FORECASTING THE PRICE OF REFINED SUGAR IN THE PHILIPPINES USING ARIMAX MODEL

Alanan, Baldeo, Hilario, Rivera, Santos

Bulacan State University
BS Mathematics with Specialization in Computer Science

May 2025

Introduction

Sugar plays a vital role in the Philippine economy, serving as a key ingredient in household consumption and food manufacturing, while also supporting over 1.5 million jobs and contributing 2% to the national GDP. Most sugarcane is produced in Negros Occidental, the country's "Sugar Bowl." However, the industry faces challenges including climate disruptions, declining productivity, and rising demand. These factors, along with typhoons and international price volatility, have led to production shortfalls and increased reliance on imports. As refined sugar prices continue to rise, this study aims to forecast future price movements using mathematical modeling that incorporates external factors to inform decisions in agriculture, industry, and policy-making.

Statement of the Problem

The price of refined sugar is influenced by various factors such as climate conditions, sugar production, government policies, and global market trends. Changes in sugar prices have a direct impact on Filipino households, food manufacturers, and businesses that rely on sugar as a primary ingredient. Despite its economic and social significance, there is limited research on accurately forecasting sugar prices using mathematical models that incorporate both historical price data and external variables to forecast future price movements.

Statement of the Problem

This study aims to develop a forecasting model for the prices of refined sugar in the Philippines using the AutoRegressive Integrated Moving Average with Exogenous Variables (ARIMAX) model. The model integrates external factors such as production and withdrawals of refined sugar, global sugar prices, exchange rate from USD to PHP, temperature, total precipitation, and inflation to improve the reliability of forecast.

This study seeks to answer:

1. Which external factors influence sugar prices and how strongly?
2. Can ARIMAX provide accurate forecasts?
3. What are the projected sugar prices for early 2025 and through 2027?

Research Objectives

- Identify external factors influencing refined sugar prices.
- Evaluate ARIMAX model for forecasting accuracy.
- Forecast short- and long-term prices up to 2027.

Scope and Delimitations

- Monthly refined sugar prices in the Philippines from 2014 to 2024.
- Uses ARIMAX model with exogenous variables: climate (Negros Region), sugar production, global prices, exchange rate, inflation, and withdrawals.
- Focus is only on refined sugar; excludes washed, brown sugar, and molasses.
- Climate data limited to Negros Region, the largest sugar-producing area.
- Import data excluded due to inconsistent monthly availability; withdrawals used to proxy demand.
- Does not cover sugar-based product prices (e.g., pastries, beverages).

Preliminaries

This section introduces key concepts used in the modeling process of forecasting, particularly relevant to time series analysis.

- Time Series Analysis
- ARIMAX Model
- Stationarity (ADF Test)
- ACF and PACF
- Ljung-Box Test
- VIF
- AIC and BIC

Time Series Analysis

- A time series is a sequence of data points recorded at uniform time intervals.
- Analysis involves identifying patterns like trends and seasonality, and forecasting.
- Denoted as $\{y_t\}$, where each y_t is associated with time t.
- Observations are usually ordered chronologically.

ARIMAX Model

- Extension of ARIMA that incorporates exogenous variables.
- Requires the time series to be **stationary** (constant mean and variance).
- ARIMA combines:
  - **AR** (AutoRegressive): The value of the current observation is affected by the previous observations.
  - **I** (Integrated): Differencing to achieve stationarity
  - **MA** (Moving Average):It captures the errors from previous observations and how it affects the current observation.
- ARIMAX includes external factors to enhance forecasting.

ADF Test for Stationarity

- Tests if a time series has a unit root (non-stationary).
- Null Hypothesis H_0: Series has a unit root.
- If p-value < 0.05, reject $H_0 \Rightarrow$ series is stationary.
- Differencing can be applied to make non-stationary series stationary.

ACF and PACF

- **Autocorrelation Function**:
  Correlation between series and its lagged values.
- **Partial Autocorrelation Function**:
  Correlation with lagged values removing shorter lags.
- **Interpretation Tips:**
  - If ACF tails off and PACF cuts off $\rightarrow$ AR model.
  - If PACF tails off and ACF cuts off $\rightarrow$ MA model.
  - If both tail off gradually $\rightarrow$ consider ARMA or ARIMA.

