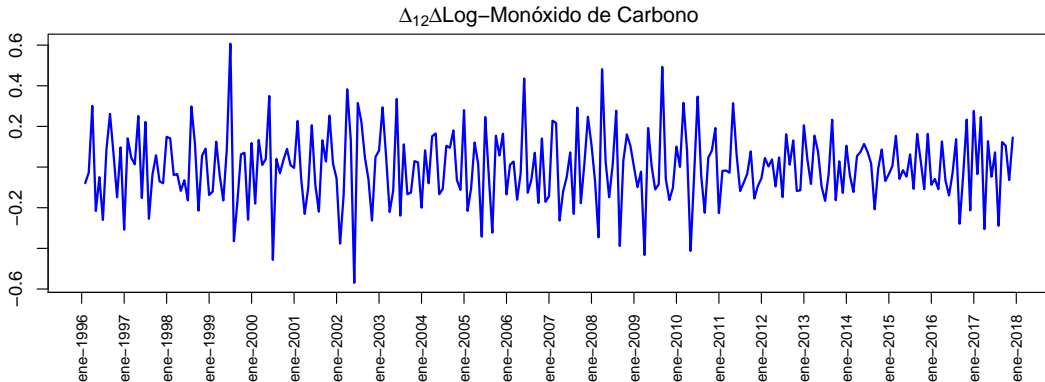
PACF para la serie mensual $\Delta\text{Log-CO}$



- Se eliminó el patrón estacional que mostraba el ACF, sigue habiendo rezagos significativos que parecen decrecer.
- PACF sigue mostrando algunos rezagos significativos que se cortan rápidamente.

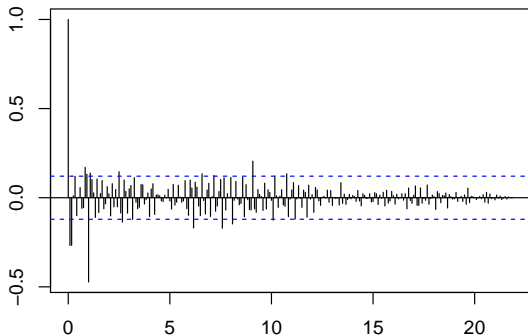
Diferencia de Orden 12 de la Diferencia

- Para eliminar la estacionalidad, diferenciamos de orden 12 la serie ya diferenciada.
- Ya no hay tendencia ni estacionalidad.
- Persisten algunos cambios en la varianza hacia el final de la serie.

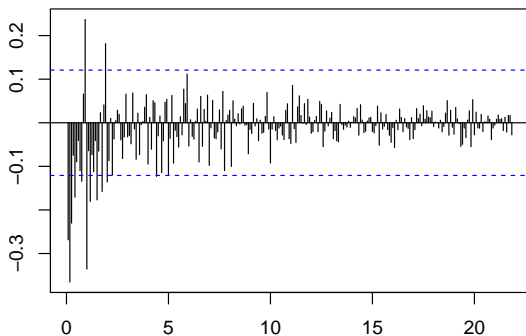


ACF y PACF Diferencia Orden 12 de la Diferencia

ACF para la serie mensual $\Delta_{12}\Delta\text{Log-CO}$



PACF para la serie mensual $\Delta_{12}\Delta\text{Log-CO}$



- ACF: se corta después del segundo rezago y muestra un valor significativo en el rezago 9.
- PACF se corta después de 4 rezagos aproximadamente (o podría decrecer).
- Esto nos hace pensar en un modelo SARIMA(p,d,q)(P,D,Q)_s o un DLM.

Index

1. Introducción

2. Datos

3. Modelos Clásicos

4. Modelo Dinámico Lineal

5. Comparación de Modelos

6. Conclusiones

Estimación de Modelos SARIMA

- Es claro que $d = 1$, $D = 1$ y $s = 12$.
- Valores de los parámetros p , q , P y Q no son tan claros.
- Creamos una función para estimar distintos modelos $SARIMA(p,d,q)(P,D,Q)_s$
 - ✓ Se hace una búsqueda exhaustiva en grid con código en paralelo.
 - ✓ Posibles valores para cada parámetro:
 - $p = 1, 2, 3, 4, 5$
 - $q = 1, 2, 3, 4, 5$
 - $P = 1, 2, 3, 4, 5$
 - $Q = 1, 2, 3, 4, 5$
 - ✓ En total se ajustaron 1296 modelos
 - ✓ 288 modelos no pudieron ser estimados con la función `arima`.
 - ✓ Se eligen los 5 modelos con menor AIC para competir y realizar las pruebas de diagnóstico.

Selección de Modelos SARIMA

Comparación AIC's para modelos SARIMA(p,d,q)×(P,D,Q)_s

Modelo							Criterio	
p	d	q	P	D	Q	s	Log-Like.	AIC
2	1	4	0	1	1	12	208.7130	-401.4260
0	1	2	0	1	1	12	204.3997	-400.7993
2	1	1	0	1	1	12	205.2593	-400.5186
0	1	3	0	1	1	12	205.2478	-400.4956
0	1	2	1	1	2	12	206.1918	-400.3836

- Los mejores modelos tienen parametros de orden chico para la parte estacional, pero mayores para la parte no estacional.
- El valor del AIC es muy similar entre los 5 modelos
- Hay mayor variabilidad en el valor de la verosimilitud.

