Bridging search behavior and market dynamics
A hybrid model for high-frequency financial data

(CFE-CMStatistics 2025)

Soudeep Deb

(Joint work with Anitha Pathlavath and Archi Roy)

Indian Institute of Management Bangalore
Bannerghatta Road, Bengaluru 560069, India.
Email: soudeep@iimb.ac.in.

December 15, 2025

Outline

Aim of the Study

We aim to analyze high-frequency financial data (applied to Bitcoin prices) using a hybrid approach which incorporates market sentiment data (captured through Google Trends).

Aim of the Study

We aim to analyze high-frequency financial data (applied to Bitcoin prices) using a hybrid approach which incorporates market sentiment data (captured through Google Trends).

Proposed approach

- Change-point detection
- Hybrid non-parametric regression (NR) + LSTM
- Captures both underlying trends and complex temporal dynamics

Comparative analysis

- Benchmarked against SVR, GARCH, and standalone NR and LSTM

Introduction to financial Markets and external events

Financial markets

- **Stock market:** Trading of shares in companies (e.g., Reliance, Infosys)
- **Currency market (Forex):** Trading of currencies (INR–USD, EUR/JPY)
- **Commodity market:** Trading of raw materials (e.g., gold, oil)
- **Bond market:** Buying and selling of debt securities

External events

- **Political:** Elections, government policies, international relations
- **Economic:** Inflation, interest rates, unemployment data
- **Natural:** Natural disasters, pandemics, climate change
- **Social:** Major social movements, tech breakthroughs

Cryptocurrency

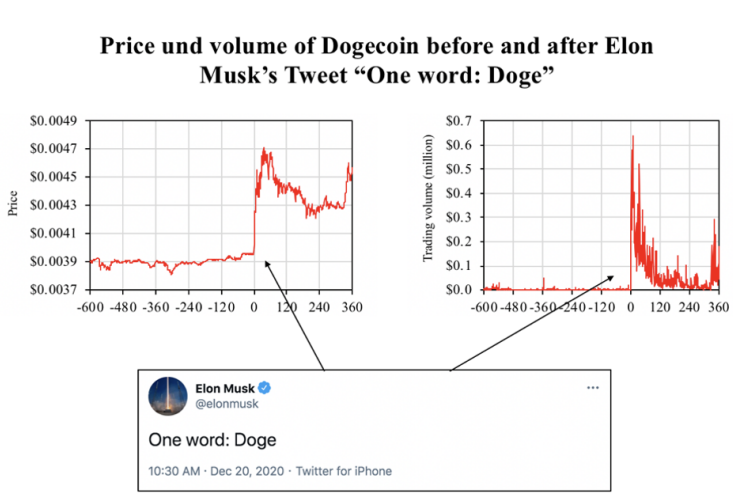
Cryptocurrency is a digital or virtual form of currency that uses cryptography for security.

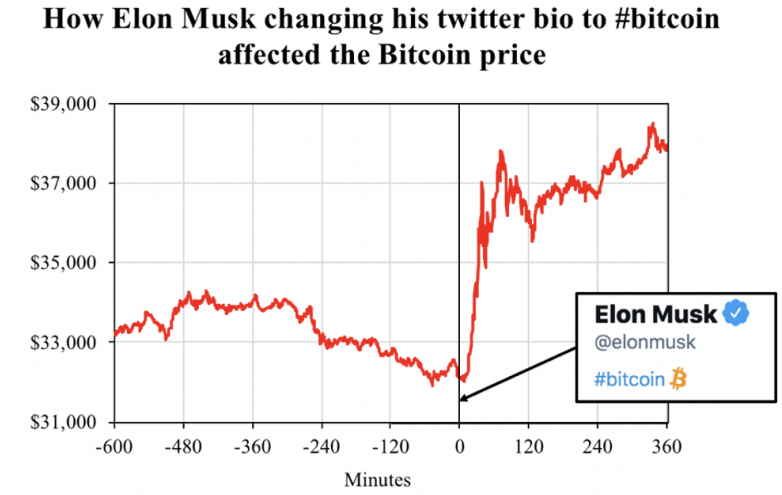
Key features

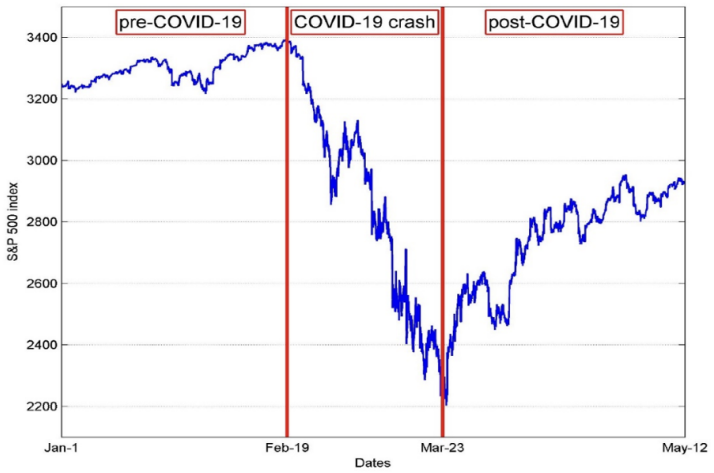
- **Decentralization:** Operates on blockchain (distributed ledger across multiple computers)
- **Security:** Cryptographic techniques secure transactions and control creation of new units
- **Anonymity:** Varying degrees of user anonymity depending on the currency

