Forecasting S & P 500 Using SARIMAX

Matthew Chen, William Gipson,Ann Kim, Liangyong Wang

Goals & Objectives

Project Goals and Objectives

Goals:

❏ Forecast daily S & P 500 prices using SARIMAX with exogenous variables

Objectives:

❏ Incorporate relevant exogenous variables

❏ Apply proper stationarity transformation and dimensionality reduction

❏ Use rolling-origin backtesting to evaluate model accuracy

Background

What is the S & P 500?

❑ List of 500 largest U.S. companies

❑ Used to measure stock market health

❑ Helps track U.S. economic trends

❑ Includes Apple, Microsoft, and others

❑ Reacts to global news and events

SARIMAX: Seasonal ARIMA with Exogenous Variables

Model Structure: SARIMA (p, d, q) x (P, D, Q)s + X

- ❏ **AR (p)**: Autoregressive terms
- ❏ **I (d)**: Differencing
- ❏ **MA (q)**: Moving average terms
- ❏ **Seasonal (P, D, Q)**: Seasonal AR, differencing, and MA terms with period s
- ❏ **X**: Exogenous variables

When to Use SARIMAX:

- ❏ **trend and/or seasonality are detected**
- ❏ **External factors are detected**

Common SARIMAX Applications:

- ❏ **Forecasting retail sales**
- ❏ **Modeling electricity demand**

Why Use SARIMAX for Forecasting the S&P 500?

❑ Forecasting its movement is important for investors and policymakers.

❑ Traditional ARIMA models do not account for external factors.

❑ SARIMAX allows the use of exogenous variables.

Examples Used in Our Project

Summary of Reference Paper (Erlemann et al., 2025)

Title: SARIMAX-Based Framework for S&P 500 Forecasting: Incorporating Economics Indicators

Published: May 2025 (Preprint on ResearchGate)

❑ S&P 500 is inherently non-stationary

❑ Acknowledge technological impact to the economy

Why We Use This Paper

Relevance to Our Project:

❑ Same target: S&P 500 daily forecasting

❑ Same model: SARIMAX with external variables

Benefits

❑ Clear, tested pipeline to follow

❑ Benchmarks we can compare against

❑ Real-world exogenous features + strong validation method

How Our Work Goes Beyond the Paper

- ❑ Parallel Processing
- ❑ Residual diagnostic
- ❑ Kalman smoothing
- ❑ Mean Absolute Percentage Error (MAPE)
- ❑ Forecast visualization

Data Components & Modeling

Exogenous Variables Used in SARIMAX Model

Category	**Exogenous Variable Included**
Financial Indicators	Gold, Crude Oil, Copper, Bitcoin, VIX, S&P GSCI
Treasury Bond & Currency Exchange	Treasury Yields (3M, 10Yr), forex(foreign currency exchange): EUR — USD, GBP—USD, JPY —USD, AUD—USD, CAD—USD
Stock Indices	SSE(Shanghai), STOXX 600 (Europe), MOEX (Moscow)
Google Trends	“SP500”, “ETF”, “Index Fund”, “SPX”
Unemployment	United States Unemployment Rate

Target Variable Used in SARIMAX Model

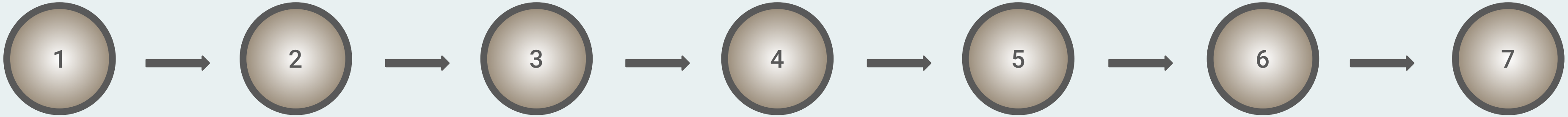
❑ Target Variable: daily closing value of the S&P 500 index

❑ Compare Predicted vs. Actual time series models of the target variable

❑ GOAL: small differences between the predicted and actual time series model

❑ Accuracy Analysis Measure: R^2, MAE, RMSE, MAPE

SARIMAX Pipeline



Data Extraction

Data Processing

ADF Stationarity Test + Differencing

PCA

SARIMAX + AIC Model Selection

Forecasting + Residual Calculation

Accuracy Analysis

Handling Missing data & Imputation

Why it mattered

- ❑ Multiple macro-financial series came from different sources → uneven date coverage & sporadic gaps
- ❑ Stationary models (ADF + SARIMAX) need complete, numeric inputs

