Time Series Analysis and Forecasting

Chapter 10: Comprehensive Review

Daniel Traian PELE

Bucharest University of Economic Studies
IDA Institute Digital Assets
Blockchain Research Center
AI4EFin Artificial Intelligence for Energy Finance
Romanian Academy, Institute for Economic Forecasting
MSCA Digital Finance

Learning Objectives

By the end of this chapter, you will be able to:

- Apply the complete forecasting workflow from data to evaluation
- Select appropriate models based on data characteristics
- Evaluate forecast accuracy using proper metrics and cross-validation
- Integrate knowledge from all previous chapters in practice

Outline

The Scientific Approach to Forecasting

Research Question

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

The Fundamental Problem

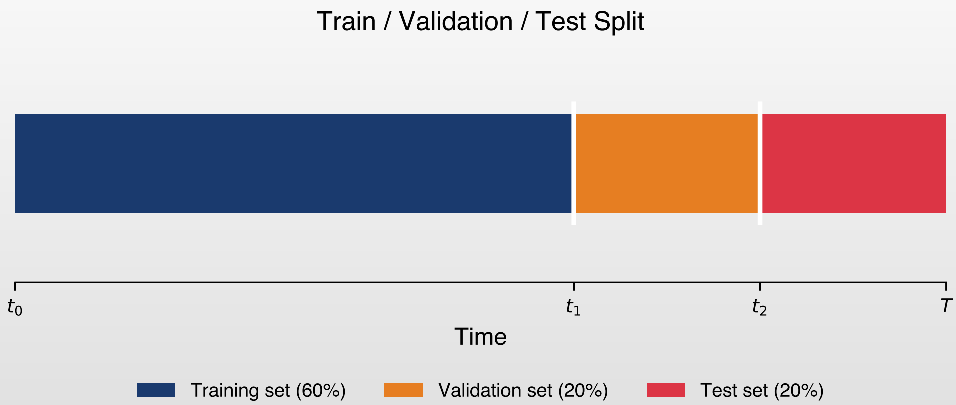
- In-sample fit $\neq$ Out-of-sample performance
- Models can "memorize" training data without learning patterns
- **Solution**:
  - Proper train/validation/test methodology

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

Validation Set

- Compare models

Test Set

- **Held out**

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

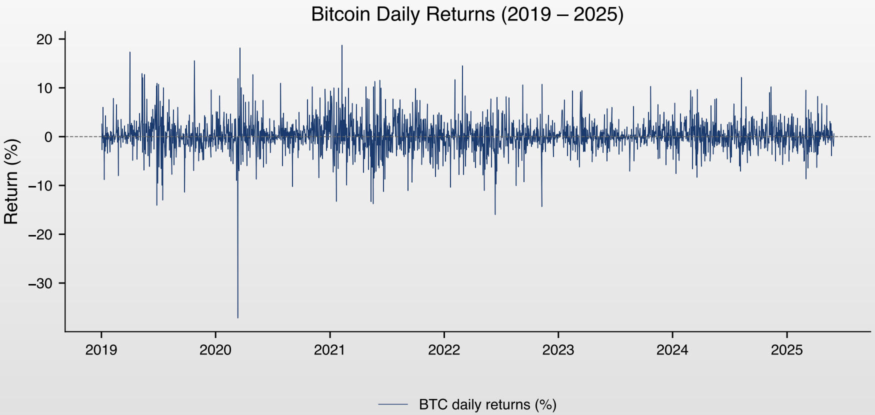
Key Insight

Financial returns are typically:

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

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

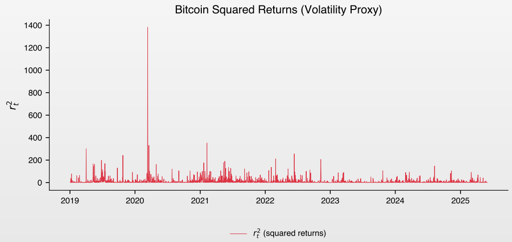
Bitcoin: Volatility Clustering



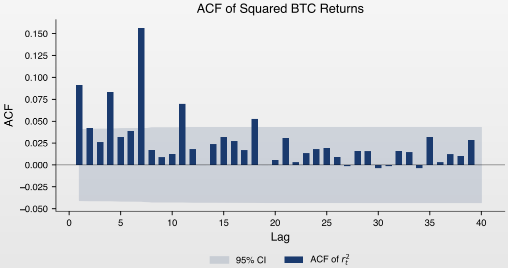
Observation

Large returns follow large returns, small follow small—**volatility clustering**.

Bitcoin: Evidence for GARCH



Squared returns r_t^2 proxy for volatility.



Significant ACF at multiple lags.

Why GARCH?

Significant ACF in r_t^2 means **past volatility predicts future volatility**.

GARCH Model Specification

Definition 2 (GARCH(p,q) Model)

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

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

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

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**: Log-variance, asymmetric

Interpretation

- α: Shock impact (ARCH effect)
- β: Volatility persistence
- $\alpha + \beta \approx 1$: High persistence

GARCH: Stationarity and Unconditional Variance

Theorem 1 (Covariance Stationarity of GARCH(1,1))

If $\alpha_1 + \beta_1 < 1$, then $\{\varepsilon_t\}$ is covariance stationary with:

$$\bar{\sigma}^2 = \mathbb{E}[\sigma_t^2] = \frac{\omega}{1 - \alpha_1 - \beta_1}$$

Derivation

Take expectations of both sides of the variance equation:

$$\begin{aligned} \mathbb{E}[\sigma_t^2] &= \omega + \alpha_1 \mathbb{E}[\varepsilon_{t-1}^2] + \beta_1 \mathbb{E}[\sigma_{t-1}^2] \\ \bar{\sigma}^2 &= \omega + (\alpha_1 + \beta_1)\bar{\sigma}^2 \quad \text{(stationarity)} \\ \bar{\sigma}^2 &= \frac{\omega}{1 - \alpha_1 - \beta_1} \end{aligned}$$

Multi-Step Forecasts Converge to $\bar{\sigma}^2$

Bitcoin: Data Split and Stationarity

Data Split

Set	Period	N
Training (70%)	2019-01 to 2023-03	1,543
Validation (20%)	2023-03 to 2024-06	441
Test (10%)	2024-06 to 2025-01	221
Total		**2,205**

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**.

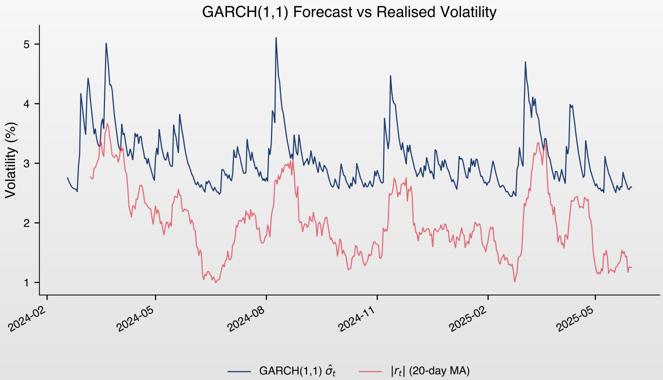
Model	AIC	BIC	Val MAE	Selection
GARCH(1,1)	6,994.8	7,020.6	**2.638**	**Best**
GARCH(2,1)	6,993.7	7,024.6	2.640	
GJR-GARCH(1,1)	6,983.7	7,014.6	2.669	
EGARCH(1,1)	—	—	—	Failed*