Diagnóstico de Residuales: Prueba Ljung-Box

Prueba Ljung-Box para modelos SARIMA(p,d,q)x(P,D,Q)_s

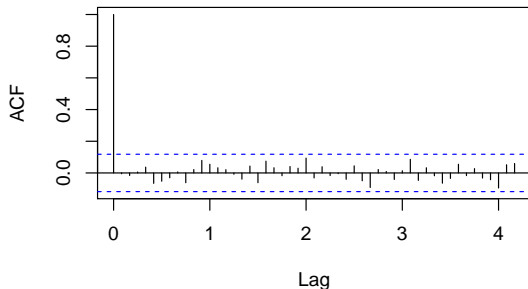
Modelo							Ljung-Box
p	d	q	P	D	Q	s	p-value
2	1	4	0	1	1	12	0.9881
0	1	2	0	1	1	12	0.8495
2	1	1	0	1	1	12	0.9885
0	1	3	0	1	1	12	0.9883
0	1	2	1	1	2	12	0.9021

- Los cinco modelos pasan la prueba.
- No hay evidencia de errores autocorrelacionados.

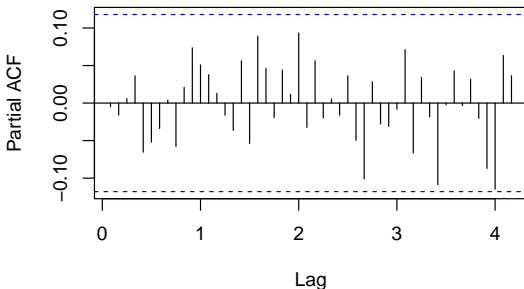
Diagnóstico de Residuales: ACF y PACF

- ACF: rezagos de los residuales no significativos para todos los modelos.
- PACF se tiene un rezago “marginamente” significativo (el rezago 4) en los modelos 2, 3 y 4.

