



Precio Promedio de la Leche Cruda en Finca Córdoba, Colombia: Enero 2013 - Julio 2024

Luis David Bassa Llorente

Universidad de Córdoba

30 de noviembre de 2024

Reacreditados Institucionalmente, resolución N° 000020 del 11 de enero de 2023 por el Ministerio de Educación Nacional, certificados en: ISO: 9001 – ISO: 45001 e ISO: 14001 ICONTEC

Unicórdoba, calidad, innovación e inclusión para la transformación del territorio

Contenido

- 1 Introducción
- 2 Planteamiento del problema
- 3 Objetivo general
- 4 Estado del Arte
- 5 Datos Empleados
- 6 Resultados del modelado
- 7 Conclusión
- 8 Bibliografía

Introducción

En Colombia la leche es un alimento básico de la canasta familiar e indispensable para mejorar la calidad de vida. La importancia de la leche en la dieta se justifica por su composición química que incluye, además de agua, un 87 %, un 13 % de nutrientes de alto valor: grasas, proteínas, carbohidratos, vitaminas y minerales [1], por lo que la leche se considera un alimento completo recomendado para ser consumido por todas las personas. Además del consumo humano y animal, la leche y los productos lácteos son una materia prima importante para la industria alimentaria, que es uno de los sectores más importantes en la economía del país y más en la región de la costa más que todo en Córdoba y sucre que una región ganadera.

Planteamiento del problema

Este trabajo de investigación tiene como objetivo centrarse en la capacidad de modelar y predecir el precio promedio de la leche cruda en fincas de Córdoba, Colombia. Se propondrán diversos modelos de series temporales con fines de pronóstico de corto alcance, tales como el Arima y Sarima, todo ellos mediante la fusión de la metodología de Box-Jenkis, y un proceso de validación cruzada, que permite evaluar el desempeño de los modelos escogidos.

Adicionalmente se verá cual de los modelos escogidos es el mejor en la capacidad de predecir los horizontes 12 empleados en la base de datos mensual del promedio de la leche cruda en finca, recopilada de enero del 2023 a julio del 2024 DANE.

Objetivo general

El objetivo principal en este trabajo de investigación es desarrollar un modelo de pronóstico que sea capaz de predecir con una alta precisión los pronósticos a futuros. Con base a estos modelo informar a los entes encargados, para tomas de decisiones politicas económicas eficaces a futuro con respecto al precio de la leche cruda en finca, adicionalmente para investigaciones futuras y ver que variables externas influyen en el precio de la leche.

Estado del Arte

Datos Empleados

La serie temporal empleada en este análisis corresponde al precio de la leche cruda en finca. La base de datos incluye una serie histórica que abarca desde enero de 2013 hasta diciembre de 2019, Datos complementada con los datos de los años 2020, 2021, 2022, 2023 y parte del año 2024, hasta julio. Las variables originales incluyen mes, año, nombre del departamento, código del departamento, nombre del municipio, código del municipio y precio promedio por litro.

Datos Empleados

El enfoque de esta investigación se centra en el departamento de Córdoba. Para lograrlo, en el proceso de depuración de la base de datos se seleccionaron únicamente las observaciones correspondientes a este departamento. Posteriormente, se calculó el precio promedio departamental a partir de los precios de sus municipios. En el caso del año 2020 Datos , debido a la ausencia de datos para los últimos cinco meses, estos se obtuvieron de un anexo individual del respectivo año, lo que permitió completar la serie.

El resultado final es una base depurada que sirve como insumo principal para el presente trabajo investigativo. Los datos fueron recopilados por el Departamento Administrativo Nacional de Estadística (DANE), la fuente oficial de estadísticas en Colombia. El análisis fue realizado utilizando el software estadístico R.

