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?

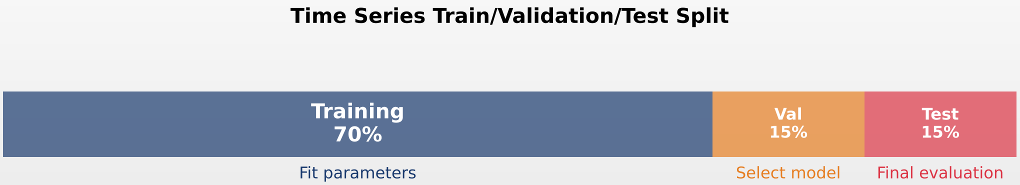
The Fundamental Problem

- 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



TSA_ch10_train_val_test_split

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

Forecast Evaluation Beyond RMSE

Alternative metrics

- **MASE** (Mean Absolute Scaled Error): $\frac{\text{MAE}_{\text{model}}}{\text{MAE}_{\text{naïve}}}$; $< 1 \Rightarrow$ beats naïve
- **DA** (Directional Accuracy): $\frac{1}{h}\sum_{t=1}^{h} 1(\operatorname{sgn}\Delta\hat{y}_t = \operatorname{sgn}\Delta y_t)$
- **QL** (Quantile Loss): asymmetric penalty α vs $1-\alpha$

$$QL_\alpha = \begin{cases} \alpha(y_t - \hat{q}_t), & y_t > \hat{q}_t \\ (1-\alpha)(\hat{q}_t - y_t), & y_t \leq \hat{q}_t \end{cases}$$

- **CRPS** (Continuous Ranked Probability Score): $\int_{-\infty}^{\infty} \left(F(x) - 1_{x \geq y}\right)^2 dx$

Forecast Evaluation: Bitcoin Results

Bitcoin Results (GARCH volatility)

Metric	Value
RMSE	2.21
MAE	1.89
MASE	0.98
Dir. Accuracy	28.7%

- MASE < 1: GARCH beats naïve
- DA 28.7%: volatility direction is hard

Interpretation

- **RMSE/MAE**: absolute volatility forecast error
- **MASE** < 1: GARCH outperforms naïve benchmark
- **DA 28.7%**: volatility direction is extremely hard to predict
- Evaluation must be done on the **test set**

Formal Forecast Comparison: Diebold–Mariano

Definition 2 (Diebold–Mariano Test)

Loss differential: $d_t = L(e_{1t}) - L(e_{2t})$, Statistic: $DM = \frac{\bar{d}}{\sqrt{\widehat{\text{Var}}(\bar{d})}} \xrightarrow{d} N(0,1)$

Hypotheses

- H_0: equal predictive accuracy
- H_1: one model is significantly better
- Large $|DM| \Rightarrow$ reject H_0

Bitcoin Result (GARCH volatility)

- Normal vs Student-t: $DM = -0.51$
- $p = 0.612$ — **do not reject** H_0
- Similar accuracy, but Student-t preferred by AIC ($\Delta = 509$)

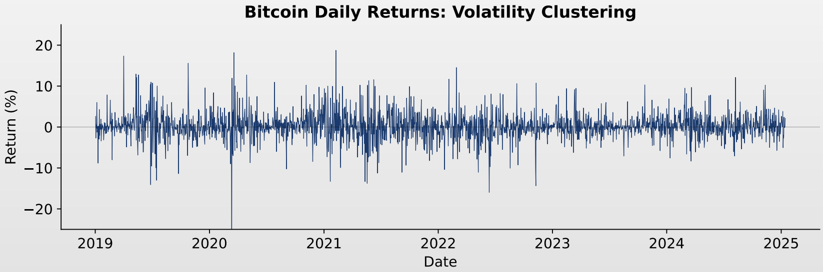
Key message

- Lower RMSE $\neq$ significant difference — formal testing is **mandatory**

Bitcoin: Volatility Clustering

Observation

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



TSA_ch10_btc_returns

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:

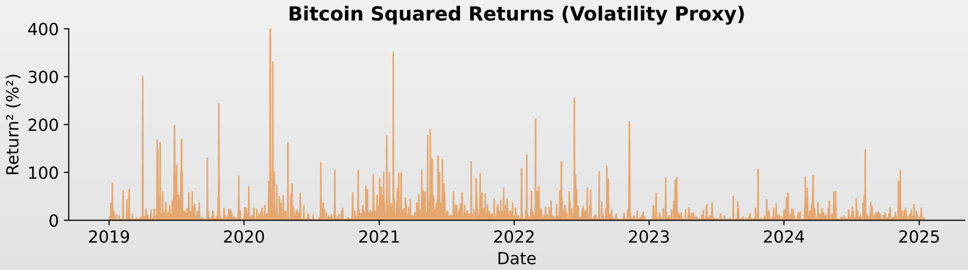
- **Unpredictable** in mean
- **Predictable** in variance

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

Bitcoin: Evidence for GARCH

Observation

- Squared returns r_t^2 exhibit significant autocorrelation $\succ$ GARCH effects
- Slow ACF decay $\succ$ high volatility persistence



TSA_ch10_btc_acf_squared

GARCH Model Specification

Definition 3 (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 $\bar{\sigma}^2$

As $h \to \infty$: $\mathbb{E}_t[\sigma_{t+h}^2] \to \bar{\sigma}^2$ at rate $(\alpha_1 + \beta_1)^h$.

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: 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

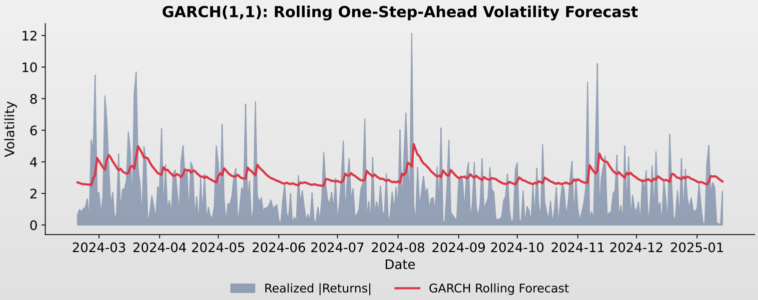
Why Stationarity Matters

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

Bitcoin: Volatility Forecast

Interpretation

- Shaded area: 95% confidence interval of the volatility forecast
- GARCH(1,1) captures Bitcoin's volatility dynamics well

