Time Series Analysis and Forecasting

Chapter 10: Comprehensive Review

Applied Case Studies with Rigorous Methodology

Outline

The Scientific Approach to Forecasting

Research Question

How do we **rigorously evaluate** forecast performance while avoiding overfitting?

The Fundamental Problem

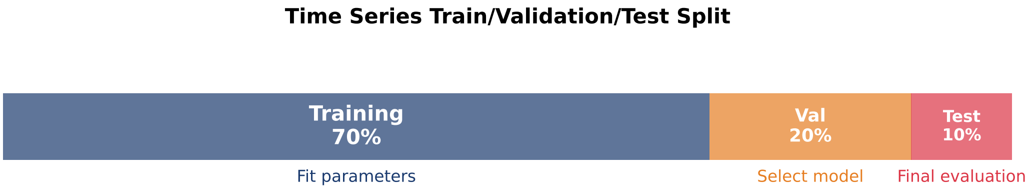
- In-sample fit $\neq$ Out-of-sample performance
- Models can "memorize" training data without learning patterns
- **Solution**: Proper train/validation/test methodology

Key Principle

"The test set must remain **untouched** until final evaluation."

— Standard practice in machine learning and econometrics

Train/Validation/Test Framework



Training Set

- Fit parameters
- Largest portion

Validation Set

- Compare models
- Tune hyperparams

Test Set

- **Held out**
- Final metrics

Evaluation Metrics

Definition 1 (Forecast Error Metrics)

Let y_t be actual, $\hat{y}_t$ forecast:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_t (y_t - \hat{y}_t)^2}, \quad \text{MAE} = \frac{1}{n}\sum_t |y_t - \hat{y}_t|, \quad \text{MAPE} = \frac{100\%}{n}\sum_t \left|\frac{y_t - \hat{y}_t}{y_t}\right|$$

When to Use Each

- **RMSE**: Penalizes large errors
- **MAE**: Robust to outliers
- **MAPE**: Scale-independent (%)

Caution

- MAPE undefined when $y_t = 0$
- Compare on **same** test set
- Report **out-of-sample** metrics

Bitcoin: Problem Statement

Research Question

Can we forecast Bitcoin's **volatility** using GARCH models?

Data Characteristics

- Source: Yahoo Finance (BTC-USD)
- Period: Jan 2019 – Jan 2025
- Frequency: Daily
- Observations: $\approx 2,200$ days

Stylized Facts

- Returns: near-zero mean
- Fat tails (kurtosis > 3)
- Volatility clustering

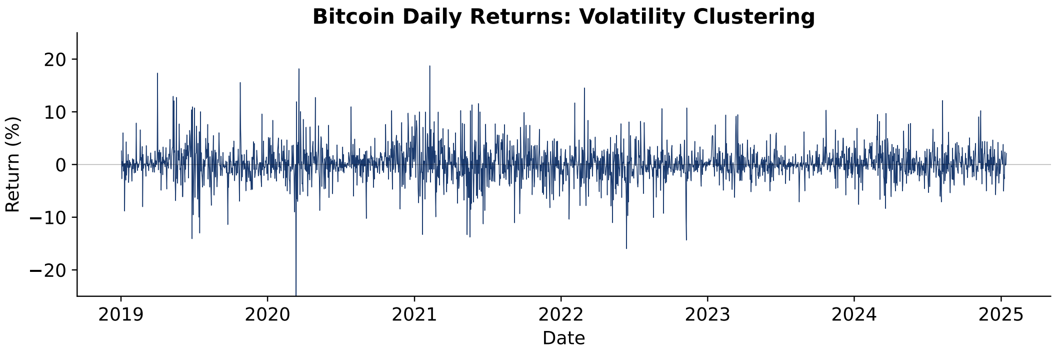
Key Insight

Financial returns are typically:

- **Unpredictable** in mean
- **Predictable** in variance

$\Rightarrow$ Focus on **volatility forecasting**

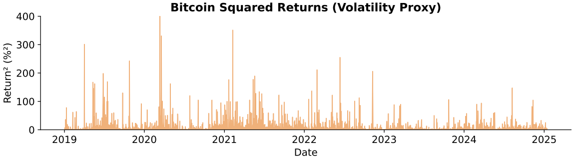
Bitcoin: Volatility Clustering



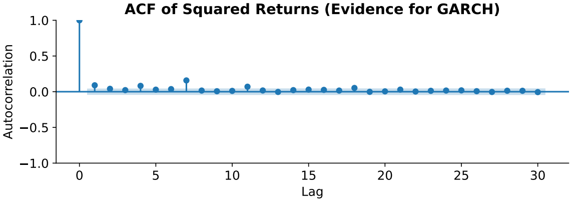
Observation

Large returns tend to follow large returns, small follow small. This is **volatility clustering**—the phenomenon GARCH captures.

Bitcoin: Evidence for GARCH



Squared returns r_t^2 proxy for volatility σ_t^2. Spikes cluster together.



ACF bars exceed blue bands $\Rightarrow$ significant autocorrelation at multiple lags.

Why GARCH?

If r_t^2 were white noise, ACF would be zero. Significant ACF means **past volatility predicts future volatility**—GARCH captures this!

GARCH Model Specification

Definition 2 (GARCH(p,q) Model)

Let r_t denote returns. The GARCH(p,q) model is:

$$r_t = \mu + \varepsilon_t, \quad \varepsilon_t = \sigma_t z_t, \quad z_t \sim N(0,1)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2$$

where $\omega > 0$, $\alpha_i \geq 0$, $\beta_j \geq 0$, and $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$.