ACF SARIMA(2,1,4)x(0,1,1)_12

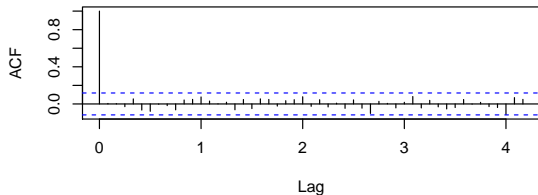


PACF SARIMA(2,1,4)x(0,1,1)_12

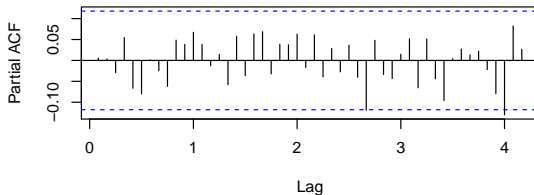


Diagnóstico de Residuales: ACF y PACF

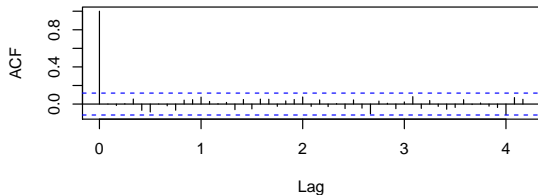
ACF SARIMA(0,1,2)x(0,1,1)_12



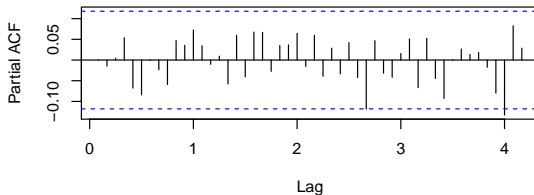
PACF SARIMA(0,1,2)x(0,1,1)_12



ACF SARIMA(2,1,1)x(0,1,1)_12

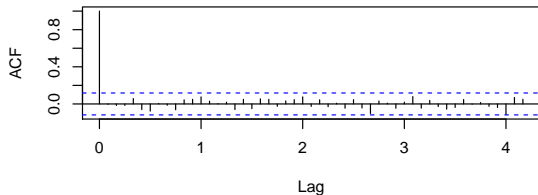


PACF SARIMA(2,1,1)x(0,1,1)_12

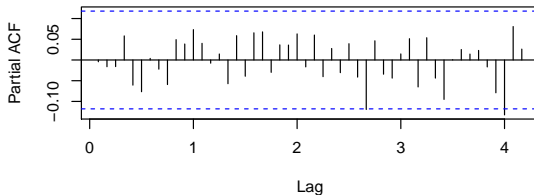


Diagnóstico de Residuales: ACF y PACF

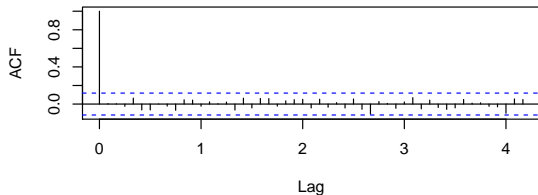
ACF SARIMA(0,1,3)x(0,1,1)_12



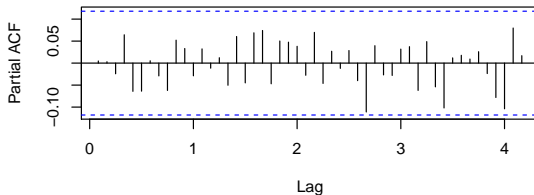
PACF SARIMA(0,1,3)x(0,1,1)_12



ACF SARIMA(0,1,2)x(1,1,2)_12

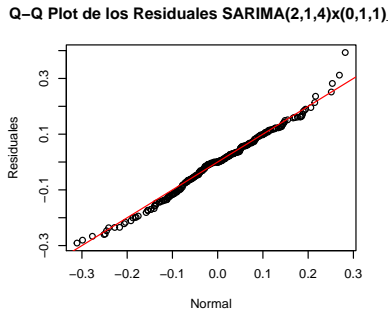
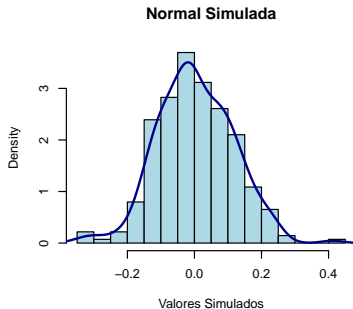
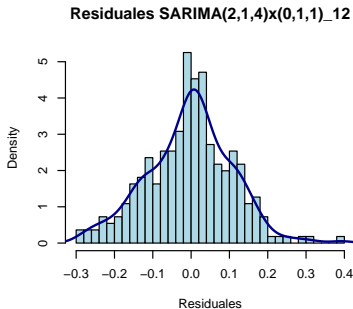


PACF SARIMA(0,1,2)x(1,1,2)_12



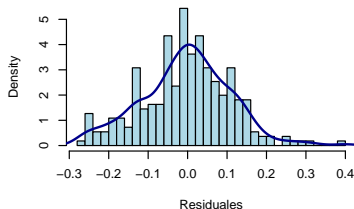
Diagnóstico de Residuales: Histogramas y Q-Q Plots

- La distribución de los residuales parecen tener cola derecha ligeramente más pesada.
- Los modelos 2, 3 y 4 parecen ser aquellos para los que el histograma de los residuales es más distinto al histograma de una normal simulados
- Los 5 Q-Q plots muestran mayor diferencia respecto a la línea de 45° en la cola derecha.

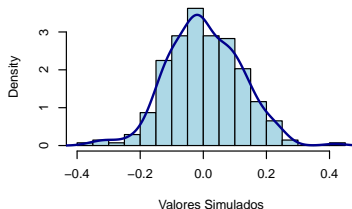


Diagnóstico de Residuales: Histogramas y Q-Q Plots

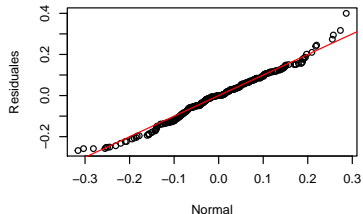
Residuales SARIMA(0,1,2)x(0,1,1)_12



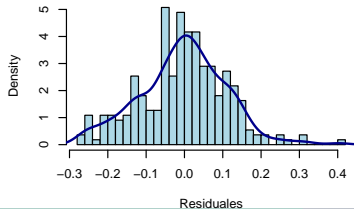
Normal Simulada



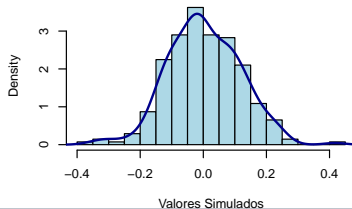
Q-Q Plot de los Residuales SARIMA(0,1,2)x(0,1,1).



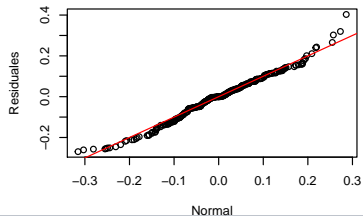
Residuales SARIMA(2,1,1)x(0,1,1)_12



Normal Simulada

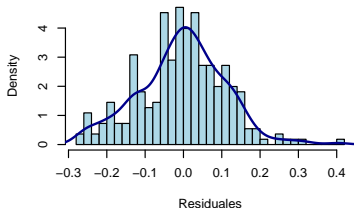


Q-Q Plot de los Residuales SARIMA(2,1,1)x(0,1,1).

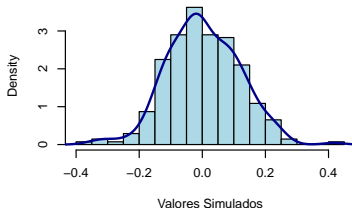


Diagnóstico de Residuales: Histogramas y Q-Q Plots

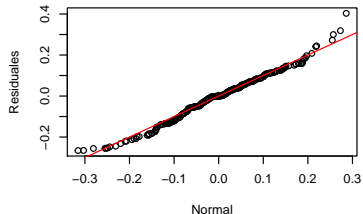
Residuales SARIMA(0,1,3)x(0,1,1)_12



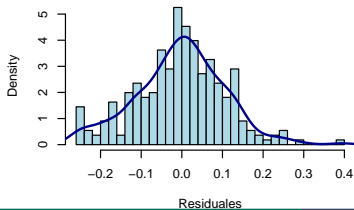
Normal Simulada



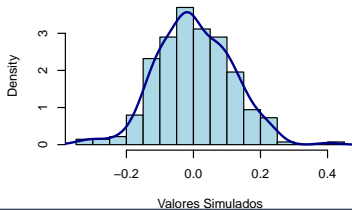
Q-Q Plot de los Residuales SARIMA(0,1,3)x(0,1,1).



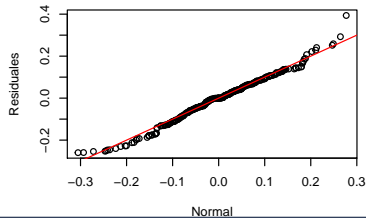
Residuales SARIMA(0,1,2)x(1,1,2)_12



Normal Simulada



Q-Q Plot de los Residuales SARIMA(0,1,2)x(1,1,2).