Ljung-Box Statistics Test

Test whether residuals from a time series model exhibit autocorrelation (i.e., differ from white noise).

Hypotheses:

- H_0: Residuals are independently distributed (white noise)
- H_a: Residuals are autocorrelated (model has lack of fit)

Decision Rule:

- If **p-value** $>$ **0.05**: Fail to reject $H_0 \rightarrow$ residuals resemble **white noise**
- If **p-value** $\leq$ **0.05**: Reject $H_0 \rightarrow$ residuals exhibit **autocorrelation**

A model with white noise residuals is considered well-fitted for forecasting.

Variance Inflation Factor (VIF)

- Detects multicollinearity among exogenous variables.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- R_i^2: Coefficient of determination for i-th variable regressed on others.
- VIF > 5 (or 10): Potential multicollinearity issue.

Model Selection: AIC and BIC

Compare multiple models to identify the best-fitting model with the right balance of fit and complexity.

Akaike Information Criterion (AIC):

- Measures goodness of fit while penalizing model complexity.
- Lower AIC $\rightarrow$ Better model.

Bayesian Information Criterion (BIC):

- Similar to AIC, but imposes a stronger penalty for complexity.
- Tends to prefer simpler models compared to AIC.

Both criteria aim to select models with good predictive performance, but **BIC penalizes complexity more heavily**.

Methodology

Methodology

- This study will employ a quantitative research design using time series analysis to forecast the price of refined sugar in the Philippines.
- The main forecasting technique to be used is the Autoregressive Integrated Moving Average with Exogenous Variables (ARIMAX) model.

Datasets

The dataset used in this study consists of monthly time series data from September 2014 to August 2024.

Target variable: Price of refined sugar in the Philippines
Exogenous variables:

- Climate data (temperature, precipitation) in Negros Region.
- Production and Withdrawals of refined sugar.
- Global Price of sugar.
- Exchange rate from PHP to USD.
- Inflation rate.

Procedures

Analytical Procedure

The analytical procedure for the model development and validation is done in Python using various libraries such as `pandas`, `statsmodels`, `sklearn`, `pmdarima`, `matplotlib`.

Data Preparation

- Making sure all of the data are aligned in monthly intervals and handling the missing data points.
- The missing data points are filled using the `forward_fill()` function from `pandas` library in Python.
- Split the dataset into training and testing sets

Stationarity Test

- The Augmented Dickey-Fuller (ADF) test is conducted on each variable to test for stationarity.
- The `adfuller()` function from the `statsmodels` library in Python is used to perform the test.
- If the variable is not stationary (p-value > 0.05), perform differencing to eliminate the presence of trend or seasonality.
- If it is still non-stationary, differentiate it again until it achieves stationarity.

Model Identification

- Finding the appropriate order of ARIMAX(p, d, q) parameters through the plot of Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).
- The AR components are determined using the PACF plot, where lags with significant spikes are the possible order for the AR component.
- The MA components are determined using the ACF plot, where lags with significant spikes are the possible order for the MA component.
- The Integrated I component is determined based on the order of differencing required to achieve stationarity

Model Estimation and Selection

- All possible ARIMAX models are fitted using the training set and evaluated for accuracy with the test set.
- Each model is assessed based on mean absolute percentage error (MAPE) to measure forecasting accuracy.
- Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used to penalize model complexity.
- The best ARIMAX model is selected based on the lowest values of AIC, BIC, and MAPE.