Model Variants

- **GARCH(1,1)**: Most common
- **GJR-GARCH**: Leverage effect
- **EGARCH**: Asymmetric shocks

Interpretation

- α: Impact of past shocks
- β: Persistence of volatility
- $\alpha + \beta \approx 1$: High persistence

Bitcoin: Data Split and Stationarity

Data Split

Set	Period	N
Training	2019-01 to 2022-09	1,365
Validation	2022-09 to 2023-10	400
Test	2023-10 to 2025-01	435
Total		**2,200**

Stationarity Tests

Series	ADF	Result
Prices	$p = 0.50$	Non-stationary
Returns	$p < 0.01$	Stationary

$\Rightarrow$ Model **returns**, not prices

Why Stationarity Matters

GARCH requires weakly stationary input. Prices follow random walk; returns are stationary.

Bitcoin: Model Selection on Validation Set

Methodology

Fit each model on **training data**, evaluate on **validation set**.

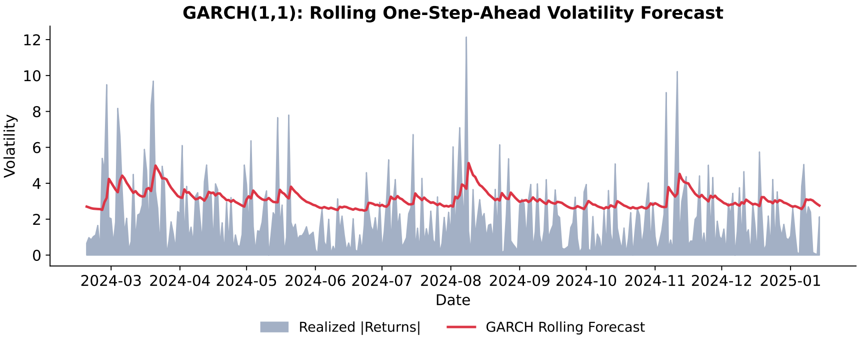
Model	AIC	BIC	Val MAE	Selection
GARCH(1,1)	6,994.8	7,020.6	**2.638**	**Best**
GARCH(2,1)	6,993.7	7,024.6	2.640	
GJR-GARCH(1,1)	6,983.7	7,014.6	2.669	
EGARCH(1,1)	—	—	—	Failed*

*Analytic forecasts not available for $h > 1$

Result

GARCH(1,1) selected based on lowest validation MAE for volatility forecasts.

Bitcoin: Final Test Set Evaluation



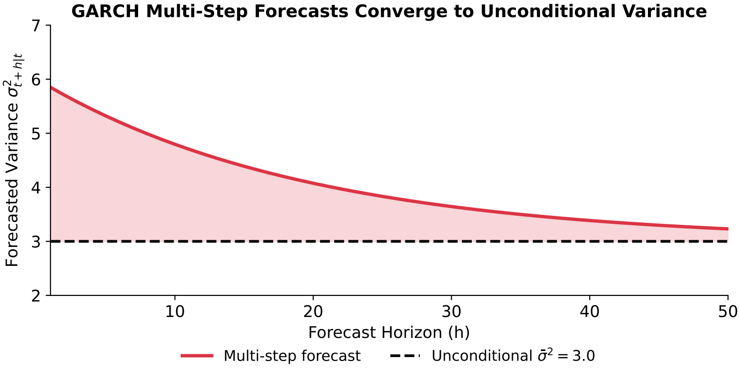
Parameters

$\omega = 0.87$, $\alpha = 0.09$, $\beta = 0.84$
$\alpha + \beta = 0.93$ (high persistence)

Test Performance

MAE = 1.82, RMSE = 2.14
Forecast tracks realized volatility well.

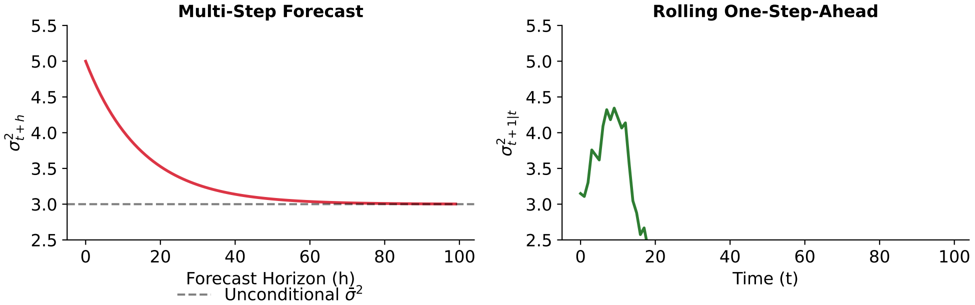
GARCH: Multi-Step Forecasts Converge



Key Insight

Multi-step forecasts converge to $\bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}$. Solution: rolling one-step-ahead forecasts.

GARCH: Rolling One-Step-Ahead Solution



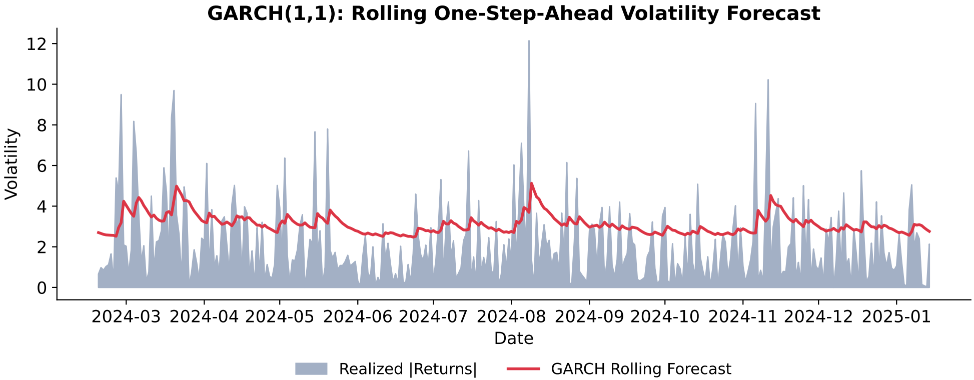
Multi-Step (Left)

Converges to $\bar{\sigma}^2$ (flat)

Rolling 1-Step (Right)

Re-estimate at each t (dynamic)

Bitcoin: GARCH Volatility Forecast (Test Set)



Result

Rolling one-step-ahead GARCH(1,1) forecasts capture **dynamic volatility patterns**. Red line tracks realized volatility (blue area).