ECM Para Predicciones

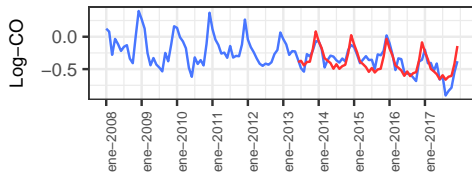
- Se reserva el 20% final de la muestra para evaluar el ECM de las predicciones.
- Todos los modelos presentan un ECM muy pequeño.
- El primer modelo ($SARIMA(0, 1, 4) \times (0, 1, 1)_{12}$) el que da el menor ECM para las predicciones del CO.
 - ✓ *La diferencia en ECM entre el primer y quinto modelo es muy pequeña, pero el quinto modelo tiene 2 parámetros menos a estimar.*

Comparación ECM's para modelos $SARIMA(p,d,q)X(P,D,Q)_s$

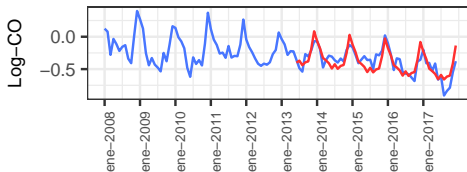
Modelo							Criterio
p	d	q	P	D	Q	s	ECM
2	1	4	0	1	1	12	0.014716
0	1	2	0	1	1	12	0.014900
2	1	1	0	1	1	12	0.014955
0	1	3	0	1	1	12	0.014959
0	1	2	1	1	2	12	0.014728

ECM Para Predicciones

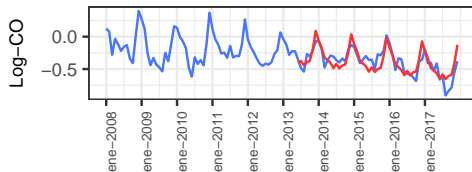
Residuales SARIMA(2,1,4)x(0,1,1)_12



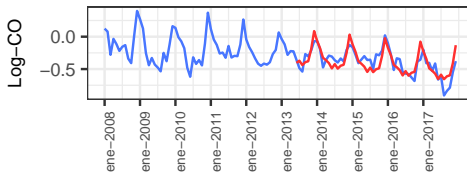
Residuales SARIMA(0,1,2)x(0,1,1)_12



Residuales SARIMA(2,1,1)x(0,1,1)_12

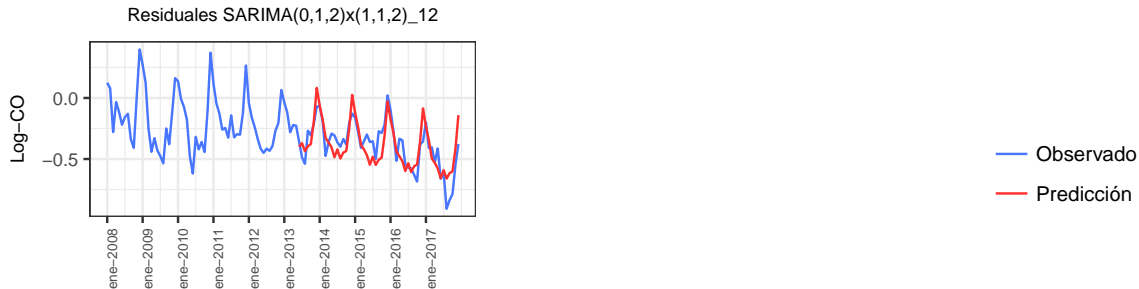


Residuales SARIMA(0,1,3)x(0,1,1)_12



— Observado
— Predicción

ECM Para Predicciones



- Los 5 modelos reproducen de manera correcta la estacionalidad de la serie, pero subestiman ligeramente los valles sobre todo a finales del 2017.

Elección del Modelo

- Tomando en cuenta el ACF, PACF, AIC, análisis de residuales y ECM obtenidos, así como la idea de tener un modelo parsimonioso, se decide utilizar el quinto modelo; es decir, un $SARIMA(0, 1, 2)X(1, 1, 2)_{12}$.

- El modelo final está dado por:

$$(1 - \Phi_1 B^{12})(1 - B)(1 - B^{12})CO_t = (1 + \theta_1 B^{12} + \theta_2 B^{24})(1 + \Theta_1 B^{12} + \Theta_2 B^{24})Z_t,$$

- Los valores estimados para los parámetros son:

- ✓ $\theta_1 = -0.5078$
- ✓ $\theta_2 = -0.3345$
- ✓ $\Phi_1 = -0.7622$
- ✓ $\Theta_1 = -0.0518$
- ✓ $\Theta_2 = -0.6546$
- ✓ $\sigma^2 = 0.0143$

Elección del Modelo

- Desarrollando polinomios característicos y sustituyendo el valor de los parámetros, se tiene que el mejor modelo SARIMA es

$$\begin{aligned} CO_t = & CO_{t-1} + 0.2378CO_{t-12} - 0.2378CO_{t-13} + 0.7622CO_{t-24} - 0.7622CO_{t-25} \\ & + Z_t - 0.5078Z_{t-1} - 0.3345Z_{t-2} - 0.0518Z_{t-12} + 0.0263Z_{t-13} \\ & + 0.0173Z_{t-14} + -0.6546Z_{t-24} + 0.3324Z_{t-25} + 0.2189Z_{t-26} \end{aligned} \quad (1)$$

donde Z_t es ruido blanco con media cero y varianza $\sigma^2 = 0.0143$.

