DS3000 / DS5110 : Foundations

Time Series Analysis

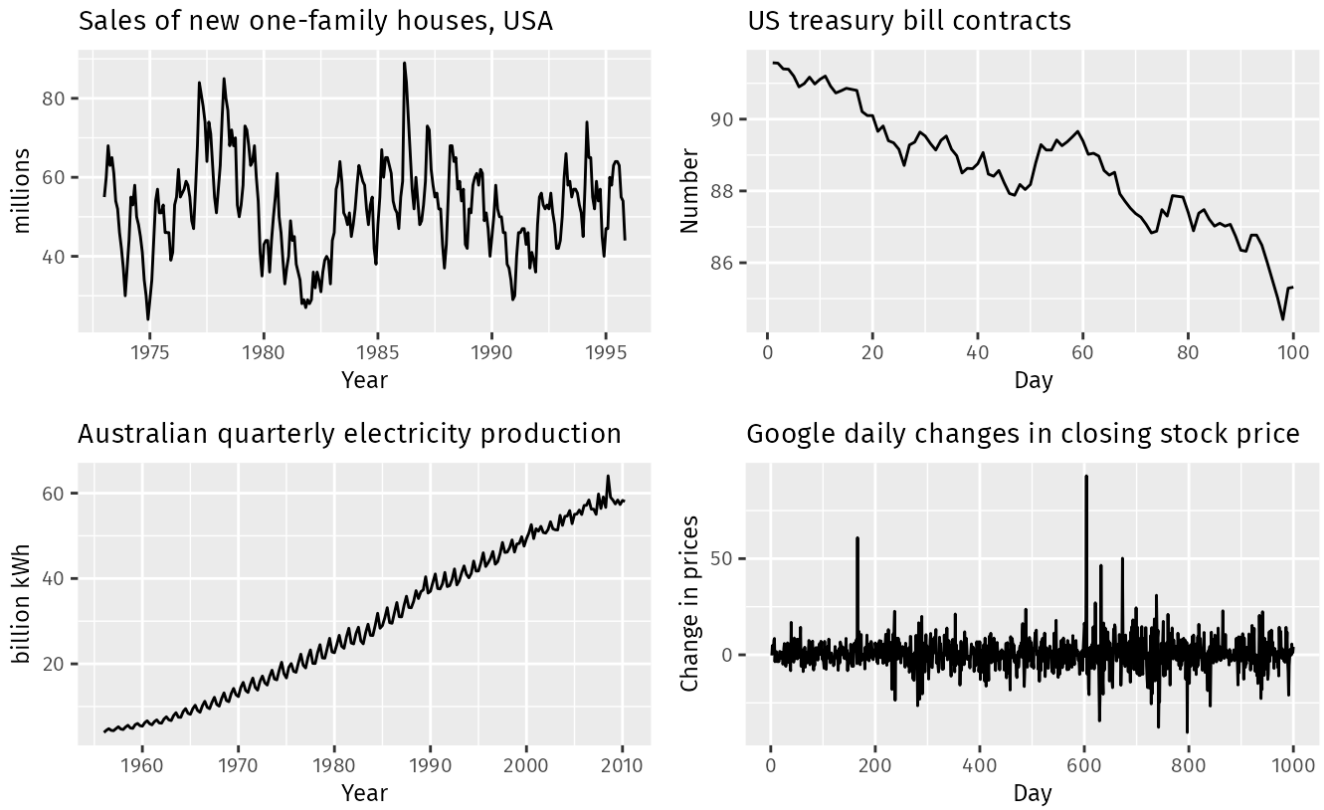
Prof. Rachlin
Northeastern University

What is a *time series*?

A *time series* is an ordered sequence of temporal measurements, acquired at regular intervals over some extended period.