Imputation strategy

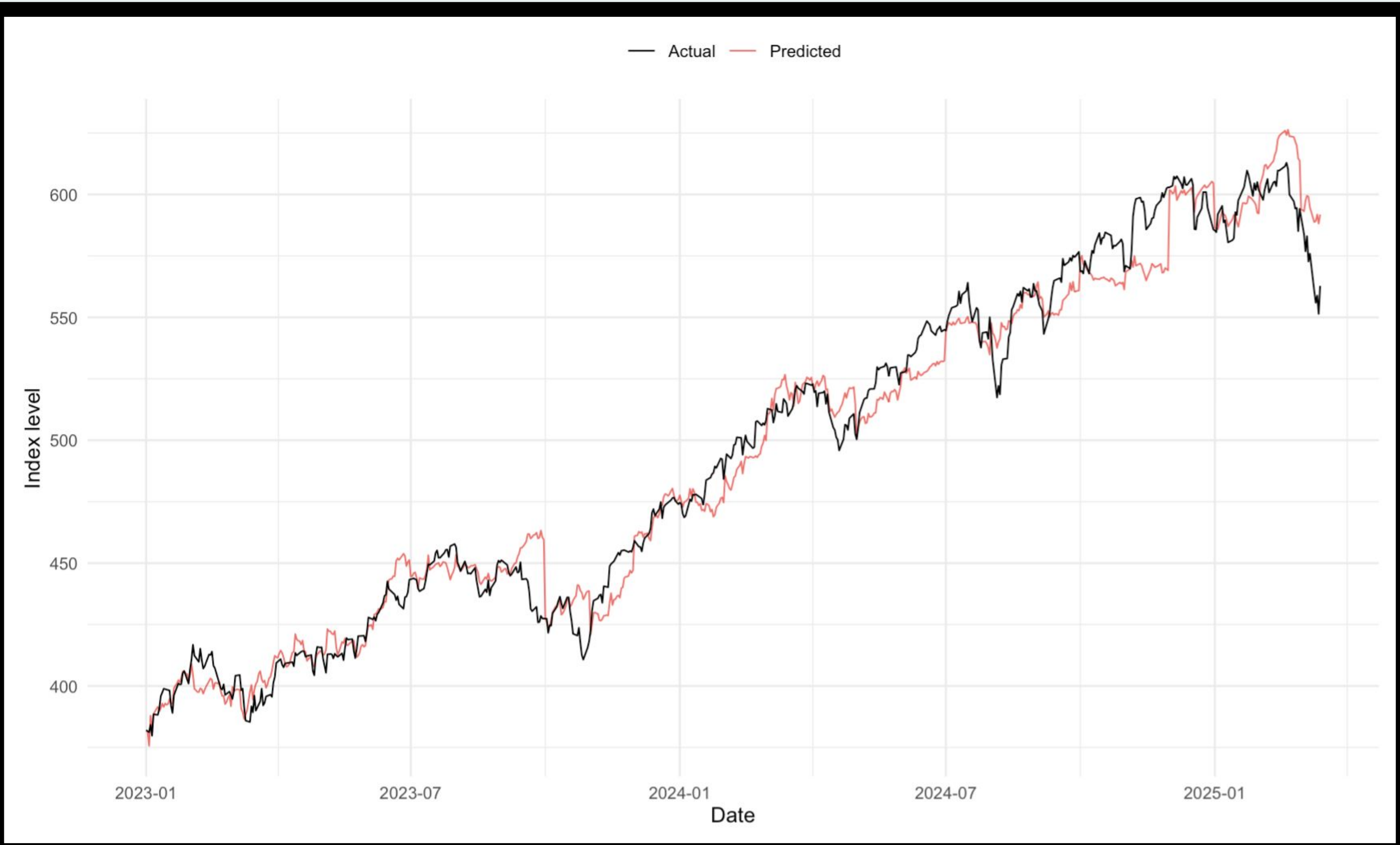
- ❑ **Primary fill**:
  - ❑ Kalman smoother built on an automatically-selected ARIMA for each series
  - ❑ Reconstructs values consistent with each series' own dynamics
- ❑ **Edge repair**:
  - ❑ **Forward LOCF** – pushes last known value forward
  - ❑ **Backward LOCF** – back-fills at the very start
  - ❑ Guarantees *no leading or trailing NA* in any regressor