Index

1. Introducción

2. Datos

3. Modelos Clásicos

4. Modelo Dinámico Lineal

5. Comparación de Modelos

6. Conclusiones

Selección del DML

- La serie del Log-CO tiene tendencia decreciente y estacionalidad.
- Por lo tanto, se necesita un DLM polinomial de segundo orden junto con un DLM estacional.
- Para el DLM polinomial se tiene el siguiente modelo Espacio-Estado:

$$\begin{aligned}Y_t^{(1)} &= A_1 X_t^{(1)} + V_t^{(1)} \\ X_{t+1}^{(1)} &= G_1 X_t^{(1)} + W_t^{(1)}\end{aligned}$$

donde $A_1 = [1, 0]$, $V_t^{(1)} = V^{(1)}$, $W_t^{(1)} = \text{diag}(W_1, W_2)$ y $G_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

- Para el DLM estacional con periodo $s = 12$, se tiene el siguiente modelo Espacio-Estado:

$$\begin{aligned}Y_t^{(2)} &= A_2 X_t^{(2)} + V_t^{(2)} \\ X_{t+1}^{(2)} &= G_2 X_t^{(2)} + W_t^{(2)}\end{aligned}$$

donde $A_2 = [1, 0_{1 \times 10}]$, $V_t^{(2)} = 0$, $W_t^{(2)} = \text{diag}(\sigma_w^2, 0_{1 \times 10})$ y $G_2 = \begin{bmatrix} -\mathbb{1}_{10 \times 10} & 0 \\ I_{10 \times 10} & 0_{10 \times 1} \end{bmatrix}$.

Estimación de V y W vía Máxima Verosimilitud

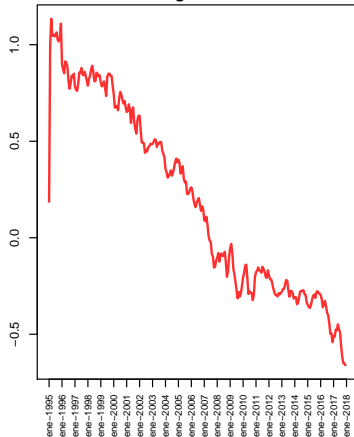
- Se estiman los valores máximo verosímiles de las matrices V y W en ambos DLM'S.
- Después de revisar la convergencia, se obtienen los siguientes estimadores:
 - ✓ $\hat{V}^{(1)} = 0.0101$
 - ✓ $\hat{W}_1 = 9 \times 10^{-4}$
 - ✓ $\hat{W}_2 = 0$
 - ✓ $\hat{\sigma}_W^2 = 0$.
- Una vez conocidas las estimaciones de V y W se puede proceder a estimar el filtrado, suavizado y la predicción del modelo DLM.

Filtrado

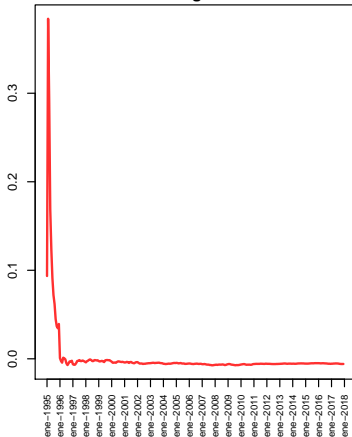
- El Filtro de Kalman es que es sensible a condiciones iniciales; este efecto se puede notar en el inicio del filtrado de los tres estados
 - ✓ *Sin embargo, el filtro corrige rápidamente las trayectorias.*
- El nivel sigue muy de cerca la dinámica de la serie observada.
- La pendiente decrece a lo largo de los primeros dos años y después se mantiene relativamente constante.
- La dinámica de la componente estacional parece ser constante a lo largo de la muestra.

Filtrado

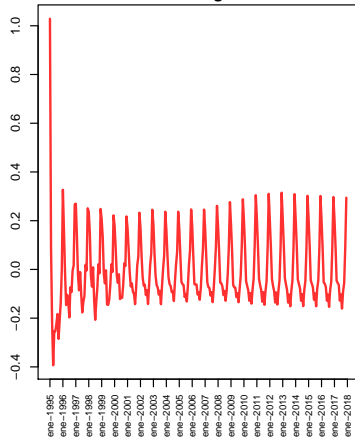
Nivel Log-CO Filtrado



Pendiente Log-CO Filtrado

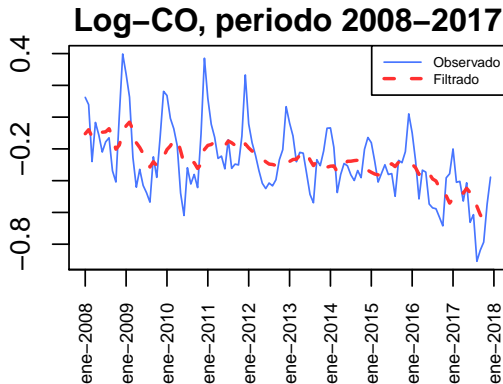
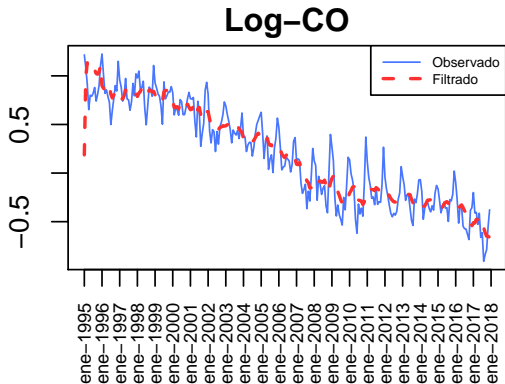