Análisis de las componentes de la serie-Tendencia

Evolución temporal del Precio promedio de la leche cruda en finca de enero del 2013 hasta julio del 2024, se observa un patron de tendencia en la serie

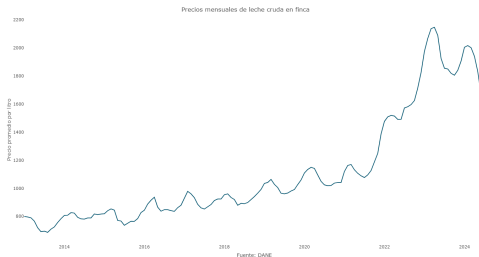
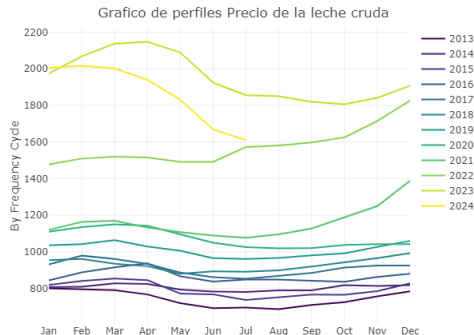


Figura 1: Serie temporal del Precio promedio de la leche curda en finca

Análisis de las componentes de la serie-Estacionalidad

Se puede apreciar un patron de estacionalidad en la serie temporal del Precio promedio de la leche cruda en finca de enero del 2013 hasta julio del 2024



Prueba de estacionariedad: KPSS Test for Level Stationarity

La prueba KPSS es adecuada para verificar la presencia de una raíz unitaria en una serie temporal, corrigiendo automáticamente posibles autocorrelaciones y heterocedasticidades en los errores.

- **Hipótesis nula (H_0):** La serie temporal es estacionaria en nivel.
- **Hipótesis alternativa (H_1):** La serie temporal no es estacionaria en nivel.

Después de realizar la prueba, obtuvimos un valor **p de 0.01**. Basándonos en este resultado, rechazamos H_0 y concluimos que la serie no es estacionaria en nivel, por lo tanto la diferenciación es necesaria.

Transformación Box- Cox.

Observamos en la Figura3 que es necesario realizar la Transformación Box-Cox ya que la varianza no es constante y apreciamos una tendencia creciente lo que significa que la media tampoco es constante

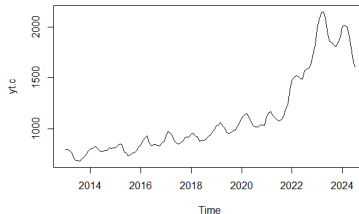


Figura 3: Gráfico de correlación temporal de la serie

Validación Cruzada y Diferenciación

La división de los datos es un paso fundamental para entrenar y evaluar el modelo. En este caso, se utilizarán los datos de entrenamiento para ajustar el modelo y verificar su desempeño. Para estabilizar la media y varianza de la serie, se aplicarán la transformación Box-Cox y técnicas de diferenciación regular y estacional, con el objetivo de eliminar tendencias y estacionalidades presentes en los datos. Esto permitirá facilitar el ajuste de un modelo eficiente para realizar pronósticos precisos.

El análisis se centrará en un horizonte de pronóstico de 12 meses. Una vez diferenciada la serie, se verificará su estacionariedad para garantizar que cumpla con los supuestos necesarios. Posteriormente, se seleccionará el modelo más adecuado para las predicciones, evaluando su desempeño en un horizonte de 12 meses, con el fin de determinar en qué caso el modelo ofrece un mejor ajuste y resultados más confiables

Metodología Box-Jenkins

Para el siguiente trabajo se utilizó la propuesta dada Box y Jenkins en 1973 que comprende un conjunto de métodos y procedimientos para la construcción de modelos de la familia **ARIMA** y sus derivados.
 La metodología Box-Jenkins reconoce las siguientes etapas para un ajuste de un modelo arima

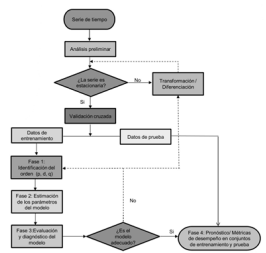


Figura 4. Metodología de box-jenkins

Modelo ARIMA

Fase 1. Indentificacion del orden del modelo ARIMA(p,d,q)

Dado que anteriormente se observo en la prueba de **KPSS** y en la Figura3 que la series no es estacionaria y presenta varianza, media no constate, aplicamos diferenciacion regular($d = 1$) y una transformacion Boxcox a nuestros datos de entranamiento utilizando las funciones ndiff y BoxCox.