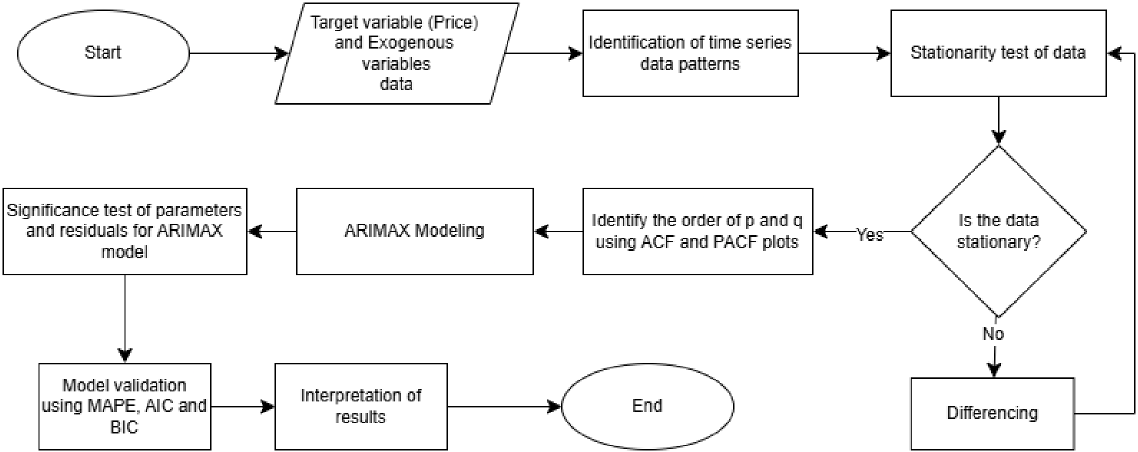
Model Diagnostics and Interpretation

- The residuals of the selected model are examined using the Ljung-Box test to ensure that the residuals resemble white noise.
- The coefficients of the variables are interpreted based on their significance (p-values).
- Variance Inflation Factor (VIF) may be calculated to assess multicollinearity among exogenous variables

Forecasting

- Once the best-performing ARIMAX model is selected, it is used to forecast the monthly retail price of refined sugar in the Philippines.
- For forecasting each exogenous variable, the ARIMA model is used along with the `auto_arima()` function from the `pmdarima` library in Python to find the order of parameters.
- Forecasts are generated for the short-term (first quarter of 2025) and long-term period (from 2025 to 2027).

Flowchart



Start
Target variable (Price) and Exogenous variables data
Identification of time series data patterns
Stationarity test of data
Is the data stationary?
Yes
No
Differencing
Identify the order of p and q using ACF and PACF plots
ARIMAX Modeling
Significance test of parameters and residuals for ARIMAX model
Model validation using MAPE, AIC and BIC
Interpretation of results
End

Error Metrics

- **Mean Absolute Error (MAE)**: The average of the absolute differences between the predicted and actual values. It provides a measure of how close predictions are to the actual outcomes.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

- **Mean Absolute Percentage Error (MAPE)**: The average of the absolute percentage errors between the predicted and actual values. It is useful for understanding the accuracy of the model in relative terms.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100$$

Error Metrics (cont.)

- **Root Mean Square Error (RMSE)**: The square root of the average of the squared differences between predicted and actual values. It gives more weight to larger errors and is sensitive to outliers.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

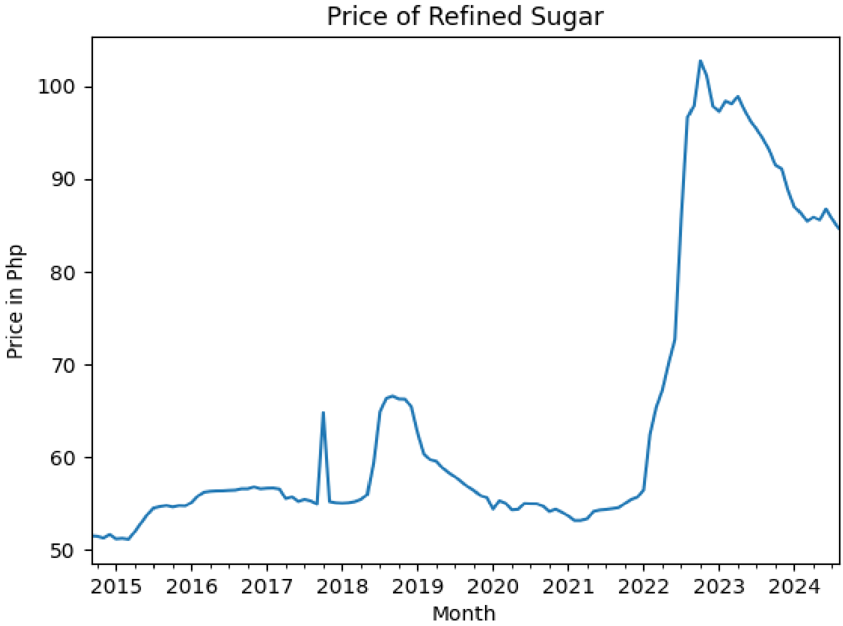
Among these, MAPE is prioritized as the primary evaluation metric because of its interpretability in comparing the forecast accuracy.