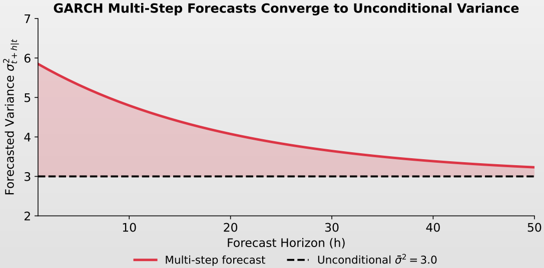


TSA_ch10_garch_forecast

GARCH: Multi-Step Forecasts Converge

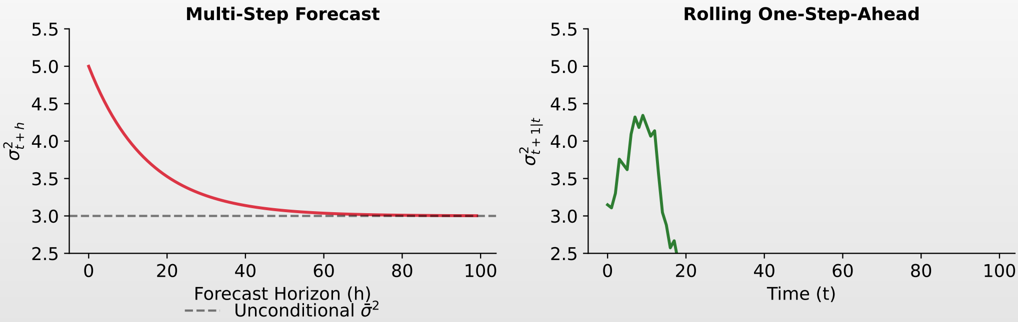
Key Insight

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



TSA_ch10_garch_convergence

GARCH: Rolling One-Step-Ahead Solution



TSA_ch10_rolling_vs_multistep

GARCH: Innovation Distributions

Model

$$r_t = \mu + \sigma_t z_t$$

- Options for z_t: $\mathcal{N}(0,1)$ (normal) or t_ν (fat tails)

Bitcoin: empirical evidence

- Residual kurtosis: **13.81** (Normal = 3)
- Skewness: -0.29
- Jarque-Bera: 9085, $p < 0.001$
- Normality **underestimates** tail risk

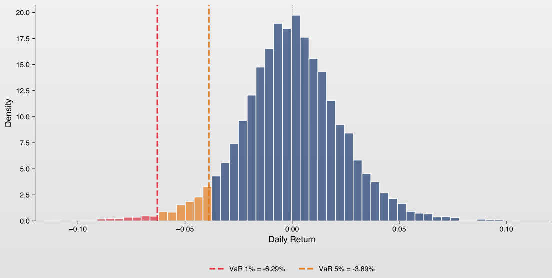
Student-t: the right choice

- $\hat{\nu} = 2.96$ degrees of freedom
- AIC Normal: 9769 vs Student-t: **9260**
- ΔAIC = 509 — **overwhelming** evidence
- Fat tails = **more realistic** VaR estimates

VaR and ES: Graphical Illustration

Interpretation

- VaR 1% = loss exceeded only in 1% of cases
- Red area = extreme losses (beyond VaR)

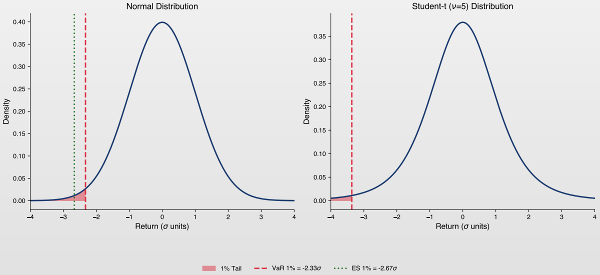


TSA_ch5_var_plot

VaR vs Expected Shortfall: Normal vs Student-t

Interpretation

- ES measures average loss when VaR is exceeded
- Student-t: VaR and ES are larger than under normal distribution



TSA_ch5_var_es

Value at Risk — Numerical Example

VaR Calculation

Portfolio: **1,000,000 EUR**, forecasted volatility $\hat{\sigma}_{T+1} = 1.5\%$

VaR with Normal Distribution

Level	z_α	**VaR (%)**	**VaR (EUR)**
5% (1 day)	1.645	2.47%	24,675
1% (1 day)	2.326	3.49%	34,890

Scaling for Longer Periods

$\text{VaR}_{h \text{ days}} = \text{VaR}_{1 \text{ day}} \cdot \sqrt{h}$ — assumes i.i.d. returns

Value at Risk — Student-t Distribution

Why Student-t?

Normal distribution **underestimates** tail risk. Student-t with ν degrees of freedom better captures fat tails (kurtosis > 3).

VaR 1% (1 day) Comparison: $\sigma = 1.5\%$, Portfolio = 1M EUR

Distribution	Quantile	VaR (EUR)
Normal	2.326	34,890
Student-t ($\nu = 6$)	3.143	47,145
Student-t ($\nu = 4$)	3.747	56,205

Observation

With $\nu = 6$ (typical for stocks), VaR is **35% higher** than normal!

VaR — Complete Example with GARCH

VaR Calculation Procedure

1. Estimate GARCH(1,1) model with Student-t distribution
2. Obtain volatility forecast: $\hat{\sigma}_{T+1}$
3. Calculate VaR: $\text{VaR}_\alpha = t_\alpha(\nu) \cdot \hat{\sigma}_{T+1} \cdot \sqrt{\frac{\nu-2}{\nu}}$

Example: S&P 500

- Estimated parameters: $\alpha = 0.088$, $\beta = 0.900$, $\nu = 6.4$
- Forecasted volatility: $\hat{\sigma}_{T+1} = 1.2\%$
- Portfolio: 10,000,000 EUR

VaR 1% (1 day): $\text{VaR} = 3.05 \times 0.012 \times 10{,}000{,}000 =$ **366,000 EUR**

What is VaR Backtesting?