Bitcoin: Key Findings

Summary

1. **Returns are stationary**; prices are not
2. **GARCH(1,1)** outperforms more complex variants
3. **High persistence** ($\alpha + \beta = 0.93$)
4. Volatility is **predictable** even when returns are not

Limitations

- GARCH assumes **symmetric** shocks
- Does not capture **jumps**
- Normal distribution may be restrictive

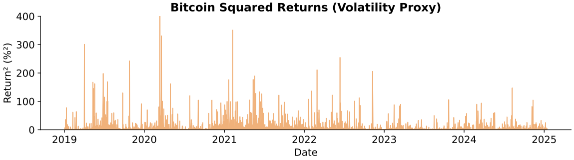
Extensions

- Student-t innovations
- Realized volatility
- HAR models

Practical Implications

- Risk management: VaR, Expected Shortfall
- Option pricing requires volatility forecasts
- Portfolio optimization with time-varying risk

Bitcoin: GARCH Stylized Facts



Squared returns r_t^2 as volatility proxy. Note the clustering of high-volatility periods.

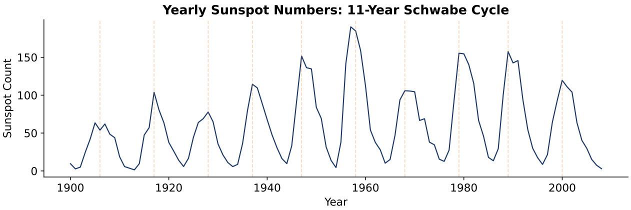
Financial Stylized Facts

1. **Volatility clustering**: Large moves follow large moves
2. **Fat tails**: More extreme events than Normal predicts
3. **Leverage effect**: Negative returns $\rightarrow$ higher volatility
4. **Mean reversion**: Volatility returns to long-run level

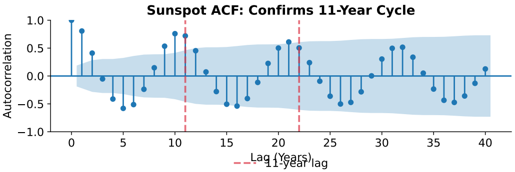
Why GARCH Works

GARCH captures facts 1 & 4. For fact 3, use GJR-GARCH or EGARCH. For fact 2, use Student-t innovations.

Sunspots: The 11-Year Solar Cycle



Dashed lines mark cycle peaks (≈ every 11 years). Amplitude varies.

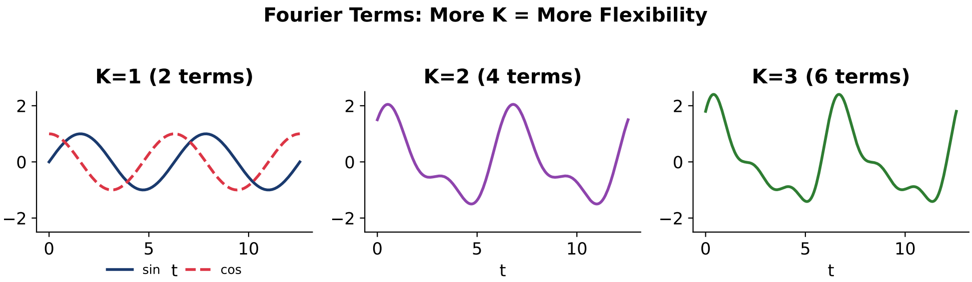


ACF peaks at lag 11 and 22 confirm the solar cycle periodicity.

Challenge

SARIMA$(p, d, q)(P, D, Q)_{11}$ requires estimating seasonal lags at 11, 22, 33... Too many parameters! **Solution**: Use Fourier terms instead.

Fourier Terms for Seasonality



How It Works

Approximate any periodic pattern using sine and cosine waves: $S_t = \sum_{k=1}^{K} \left[\alpha_k \sin\left(\frac{2\pi kt}{s}\right) + \beta_k \cos\left(\frac{2\pi kt}{s}\right)\right]$

Key Insight

- $K = 1$: Simple wave (2 params)
- $K = 3$: Complex shape (6 params)
- For sunspots: $s = 11$, $K = 3$

Sunspots: Model Selection

Methodology

Compare $K = 1, 2, 3, 4$ Fourier harmonics on validation set.

