Análisis Estadístico y Pronóstico de la Producción Mensual de Leche

El conjunto de datos utilizado en este análisis corresponde a una serie temporal mensual de producción de leche, con una extensión de 120 meses (equivalente a 10 años de observaciones continuas). Cada dato representa la cantidad de producto generada por la empresa en un mes específico, permitiendo estudiar el comportamiento histórico de la producción, así como identificar tendencias, estacionalidades y posibles anomalías.

La inspección inicial de la serie (Figura 1) revela una tendencia creciente a lo largo del periodo analizado, lo cual indica que la producción ha mostrado un aumento sostenido con el paso del tiempo. Asimismo, es evidente la presencia de un patrón estacional anual, caracterizado por oscilaciones regulares que se repiten aproximadamente cada 12 meses. Este comportamiento estacional es típico en industrias agroalimentarias, donde factores climáticos, ciclos biológicos y condiciones de demanda afectan la producción de manera periódica.

Dado que la serie exhibe tendencia y estacionalidad, puede concluirse que no es estacionaria en su forma original. Esto implica que, previo a la construcción de un modelo de pronóstico, se requiere aplicar transformaciones adecuadas (como diferenciación estacional y/o logaritmos) para estabilizar la varianza y las medias en el tiempo.

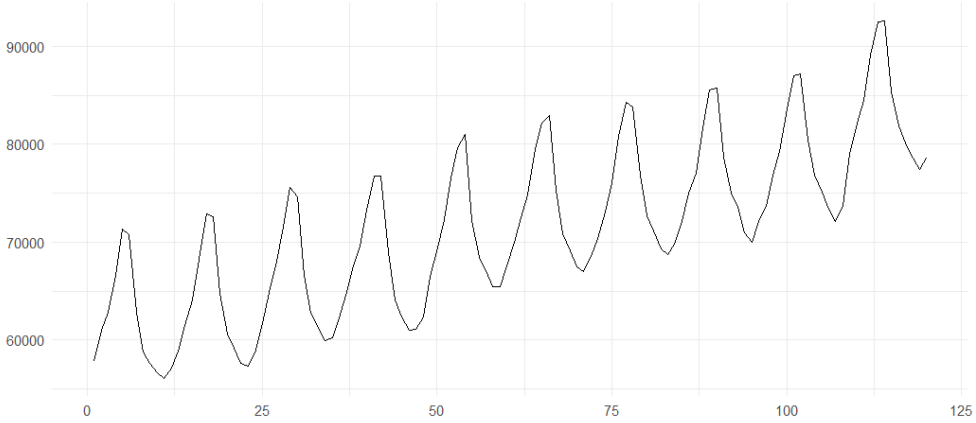

Figure 1: Serie de tiempo de la base de datos Milk

Limpieza de datos.

Antes de realizar el análisis, se efectuó una revisión exploratoria de la serie para identificar valores atípicos, datos faltantes o inconsistencias. Se emplearon visualizaciones como el gráfico de la serie temporal y diagramas de caja por año, así como una inspección de estadísticos descriptivos básicos.

A partir de esta revisión no se detectaron outliers significativos, valores faltantes ni patrones anómalos que requirieran un preprocesamiento adicional. Por ello, el análisis se llevó a cabo utilizando los datos originales sin modificaciones.

Transformación y Estacionarización de la Serie

La serie original presenta una tendencia creciente y un marcado componente estacional anual, por lo que no cumple con los supuestos de estacionariedad requeridos para la modelación mediante **SARIMA**. Con el fin de estabilizar la media y eliminar estos patrones sistemáticos, se aplicó la siguiente transformación combinada:

$$Y_t = (1 - B)(1 - B^{12})X_t$$

donde B representa el operador rezago. Esta transformación incluye:

- **Diferenciación regular** (Elimina la tendencia)

$$(1 - B)$$

- **Diferenciación estacional** (remueve el ciclo anual identificado previamente)

$$(1 - B^{12})$$

Tras aplicar ambas diferenciaciones, la serie resultante presenta una media aproximadamente constante, ausencia de estructura estacional visible y variabilidad más estable, lo cual sugiere que se logró obtener una serie razonablemente estacionaria. Esto puede observarse en la **Figura 2**, donde se muestra la serie transformada.