*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

$\omega = 0.87$, $\alpha = 0.09$, $\beta = 0.84$

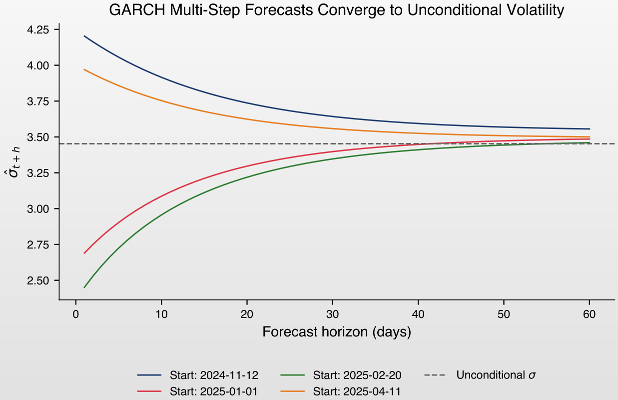
$\alpha + \beta = 0.93$ (high persistence)

Test Performance

MAE $= 1.82$, RMSE $= 2.14$

Rolling forecasts track volatility well.

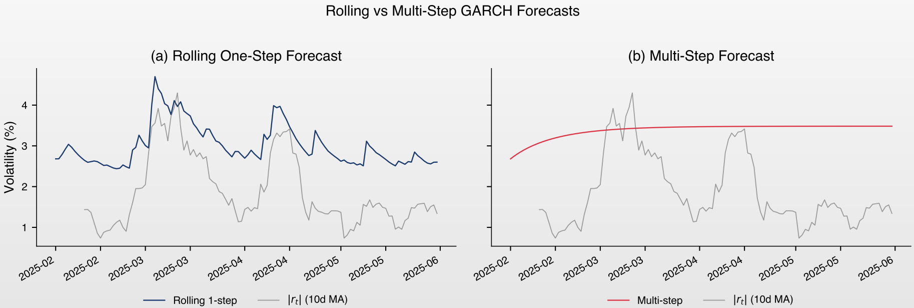
GARCH: Multi-Step Forecasts Converge



Key Insight

Multi-step forecasts converge to $\bar{\sigma}^2 = \frac{\omega}{1-\alpha-\beta}$. Use rolling 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)

TSA_ch10_rolling_vs_multistep

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

Practical Implications

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

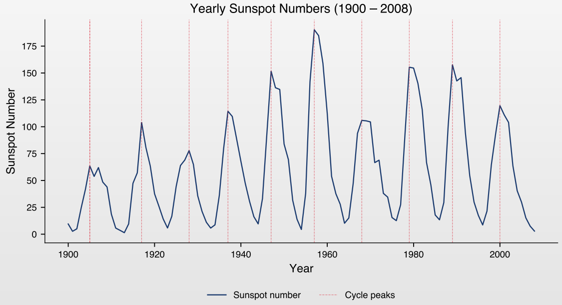
Limitations

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

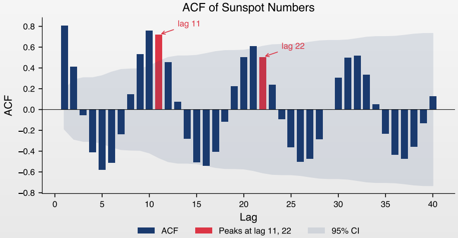
Extensions

- Student-t innovations
- Realized volatility
- HAR models

Sunspots: The 11-Year Solar Cycle



Cycle peaks every 11 years.