Figura 5: Gráfico de la serie temporal con varianza y media estabilizada.

Identificación de p y q mediante EACF para Horizonte 12

Fase 1. Identificación del orden del modelo ARIMA(p,d,q)

Identificaremos modelos (p, q) significativos según el EACF y revisaremos si algunos relevantes fueron omitidos por estar bajo el umbral de 0.1796729 o fuera del triángulo. despues de rivisar ninguno fue omitido y si los modelo posibles fuera del triangulo son significativo.

AR/MA

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	x	x	o	x	x	x	x	x	o	o	x	x	x	o	o	x
1	o	x	o	o	o	o	o	o	o	o	o	x	x	o	o	o
2	o	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o
3	x	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o
4	x	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o
5	x	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o
6	x	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o
7	x	o	x	o	o	o	o	o	o	o	o	o	o	o	o	o
8	x	o	x	o	o	o	o	o	o	o	o	o	o	o	o	o
9	x	x	x	o	o	o	o	o	o	o	o	o	o	o	o	o
10	x	o	x	o	o	o	o	o	o	o	o	o	o	o	o	o
11	x	o	x	x	o	o	o	o	o	o	o	o	o	o	o	o
12	o	x	o	x	x	o	o	o	o	o	o	o	o	o	o	o
13	o	x	o	o	o	o	o	x	o	o	o	o	o	o	o	o
14	x	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o
15	x	o	o	o	o	o	o	x	o	o	o	o	o	o	o	o

Selección de Modelos para el Horizonte de Pronóstico

Fase 2. Estimación del modelo (p,d,q)

A continuación, se presentan los modelos candidatos seleccionados manualmente mediante la matrix de EACF para el análisis de nuestra serie.

Tabla 1: Modelos candidatos para el horizonte de pronóstico $h = 12$

Horizonte	Modelos Candidatos	Significancia
$h = 12$	ARIMA(3,1,2)	No son
	ARIMA(4,1,2)	No son

Despues de hacer la multiples combinaciones se llego que los modelos candidatos no son significativo sus coeficiente se por lo q se aplica el principio de parsimonia.

Selección de Modelos para el Horizonte de Pronóstico

Fase 2. Estimación del modelo (p,d,q)

A continuación, se presenta los modelos candidatos que fueron seleccionados mediante principio de parsimonia y que sus coeficiente todo fueran significativo.

Tabla 2: Modelos seleccionados para el horizonte de pronóstico $h = 12$

Horizonte	Modelos Seleccionados	Sum Coef AR	AIC	log likelihood
$h = 12$	ARIMA(2, 1, 2)	0.008460305	N.A	1036.09
	ARIMA(0, 1, 2)	0.8070223	N.A	1025.18

- El primer modelo fue seleccionado mediante el metodo de **CSS** el segundo con **CSS** sin .
- El criterio de seleccion fue la metricas de entranimiento, mayor log likelihood.