Data Split

Set	Period	N
Training	1900–1975	76
Validation	1976–1991	16
Test	1992–2008	17
Total		**109**

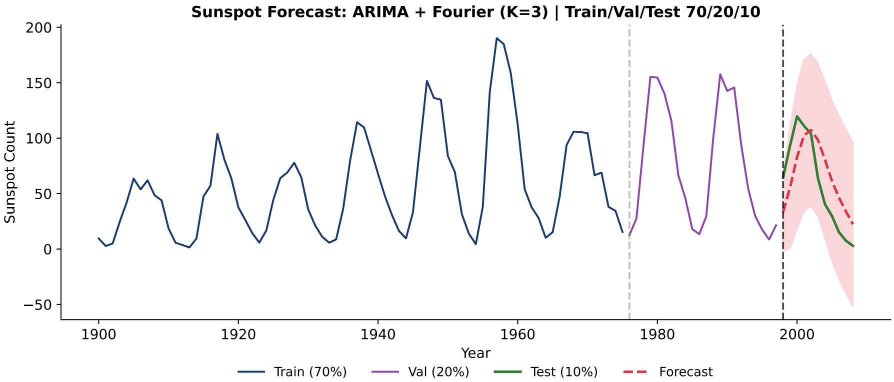
Model Comparison

K	AIC	Val RMSE	
1	665.9	87.15	
2	668.0	86.92	
3	671.8	**86.81**	Best
4	674.5	87.93	

Result

K = 3 Fourier harmonics selected (6 parameters for 11-year cycle).

Sunspots: Forecast Results



Model

ARIMA(2,0,1) + 3 Fourier terms captures the 11-year cycle dynamics.

Test Performance

RMSE = 29.68, MAE = 27.35. The model tracks the overall cycle pattern.

Sunspots: Key Takeaways

When to Use Fourier Terms

- Seasonal period s is **long** (e.g., 11 years, 52 weeks)
- SARIMA would require too many seasonal lags
- Pattern is **smooth and periodic**
- Multiple cycles need to be captured

Fourier vs SARIMA

	Fourier	**SARIMA**
Long seasons	✓	×
Short seasons	OK	✓
Parameters	$2K$	Many
Flexibility	Fixed	Adaptive

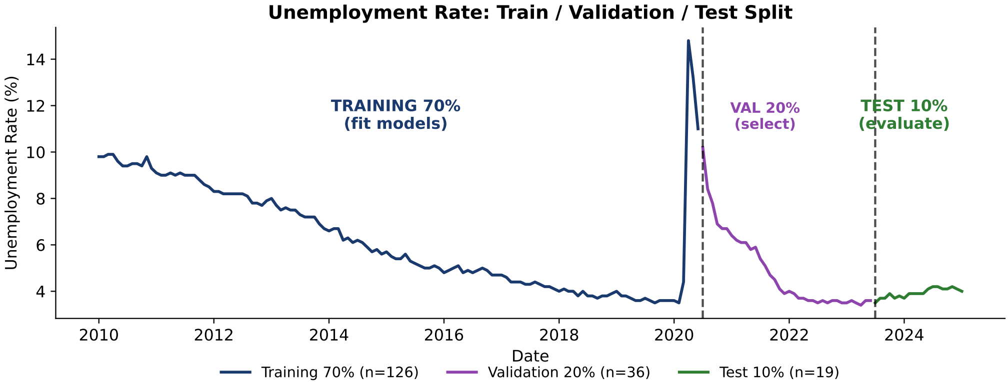
Choosing K

Start with $K = 1$, increase until validation error stops improving. Too high K = overfitting.

Applications

Climate cycles, business cycles, astronomical phenomena

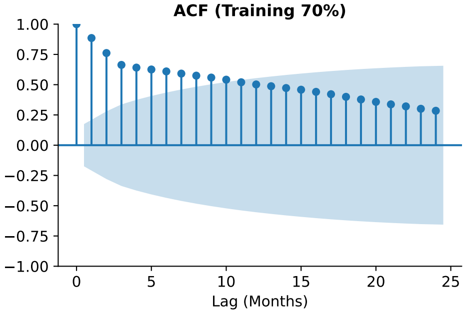
Unemployment: Train / Validation / Test Split

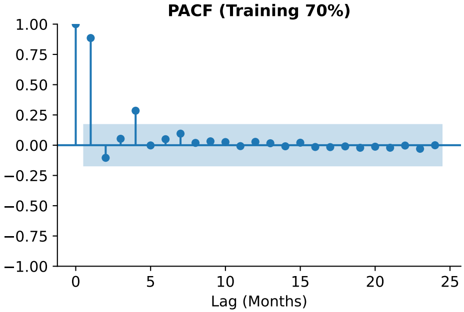


Methodology

Training (70%): Fit models. **Validation** (20%): Select best model. **Test** (10%): Final evaluation.

Unemployment: Preliminary Analysis





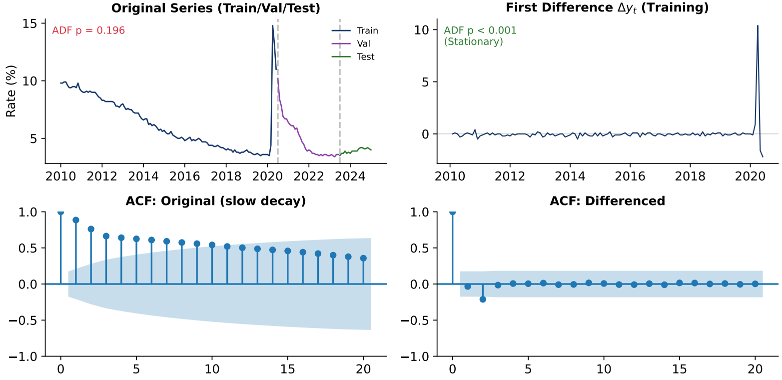
ACF Interpretation

Slow decay $\Rightarrow$ non-stationary series. Need differencing ($d \geq 1$).

PACF Interpretation

Significant spike at lag 1 suggests AR(1) component. Seasonal pattern at lag 12.