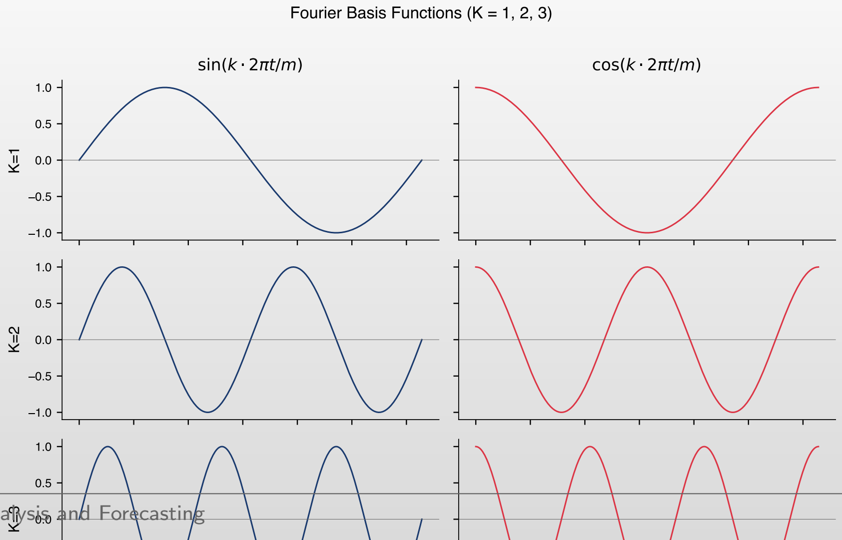
ACF peaks at lag 11, 22.

Challenge

SARIMA$_{11}$ requires too many parameters. **Solution**: Use Fourier terms.

TSA_ch10_sunspots_acf

Fourier Terms for Seasonality



Sunspots: Model Selection

Methodology

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

Data Split

Set	Period	N
Training (70%)	1900–1975	76
Validation (20%)	1976–1997	22
Test (10%)	1998–2008	11
Total		**109**

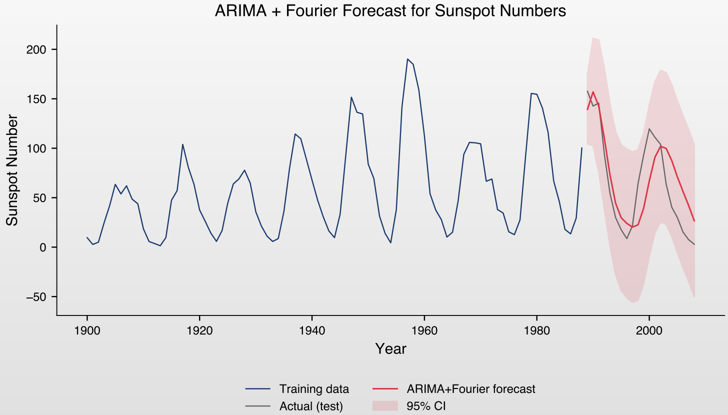
Model Comparison

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.

Test Performance

RMSE = 31.10, MAE = 25.83.

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

Choosing K

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

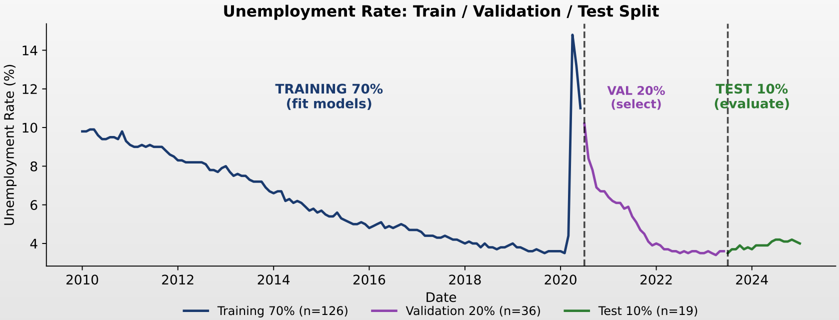
Fourier vs SARIMA

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

Applications

Climate cycles, business cycles, astronomical phenomena

Unemployment: Train / Validation / Test Split

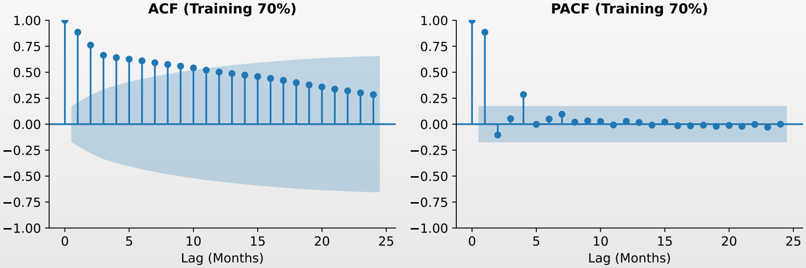


Methodology

Training: Fit models. **Validation**: Select best. **Test**: Final evaluation.