MAPE Forecast Accuracy Rating

Table: *MAPE Forecast Accuracy Rating*

MAPE (%)	Forecast Accuracy
$< 10\%$	Highly accurate forecast
10% to $< 20\%$	Good forecast
20% to $< 50\%$	Reasonable forecast
$\geq 50\%$	Inaccurate forecast

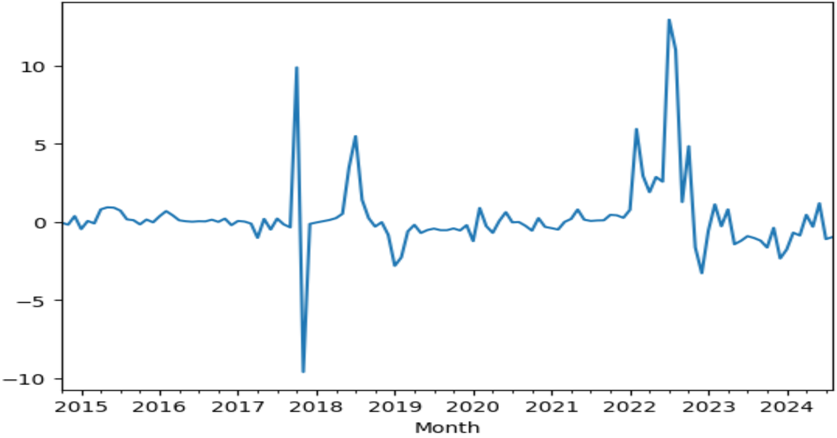
Results and Discussion



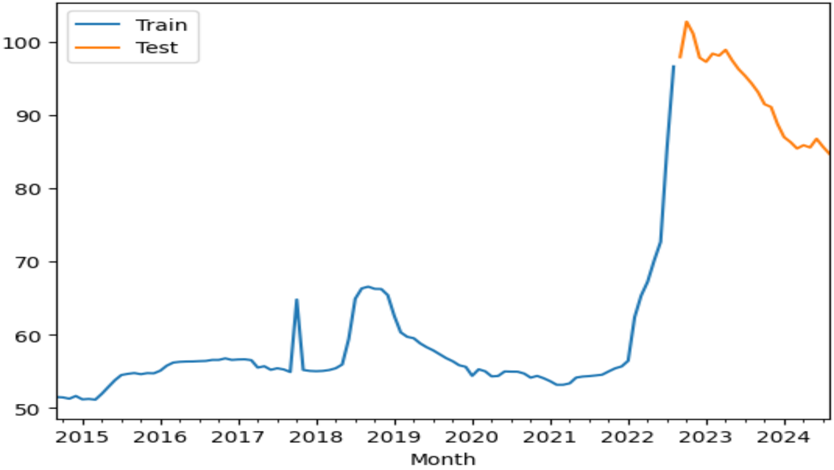
Price of Refined Sugar
100
90
80
70
60
50
Price in Php
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Month

Test for Stationarity

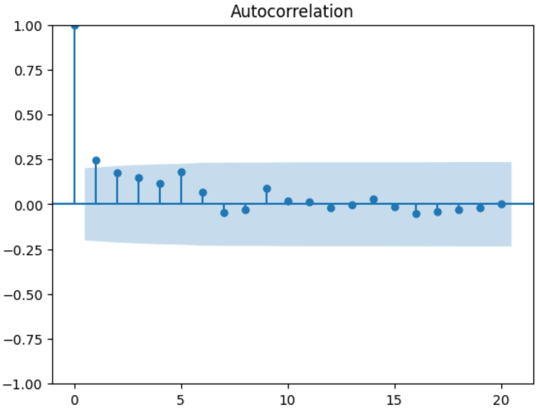
- Performing the Augmented Dicky-Fuller test on the data shows a P-value of 0.664.
- Performing first-order differencing results in p-value of 0.0000057.