Definition

- **Backtesting** = ex-post verification of VaR model quality
- Compares realized losses with the forecasted VaR threshold
  - A **violation** occurs when $r_t < -\mathrm{VaR}_t$

Backtesting Principle

- Violation indicator: $I_t = 1(r_t < -\mathrm{VaR}_{\alpha,t})$
- For a correctly specified model at level α:
  - Frequency: $\hat{p} = \frac{1}{T}\sum I_t \approx \alpha$; violations **independent**
- VaR 1% over 250 days $\Rightarrow$ expect $\sim$2.5 violations/year

Importance

- Regulatory requirement under **Basel III/IV** for banks: backtesting is mandatory

Kupiec Test (1995) — Unconditional Coverage

Hypotheses

- H_0: Violation rate equals the VaR level ($p = \alpha$)
- H_1: Violation rate differs from the VaR level ($p \neq \alpha$)

Test Statistic (Likelihood Ratio)

- **Formula**: $LR_{uc} = -2\ln\left[\frac{\alpha^x(1-\alpha)^{T-x}}{\hat{p}^x(1-\hat{p})^{T-x}}\right] \sim \chi^2(1)$
- **Notation**: x = no. violations, T = no. observations, $\hat{p} = x/T$

Example

- VaR 1%, $T = 250$ days, $x = 5$ violations: $\hat{p} = 2\%$
  - Too many violations $\Rightarrow$ model **underestimates** risk
- VaR 1%, $T = 250$ days, $x = 1$ violation: $\hat{p} = 0.4\% \Rightarrow$ acceptable

Christoffersen Test (1998) — Conditional Coverage

Motivation

- Kupiec only tests the **frequency** of violations
- Does not detect **clustering** of violations (consecutive violations)
  - If violations cluster $\Rightarrow$ model fails to capture volatility dynamics

Independence + Conditional Coverage Test

- **Formula**: $LR_{cc} = LR_{uc} + LR_{ind} \sim \chi^2(2)$
- LR_{ind} tests whether $P(I_t = 1 | I_{t-1} = 1) = P(I_t = 1 | I_{t-1} = 0)$
- A good model: violations are rare **and** uniformly distributed over time

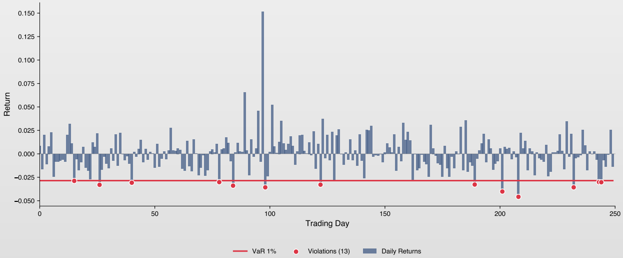
Recommendation

- Use **both** tests: Kupiec (frequency) + Christoffersen (independence)

VaR Backtesting: Visualization

Interpretation

- Red line: VaR 1% threshold estimated with GARCH(1,1)
- Red dots: 13 violations out of 250 days ($\hat{p} = 5.2\%$)
  - **Basel red zone** $\Rightarrow$ model significantly underestimates risk
  - Solutions: Student-t distribution, EGARCH model, or more conservative VaR level



TSA_ch5_backtest

VaR Backtesting: Basel Traffic Light

Basel III/IV Traffic Light Zones

Zone	Violations/250 days	Interpretation	Penalty
Green	0–4	Model acceptable	No penalty
Yellow	5–9	Needs investigation	Factor k increases
Red	≥ 10	Model inadequate	Maximum penalty

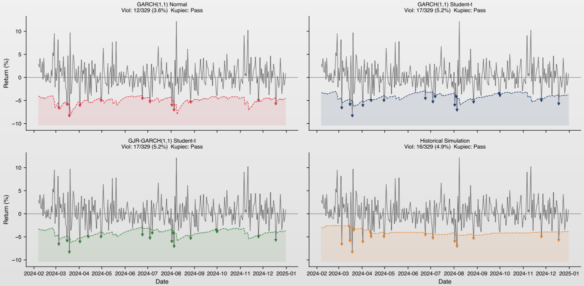
Practical Example

- Portfolio with VaR 1%: 250 days of backtesting
- 3 violations $\Rightarrow$ **Green zone** $\Rightarrow$ model acceptable
- 7 violations $\Rightarrow$ **Yellow zone** $\Rightarrow$ revision needed
- 13 violations $\Rightarrow$ **Red zone** $\Rightarrow$ model rejected

Application: Rolling VaR with Multiple Models

Methodology

- Rolling one-step-ahead VaR: $\text{VaR}^{\alpha}_{t+1} = \mu + \hat{\sigma}_{t+1} \cdot z_{\alpha}$
- 4 models compared on Bitcoin test set (329 days, 2024)



TSA_ch10_empirical_slides

VaR Backtesting: Model Comparison

Bitcoin Results — VaR 5% (T = 329 days, expected: 16.5 violations)

Model	AIC	Violations	Rate	Kupiec p	Chr. p	Conclusion
GARCH(1,1)-N	9769	12	3.6%	0.238	1.000	Too conservative
GARCH(1,1)-t	**9260**	**17**	**5.2%**	**0.890**	0.272	**Best**
GJR-GARCH(1,1)-t	9260	17	5.2%	0.890	0.272	$\gamma \approx 0$ (symmetric)
Hist. Simulation	—	16	4.9%	0.909	0.216	Good, non-parametric

Conclusions

- **Student-t**: perfect coverage (5.2% $\approx$ 5%)
- Normal: too conservative (3.6% $<$ 5%)
- GJR $\approx$ GARCH: no leverage for Bitcoin
- All pass both Kupiec **and** Christoffersen

Practical Lessons

- Innovation distribution matters: **ΔAIC = 509**
- Historical simulation: simple and effective alternative
- Formal statistical testing (Kupiec, Christoffersen) is **mandatory**

GARCH Limitations and Modern Extensions

Limitations

- Does not capture **jumps**
- Constant parameters over time
- Sensitive to chosen distribution
- Does not model different **regimes**

Extensions

- **GJR-GARCH**: leverage effect
- **EGARCH**: asymmetric shocks
- **Markov-Switching GARCH**: regimes
- Realized volatility (HAR)
- Hybrid GARCH + ML