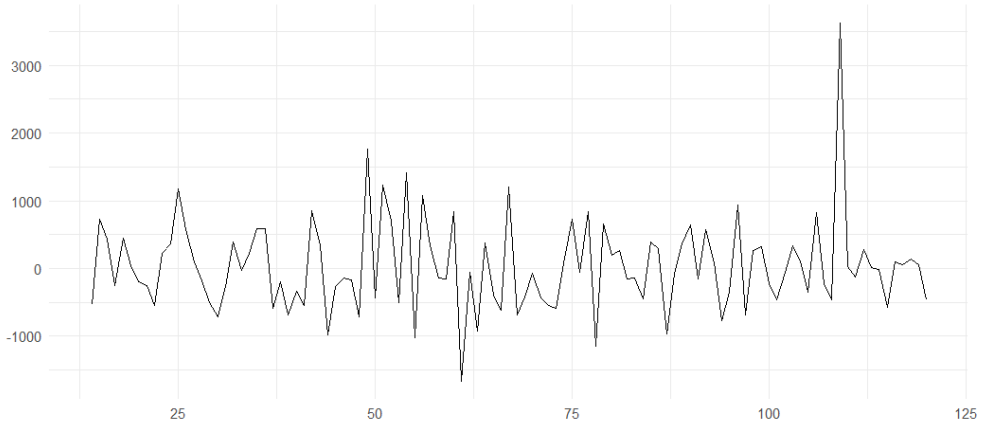

Figure 2: Serie de tiempo tras diferenciar una vez y diferenciar con lag igual a 12

Exploración de modelos

A partir de la transformación aplicada previamente, la estructura adecuada para modelar la serie es un modelo **SARIMA**, el cual permite capturar tanto la dinámica regular como la dinámica estacional anual observada en los datos. Dado que la serie fue diferenciada una vez (d = 1) y se aplicó una diferenciación estacional anual (D = 1), el espacio de modelos a considerar se define como:

$$\text{SARIMA}(p,\, 1,\, q)\,(1,\, 1,\, 1)_{12}$$

donde los órdenes p, q, P serán determinados mediante la inspección de las funciones ACF y PACF de la serie transformada.

Además del enfoque SARIMA, se considerará el modelo **Holt–Winters** como alternativa complementaria, especialmente útil para series con tendencia y estacionalidad marcadas.

Ajuste de modelos

Se estimaron diferentes modelos SARIMA explorando combinaciones razonables de los parámetros p y q, manteniendo los órdenes de diferenciación previamente establecidos. En todos los casos, los diagnósticos de los residuos indican ausencia de autocorrelación significativa y la prueba de Ljung–Box no rechaza la hipótesis nula, lo que sugiere que los modelos ajustan adecuadamente la estructura temporal de la serie.

Dado que varios modelos cumplen los criterios de diagnóstico, la selección se realizó comparando el **RMSE** obtenido por cada uno. Este indicador permitió identificar el modelo con mejor desempeño predictivo dentro del conjunto evaluado.

A continuación, se presentan los residuos de los modelos analizados (Figuras 3–5).