Estacionalidad Log-CO Filtrado

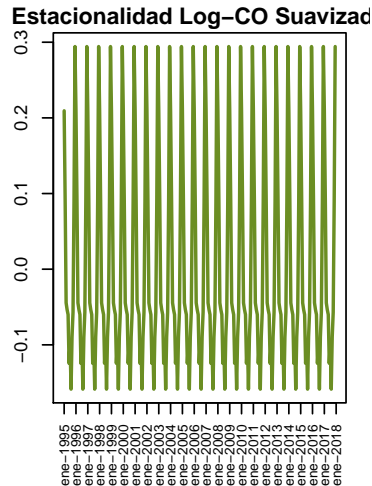
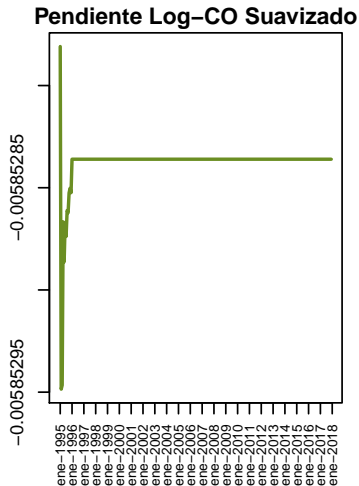
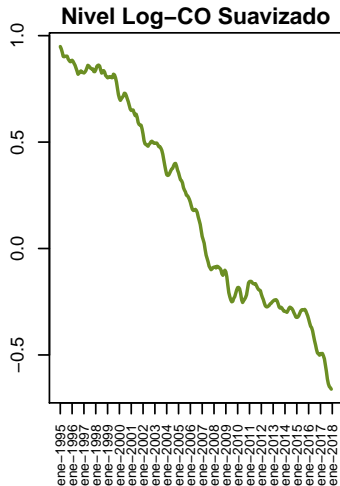


Filtrado

- Se puede observar una variabilidad mucho menor en la serie filtrada que en la observada, esto debido principalmente a que el cociente V/W es grande ($V/W_1 = 11.04$).



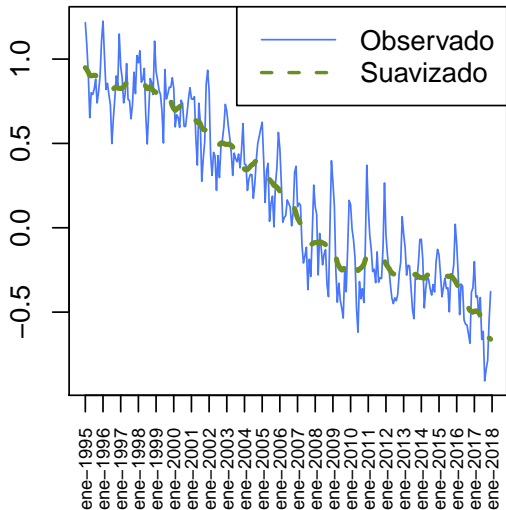
Suavizado



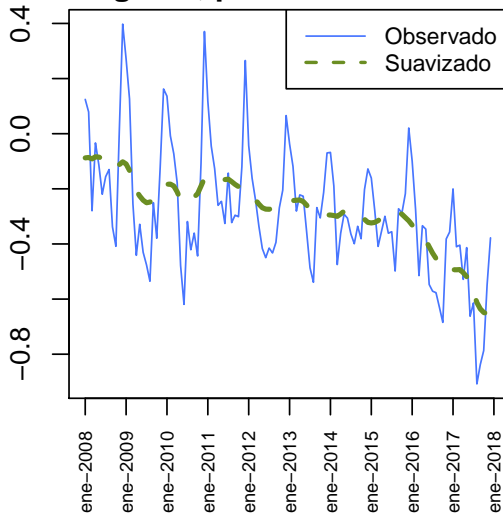
- Las tres series tienen menor variabilidad que las series filtradas pues toman en cuenta los valores futuros de la serie.

Suavizado

Log-CO



Log-CO, periodo 2008-2017

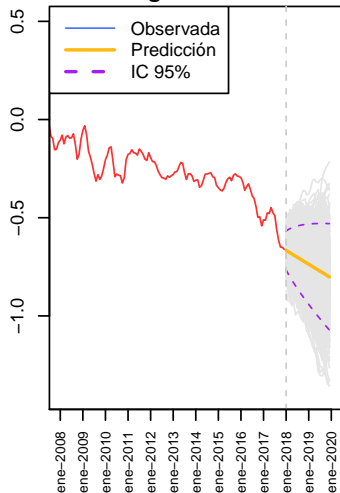


Predicción

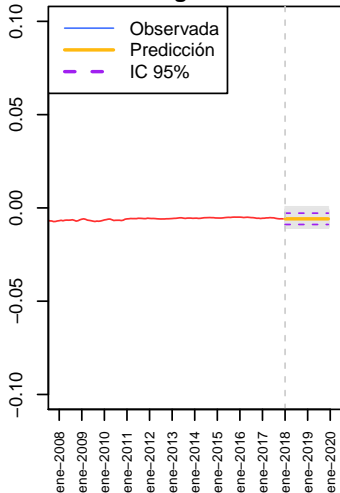
- Se realiza la predicción para los siguientes 2 años (24 observaciones) y se simulan 1000 trayectorias.
- Parece que un mayor número de trayectorias quedan fuera del intervalo de confianza para el nivel que para los otros dos estados.
- Los intervalos de confianza para la pendiente y estacionalidad son más angostos que para el nivel.
- Los valores pronosticados para los tres estados parecen continuar de manera razonable con la trayectoria de las series filtradas.