Estimación de los parámetros del modelo **ARIMA** para **h=12**

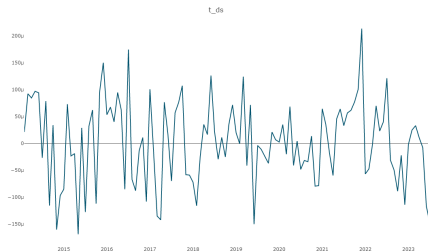
Tabla 3: Pruebas de significancia de los coeficientes para modelos

Modelo	Coefficiente	Estimación	Error Estándar	Valor-z	Pr(> z)
ARIMA(2,1,2)	ar1	1.7332	0.0134	128.872	$< 2.2e - 16$ ***
	ar2	-0.9974	0.0141	-70.801	$< 2.2e - 16$ ***
	ma1	-1.6346	0.0528	-30.943	$< 2.2e - 16$ ***
	ma2	0.9074	0.0513	17.687	$< 2.2e - 16$ ***
ARIMA(0,1,2)	ma1	0.477331	0.085213	5.6016	$< 2.124e - 08$ * **
	ma2	0.329692	0.076331	4.3192	$< 1.566e - 05$ * **

Modelo SARIMA

Fase 1. Identificación de componente regulares (p,d,q), y estacional (P,D,Q)

Dado que anteriormente se observó en la prueba de **KPSS** y en la Figura 3 que la serie no es estacionaria y presenta varianza, media no constante, aplicamos diferenciación regular ($d = 1$) y una estacional ($D = 1$) a nuestros datos de entrenamiento utilizando las funciones `ndiff` y `nsdiffs`.



Prueba de Estacionariedad: Dickey-Fuller aumentada (ADF)

Ahora utilizamos una prueba para verificar si ya la serie es estacionaria la recomendada es la **ADF**

- **Hipótesis nula (H_0):** La serie temporal no es estacionaria en nivel.
- **Hipótesis alternativa (H_1):** La serie temporal es estacionaria en nivel.

Después de realizar la prueba, obtuvimos un valor **p de 0.01**. Basándonos en este resultado, no rechazamos H_0 y concluimos que la serie es estacionaria en un nivel de significancia del 0.05.

Identificación de P y Q mediante ACF Y PACF para Horizonte 12

Fase 1. Identificación del orden AR MA no estacional (p,q) y estacional (P,Q)

Componente estacional (P, Q): Para identificar el orden estacional apropiado para un modelo SARIMA, se deben examinar los rezagos estacionales. Los picos significativos en los rezagos estacionales en los gráficos ACF y PACF pueden sugerir los órdenes P y Q .

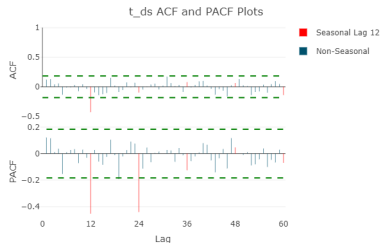


Figura 8: ACF Y PCF lag=60.

Identificación de p y q mediante EACF para Horizonte 12

Fase 1. Indentificacion del orden AR MA no estacional (p,q) y estacional (P,Q)

Componente no estacional (p, q): Para identificar el orden estacional apropiado para un modelo SARIMA, se hace de manera analoga al modelo arima usando el **EACF** de la serie diferenciada regular y estacionalmente.

AR/MA													
		0	1	2	3	4	5	6	7	8	9	10	
0		○	○	○	○	○	○	○	○	○	○	○	
1		x	○	○	○	○	○	○	○	○	○	○	
2		○	x	○	○	○	○	○	○	○	○	○	
3		x	○	x	○	○	○	○	○	○	○	○	
4		x	○	x	x	○	○	○	○	○	○	○	
5		○	○	x	x	○	○	○	○	○	○	○	
6		○	○	○	○	○	○	○	○	○	○	○	
7		x	○	○	○	○	○	○	○	○	○	○	
8		x	x	x	○	○	○	○	x	○	○	○	
9		x	x	○	○	○	○	x	○	○	○	○	
10		x	○	○	○	○	○	x	○	○	○	○	

Selección de Modelos para el Horizonte de Pronóstico

Fase 2. Estimación del modelo Sarima (p,d,q)(P,Q,D)

A continuación, se presentan los modelos candidatos seleccionados manualmente mediante la matrix de EACF para el análisis de nuestra serie.