TSA_ch10_unemployment_train_val_test

Unemployment: Preliminary Analysis



ACF Interpretation

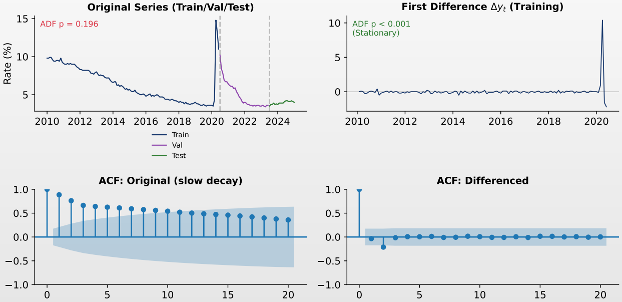
Slow decay $\Rightarrow$ non-stationary.

PACF Interpretation

Spike at lag 1 $\Rightarrow$ AR(1) component.

TSA_ch10_unemployment_acf_pacf

Unemployment: Stationarity Tests



Original: ADF p = 0.056

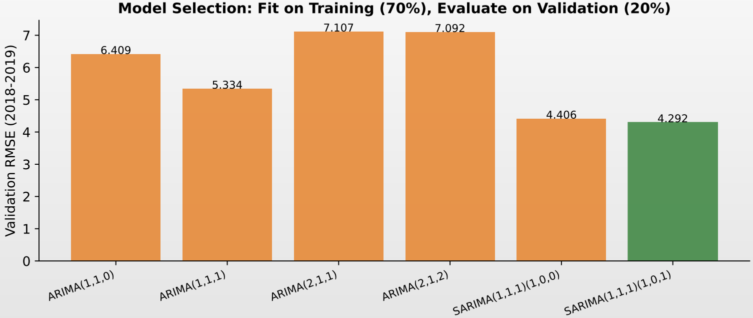
Non-stationary (slow ACF decay)

Differenced: ADF p < 0.001

Stationary $\Rightarrow$ use $d = 1$

TSA_ch10_unemployment_stationarity

Unemployment: Model Selection (Validation Set)



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

Selected by lowest validation RMSE.

TSA_ch10_sarima_model_selection

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.)

TSA_ch10_sarima_parameters

Ljung-Box Test for Residual Autocorrelation

Definition 3 (Ljung-Box Test)

For residuals $\hat{\varepsilon}_t$ with sample autocorrelations $\hat{\rho}_k$, the test statistic:

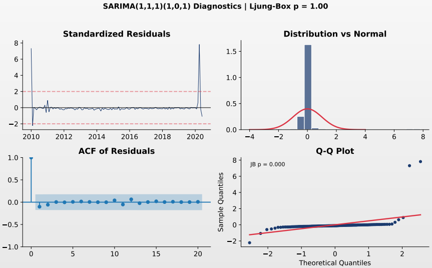
$$Q(h) = n(n+2)\sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k} \overset{H_0}{\sim} \chi^2(h-p-q)$$

where p, q are ARMA orders. H_0: Residuals are white noise.

Interpretation

- Large Q (small p-value): Reject H_0, residuals have structure
- Small Q (large p-value): Fail to reject H_0, model is adequate
- Rule of thumb: Use $h = \min(10, n/5)$ for lag order

Unemployment: SARIMA Diagnostics



Residual Checks

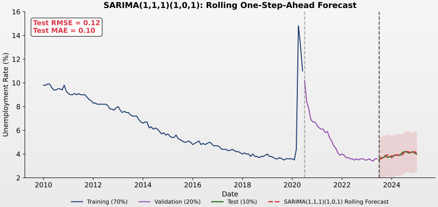
Histogram, ACF, Q-Q plot for normality.