Predicción

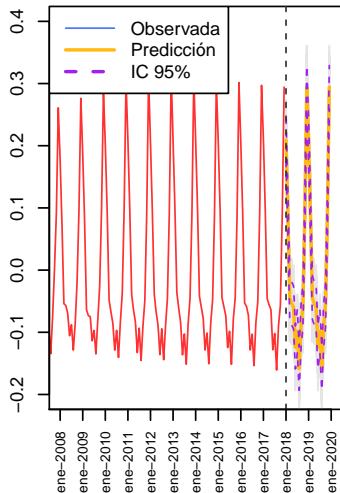
Nivel Log-CO Predicción



Pendiente Log-CO Predicción

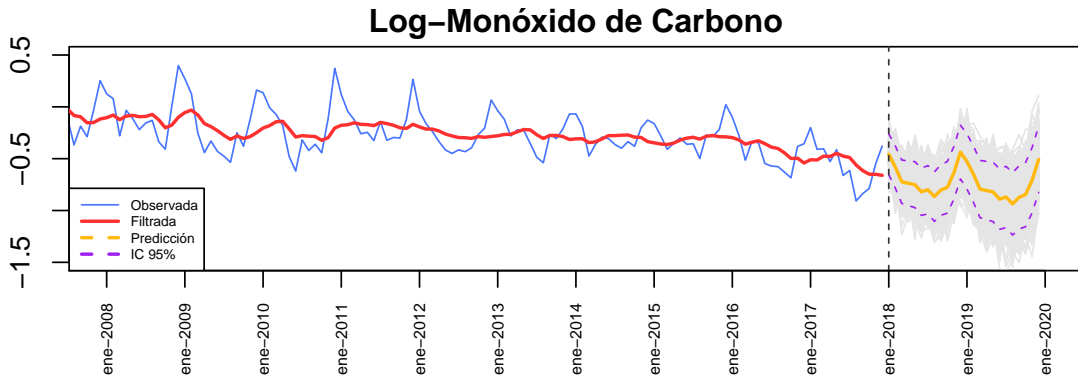


Estacionalidad Log-CO Predicción



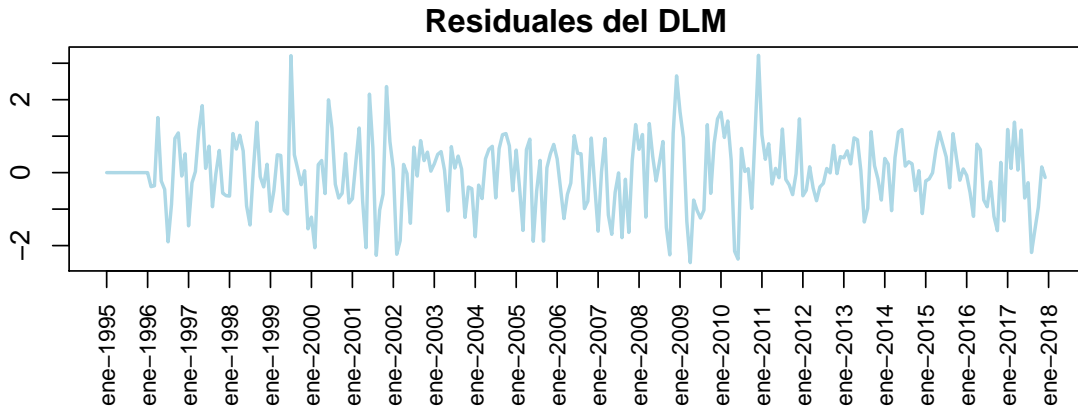
Predicción

- Se puede apreciar que el pronóstico mantiene bien la estacionalidad de la serie y es acorde a la tendencia que presentaba la serie.



Análisis de Residuales

- Se obtienen los residuales del modelo DLM estimado; éstos fluctúan al rededor del cero y parecen tener varianza constante.



Análisis de Residuales: Prueba Ljung-Box

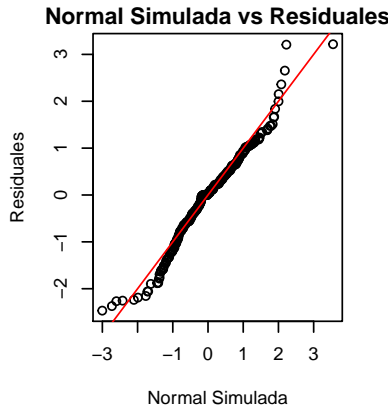
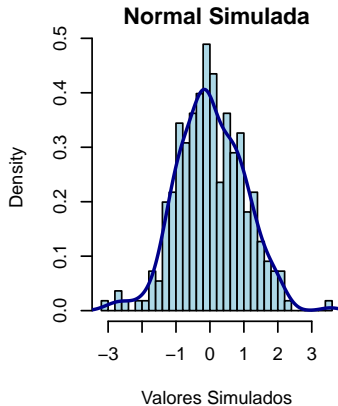
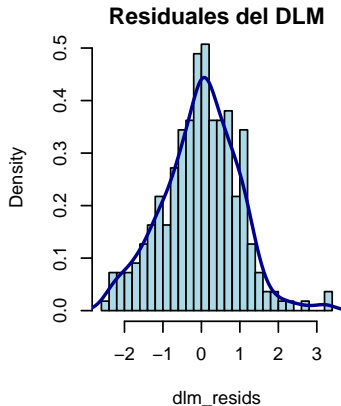
Prueba Ljung-Box Residuales