Tabla 4: Modelos candidatos para el horizonte de pronóstico $h = 12$

Horizonte	Modelos Candidatos	Significancia
$h = 12$	SARIMA(1, 1, 1)(2, 1, 1) _[12]	No son

se observo que el modelo candidato, no son significativo sus coeficiente se por lo q se aplica el principio de parsimonia.

Selección de Modelos para el Horizonte de Pronóstico

Fase 2. Estimación del modelo (p,d,q)

Se presenta el unico modelo candidato que fue seleccionado mediante el principio de parsimonia y que sus coeficiente fueran todos significativo.

Tabla 5: Modelos seleccionados para el horizonte de pronóstico $h = 12$

Horizonte	Modelos Seleccionados	Sum Coef AR	AIC	log likelihood
$h = 12$	SARIMA(0, 1, 0)(2, 1, 1) _[12]	-1.097669	N.A	954.09

- El modelo fue seleccionado mediante el metodo de **CSS**.
- El criterio de seleccion fue la metricas de entranimiento, mayor log likelihood.

Selección de Modelo con Auto.Arima

Fase 2. Estimacion del modelo (p,d,q),(P,D,Q) auto.arima

A continuación, mediante la funcion **Auto.Arima** que nos arroja el mejor modelo segun el criterio AIC y de ello escogemos el mejor de los tres metodos posibles **ML**, **CSS-ML** ,**CSS** .

Tabla 6: Modelos seleccionados para el horizonte de pronóstico $h = 12$

Horizonte	Modelos Seleccionados	Sum Coef AR	AIC	log likelihood
$h = 12$	SARIMA(0, 1, 0)(0, 1, 1) _[12]	-1882.65	0	940.19

- El modelo fue seleccionado mediante el metodo de **CSS**.
- El criterio de seleccion fue la metricas de entranimiento, tuviera las menores metricas.

Estimación de los parámetros del modelo **SARIMA** para **h=12**

Tabla 7: Pruebas de significancia de los coeficientes para los modelos

Modelo	Coefficiente	Estimación	Error Estándar	Valor-z	Pr(> z)
SARIMA(0,1,0)(2,1,1)_[12]	sar1	-0.479215	0.108430	-4.4196	9.889e-06 ***
	sar2	-0.351642	0.087307	-4.0277	5.633e-05 ***
	sma1	-0.266812	0.128019	-2.0842	0.03715 *

Estimación de los parámetros del modelo de la funcion **Auto.Arima** para **h=12**

Tabla 8: Pruebas de significancia de los coeficientes para el modelo con la Funcion **Auto.Arima**

Modelo	Coefficiente	Estimación	Error Estándar	Valor-z	Pr(> z)
SARIMA(0,1,0)(0,1,1)_[12]	sma1	-0.802609	0.098476	-8.1503	3.63e-16 ***

Fase 3: Evaluación y diagnóstico de los modelos

Prueba de hipótesis sobre la media de los residuos: One Sample t-test

Prueba de Hipótesis

- **Hipótesis nula (H_0):** La media de los residuos es igual a cero ($\mu = 0$).
- **Hipótesis alternativa (H_1):** La media de los residuos es diferente de cero ($\mu \neq 0$).

Tabla 9: P-valores de la prueba t para los modelos ARIMA-SARIMA

Modelos ARIMA-SARIMA	P-valor
ARIMA(2,1,2)	0.006548
ARIMA(0,1,2)	0.2004
SARIMA(0,1,0)(2,1,1) _[12]	0.1406
Auto.ARIMA(0,1,1)(0,1,0) _[12]	0.6375

Evaluación y diagnóstico de los modelos

Prueba de Normalidad de los Residuos: Kolmogorov-Smirnov test

Prueba de Hipótesis

- **Hipótesis nula (H_0):** Los residuos siguen una distribución normal.
- **Hipótesis alternativa (H_1):** Los residuos no siguen una distribución normal.

