

COMPLETE STEP-BY-STEP GUIDE — TIME SERIES ANALYSIS

Word version with formulas (clean rendering) - usable by a beginner

General Scope of Time Series Analysis

Time series methods are often associated exclusively with financial data, such as stock prices, interest rates, or exchange rates. This perception is restrictive. In reality, the fundamental criterion for applying time series analysis is not the domain being studied, but the presence of a temporal dimension in the data.

Any variable observed in an ordered manner over time can be analyzed as a time series, whether it originates from economics, healthcare, energy, climate studies, industrial processes, or digital platforms. For example, daily hospital admissions, hourly electricity consumption, daily temperature measurements, or daily product sales all fall within the scope of time series analysis.

Thus, time series analysis constitutes a general methodological framework designed to analyze, model, and forecast the evolution of phenomena observed over time, independently of their financial or non-financial nature.

1) Objective of the Work

Before any modeling, define what you want to obtain: (1) forecast the target variable, (2) understand the dynamics (trend/seasonality), or (3) detect anomalies. You must also define a forecast horizon h (e.g., 7 days, 6 months).

2) Understanding the Structure of the Data

A time series is a sequence of observations indexed over time:

y_1, y_2, \dots, y_t

What you must check: (a) a date/time column, (b) a value column, (c) sorted dates, and (d) a clear frequency (daily, monthly, etc.).

3) Cleaning & Preparation (Why It Is Essential)

3.1 Sorting, checking frequency, and gaps

Why: ARIMA/SARIMA models and statistical tests assume a regular time step. If you have missing dates or duplicates, you must correct them before testing or modeling.

3.2 Missing values

Why: NA values can break statistical tests (ADF/KPSS) and lead to errors or biased parameters. Expected result: a series without gaps, or a documented strategy (deletion/interpolation).

3.3 Variance stabilization (if amplitude increases with level)

When: if the series “grows” (higher variability when the level is high).

Common transformation:

$$z_t = \log(y_t), \quad t = 1, 2, \dots, T$$

4) Decomposition: Trend, Seasonality, Noise (How to Recognize a Seasonal Series)

Why: to determine whether a seasonal model (SARIMA) is required and which transformations to apply. Seasonality is a pattern that repeats every s periods (e.g., monthly with annual cycle $\rightarrow s = 12$).

Two classical forms:

Additive: $y_t = T_t + S_t + e_t$

Multiplicative: $y_t = T_t \times S_t \times e_t$

Interpretation: additive model if seasonal amplitude is approximately constant.

Multiplicative if seasonal amplitude increases with the level (often after log transform, the model becomes additive).

5) Stationarity (Key Step Before ARIMA)

A series is stationary if its mean and variance do not change over time, and if dependence depends only on the lag.

Why: AR/MA/ARIMA models assume a stable structure over time.

4.1 Deciding if the series is stationary: tests + logic

ADF (Augmented Dickey-Fuller): $H_0 = \text{non-stationary}$. If $p\text{-value} < 0.05 \rightarrow \text{reject } H_0 \rightarrow \text{stationary}$.

KPSS: $H_0 = \text{stationary}$. If $p\text{-value} < 0.05 \rightarrow \text{reject } H_0 \rightarrow \text{non-stationary}$.

Best practice: use ADF and KPSS together (they complement each other).

6) Making the Series Stationary: Differencing (d) and Seasonal Differencing (D)

5.1 Simple differencing (often removes trend)

Formula:

$$\Delta y_t = y_t - y_{t-1}$$

If the trend remains, a second difference can be applied:

$$\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$$

The number of simple differences corresponds to parameter d in $ARIMA(p,d,q)$.

5.2 Seasonal differencing (removes seasonality)

If the series is seasonal with period s (e.g., $s = 12$):

$$\Delta_s y_t = y_t - y_{t-s}$$

The number of seasonal differences corresponds to parameter D in $SARIMA$.

7) Identifying p and q : ACF / PACF (Explaining the Famous “Spikes”)

ACF: correlation between the series and its lags.

Definition: $\text{Corr}(y_t, y_{t-k})$.

PACF: “direct” correlation with lag k after removing the effect of intermediate lags.

Practical rules:

- PACF cuts off at $p \rightarrow AR(p)$ candidate
- ACF cuts off at $q \rightarrow MA(q)$ candidate
- ACF and PACF decay slowly $\rightarrow ARMA/ARIMA$ candidate

Expected result: 2 to 4 candidate models (not 20).

8) Models and Formulas (AR, MA, ARIMA, SARIMA)

8.1 AR(p):

$$y_t = \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t$$

8.2 MA(q):

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q}$$

8.3 ARIMA(p,d,q) (operator form):

$$\Phi(L)(1-L)^d y_t = \Theta(L) \varepsilon_t$$

where $\Phi(L)$ and $\Theta(L)$ are lag polynomials.

8.4 SARIMA(p,d,q)(P,D,Q,s):

$$\Phi(L)\Phi_s(L^s)(1-L)^d(1-L^s)^D y_t = \Theta(L)\Theta_s(L^s) \varepsilon_t$$

This form adds seasonal AR/MA structure and seasonal differencing.

9) Choosing the Best Model: AIC/BIC + Time Validation

AIC/BIC: smaller is better (trade-off between fit and complexity).

Important: time-based validation (train = past, test = future).

Evaluation on the test set using MAE / RMSE / MAPE.

10) Diagnostics (Residuals = White Noise)

After estimation, residuals should resemble white noise: mean close to zero, no autocorrelation, stable variance.

Ljung-Box test: H_0 = residuals are independent (p-value > 0.05 → OK).

11) Forecasting

Forecasts \hat{y}_{t+h} are produced with confidence intervals.

Short horizons are generally more reliable.

Always comment on uncertainty (intervals).