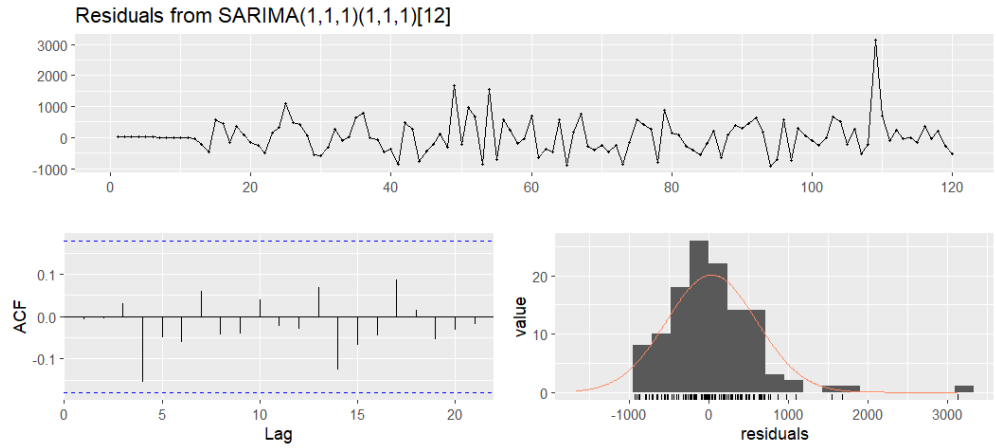

Figure 3: Residuales del modelo SARIMA(1,1,1)(1,1,1)12

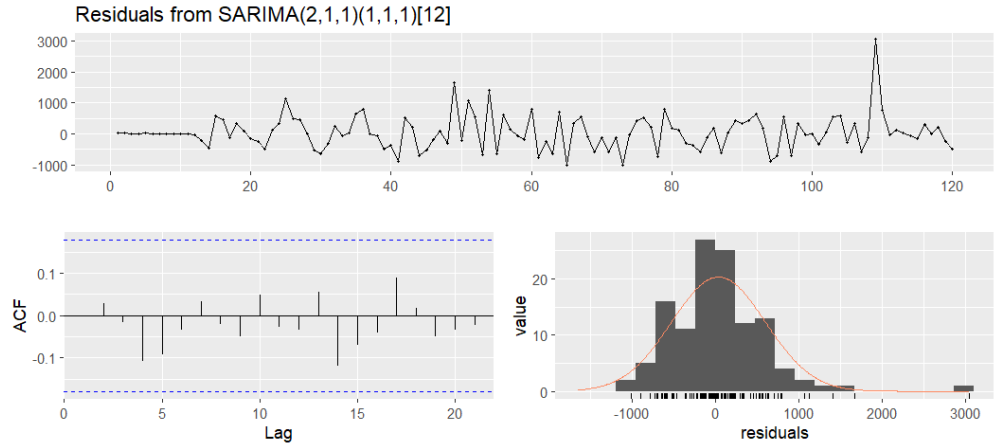

Figure 4: Residuales del modelo SARIMA(2,1,1)(1,1,1)12

Selección del modelo.

Para seleccionar el modelo óptimo, utilizamos el criterio de minimización de la raíz del error cuadrático medio (RMSE) evaluada en un conjunto de prueba que corresponde al 20% final de la serie temporal, habiendo entrenado cada modelo con el 80% inicial.

Este criterio es más pragmático que el AIC, ya que permite comparar modelos de naturaleza distinta y proporciona una estimación más directa del desempeño predictivo futuro.

Modelo	p	q	RMSE	AIC
SARIMA 1	1	1	2142.196	1299.884
SARIMA 2	2	1	2145.579	1301.842
SARIMA 3	1	2	2145.579	1301.84

Modelo	p	q	RMSE	AIC
SARIMA 4	2	2	2187.612	1298.838
Holt-Winters	-	-	3144.533	-

El modelo SARIMA(1,1,1)(1,1,1)12 minimizó el RMSE, por lo que fue seleccionado para el análisis posterior.

La siguiente gráfica muestra la comparación entre las predicciones de los modelos SARIMA y Holt-Winters frente a los datos reales (**Figura 6**):

Validación del modelo.

Una vez elegido el modelo SARIMA(1,1,1)(1,1,1)12 hay que re-entrenarlo en el 100% de los datos para hacer las predicciones finales.

Las siguientes gráficas muestran los residuales del modelo (**Figura 7**); el ACF de los residuales (**Figura 8**); y el PACF de los residuales (**Figura 9**):

Los residuales presentan un comportamiento adecuado, permaneciendo dentro de las bandas de confianza. La prueba Ljung-Box no rechaza H_0 con un p-value $= 0.9488$, confirmando que los residuales no presentan autocorrelación. Por tanto, el modelo es válido para realizar predicciones.

Predicción a 1 año.

Según el modelo, la producción proyectada de la empresa dentro de 12 meses será de **81,855 galones de leche**.

Riesgo de la predicción.

Como consultora, es fundamental comunicar al cliente no solo la predicción puntual, sino también una medida de incertidumbre asociada.

Para nuestro caso, las bandas de confianza son pequeña, para 12 meses tenemos un intervalo de $\pm$ 3629 galones alrededor de la media (el intervalo de confianza es $(78225, 85484)$). Hay un 95% de probabilidad de que no se cumpla la predicción que se está dando, 2.5% que esté arriba y 2.5% que esté abajo (caso desfavorable.)

Conclusiones

El análisis comparativo entre los modelos evaluados permitió identificar que el SARIMA(1,1,1)(1,1,1)_{12} obtuvo el menor RMSE, por lo que fue seleccionado como el modelo final. Los diagnósticos de residuos y la prueba de Ljung–Box indican que el modelo reproduce adecuadamente la dinámica temporal de la serie y que los errores no presentan autocorrelación, validando su capacidad predictiva.