Ljung-Box: $p = 0.66$

Fail to reject $H_0 \Rightarrow$ No remaining autocorrelation.

TSA_ch10_sarima_diagnostics

Unemployment: SARIMA Rolling Forecast



Problem: Structural Break

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

TSA_ch10_sarima_forecast

Prophet Model

Definition 4 (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 (70%)	2010-01 to 2020-06	126
Validation (20%)	2020-07 to 2023-06	36
Test (10%)	2023-07 to 2025-01	19
Total		**181**

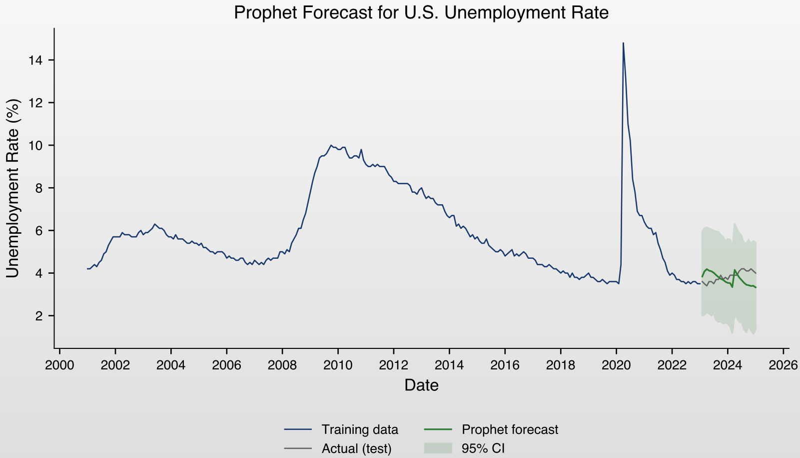
Scale Comparison

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.

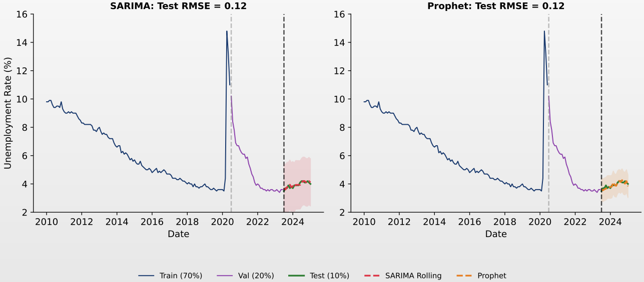
Unemployment: Prophet Forecast Results



Key Finding

Prophet adapts via changepoint detection. **Test RMSE = 0.58.**

Unemployment: SARIMA vs Prophet Comparison



SARIMA: RMSE = 0.12

Rolling forecast performs well.

Prophet: RMSE = 0.58

Higher error due to structural break.

TSA_ch10_prophet_vs_sarima_unemployment

Prophet: When to Use It

Ideal Use Cases

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

Caveat: Structural Breaks

Prophet handles breaks via changepoints, but **SARIMA outperformed** it on unemployment (0.12 vs 0.58). Always validate!

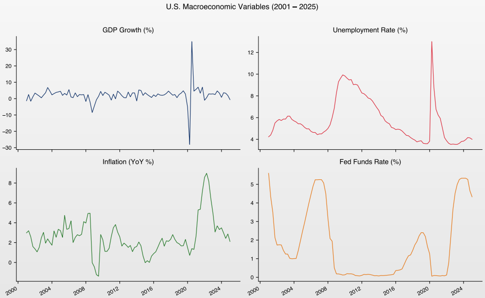
Prophet vs ARIMA

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

Key Parameters

`changepoint_prior_scale`: flexibility
`seasonality_prior_scale`: smoothness

VAR: Multivariate Economic Data



Relationships

GDP ↔ Unemployment (Okun)

Why VAR?