Examples include:

- temperature / climate data
- IoT
- financial markets, GDP
- physical processes
- seasonal migrations
- sales
- median housing prices

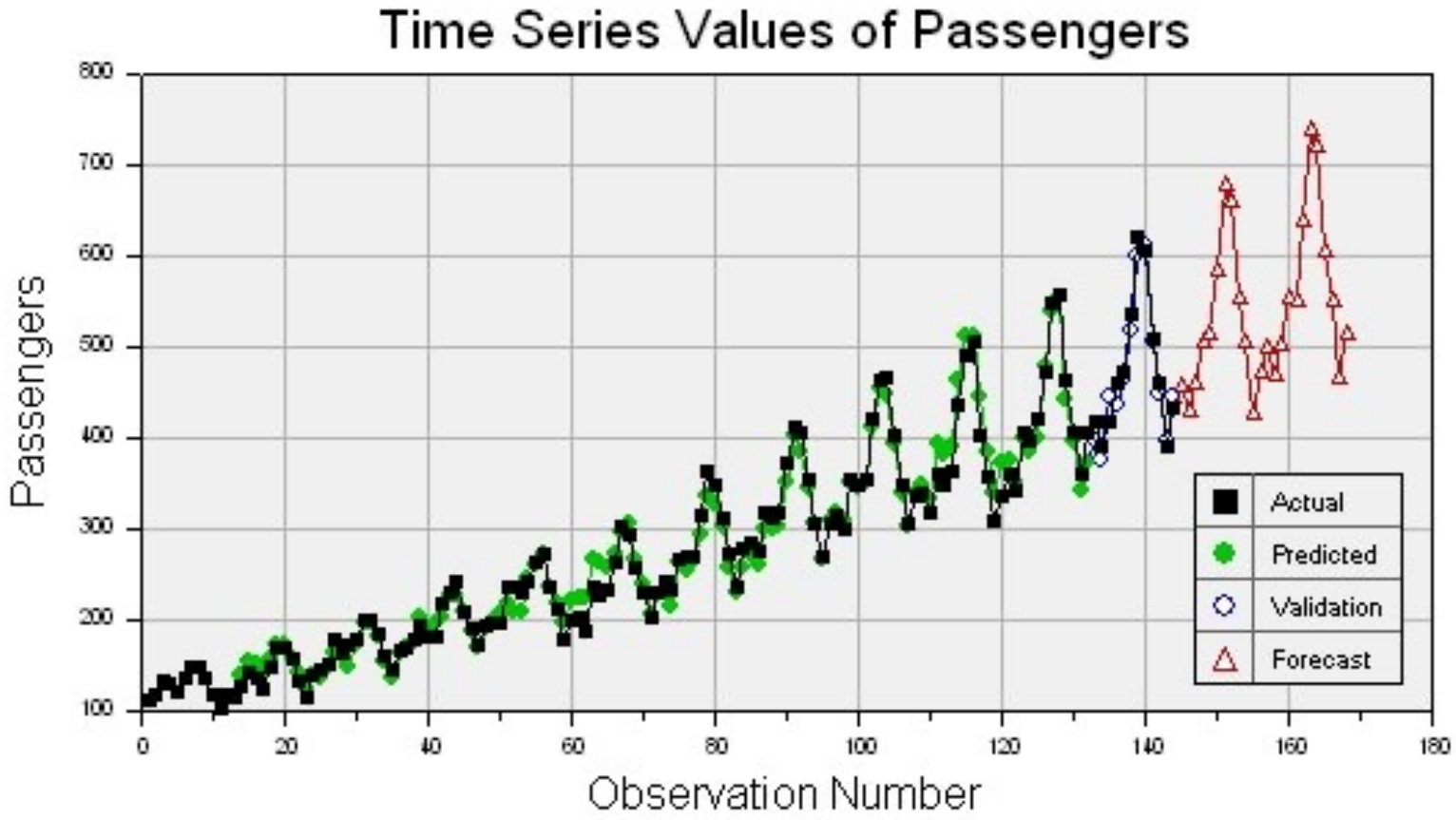


The Goal of Time-Series Analysis

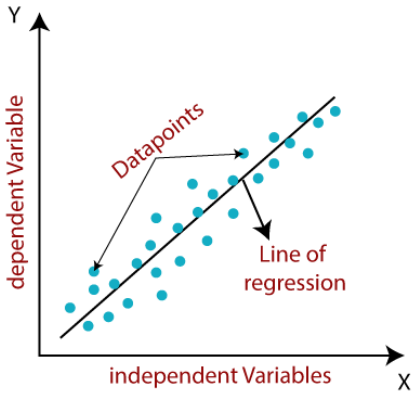
In time-series analysis has two main goals:

1. to understand the underlying structure of the data by defining a *model* that accurately reproduces our data while accounting for observational errors or other random factors.

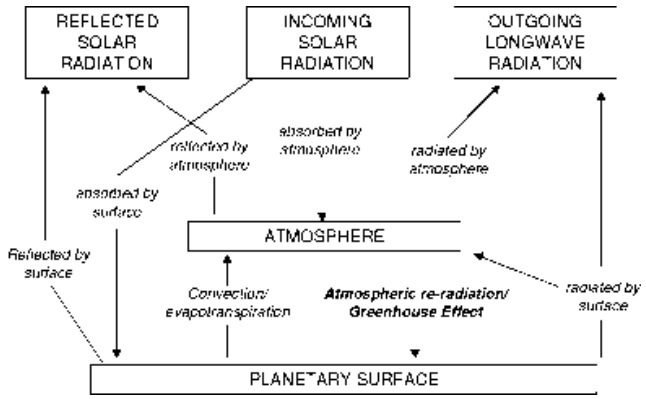
2. to be able to extrapolate the data into the future by making predictions based on the model. (This assumes that the past is a good indicator of the future.)