Con este modelo, se proyecta que la producción dentro de 12 meses será aproximadamente de 81,855 galones, acompañada de un intervalo de confianza estrecho, lo que refleja una predicción confiable. En conjunto, los resultados proporcionan una estimación robusta que puede apoyar la planeación operativa y la toma de decisiones de la empresa.

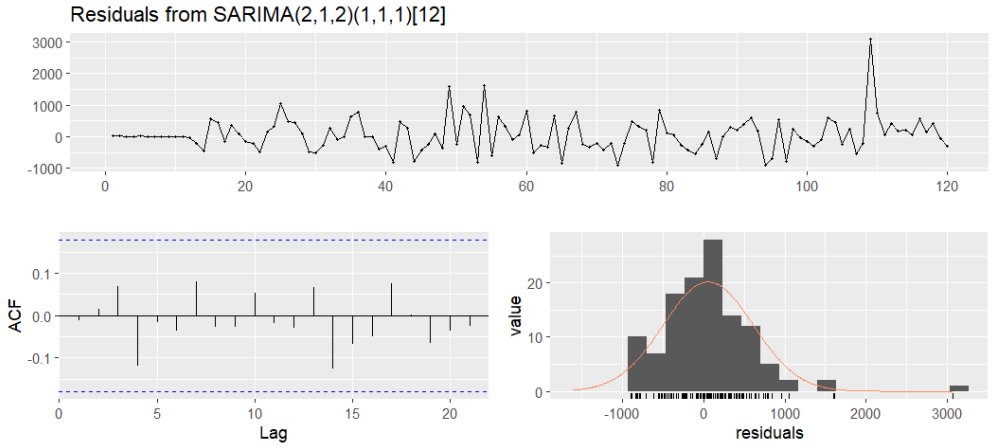

Figure 5: Residuales del modelo SARIMA(2,1,2)(1,1,1)12