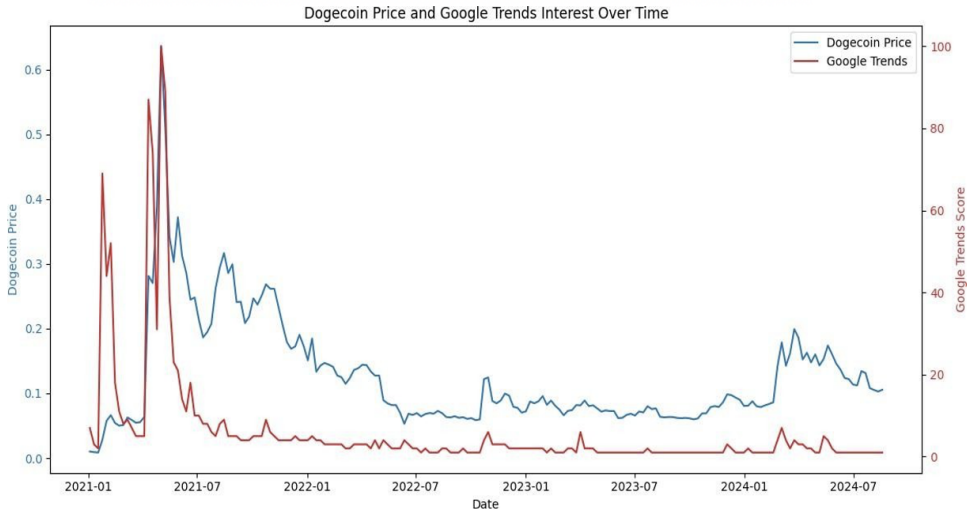
Popular cryptocurrencies: Bitcoin, Ethereum, Dogecoin, Litecoin, Ripple