Key message

- GARCH is a **starting point**, not the end of risk modeling

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

Practical Implications

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

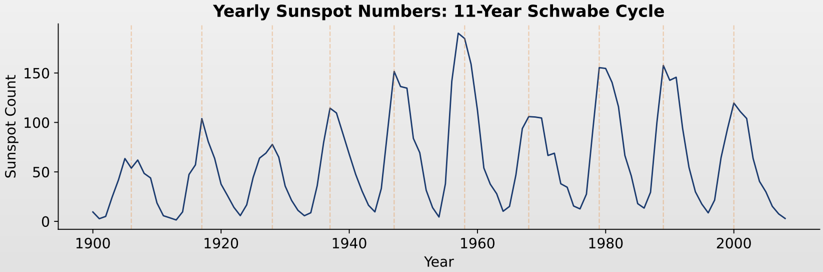
Extensions

- Student-t innovations
- Realized volatility
- HAR models

Sunspots: The 11-Year Solar Cycle

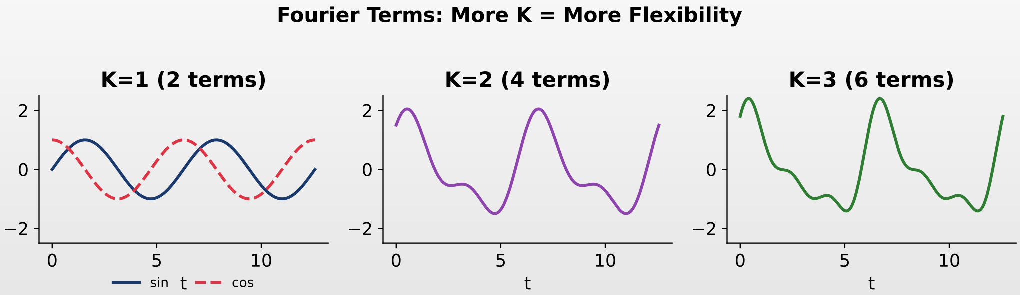
Observation

- Clear ≈ 11-year solar cycle; variable amplitude across cycles
- Periodic ACF ≻ long seasonality, ideal for Fourier terms



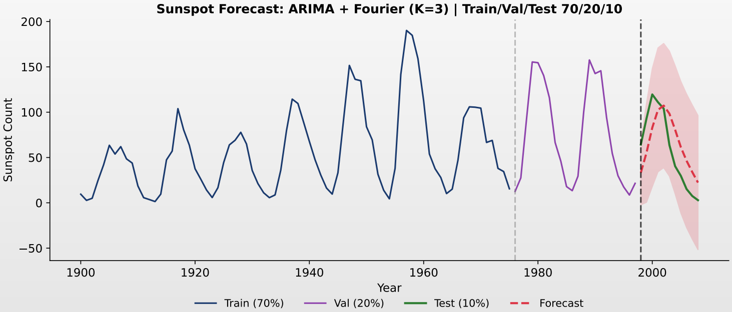
TSA_ch10_sunspots_acf

Fourier Terms for Seasonality



TSA_ch10_fourier_terms

Sunspots: Forecast Results



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).

Overfitting in Choosing K

Overfitting risk

- K too large = memorizing historical cycle
- Model fits noise, not signal
- Test performance **degrades**

Fourier ≈ periodic regression

- Each harmonic adds 2 parameters (sin, cos)
- $K = 3$: 6 extra parameters
- $K = 6$: 12 parameters — overfitting risk

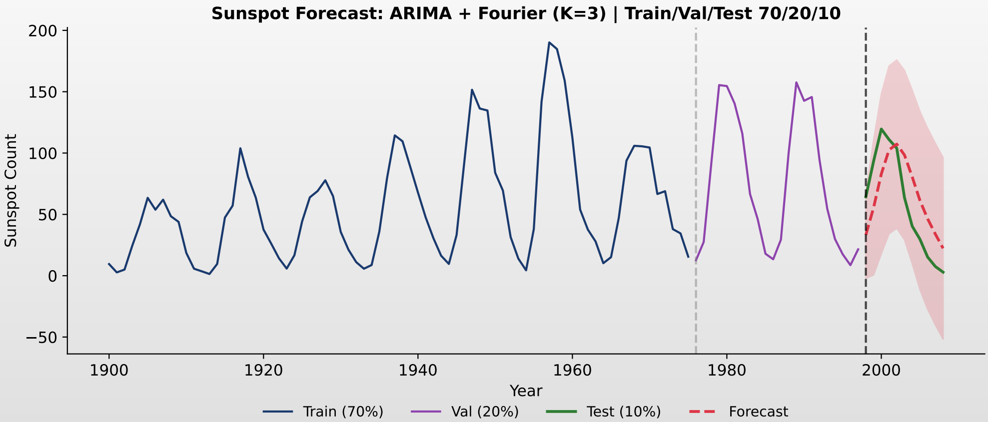
Solution: validation

- Select K on **validation** set
- Evaluate on **test** — untouched
- Trade-off: complexity vs generalization