Rezago	Valor.p
1	0.000258
2	0.000182
3	0.000168
4	0.000447
5	0.000213
6	0.000045
7	0.000079
8	0.000122
9	0.000117
10	0.000228
11	0.000226
12	0.000089

- En todos los casos se rechaza la hipótesis nula de no correlación de los residuales

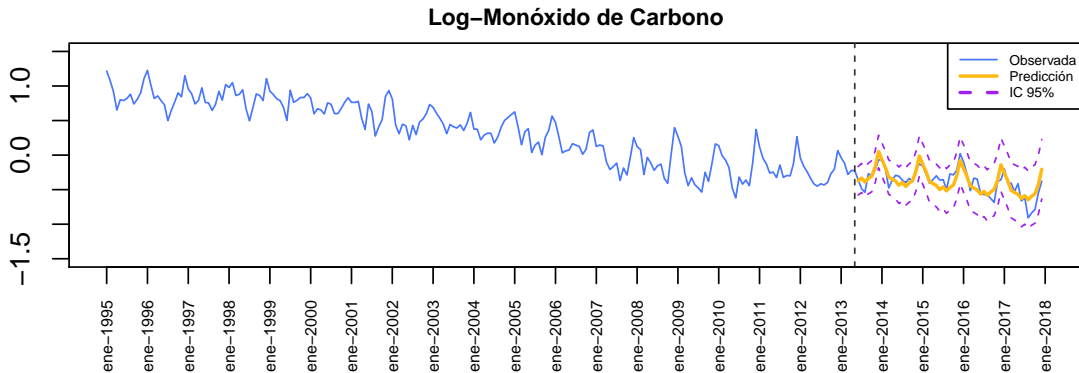
Análisis de Residuales: Histograma y Q-Q plot

■ Los residuales sí parecen tener una forma de normal.



ECM Para Predicciones

- Se reserva el 20% final de la muestra para evaluar el ECM de las predicciones del DLM.
- El modelo está sobrestimando, pero todas las observaciones quedaron dentro del IC.
- El ECM para este modelo es de 0.0125.



Index

1. Introducción

2. Datos

3. Modelos Clásicos

4. Modelo Dinámico Lineal

5. Comparación de Modelos

6. Conclusiones

Comparación de modelos

Comparación entre el modelo clásico y el DLM		
	<i>SARIMA</i> (0, 1, 2) <i>X</i> (1, 1, 2) ₁₂	DLM (n=2 + Est.)
Ajuste	✓	✓
No correl. residuales	✓	✗
Residuales Normales	✓	✓
ECM Predicciones	0.0147	0.0125

- El modelo clásico pasa todas las pruebas de validación mientras que el DLM tiene errores autocorrelacionados.
- El modelo clásico tiene mayor ECM para el pronóstico del último 20% de la muestra que el DLM, aunque la diferencia entre ambos es mínima.

Index

1. Introducción

2. Datos

3. Modelos Clásicos

4. Modelo Dinámico Lineal

5. Comparación de Modelos

6. Conclusiones

Consideraciones Finales

- En este trabajo se utilizaron modelos clásicos y modelos Dinámicos Lineales para ajustar y estimar un modelo para el logaritmo del promedio mensual del Monóxido de Carbono (Log-CO) en la Ciudad de México.
- La serie Log-CO presenta tendencia decreciente y estacionalidad.
- Dentro de los modelos clásicos se realizó una búsqueda en grid con código en paralelo para distintas combinaciones de los parámetros de un $SARIMA(p,d,q)(P,D,Q)_s$.
 - ✓ Se estimaron 1296 modelos.
 - ✓ Se compararon los 5 modelos con menor AIC.
 - ✓ Únicamente 2 de los 5 modelos pasaron todas las pruebas de diagnóstico.
 - ✓ Los 5 modelos tienen un ECM pequeño (0.0149 aprox.)
 - ✓ Se eligió el modelo $SARIMA(0,1,2)X(1,1,2)_{12}$ como el mejor modelo clásico.

Consideraciones Finales

- Para el DLM se utilizó un modelo polinomial de orden 2 con estacionalidad.
 - ✓ *Las matrices de varianzas y covarianzas se estimaron vía Máxima Verosimilitud.*
 - ✓ *Se obtuvo el filtrado, suavizado y predicciones (a 24 meses) para la serie del Log-CO.*
 - ✓ *Los residuales del modelo pasan todas las pruebas de diagnósticos excepto la de autocorrelación.*
 - ✓ *El modelo tiene un ECM de 0.0125.*
- Entre el modelo $SARIMA(0, 1, 2)X(1, 1, 2)_{12}$ y el DLM polinomial con estacionalidad se eligió el modelo clásico para ajustar el Log-CO.

15 de mayo del 2018