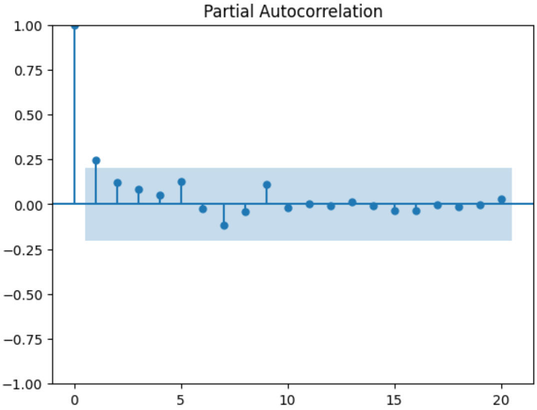
Train-Test split



Order of p and q



$q = [1]$



$p = [1]$

Selected Model

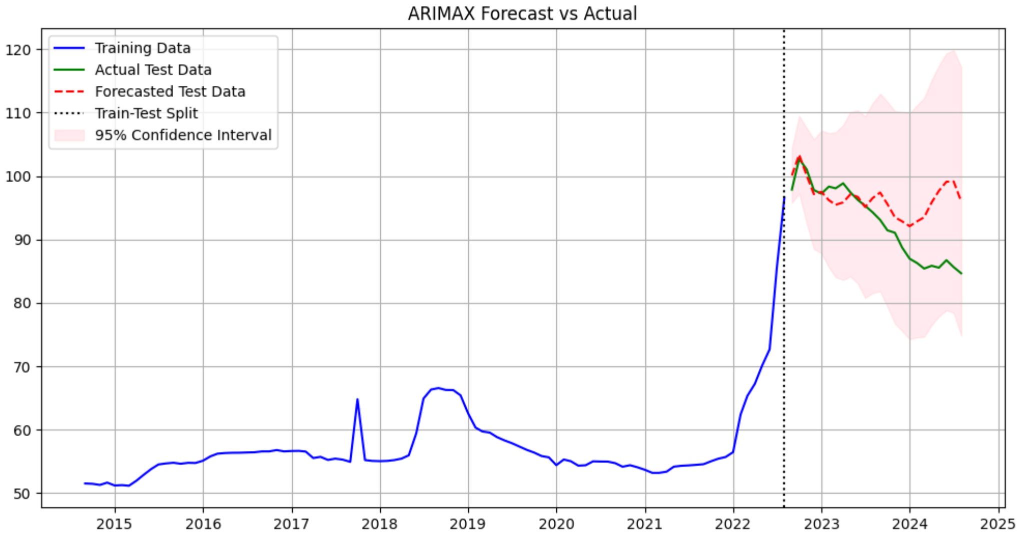
ARIMAX(1,1,1)

Model evaluation

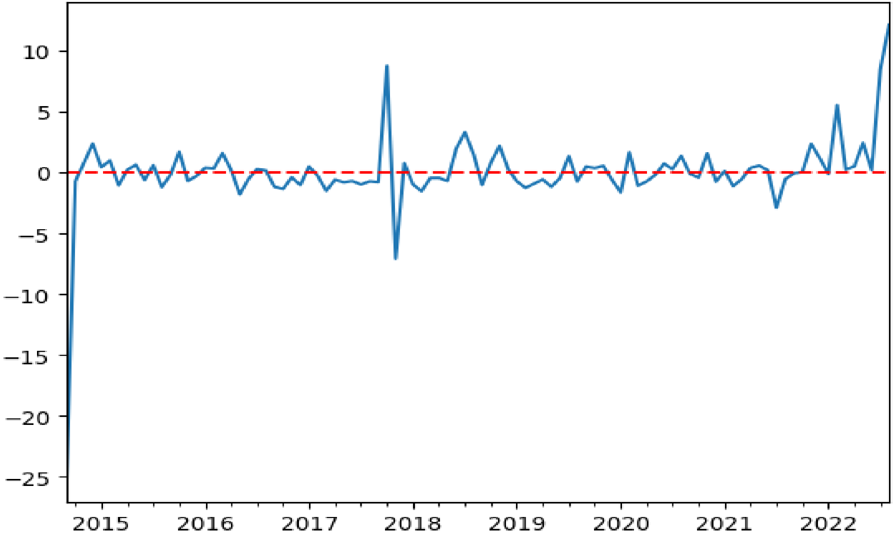
Table: *Performance Metrics of ARIMAX(1,1,1) Model*