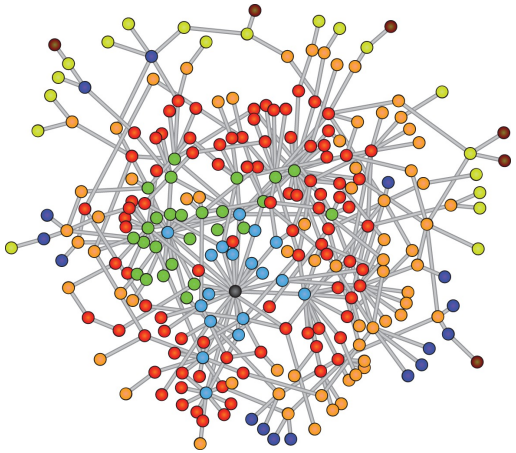
The models may take many forms



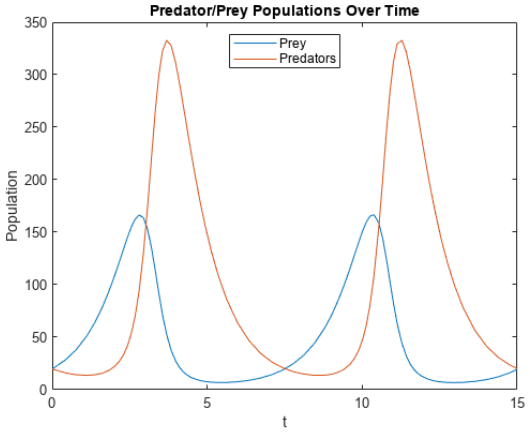
Linear Models



Conceptual Models



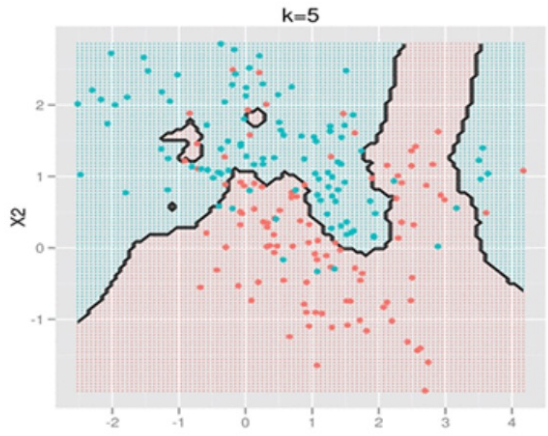