Our results

- $K = 3$ minimizes Val RMSE
- $K = 4$ increases error → overfitting

Sunspots: Forecast Results



TSA_ch10_sunspot_forecast

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

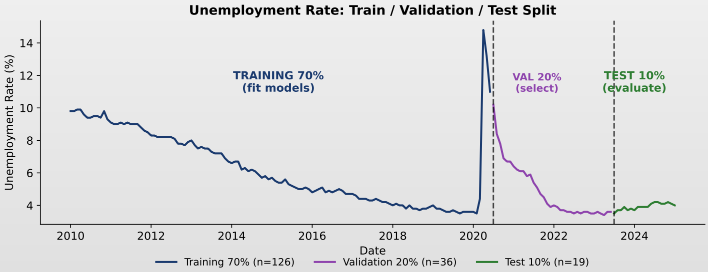
Applications

Climate cycles, business cycles, astronomical phenomena

Unemployment: Train / Validation / Test Split

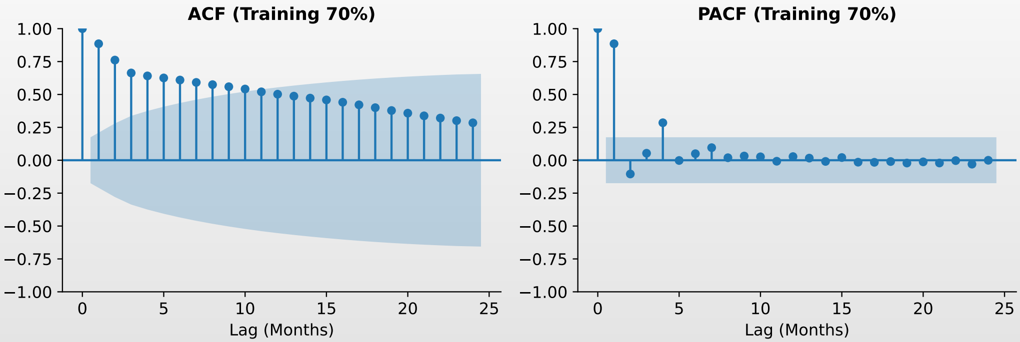
Methodology

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



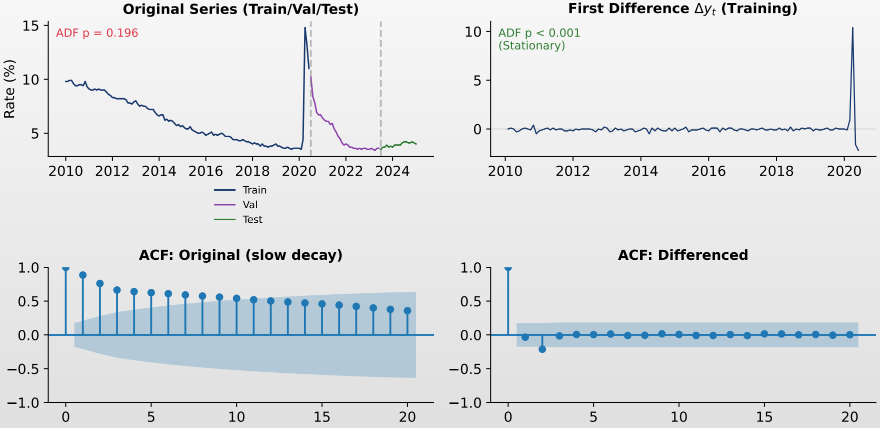
TSA_ch10_unemployment_train_val_test

Unemployment: Preliminary Analysis



TSA_ch10_unemployment_acf_pacf

Unemployment: Stationarity Tests



TSA_ch10_unemployment_stationarity

Structural Breaks: Formal Approach

Classical methods

- **Chow Test**: break at known point
- **Bai–Perron**: multiple unknown breaks
- **CUSUM**: sequential detection

Problem

- ADF can confuse **break** with **unit root**
- Zivot–Andrews test: ADF with endogenous break

Result: Unemployment at COVID (March 2020)

- Chow Test: $F = 21.73$, $p < 0.001$
- Structural break **confirmed**
- SARIMA: constant parameters — risk
- Prophet: detects changepoints automatically

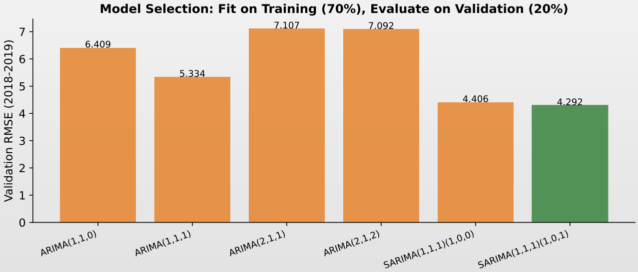
Key message

- Model must be adapted to **parameter stability**

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)(1,0,0)$_{12}$ 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.)

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	***

TSA_ch10_sarima_parameters

Ljung-Box Test for Residual Autocorrelation

Definition 4 (Ljung-Box Test)

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

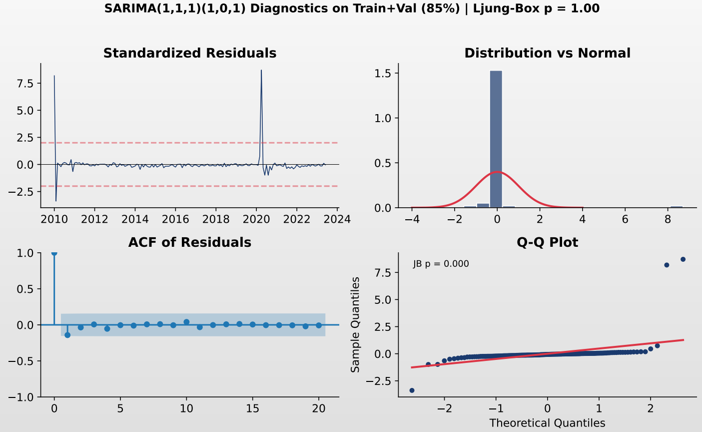
$$Q(h) = n(n+2)\sum_{k=1}^{h} \frac{\hat{\rho}_k^2}{n-k} \overset{H_0}{\sim} \chi^2(h-p-q)$$

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

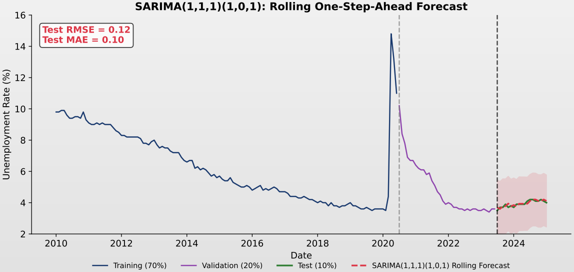


TSA_ch10_sarima_diagnostics

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 5 (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

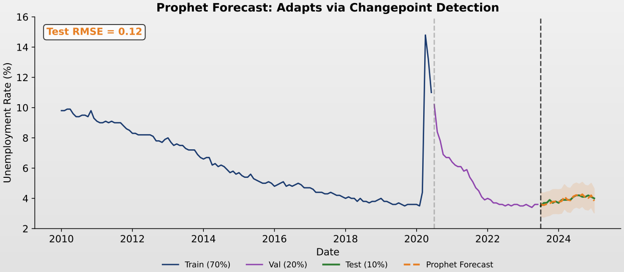
Advantages

- Handles missing data
- Interpretable components
- Robust to outliers

Unemployment: Prophet Forecast Results

Key Finding

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



TSA_ch10_unemployment_forecast

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**

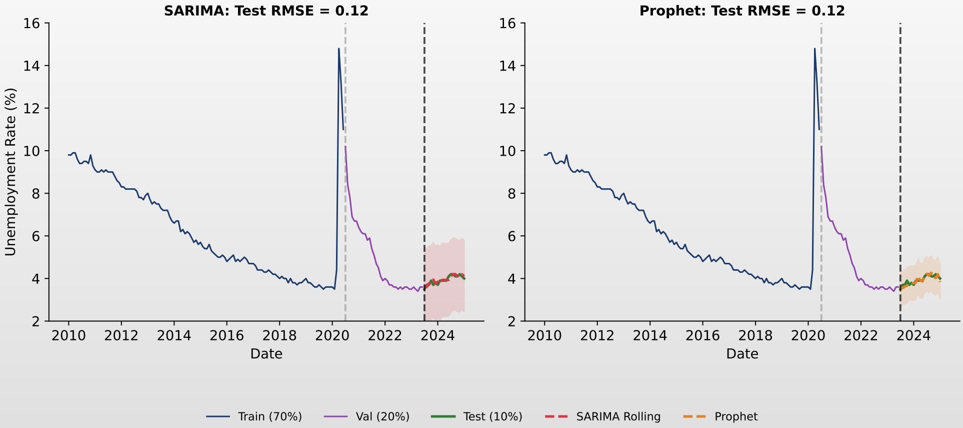
Scale Comparison

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: SARIMA vs Prophet Comparison



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!