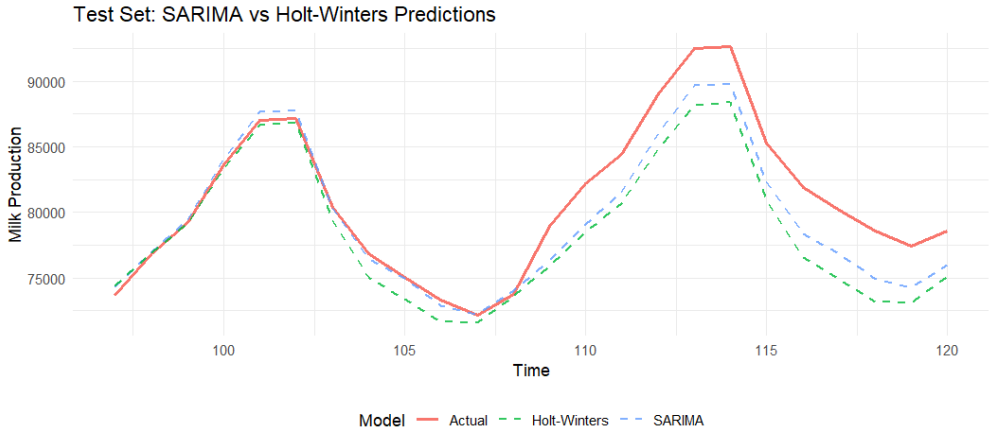

Figure 6: SARIMA vs Holt-Winters

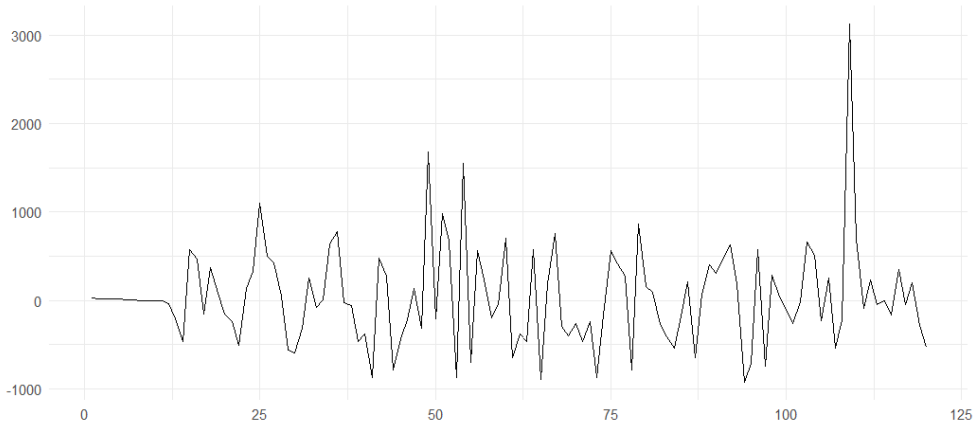

Figure 7: Resiudales modelo seleccionado

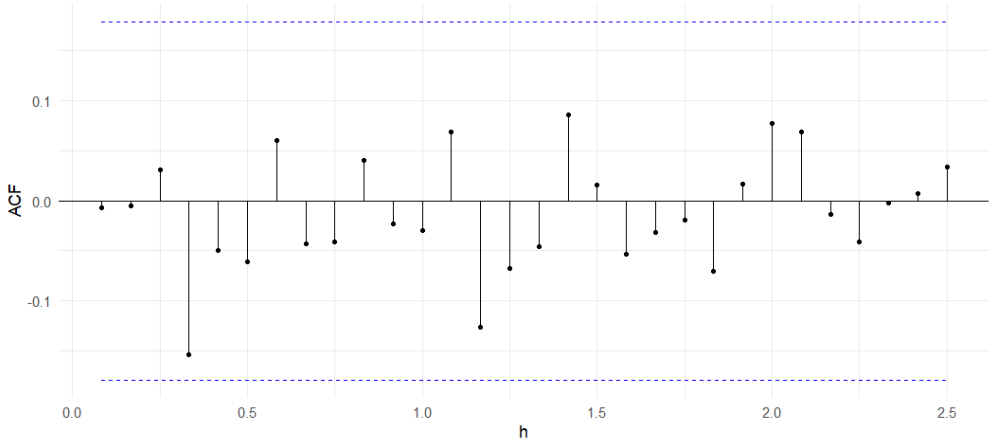

Figure 8: ACF modelo seleccionado

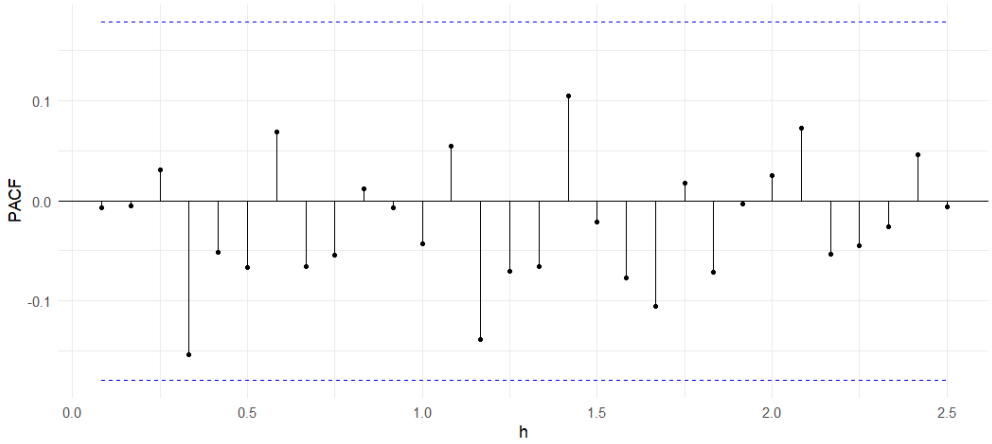

Figure 9: PACF modelo seleccionado

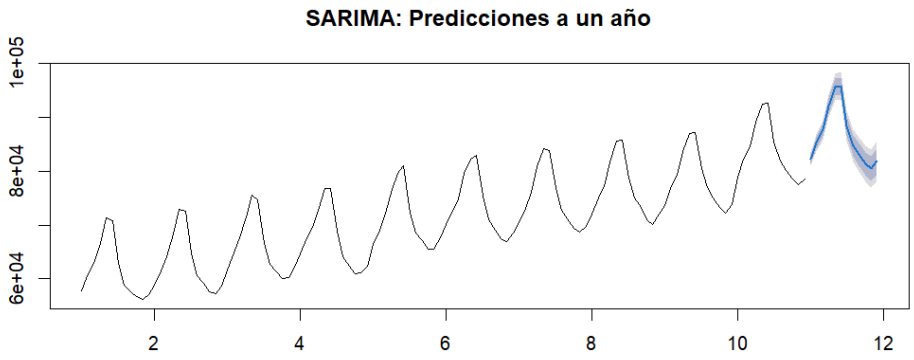

Figure 10: Predicción a 12 meses con base en el modelo seleccionado