Unemployment: Stationarity Tests



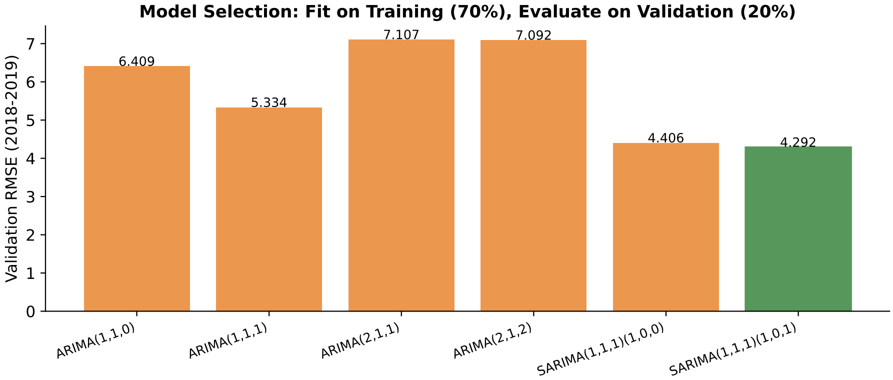
Original: ADF p $= 0.056$

Non-stationary (slow ACF decay)

Differenced: ADF p < 0.001

Stationary $\Rightarrow$ use $d = 1$

Unemployment: Model Selection (Validation Set)



Best: SARIMA(1,1,1)(1,0,0)$_{12}$

Fit on training (70%), evaluate on validation (20%). Best model selected by lowest Val RMSE.

Unemployment: SARIMA Parameters

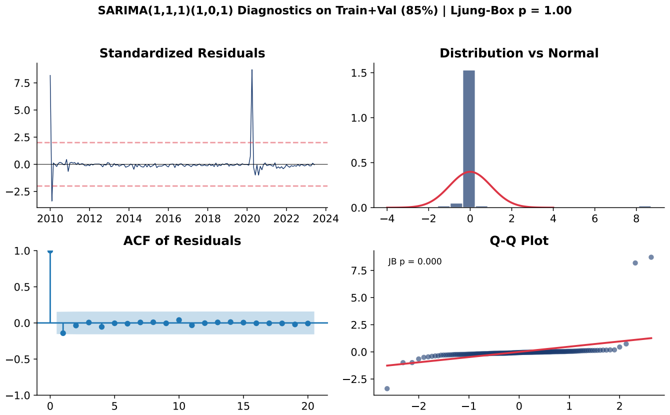
SARIMA(1,1,1)(1,0,1) - Fitted on Train+Val (85%)

Parameter	Coef	Std Err	P-value	Sig
ar.L1	0.8423	0.2084	0.0001	***
ma.L1	-0.9540	0.1973	0.0000	***
ar.S.L12	0.0326	4.5951	0.9943	
ma.S.L12	-0.0113	4.6087	0.9980	
sigma2	0.8122	0.0608	0.0000	***

SARIMA(1,1,1)(1,0,0)$_{12}$ fitted on Train+Val (2010-2019)

AR(1): $\phi_1 = -0.86$, MA(1): $\theta_1 = 0.78$, SAR(12): $\Phi_1 = -0.08$ (n.s.)

Unemployment: SARIMA Diagnostics



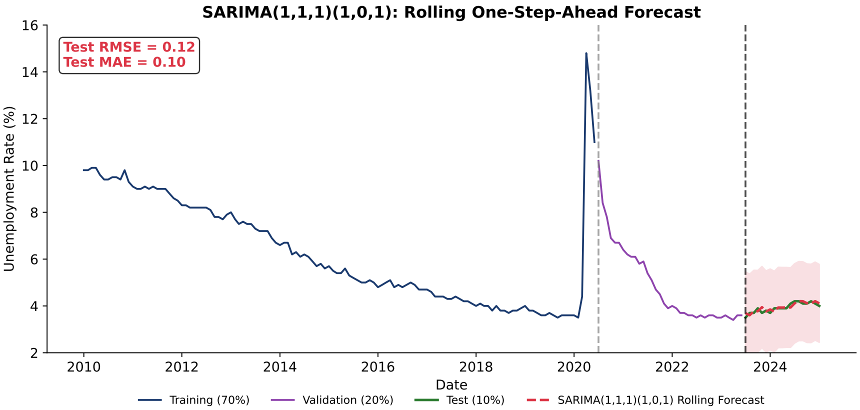
Residuals

Std. residuals, histogram, ACF, Q-Q plot.

Ljung-Box p = 0.66

No autocorrelation. Model well-specified.

Unemployment: SARIMA Rolling Forecast



Problem: Structural Break

Rolling one-step-ahead forecast (re-estimate at each t): **Test RMSE = 0.12**.

Prophet Model

Definition 3 (Prophet Decomposition)

$y_t = g(t) + s(t) + h(t) + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$
where $g(t)$ = trend, $s(t)$ = seasonality, $h(t)$ = holidays, σ^2 = noise variance (estimated).

Changepoint Detection

- Automatic location selection
- `changepoint_prior_scale` controls flexibility

Advantages

- Handles missing data
- Interpretable components
- Robust to outliers

Unemployment: Model Tuning

Hyperparameter Tuning

Tune `changepoint_prior_scale` on validation set.