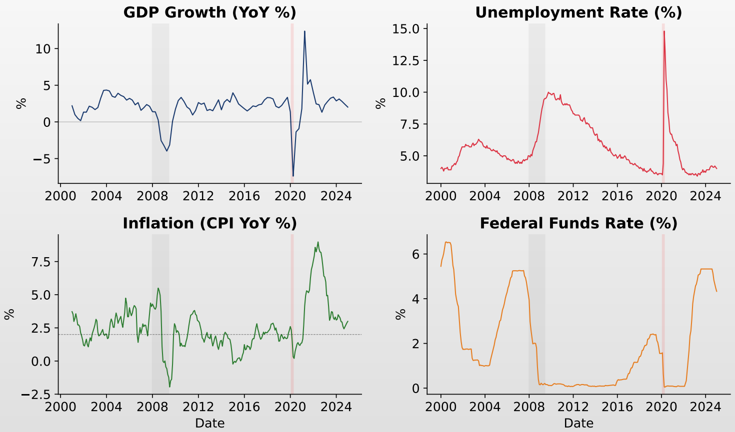
Prophet vs ARIMA

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

Key Parameters

`changepoint_prior_scale`: flexibility
`seasonality_prior_scale`: smoothness

VAR: Multivariate Economic Data



TSA_ch10_economic_vars

VAR Model Specification

Definition 6 (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 7 (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.

VAR Model Stability

Stability condition

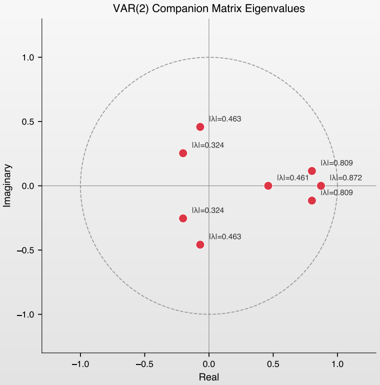
- All eigenvalues of the **companion matrix**:

$$|\lambda_i| < 1, \quad \forall i$$

VAR(2) Results — economic data

$\lvert\lambda_1\rvert, \lvert\lambda_2\rvert$	0.324
$\lvert\lambda_3\rvert, \lvert\lambda_4\rvert$	0.463
$\lvert\lambda_5\rvert$	0.461
$\lvert\lambda_6\rvert$	**0.872**
$\lvert\lambda_7\rvert, \lvert\lambda_8\rvert$	0.810

- Max $|\lambda| = 0.872 < 1$ — **stable**



TSA_ch10_empirical_slides

VAR vs VECM: Cointegration

Problem

- If variables are $I(1) \succ$ VAR on levels produces spurious regressions

Definition 8 (VECM)

$$\Delta y_t = \Pi y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta y_{t-i} + u_t, \quad \Pi = \alpha\beta'$$

Johansen Test — economic data

r	Trace	CV 5%	Reject?
0	**64.09**	47.85	**Yes**
1	24.03	29.80	No
2	11.89	15.49	No
3	1.28	3.84	No

- **1 cointegrating relation** found
- VECM more appropriate than VAR on levels

Key message

- VAR on differences: loses long-run relationship; VECM: preserves it through $\Pi = \alpha\beta'$

Granger Causality: Empirical Results

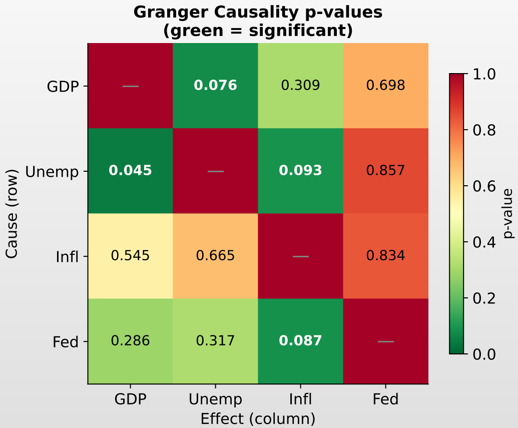
Interpretation

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

Granger Causality: Empirical Results



TSA_ch10_granger_heatmap

Granger Causality: Formal Definition

Definition 9 (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.

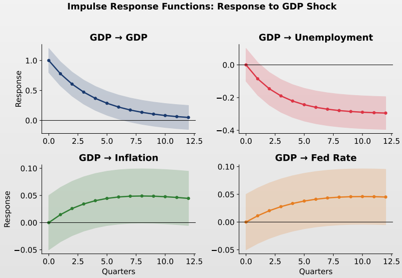
Test Procedure

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

Impulse Response Functions (IRF)

Effects

- $\uparrow$ GDP $\succ$ $\downarrow$ Unemployment (Okun), $\uparrow$ Inflation (demand), Fed raises rate