Tabla 10: P-valores de la prueba de Kolmogorov-Smirnov para normalidad de los residuos

Modelo ARIMA-SARIMA	P-valor
ARIMA(2,1,2)	2.2e-16
ARIMA(0,1,2)	2.2e-16
SARIMA(0,1,0)(2,1,1) _[12]	2.2e-16
Auto.ARIMA(0,1,0)(0,1,1) _[12]	2.2e-16

Evaluación y diagnóstico de los modelos

Independencia de los Residuos: Ljung-Box

Prueba de Hipótesis

- **Hipótesis nula (H_0):** No hay autocorrelación en los residuos hasta el lag especificado.
- **Hipótesis alternativa (H_1):** Existe autocorrelación en los residuos hasta el lag especificado.

Tabla 11: P-valores de la prueba de Box-Ljung Test

Modelo ARIMA-SARIMA	P-valor
ARIMA(2,1,2)	0.2293
ARIMA(0,1,2)	0.1129
SARIMA(0,1,0)(2,1,1) _[12]	0.3137
Auto.ARIMA(0,1,0)(0,1,1) _[12]	0.6421

Evaluación y diagnóstico de los modelos

Prueba sobre la homocedasticidad de los residuos: ARCH LM-test

Prueba de Hipótesis

- **Hipótesis nula (H_0):** Los residuos son homocedásticos (la varianza de los residuos es constante).
- **Hipótesis alternativa (H_1):** Los residuos son heterocedásticos.

Tabla 12: P-valores de la prueba ARCH para los modelos ARIMA-SARIMA

Modelo ARIMA-SARIMA	P-valor
ARIMA(2,1,2)	0.2879
ARIMA(0,1,2)	0.01826
SARIMA(0,1,0)(2,1,1) _[12]	0.08888
Auto.ARIMA(0,1,0)(0,1,1) _[12]	1

Evaluación y diagnóstico de los modelos ARIMA

Dependencias No Lineales: Prueba BDS

Prueba de Hipótesis

- **Hipótesis nula (H_0):** Los residuos de la serie temporal siguen un proceso *i.i.d.*
- **Hipótesis alternativa (H_1):** Los residuos de la serie temporal *no* siguen un proceso *i.i.d.*

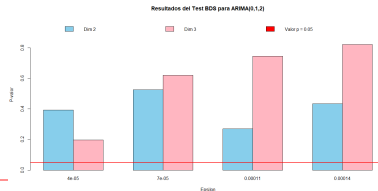
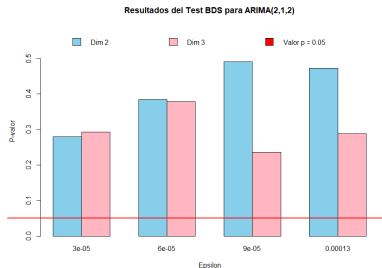


Figura 11: P-valores del Test BDS

Evaluación y diagnóstico de los modelos

Dependencias No Lineales: Prueba BDS

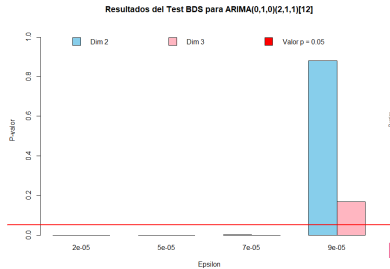


Figura 12: P-valores del Test BDS

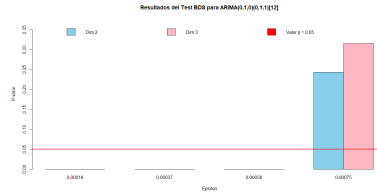


Figura 13: P-valores del Test BDS: autoARIMA

Selección del mejor modelo manual

Para la selección de los modelos finales de la parte de **ARIMA** se tuvieron en cuenta, los modelos que tuvieran las menores métricas de entrenamiento y que tuvieran el mayor log likelihood dado que los dos no cumplen algunos de los supuestos.