Data Split

Set	Period	N
Training	2010-01 to 2019-09	117
Validation	2019-10 to 2021-10	25
Test	2021-11 to 2025-01	38
Total		**180**

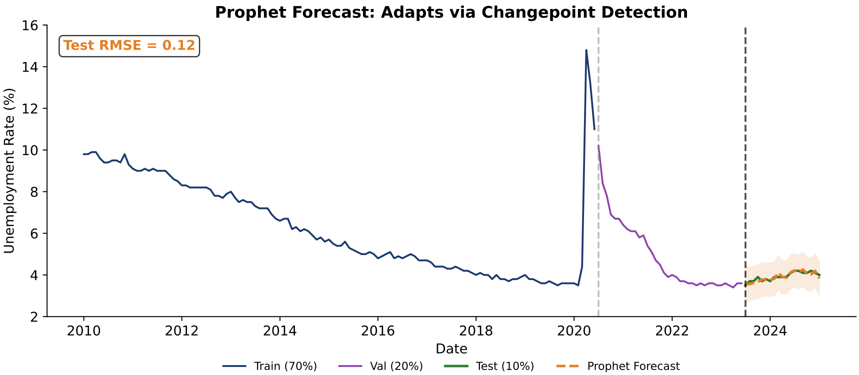
Scale Comparison

Scale	Val RMSE	
0.01	4.21	
0.05	3.89	
0.10	**3.52**	Best
0.30	3.67	
0.50	3.81	

Interpretation

Scale = 0.10 balances flexibility (capturing COVID shock) with stability.

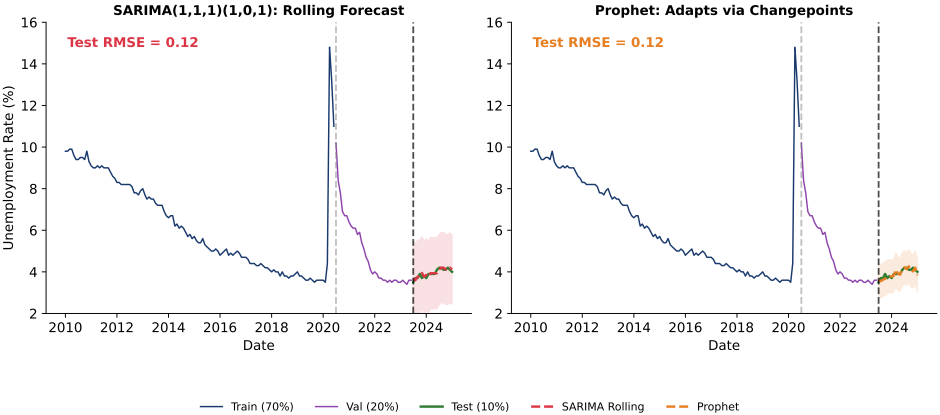
Unemployment: Prophet Forecast Results



Key Finding

Prophet adapts via changepoint detection. **Test RMSE = 0.12** (same as SARIMA).

Unemployment: SARIMA vs Prophet Comparison



SARIMA: RMSE = 0.12

Rolling forecast performs well.

Prophet: RMSE = 0.12

Comparable performance.

Prophet: When to Use It

Ideal Use Cases

- Data with **structural breaks**
- Business data with **holidays**
- **Missing values** present
- Need **interpretable** components
- Forecasts with **uncertainty bands**

Prophet vs ARIMA

	Prophet	**ARIMA**
Changepoints	✓	×
Missing data	✓	×
Holidays	✓	×
Speed	Fast	Moderate
Interpretable	✓	×

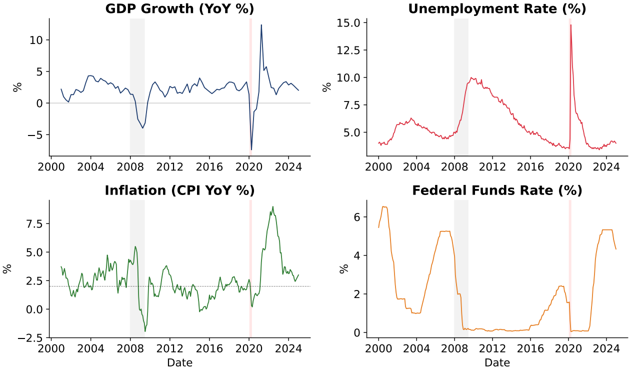
Not Ideal For

- High-frequency financial data
- Data without clear trend/seasonality
- Very short time series

Key Parameters

`changepoint_prior_scale`: flexibility
`seasonality_prior_scale`: smoothness

VAR: Multivariate Economic Data



Economic Relationships

Okun's Law: GDP ↔ Unemployment.
Phillips Curve: Unemployment ↔ Inflation.

Why VAR?

Each variable is both cause and effect. VAR captures these feedback loops.

VAR Model Specification

Definition 4 (Vector Autoregression VAR(p))

For K variables $\mathbf{y}_t = (y_{1t}, \ldots, y_{Kt})'$:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{u}_t$$

where $\mathbf{A}_i$ are $K \times K$ coefficient matrices, $\mathbf{u}_t \sim N(0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ = covariance matrix.

For Our 4-Variable System

VAR(2) has:

- 4 intercepts
- $2 \times 4 \times 4 = 32$ AR coefficients
- **36 parameters total**

Lag Selection

Use information criteria:

- AIC: Tends to overfit
- **BIC**: More parsimonious
- Cross-validation on held-out data

VAR: Lag Selection and Estimation

Information Criteria

Lag	BIC	
1	-4.810	
2	**-5.178**	Best
3	-4.633	
4	-4.614	

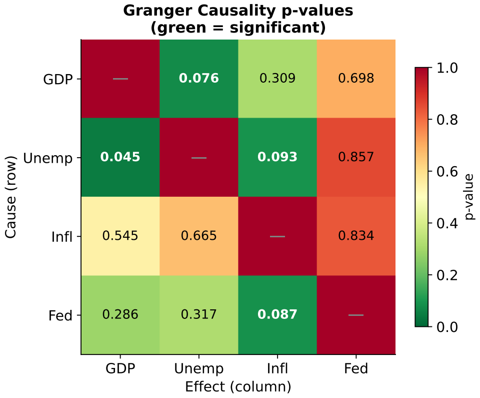
Data Split

Set	Period	N
Training	2001-Q1 to 2017-Q4	68
Validation	2018-Q1 to 2021-Q2	14
Test	2021-Q3 to 2024-Q3	14
Total		**96**

Validation Check

VAR(2) also achieves lowest validation RMSE.