Findings from Exploratory Data Analysis

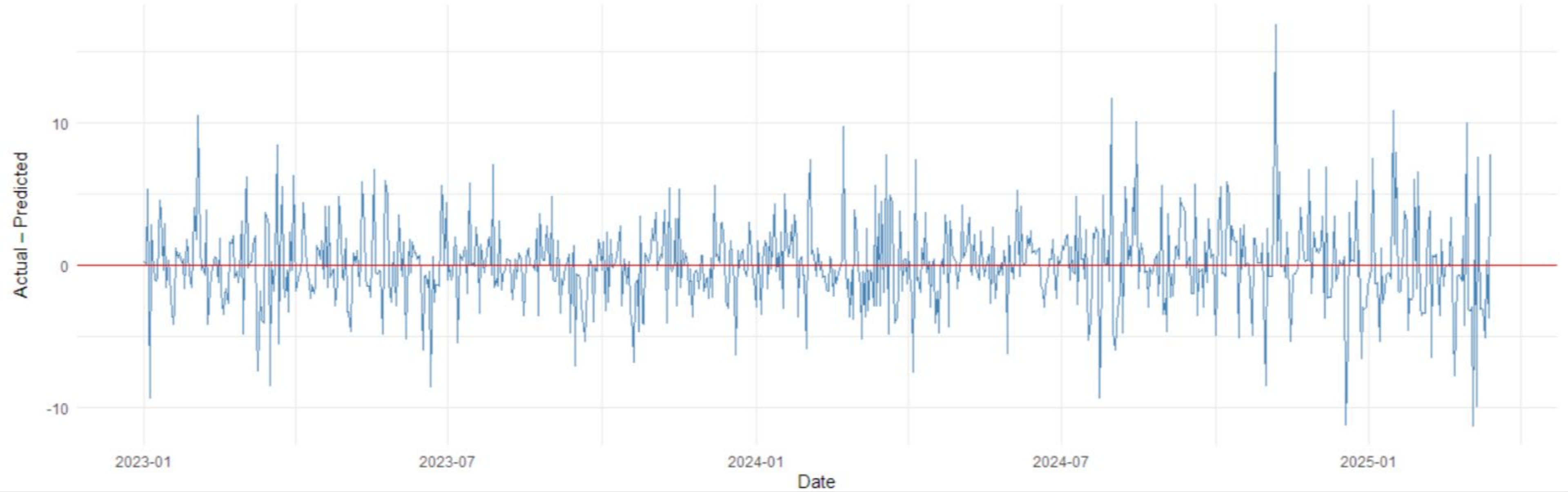
- ❑ **Preprocessing**:
    - ❑ Merged all datasets by date
    - ❑ Removed all-NA or constant columns
    - ❑ Imputed missing values using Kalman smoothing and LOCF
- ❑ **Stationarity Check**:
    - ❑ Applied ADF tests on all series
    - ❑ Differenced non-stationary training time series
- ❑ **Dimensionality Reduction**:
    - ❑ Used PCA to reduce the number of exogenous features while retaining 95% of the variance
- ❑ **Initial Observations**:
    - ❑ Target variable shows an increasing general trend