- Para el Modelo ARIMA)
el modelo escogido fue: **ARIMA** (2,1,2).
- Para el Modelo SARIMA
el modelo escogido fue: **SARIMA** (0,1,0)(2,1,1)_[12].

Evaluación del desempeño del modelo seleccionado: ARIMA, SARIMA y autoARIMA

Tabla 13: Métricas de desempeño del modelo **ARIMA (2,1,2)**

Conjunto de datos	ME	RMSE	MAE	MPE	MAPE	Theil's U
Entrenamiento	9.3347	29.0628	20.7717	0.6767	1.9124	0.7453
Prueba	-125.1864	157.6333	134.3641	-6.8804	7.3786	2.1960

Tabla 14: Métricas de desempeño del modelo **SARIMA (0,1,0)(2,1,1)₁₂**

Conjunto de datos	ME	RMSE	MAE	MPE	MAPE	Theil's U
Entrenamiento	2.3085	23.7355	14.6951	0.1413	1.2466	0.5789
Prueba	-399.2031	466.0183	399.2031	-21.7070	21.7070	6.4291

Tabla 15: Métricas de desempeño del modelo **autoARIMA (0,1,0)(0,1,1)₁₂**

Conjunto de datos	ME	RMSE	MAE	MPE	MAPE	Theil's U
Entrenamiento	6.0040	35.4710	23.2406	0.5148	2.2611	0.9845
Prueba	-429.9347	483.0215	429.9347	-23.2373	23.2373	6.6198

Conclusión

teniendo en cuenta estas metricas de desempeño se puede concluir que el modelo tiene un desempeño moderadamente bueno en la predicción del precio promedio de la leche cruda de granja en Córdoba, Colombia. En el conjunto de entrenamiento, el modelo mostró un buen ajuste con un MAPE del 1,91 , lo que refleja su capacidad para capturar patrones históricos lineales. Sin embargo, limitaciones como el sesgo de subestimación (ME negativo) y la U de Theil mayor que 2 son evidentes en el conjunto de prueba, lo que indica que la eficiencia de predicción a corto plazo de este modelo es menor que la de la predicción ingenua.



Estos resultados en enfatiza que si bien los modelos ARIMA pueden usarse para identificar tendencias y patrones lineales en series de tiempo, su capacidad para manejar dinámicas no lineales complejas se ve truncada. como se pudo obeservar en la literatura del articulo del estado del arte, estos resultados dan en la necesidad de introducir métodos más avanzados, como modelos híbridos o métodos basados en redes neuronales, que puedan capturar patrones ocultos y modelar el impacto de factores externos. En general,Estos hallazgos representan un primer paso valioso para comprender la dinámica de los precios de la leche cruda en la región de cordoba , proporcionando información crítica para los productores lecheros de la region, comunicar a los entes, para asi implementar políticas efectivas y para investigaciones futuras destinadas a optimizar la previsibilidad y confiabilidad del modelo.

Bibliografía

[1] Johnson, A. H. (1974). *The composition of milk*. En *Fundamentals of dairy chemistry* (pp. 1–57). The AVI Publishing Company Inc., Westport, CT.

[2] Departamento Administrativo Nacional de Estadística (DANE) (2023). *Promedio de leche cruda en finca: Serie histórica 2013 a 2024*. Consultado el 3 octubre, Disponible en: <https://www.dane.gov.co/index.php/estadisticas-por-tema/agropecuario/sistema-de-informacion-de-precios-sipsa/boletin-mensual-precios-de-leche-cruda-en-finca>Link

[3] Departamento Administrativo Nacional de Estadística (DANE) (2023). *Promedio de leche cruda en finca: Anexo histórico - Boletín mensual para completar datos 2020*. Consultado el 3 octubre, Disponible en: <https://www.dane.gov.co/index.php/estadisticas-por-tema/agropecuario/sistema-de-informacion-de-precios-sipsa/mayoristas-boletin-mensual-1/boletin-mensual-precios-de-leche-en-finca-historicos>Link