Each variable is cause and effect.

TSA_ch10_economic_vars

VAR Model Specification

Definition 5 (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

Information Criteria for Model Selection

Definition 6 (Akaike and Bayesian Information Criteria)

For a model with log-likelihood $\mathcal{L}$, k parameters, and n observations:

$$\text{AIC} = -2\mathcal{L} + 2k$$
$$\text{BIC} = -2\mathcal{L} + k\ln(n)$$

AIC

- Asymptotically efficient
- May overfit with small n
- Minimizes prediction error

BIC

- Consistent (finds true model)
- Heavier penalty: $\ln(n) > 2$ if $n > 7$
- More parsimonious

VAR: Lag Selection and Estimation

BIC by Lag Order

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

Data Split

Set	Period	N
Training (70%)	2001-Q1 to 2017-Q4	67
Validation (20%)	2018-Q1 to 2022-Q4	20
Test (10%)	2023-Q1 to 2025-Q1	10
Total		**97**

Validation Check

VAR(2) also achieves lowest validation RMSE.

Granger Causality: Formal Definition

Definition 7 (Granger Causality)

X **Granger-causes** Y if, for some $h > 0$:

$$\text{MSE}\left[\mathbb{E}[Y_{t+h}|\mathcal{F}_t^{X,Y}]\right] < \text{MSE}\left[\mathbb{E}[Y_{t+h}|\mathcal{F}_t^{Y}]\right]$$

where $\mathcal{F}_t^{X,Y}$ includes past values of both X and Y, while $\mathcal{F}_t^{Y}$ includes only past Y.

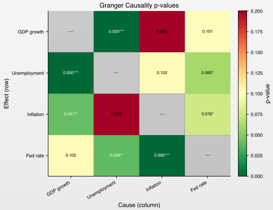
Important Caveat

Granger causality is **predictive causality**, not true causality. "X Granger-causes Y" means X contains useful information for forecasting Y, not that X causes Y in a structural sense.

Test Procedure

Use F-test (or Wald test) to test H_0: coefficients on lagged X are jointly zero in the Y equation.

Granger Causality: Empirical Results



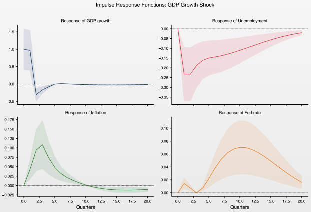
Interpretation

Each cell shows p-value for testing whether the row variable Granger-causes the column variable. Green: $p < 0.10$. Read: row causes column.

Economic Findings

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

Impulse Response Functions (IRF)



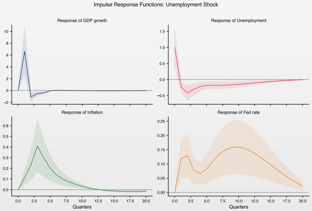
What is IRF?

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

GDP Shock Effects

- **Unemp** ↓: Okun's Law
- **Inflation** ↑: Demand-pull
- **Fed Rate** ↑: Taylor Rule

IRF: Unemployment Shock

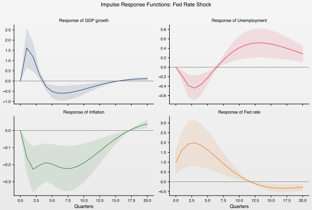


Effects

$\uparrow$ Unemp $\Rightarrow$ $\downarrow$ GDP, $\downarrow$ Infl, Fed cuts rates.

TSA_ch10_irf_unemp_shock

IRF: Fed Rate Shock

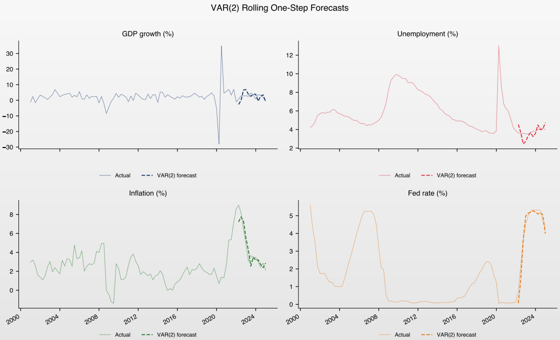


Monetary Policy

Rate hike $\Rightarrow$ GDP $\downarrow$, Unemp $\uparrow$, Infl $\downarrow$.