Model	MAE	MAPE	RMSE	AIC	BIC
ARIMAX(1,1,1)	4.58	5.21%	6.25	443.81	469.35

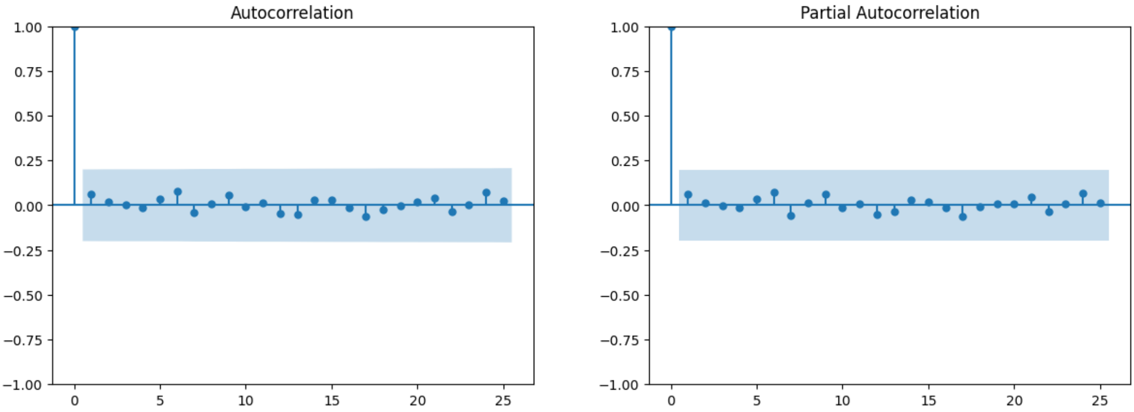
Model evaluation



Residual diagnostics



Residual diagnostics



Residual diagnostics

Ljung-Box test

Lag	p-value	Lag	p-value
1	0.524	7	0.985
2	0.805	8	0.994
3	0.933	9	0.995
4	0.979	10	0.998
5	0.990	11	0.999
6	0.977	12	0.999

The p-value for all 12 lags are higher than 0.05. Hence, the residuals shows no autocorrelation and it resembles white noise, increasing the validity of the model.

Test for Multicollinearity

Variance Inflation Factor

Variable	VIF
Production	2.99
Withdrawals	1.77
Global Price	1.44
Exchange Rate	2.53
Temperature	1.2
Precipitation	2.56
Inflation	2.23

ARIMAX(1,1,1) Model Summary

Variable	Coefficient	P-value	Interpretation
Production	-2.006e-07	0.832	Not significant
Withdrawals	2.414e-07	0.811	Not significant
Global Price	-10.4240	0.000	Highly significant
Exchange Rate	1.4639	0.000	Highly significant
Temperature	0.2915	0.521	Not significant
Precipitation	0.0041	0.186	Somewhat significant
Inflation	0.8710	0.382	Not significant
AR(1), MA(1)	0	1.000	Not contributing to the model

ARIMAX(1,1,1) Model Summary

The results of the summary shows that the following must be consider:

1. The coefficients for AR(1) and MA(1) are both zero and not contributing to the model.
2. Some of the variables are not significant based on their p-value, leaving Global Price and exchange rate with high significance.
3. The coefficient of the Global Price is negative, which is counterintuitive.

1. The coefficients for AR(1) and MA(1) are both zero and not contributing to the model

Table: *Performance Metrics of ARIMAX(1,1,1) vs ARIMAX(0,1,0)*

Model	MAE	MAPE	RMSE	AIC	BIC
ARIMAX(1,1,1)	4.58	5.21%	6.25	443.81	469.35
ARIMAX(0,1,0)	4.58	5.21%	6.25	778.34	798.77

2. Some of the variables are not significant based on their p-value, leaving Global Price and exchange rate with high significance.