Results

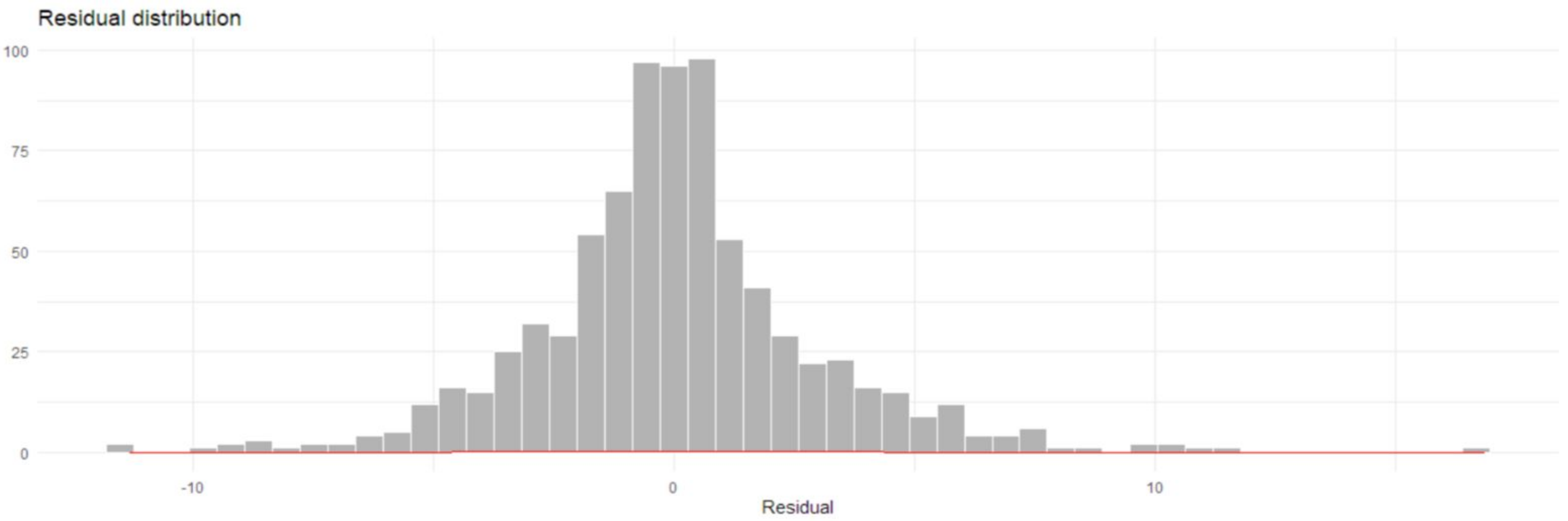
S & P 500: Actual vs SARIMAX



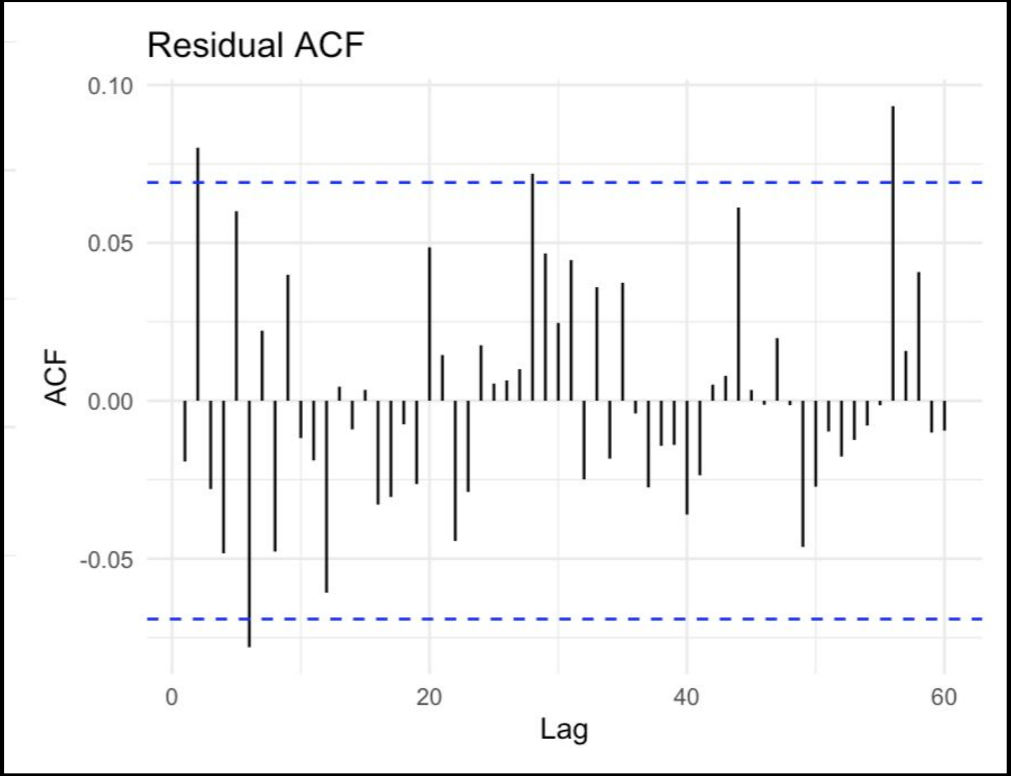
SARIMAX Model Residual Diagnostic Checks: Residuals Over Time



SARIMAX Model Residual Diagnostic Checks: Residual Distribution



SARIMAX Model Residual Diagnostic Checks: ACF Plot



SARIMAX Performance Summary

- ❑ **What is the S&P 500 Index?**
  - ○ A score that shows how 500 big U.S. companies are doing overall
  - ○ Changes every second during market hours (Mon–Fri, 9:30am–4:00pm EST)
  - ○ Today’s values are around 4,000–5,500
- ❑ **MAE (Mean Absolute Error):**
  - ○ Shows average mistake size in points
  - ○ MAE = 8.375 → Forecast is off by about 8 index points on average
- ❑ **RMSE (Root Mean Squared Error):**
  - ○ Like MAE but gives more weight to bigger mistakes
  - ○ RMSE = 11.084 → Slightly higher average error when larger mistakes are penalized more
- ❑ **R^2 (R-squared)**
  - ○ Proportion of variance explained by the model
  - ○ R^2 = 0.974 → Model explains 97.4% of the S&P 500’s variation
- ❑ **MAPE (Mean Absolute Percentage Error)**
  - ❑ Average percentage error
  - ❑ MAPE = 1.67% → Forecasts are off by just 1.67%, indicating excellent accuracy

Future Directions

❏ Incorporate tariff data into the SARIMAX model to improve performance (e.g., average percentage of tariffs)

❏ Compare the SARIMAX model to auto.arima() in R and apply variance stabilization transformations using the Box-Cox transformation parameter if necessary

Questions?