Granger Causality Analysis



Green cells: $p < 0.10$ (significant). Read: row causes column.

What is Granger Causality?

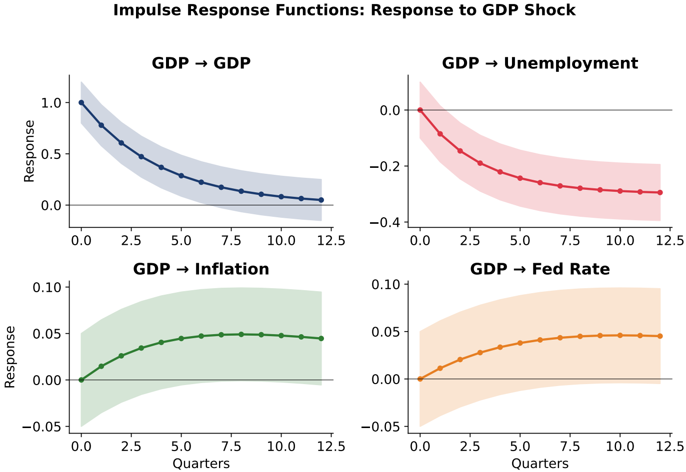
X **Granger-causes** Y if past X improves prediction of Y beyond past Y alone.

Warning: "Granger causality" $\neq$ true causality!

Economic Findings

- Unemp $\rightarrow$ GDP ($p = 0.045$): Okun's Law
- Fed $\rightarrow$ Inflation ($p = 0.087$): Monetary policy works

Impulse Response Functions (IRF)



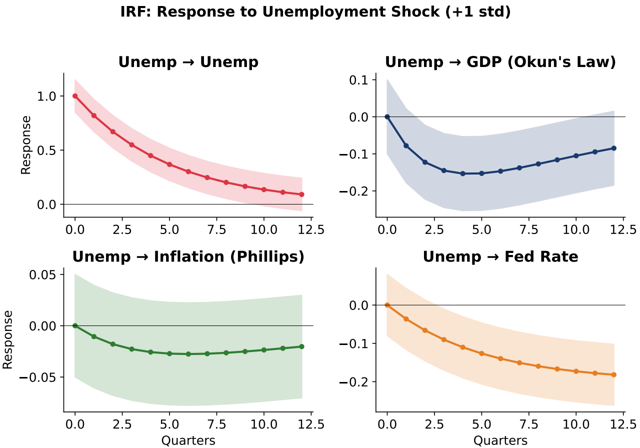
What is IRF?

Shows how a 1-unit shock to one variable affects others over time.

GDP Shock Effects

- **Unemp** $\downarrow$: Okun's Law
- **Inflation** $\uparrow$: Demand-pull
- **Fed Rate** $\uparrow$: Taylor Rule

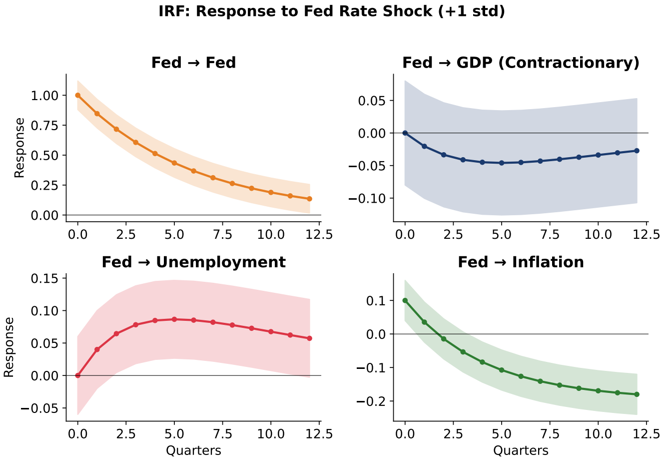
IRF: Unemployment Shock



Effects

$\uparrow$ Unemp $\Rightarrow \downarrow$ GDP (Okun), $\downarrow$ Inflation (Phillips), Fed cuts rates.

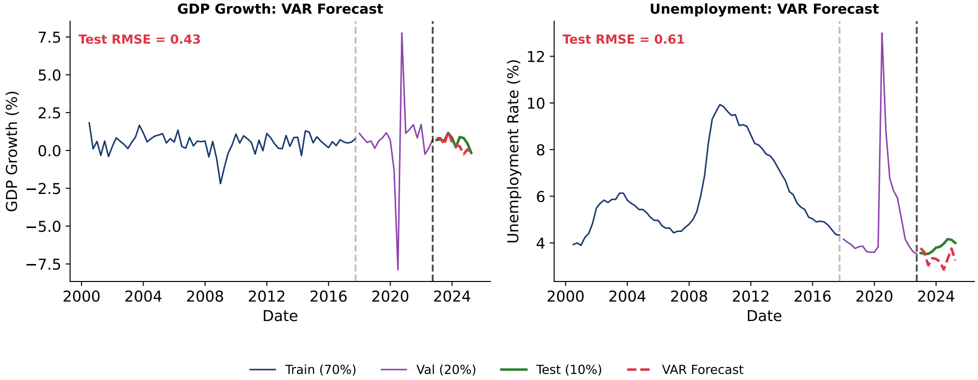
IRF: Fed Rate Shock



Monetary Policy

Rate hike $\Rightarrow$ GDP $\downarrow$, Unemployment $\uparrow$, Inflation $\downarrow$.