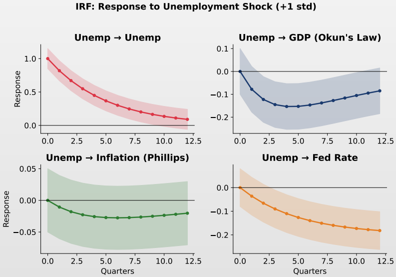


TSA_ch10_irf_gdp_shock

IRF: Unemployment Shock

Effects

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

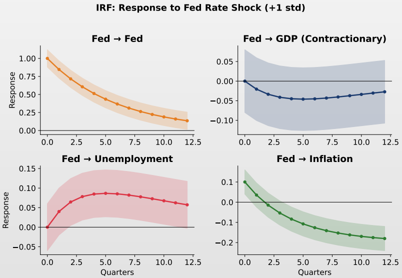


TSA_ch10_irf_unemp_shock

IRF: Fed Rate Shock

Monetary Policy

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

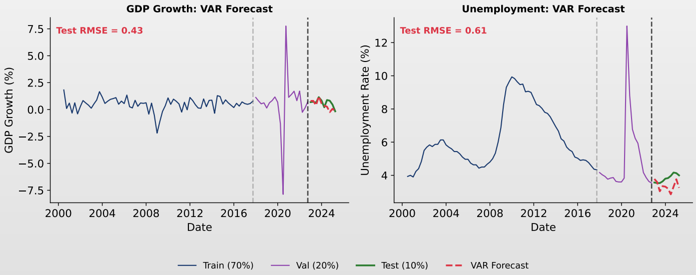


TSA_ch10_irf_fed_shock

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%**

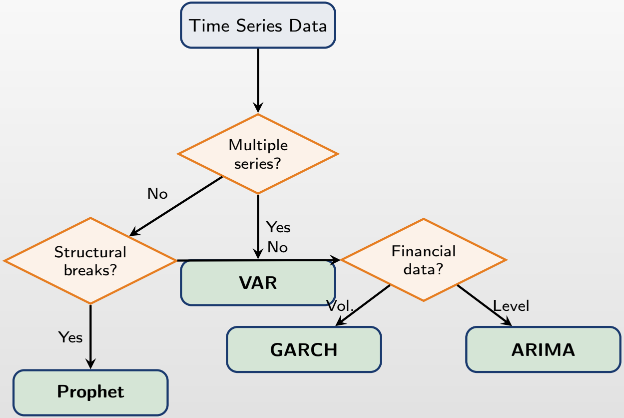
Strengths

- Cross-variable dynamics
- Good directional accuracy

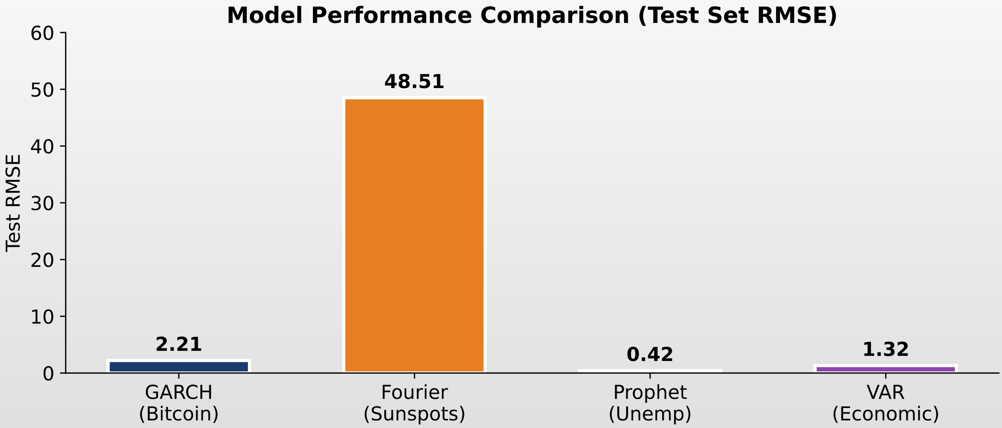
Limitations

- Many parameters
- Sensitive to lag selection

Model Selection Framework



Summary: Model Comparison



TSA_ch10_model_comparison

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

Forecasting vs Causality vs Decision

Objective	Model	Focus
Pure prediction	ARIMA / ML	Out-of-sample accuracy
Financial risk	GARCH	Volatility, VaR
Macro dynamics	VAR	Multivariate interactions
Structural relations	SVAR / VECM	Causal identification
Regimes	Markov Switching	Regime changes

Key Message

- There is no universal model
- There is **fit between model and problem**

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

The Role of AI in Time Series Modeling

AI can

- Generate code for estimation and forecasting
- Select models (AutoML, grid search)
- Combine forecasts (ensemble)
- Detect anomalies and patterns

But cannot

- Replace statistical validation
- Automatically detect **data leakage**
- Guarantee correct economic interpretation
- Verify model assumptions

Principle

- AI is a **tool**, not an authority
- Statistical validation remains the researcher's responsibility

AI Exercise: Critical Thinking

Prompt to test in ChatGPT / Claude / Copilot

"Download monthly US Retail Sales from FRED (series RSXFS) for 2010-01 to 2024-12 (180 observations). Perform a complete time series analysis: decomposition, stationarity tests, model selection (compare ETS, SARIMA, and Prophet), 12-month forecast, and evaluation using RMSE/MAE/MASE on a 70/15/15 temporal split. Give me publication-quality Python code."

Exercise:

1. Run the prompt in an LLM of your choice and critically analyze the response.
2. Does it follow the correct workflow? (plot → decompose → test → model → diagnose → forecast)
3. Does it compare multiple models (ETS, ARIMA, SARIMA) with proper benchmarks?
4. Is the train/test split done properly? Is there any data leakage?
5. Does it discuss limitations and assumptions of the chosen model?

Warning: AI-generated code may run without errors and look professional. *That does not mean it is correct.*

Question 1

Question

- Which model would you choose to forecast the volatility of financial returns?

Answer Choices

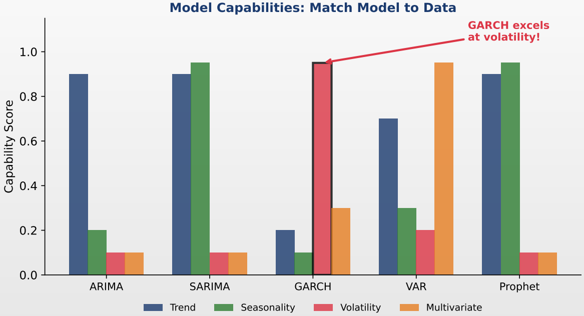
(A) ARIMA — captures trends and autocorrelations

(B) GARCH — models conditional variance

(C) Prophet — detects changepoints and seasonality

(D) VAR — multivariate model for interdependencies

Question 1: Answer



Answer: (B)

- GARCH captures volatility clustering and time-varying risk. ARIMA models the level, Prophet handles seasonality, VAR captures cross-series dynamics — none model variance directly.

TSA_ch10_quiz1_model_selection

Question 2

Question

A SARIMA model achieves RMSE $= 0.05$ on training but RMSE $= 2.30$ on test. What does this indicate?