Table: *Performance Metrics of ARIMAX(1,1,1) with all exogenous variable vs only the significant variables*

ARIMAX(1,1,1)	MAE	MAPE	RMSE	AIC	BIC
With all Exogenous variables	4.58	5.21%	6.25	443.81	469.35
Only the significant variables	17.32	18.51%	17.72	668.23	688.66

3. The coefficient of the Global Price is negative, which is counterintuitive.

Table: *Performance Metrics of ARIMAX(1,1,1) with global price vs no global price variable*

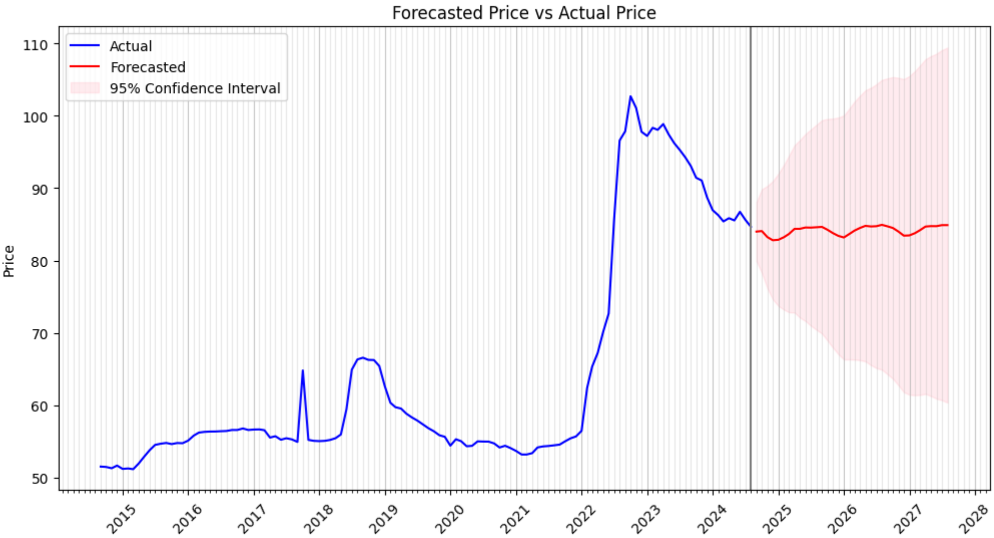
ARIMAX(1,1,1)	MAE	MAPE	RMSE	AIC	BIC
With global price	4.58	5.21%	6.25	443.81	469.35
No global price	4.95	5.62%	6.60	442.03	465.02

Key Takeaways from Modeling process

- Residuals resemble white noise (Ljung-Box Test)
- No multicollinearity (checked via VIF)
- Global price variable excluded to improve interpretability

Forecasting

Forecasting the Price of Refined sugar for 2025 to 2027



Forecast Result Interpretation

- The 3-year forecast shows a relatively stable trend with mild fluctuations and a slight upward tendency over time.
- The model is 95 percent confident that the actual price will lie within the confidence interval.
- The forecasted price of refined sugar was ranging from 82.80 to 84.93 pesos per kilogram.

Forecasted Prices vs the Latest Actual Prices

Retrieved from Department of Agriculture price monitoring (*Bantay Presyo*)

Date	Actual Price	Forecasted Price
September 2024	84	83.99
October 2024	83.15	84.07
November 2024	83.13	83.23
December 2024	82.83	82.80
January 2025	83.65	82.86
February 2025	83.07	83.22
March 2025	83.14	83.67
April 2025	83.49	83.37

Forecasted Prices vs the Latest Actual Prices

Table: *Evaluation of the Forecasted price*

MAE	MAPE	RMSE
0.43	0.51%	0.56

Conclusion

- Developed an ARIMAX(1,1,1) model to forecast monthly retail prices of refined sugar in the Philippines.
- Integrated key exogenous variables: production, withdrawals, exchange rate, regional temperature, precipitation, and inflation rate.
- Achieved high predictive accuracy:
  - MAPE: 5.62% (test set), 0.51% (short-term validation).
  - MAE: 0.43, RMSE: 0.56 (short-term validation).
- Global price and exchange rate were statistically significant, but excluding global price improved model performance.
- ARIMAX(1,1,1) outperformed other variants with the lowest AIC and BIC.
- Three-year forecast shows stable prices with mild fluctuations and a slight upward trend (82.80 to 84.93 pesos/kg).

Recommendations

- **Larger datasets:** Use longer historical data (e.g., 20 years) and include additional exogenous variables like transportation costs, energy prices, and labor rates.
- **Alternative models:** Explore models like SARIMAX, VARMA, VARMAX, or machine learning approaches (e.g., Random Forest, XGBoost, LSTM) for better performance.
- **Statistical interpretation:** Investigate counterintuitive results, such as the negative coefficient for global sugar prices, to improve understanding.

Questions?