Motivating example









Outline

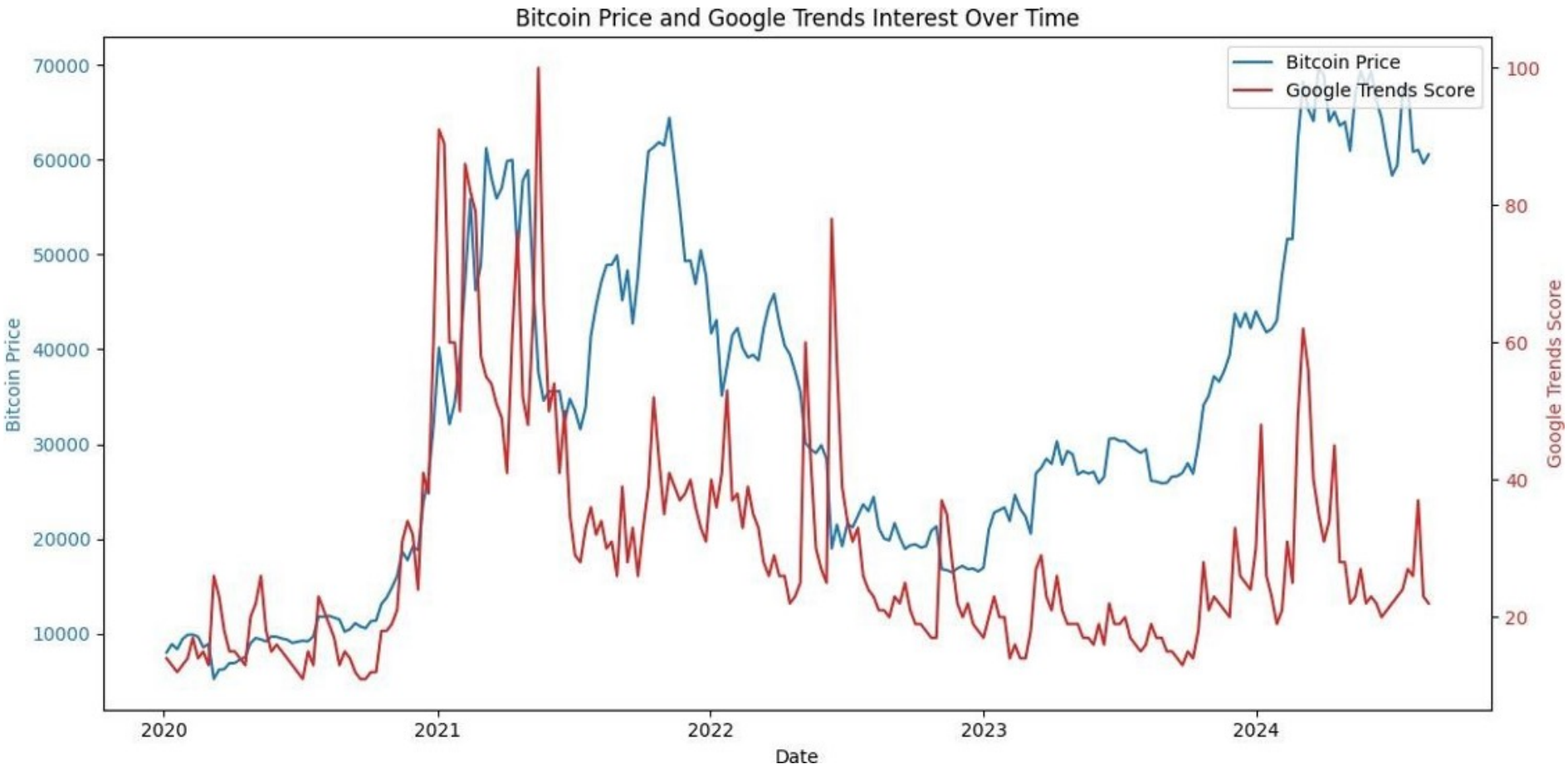
Data

- Bitcoin price is the key response variable
- Google Trends data is used as a covariate in the NR module
- Frequency: Hourly
- Period: January 2019 – December 2023
- Total observations: $\sim$ 43,800 data points

Google Trends Score

- **Definition:** Measures how often a term is searched on Google in a relative scale (scored 0–100, with 100 indicating peak popularity)
- **Data source:** Aggregated search data, available by time and region
- **Purpose in model:** Proxy for public interest in Bitcoin
- **Goal:** Identify potential links between public interest and Bitcoin price changes

Bitcoin price and Google trends



EDA: General findings

	Price (log-transformed)	Google Trends
Mean	9.8477	32.8744
Standard deviation	0.7679	16.9115
Range	(8.118, 11.136)	(8.0, 100.0)
Quartiles	(9.174, 9.9872, 10.503)	(20.5, 27.85, 40.66)
Skewness	−0.3730	1.424
Kurtosis	−0.8935	1.97816
Correlation	0.67	

EDA: Granger Causality

Granger causality is a statistical hypothesis test used to determine whether one time series can predict another.

The test compares:

1. **Restricted model:** Uses only past values of Bitcoin prices
2. **Unrestricted model:** Uses past values of Bitcoin prices and lagged Google Trends data

EDA: Granger Causality

Granger causality is a statistical hypothesis test used to determine whether one time series can predict another.

The test compares:

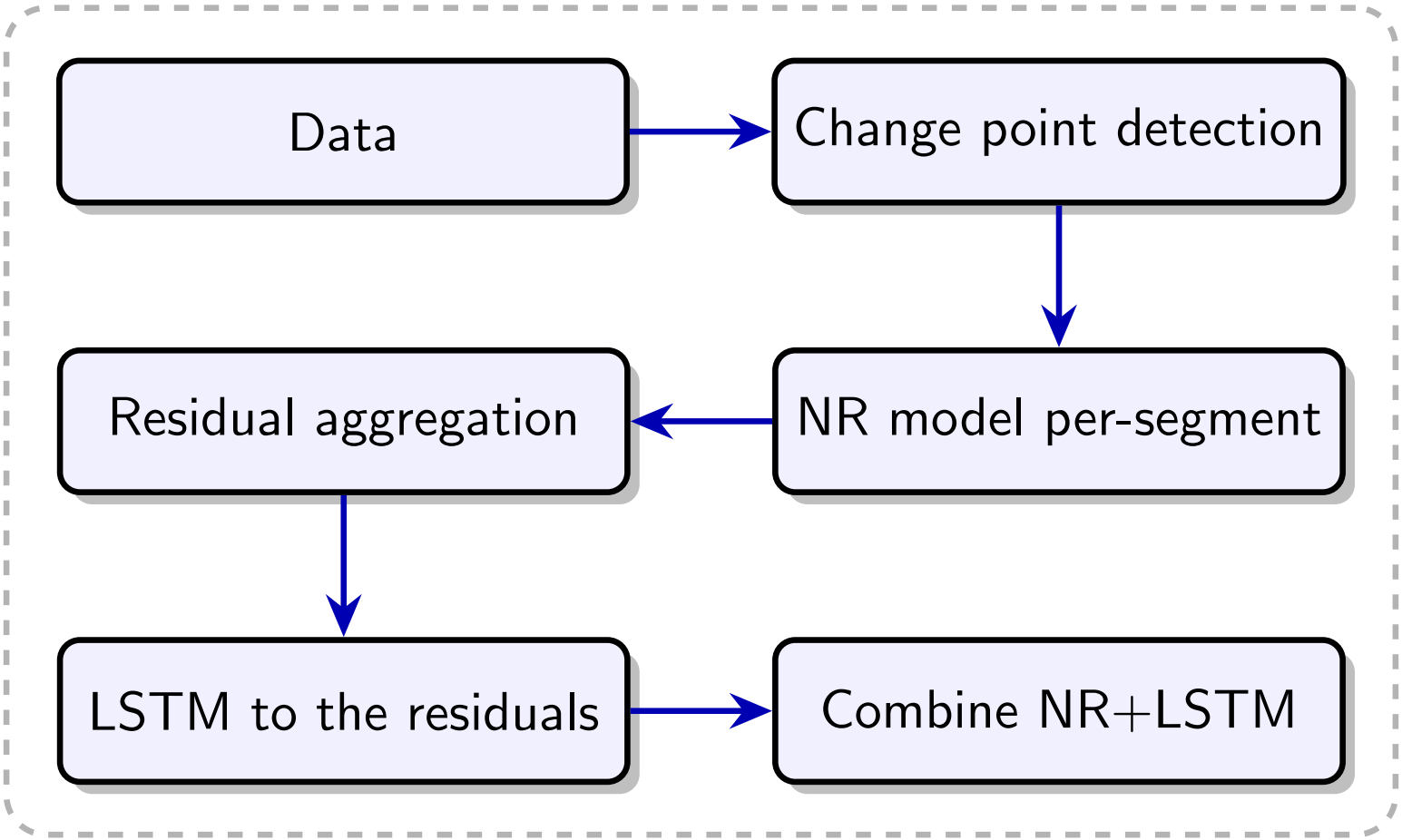
1. **Restricted model:** Uses only past values of Bitcoin prices
2. **Unrestricted model:** Uses past values of Bitcoin prices and lagged Google Trends data

We find the following (with 6 hour lag):

- Bitcoin price $\rightarrow$ Google Trends: p-value is 0.29 (not significant)
- Google Trends $\rightarrow$ Bitcoin price: p-value is 0.004 (significant)

Outline

Algorithm Pipeline



Why changepoint detection

- Use existing algorithm (PELT) to detect change points to identify **structural shifts** before applying NR–LSTM
- Split the dataset into segments based on detected change points
- Captures different patterns across regimes
- Accounts for market-behavior shifts, improving prediction accuracy
- Handles non-stationarity more reliably

Why NR?

High-frequency financial data often show **nonlinearity**, **high volatility**, **long-term dependency**, and **irregular patterns**.

Why nonparametric regression (NR)?

- Traditional models (e.g., linear regression, ARIMA) assume a fixed functional form
- NR makes **no strong functional-form assumptions** and estimates relationships directly from data (kernel smoothing)
- Captures smooth nonlinear trends and flexible relationships between variables

Why LSTM?

Bitcoin prices are often **highly nonlinear**, **nonstationary**, and exhibit **long-term dependencies**.

Why LSTM works well?

- LSTM is an RNN architecture designed to retain past information via memory cells and gates
- Learns relationships over long time intervals; mitigates vanishing gradients
- Effective for sequential time-series, volatility clustering, and nonlinear temporal behavior (e.g., sudden jumps)

Hybrid NR–LSTM: Philosophy

To leverage strengths of both models, we combine **NR** and **LSTM**.

Framework (trend + residual)

- Split series into a **trend component** and a **residual component**
- Model the trend using NR
- Model residual (unexplained fluctuations) using LSTM
- Combine both to obtain the final hybrid forecast

Hybrid NR–LSTM: Notation

We denote $\{Y_t\}$ as the hourly Bitcoin price and $\{X_t\}$ as the Google Trends score. The generic heteroskedastic nonparametric regression is developed using the model:

$$Y_t = \mu(X_{t-6}) + \sigma(X_{t-6})\,\varepsilon_t,$$

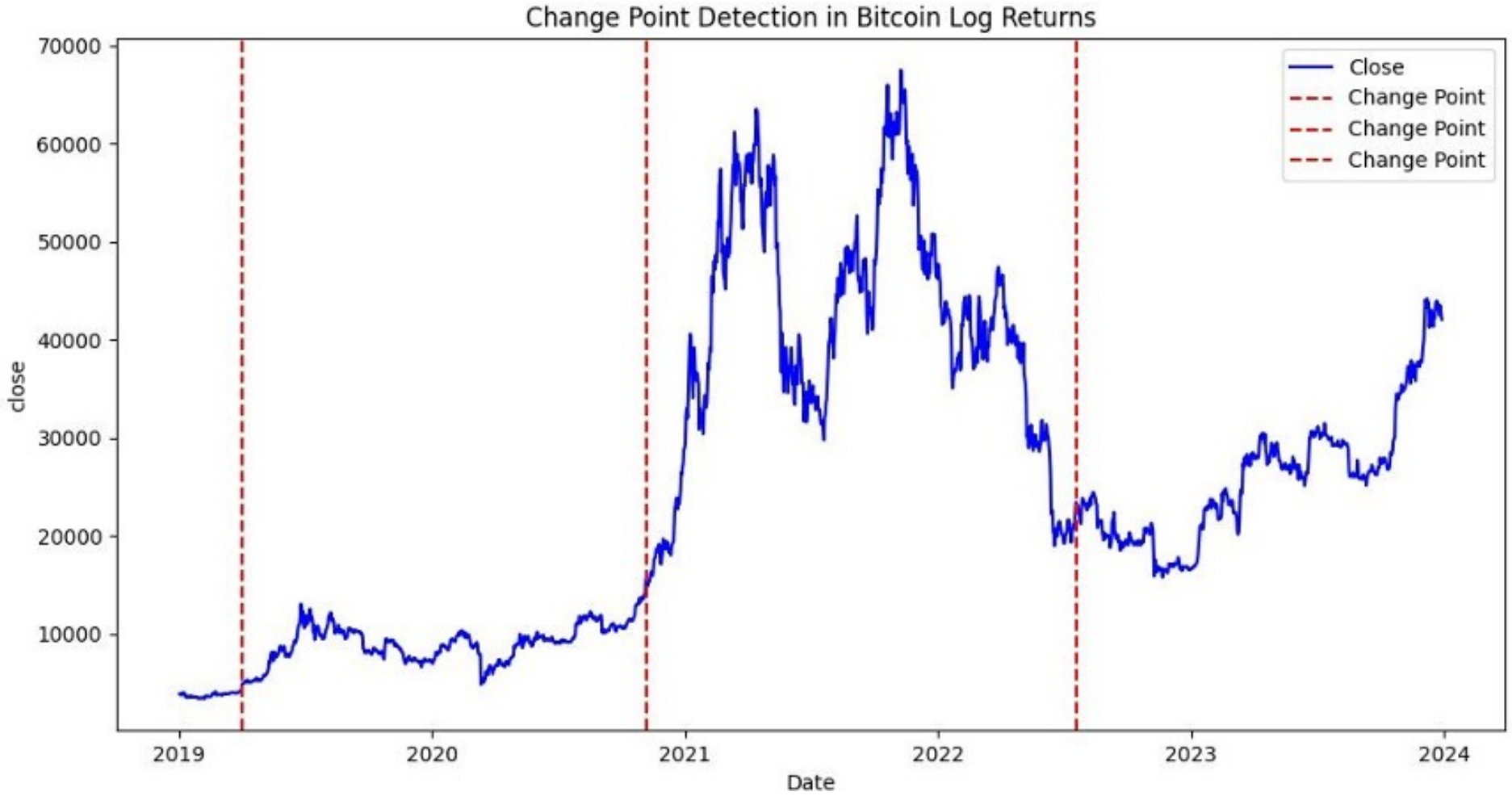
where $\mu(\cdot)$ and $\sigma(\cdot)$ are unknown smooth functions estimated using a nonparametric Nadaraya–Watson kernel estimator.

After estimating $\mu(\cdot)$ and $\sigma(\cdot)$, we extract residual series and apply LSTM to the residuals. The final prediction takes the form

$$\hat{Y}_t = \hat{\mu}\,(X_{t-6}) + \hat{\sigma}\,(X_{t-6})\;\hat{\varepsilon}_t^{(\mathrm{LSTM})}.$$