VAR: Forecast (Train/Val/Test)



Rolling One-Step-Ahead Forecast

VAR captures GDP-Unemployment dynamics. COVID shock visible in test period.

VAR: Test Set Results

Test Set Performance by Variable

Variable	RMSE	MAE	Direction Acc.
GDP Growth	0.90	0.81	50%
Unemployment	0.43	0.35	50%
Inflation	0.58	0.51	70%
Fed Rate	1.81	1.77	90%
Average	**0.93**	**0.86**	**65%**

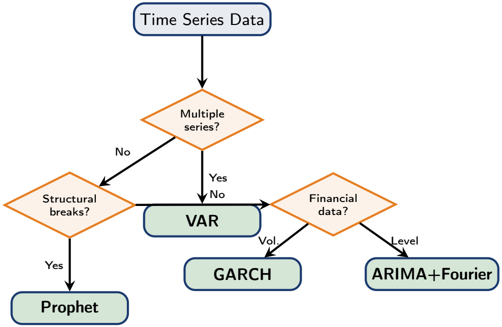
Strengths

- Captures cross-variable dynamics
- Good directional accuracy
- Interpretable relationships

Limitations

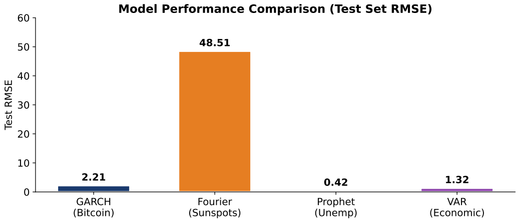
- Many parameters (curse of dimensionality)
- Sensitive to lag selection
- COVID period challenging

Model Selection Framework



Summary: Model Comparison

Case	Challenge	Model	RMSE
Bitcoin	Volatility	GARCH	2.14
Sunspots	Seasonality	Fourier	29.68
Unemp	Break	Prophet	0.42
Economic	Multi-var	VAR	0.93



Key Principle

Match the model to the data characteristics. No single model dominates—choose based on:

- Nature of the forecasting problem (level vs. volatility)
- Data properties (seasonality, breaks, multiple series)
- Interpretability requirements

Comprehensive Model Comparison

Feature	GARCH	Fourier	Prophet	VAR
Target	Volatility	Level	Level	Multiple
Seasonality	No	Yes (long)	Yes (multi)	No
Structural breaks	No	No	Yes	No
Multiple series	No	No	No	Yes
Interpretable	Medium	High	High	High
Parameters	Few	$2K$	Auto	Many
Missing data	No	No	Yes	No
Best for	Finance	Cycles	Business	Macro

Our Results

- GARCH: MAE=1.82 (volatility)
- Fourier: RMSE=29.68 (cycles)
- Prophet: RMSE=0.12 (breaks)
- VAR: Avg RMSE=0.93 (multi)

Key Insight

Each model excels in its domain. The art is matching the model to the data characteristics.

Best Practices for Applied Forecasting

Methodology

1. **Explore** data thoroughly
2. **Test** for stationarity
3. **Split** train/validation/test
4. **Compare** models on validation
5. **Report** test set metrics

Practical Tips

- Start simple (random walk, naive)
- Add complexity only if needed
- Visualize forecasts vs actuals
- Check residuals for patterns
- Report confidence intervals

Common Mistakes

- Peeking at test data
- Over-fitting to training set
- Ignoring model assumptions
- Not reporting uncertainty

Remember

"All models are wrong, but some are useful."

— George E. P. Box

Key Takeaways

1. **Rigorous Methodology**
    - Train/validation/test split prevents overfitting
    - Test set must remain untouched until final evaluation
2. **Match Model to Data**
    - Financial volatility $\rightarrow$ GARCH
    - Long seasonality $\rightarrow$ Fourier terms
    - Structural breaks $\rightarrow$ Prophet
    - Multiple series $\rightarrow$ VAR
3. **Interpret Results Carefully**
    - Granger causality $\neq$ true causality
    - Out-of-sample performance matters most
    - Simpler models often work better

References

- Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2015). *Time Series Analysis: Forecasting and Control*. 5th ed., Wiley.
- Hamilton, J.D. (1994). *Time Series Analysis*. Princeton University Press.
- Tsay, R.S. (2010). *Analysis of Financial Time Series*. 3rd ed., Wiley.
- Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. 3rd ed., OTexts.
- Taylor, S.J., & Letham, B. (2018). Forecasting at Scale. *The American Statistician*, 72(1), 37-45.
- Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, 31(3), 307-327.
- Sims, C.A. (1980). Macroeconomics and Reality. *Econometrica*, 48(1), 1-48.

Data Sources

Real Data Used in This Chapter

- **Bitcoin**: Yahoo Finance (BTC-USD), 2019–2025
- **Sunspots**: Statsmodels Wolfer dataset, 1900–2008
- **US Unemployment**: Federal Reserve FRED (UNRATE), 2010–2025
- **Economic Variables**: FRED (GDPC1, UNRATE, CPIAUCSL, FEDFUNDS), 2000–2025

Reproducibility

All analyses can be reproduced using the accompanying Jupyter notebook:
`chapter10_lecture_notebook.ipynb`

Thank You

Questions?

Prof. Daniel Traian Pele, PhD

danpele@ase.ro

Bucharest University of Economic Studies