Answer Choices

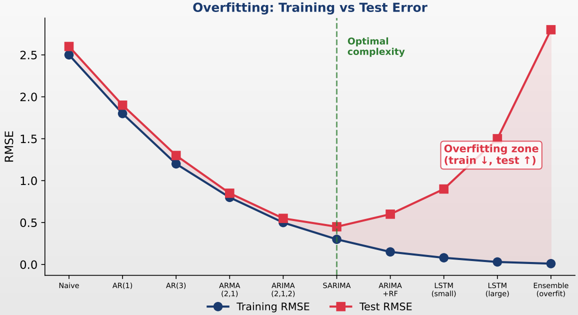
(A) The model is excellent — low training error confirms quality

(B) The model suffers from overfitting — it memorizes noise

(C) The test set is faulty and should be replaced

(D) The difference is normal — all models have higher test error

Question 2: Answer



Answer: (B)

- A 46× ratio between test and training RMSE signals severe overfitting. The model fits noise in the training data and fails to generalize. Solution: simpler model, proper validation.

TSA_ch10_quiz2_overfitting

Question 3

Question

- Why is it important to separate data into train/validation/test sets?

Answer Choices

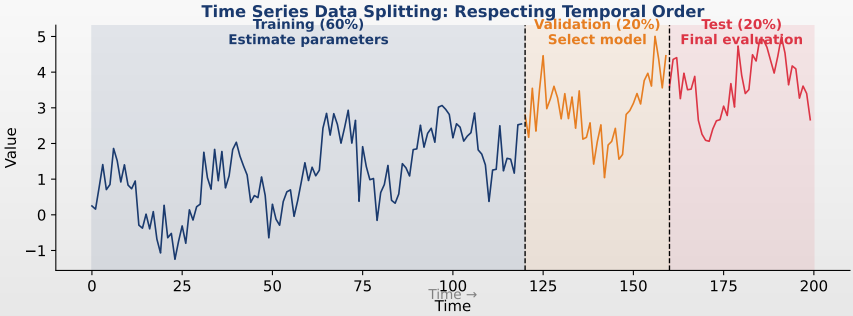
(A) To have more training data

(B) To prevent overfitting and evaluate correctly

(C) It is just a convention with no real importance

(D) To reduce computation time

Question 3: Answer



Answer: (B)

- Train: estimate parameters. Validation: select model/hyperparameters. Test: final unbiased evaluation. Mixing these roles leads to optimistic performance estimates.

TSA_ch10_quiz3_train_val_test

Question 4

Question

- Is Granger causality equivalent to true (structural) causality?

Answer Choices

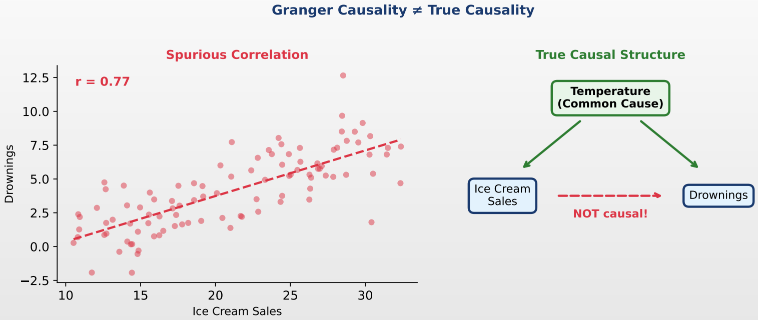
(A) Yes — if X predicts Y, then X causes Y

(B) No — it only tests predictive content, not causation

(C) It depends on the number of lags selected

(D) Yes, if the p-value is below 0.05

Question 4: Answer



Answer: (B)

- Granger causality tests whether past X improves forecasts of Y. Spurious correlations (e.g., ice cream sales and drownings) can pass the test due to common causes.

TSA_ch10_quiz4_granger_causality

Question 5

Question

- What model do you use for a series with long seasonality (e.g., $s = 365$ days)?

Answer Choices

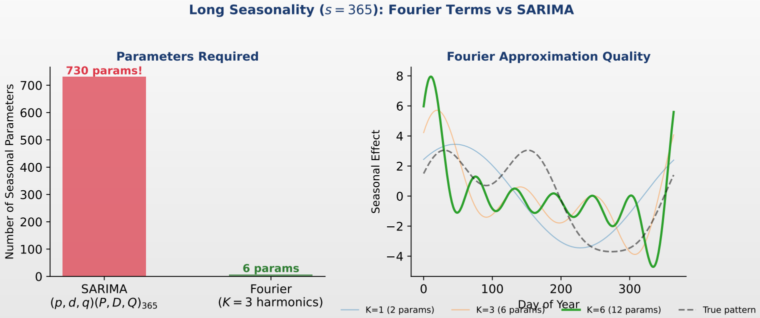
(A) SARIMA$(p, d, q)(P, D, Q)_{365}$

(B) GARCH — models variation

(C) ARIMA + Fourier terms or Prophet/TBATS

(D) VAR with 365 lags

Question 5: Answer



Answer: (C)

- SARIMA$_{365}$ needs $\sim$730 seasonal parameters — infeasible. Fourier terms with $K = 3$ use only 6 parameters. Prophet and TBATS handle multiple seasonalities automatically.

TSA_ch10_quiz5_long_seasonality

Bibliography I

Fundamental Textbooks (common references across all chapters)

- Hamilton, J.D. (1994). *Time Series Analysis*, Princeton University Press.
- Hyndman, R.J., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*, 3rd ed., OTexts.
- Shumway, R.H., & Stoffer, D.S. (2017). *Time Series Analysis and Its Applications*, 4th ed., Springer.

Domain-Specific References

- Tsay, R.S. (2010). *Analysis of Financial Time Series*, 3rd ed., Wiley. (GARCH, VAR)
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*, Springer. (VAR, VECM)
- Francq, C., & Zakoïan, J.-M. (2019). *GARCH Models*, 2nd ed., Wiley. (Volatility)

Bibliography II

Modern Approaches and Forecasting Competitions

- Petropoulos, F., et al. (2022). Forecasting: Theory and Practice, *International Journal of Forecasting*, 38(3), 845–1054.
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition, *International Journal of Forecasting*, 36(1), 54–74.
- Taylor, S.J., & Letham, B. (2018). Forecasting at Scale, *The American Statistician*, 72(1), 37–45.

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.

Thank You!

Questions?

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

Quantlet Quantinar