Outline

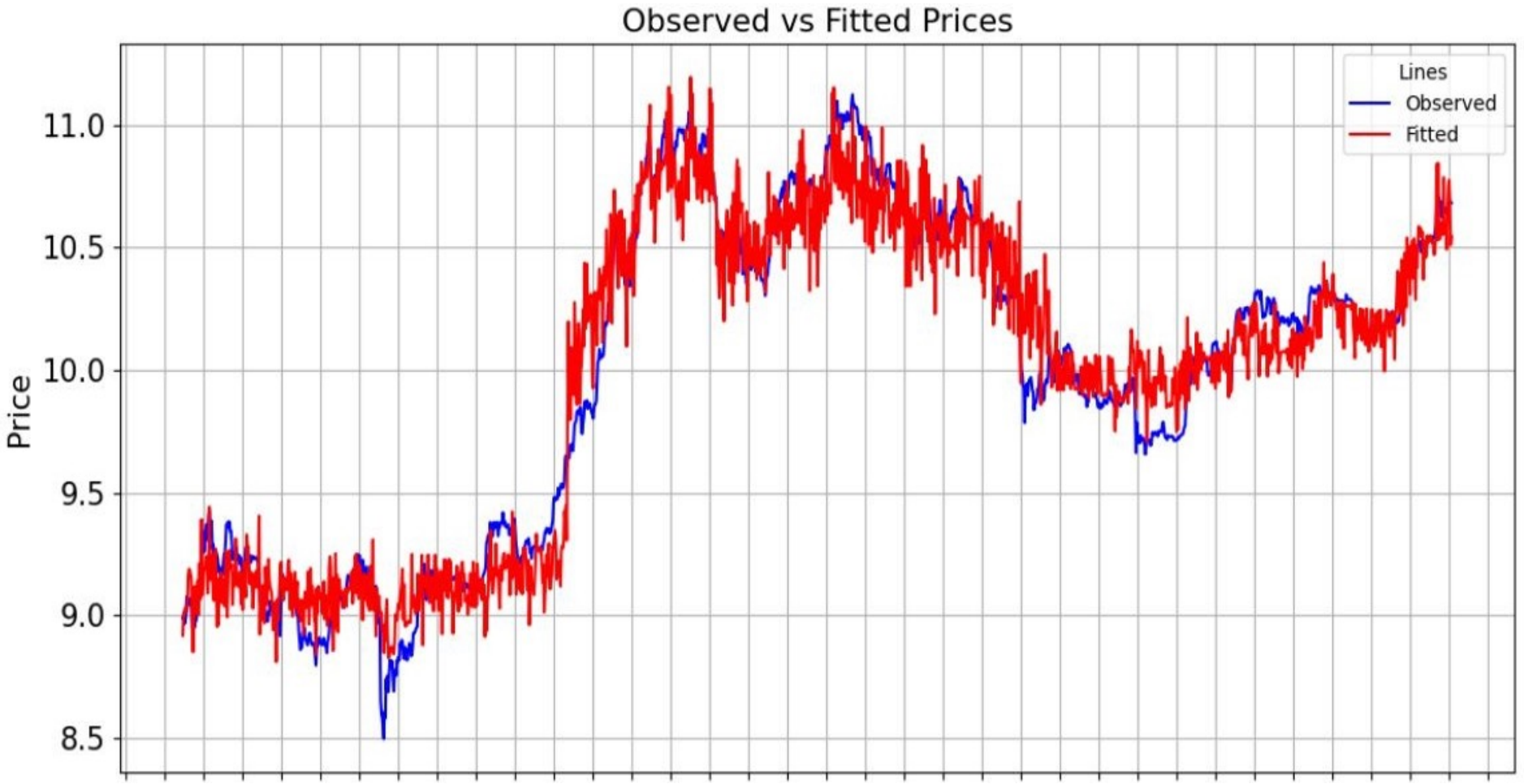
Detected segments



Insights obtained from the model

- The three segments show different patterns for the two unknown functions.
- Error structures are modeled better through the LSTM module.
- Forecast of both price and volatility incorporate Google Trends information.

How good is the fitted model?



Robust forecasting strategy

- One-month expanding window approach, starting with data from 2019–2022
- After each prediction, expand training window by one month (include most recent data)
- Predict next day's Bitcoin price; repeat for entire 2023 dataset
- Track MAE, RMSE, MAPE at each step
- Expanding window adapts to recent market changes and improves performance over time