TSA_ch10_irf_fed_shock

VAR: Forecast (Train/Val/Test)



Rolling One-Step-Ahead Forecast

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

TSA_ch10_var_forecast

VAR: Test Set Results

Test Set Performance by Variable

Variable	RMSE	MAE	Dir. Acc.
GDP Growth	1.33	0.99	50%
Unemployment	0.64	0.52	50%
Inflation	1.56	1.12	60%
Fed Rate	2.59	2.45	80%
Average	**1.53**	**1.27**	**60%**

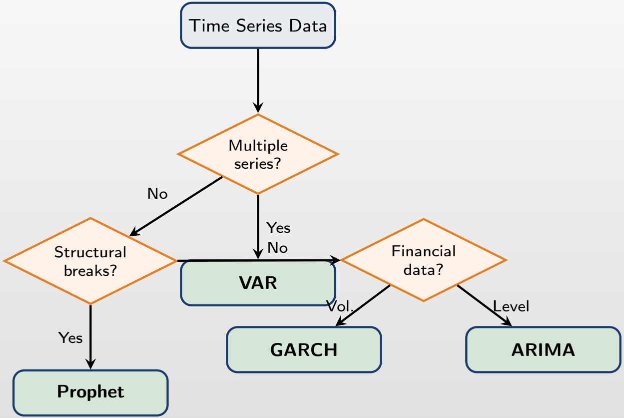
Strengths

- Cross-variable dynamics
- Good directional accuracy

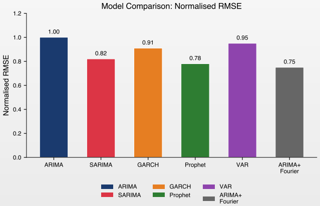
Limitations

- Many parameters
- Sensitive to lag selection

Model Selection Framework



Summary: Model Comparison



Results

Case	Challenge	Model	RMSE
Bitcoin	Volatility	GARCH	2.15
Sunspots	Seasonality	Fourier	31.10
Unemp	Break	SARIMA	0.12
Economic	Multi-var	VAR	1.53

Key Principle

Match model to data characteristics—no single model dominates

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=31.10 (cycles)
- SARIMA: RMSE=0.12 (breaks)
- VAR: Avg RMSE=1.53 (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
2. **Test** stationarity
3. **Split** train/val/test
4. **Compare** on validation
5. **Report** test metrics

Common Mistakes

- Peeking at test data
- Over-fitting
- Ignoring assumptions

Practical Tips

- Start simple (naive)
- Add complexity if needed
- Check residuals
- Report CIs

Remember

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

Key Takeaways

What We Learned

- Model selection depends on data characteristics: stationarity, seasonality, volatility
- The Box-Jenkins methodology provides a systematic framework for time series modeling
- Proper evaluation requires out-of-sample testing and time series cross-validation

Important

No single model wins everywhere. Match the model to the data: ARIMA for trends, SARIMA for seasonality, GARCH for volatility, VAR/VECM for multivariate dynamics, Prophet/TBATS for complex patterns. Always validate out-of-sample!

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`

Online Resources and Code

- **Quantlet**: https://quantlet.com → Code repository for statistics
- **Quantinar**: https://quantinar.com → Learning platform for quantitative methods
- **GitHub TSA_ch10**: https://github.com/QuantLet/TSA/tree/main/TSA_ch10

Thank You!

Questions?

Course materials available at: https://danpele.github.io/Time-Series-Analysis/

Quantlet Quantinar