Comparative models

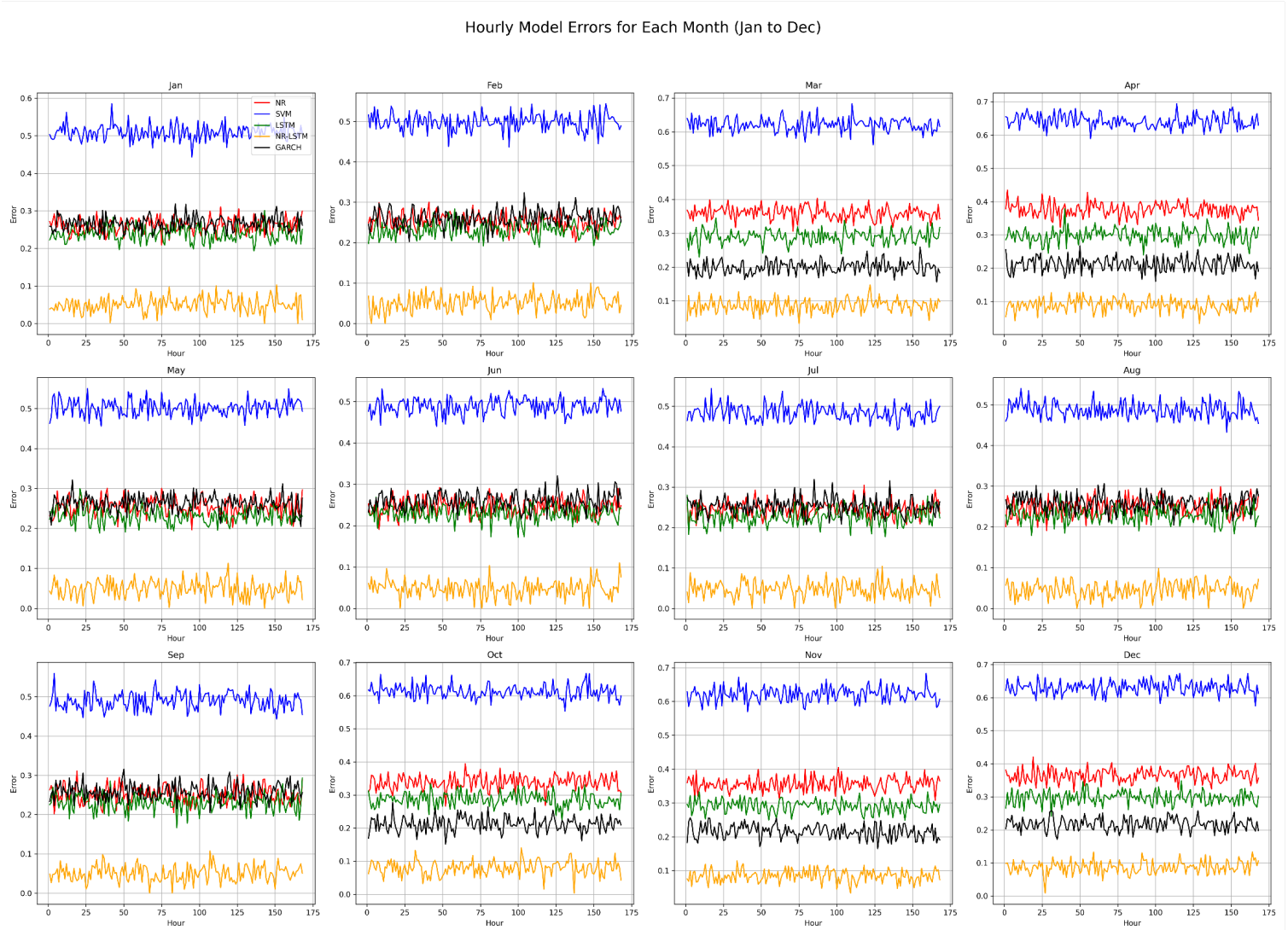
- **NR:** captures nonlinear relationships without assuming a specific functional form
- **LSTM:** models long-term dependencies in sequential data
- **SVR:** learns a hyperplane for continuous prediction; captures nonlinear relationships with margin
- **GARCH:** included as a volatility-model benchmark

Summary of forecasting results

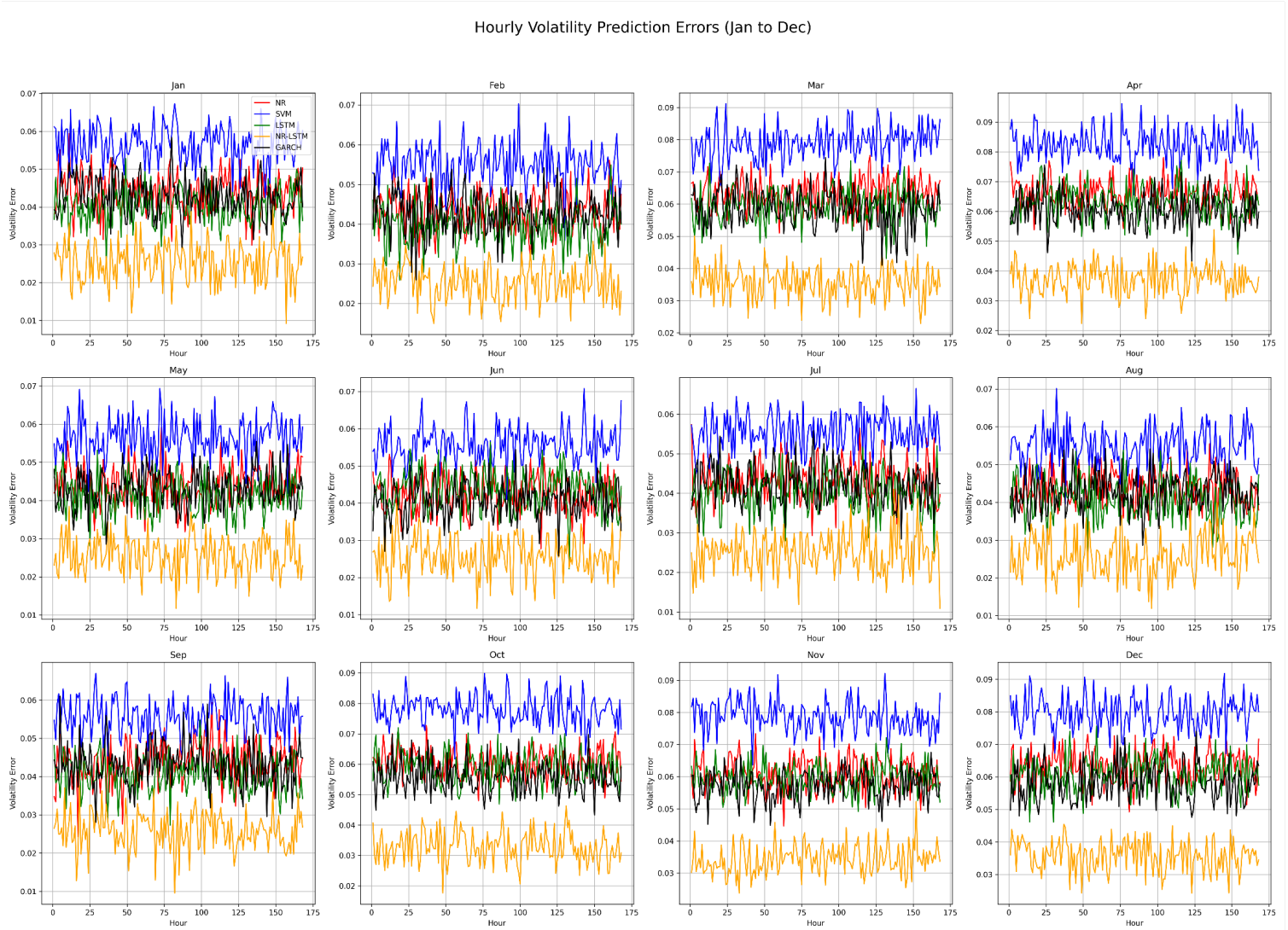
- The NR–LSTM hybrid model records the lowest RMSE: **0.0641**
- Outperforms the competitors by combining both trend and residual patterns accurately:

Method	RMSE
NR	0.2961
SVM	0.5488
LSTM	0.2621
GARCH	0.24
NR–LSTM	0.0641

Forecast accuracy for price



Forecast accuracy for volatility



Outline

Summary

- Studied hourly Bitcoin prices (2019–2023) with Google Trends as a proxy for market attention/sentiment.
- EDA indicates a strong association and Granger evidence that Trends helps predict price (6-hour lag).
- Proposed three-stage pipeline of (i) change-point detection (PELT), (ii) NR per segment (trend/scale), (iii) LSTM on residuals.
- Hybridization improves robustness under nonstationarity and nonlinear dynamics.
- Empirically, NR–LSTM achieves best RMSE (0.0641) compared with NR, SVR, LSTM and GARCH.

Future scope

- **Richer covariates:** add macro/market variables (rates, inflation, FX, equity indices) and crypto-specific signals (on-chain metrics, volume, order-book features).
- **Multi-source sentiment:** combine Google Trends with news or social-media sentiment; evaluate lead–lag structure across sources.
- **Probabilistic forecasts:** prediction intervals / quantile loss; regime-wise uncertainty calibration.
- **Model extensions:** attention/Transformer for residuals; regime-aware LSTM; jointly learn segmentation + forecasting.
- **Trading/decision evaluation:** backtesting with transaction costs, risk metrics (VaR/ES), and robustness under stress periods.

References

- Bates, J. M., Granger, C. W. (1969). The combination of forecasts. Journal of the operational research society, 20(4), 451-468.
- Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International journal of forecasting, 5(4), 559-583.
- Ozdemir, O., Yozgatligil, C. (2024). Forecasting performance of machine learning, time series, and hybrid methods for low-and high-frequency time series. Statistica Neerlandica, 78(2), 441-474.
- Roy, A., Podder, M., **Deb, S.** (2024+) Nonparametric method of structural break detection in stochastic time series regression model. Pre-print: `https://arxiv.org/abs/2410.15713`.
- Song, Y., Cai, C., Ma, D., Li, C. (2024). Modelling and forecasting high-frequency data with jumps based on a hybrid nonparametric regression and LSTM model. Expert Systems with Applications, 237, 121527.
- Yu, Y., Si, X., Hu, C., Zhang, J. (2019). A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 31(7), 1235-1270.
- Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159-175.

Thank you!

Contact: soudeep@iimb.ac.in
Webpage: https://soudeepd.github.io/