Network Models



$$\frac{dV}{dt} = aV - bVP$$

$$\frac{dP}{dt} = -cP + dVP$$

Differential Equations (Mathematical Models)



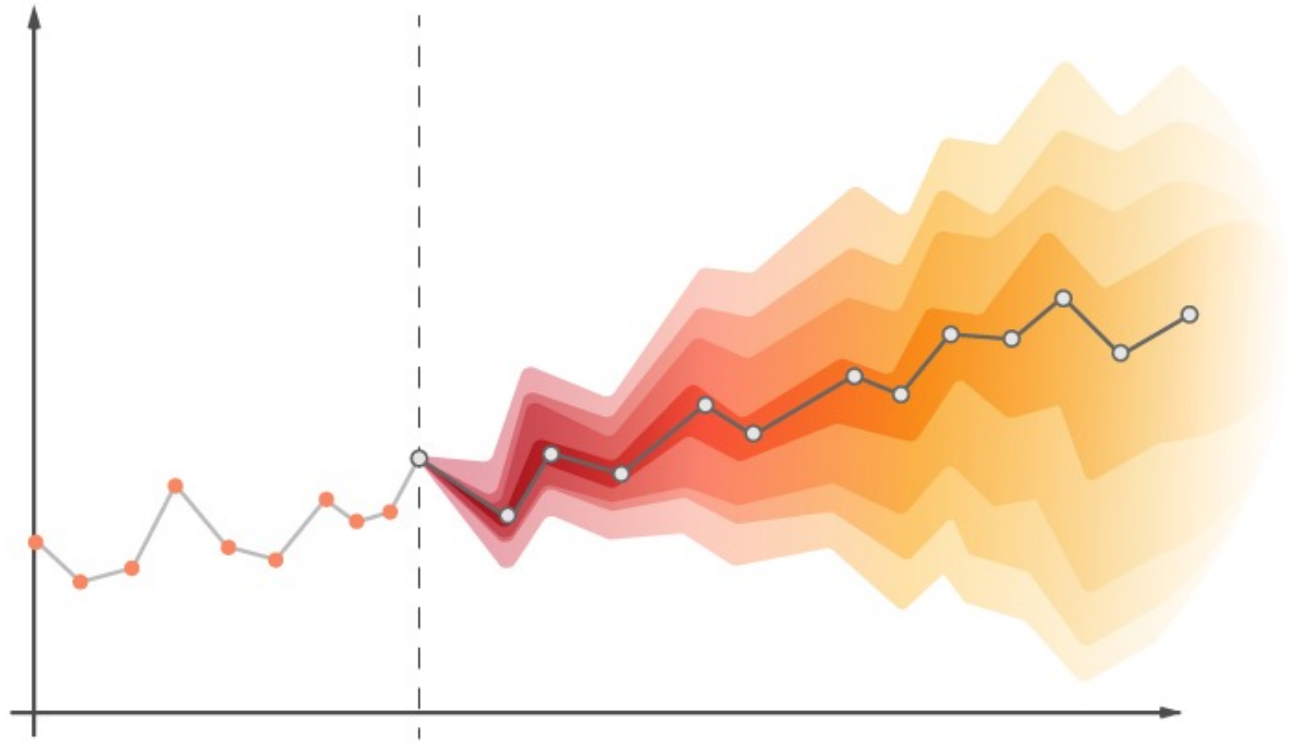
Machine Learning Models

Forecasting

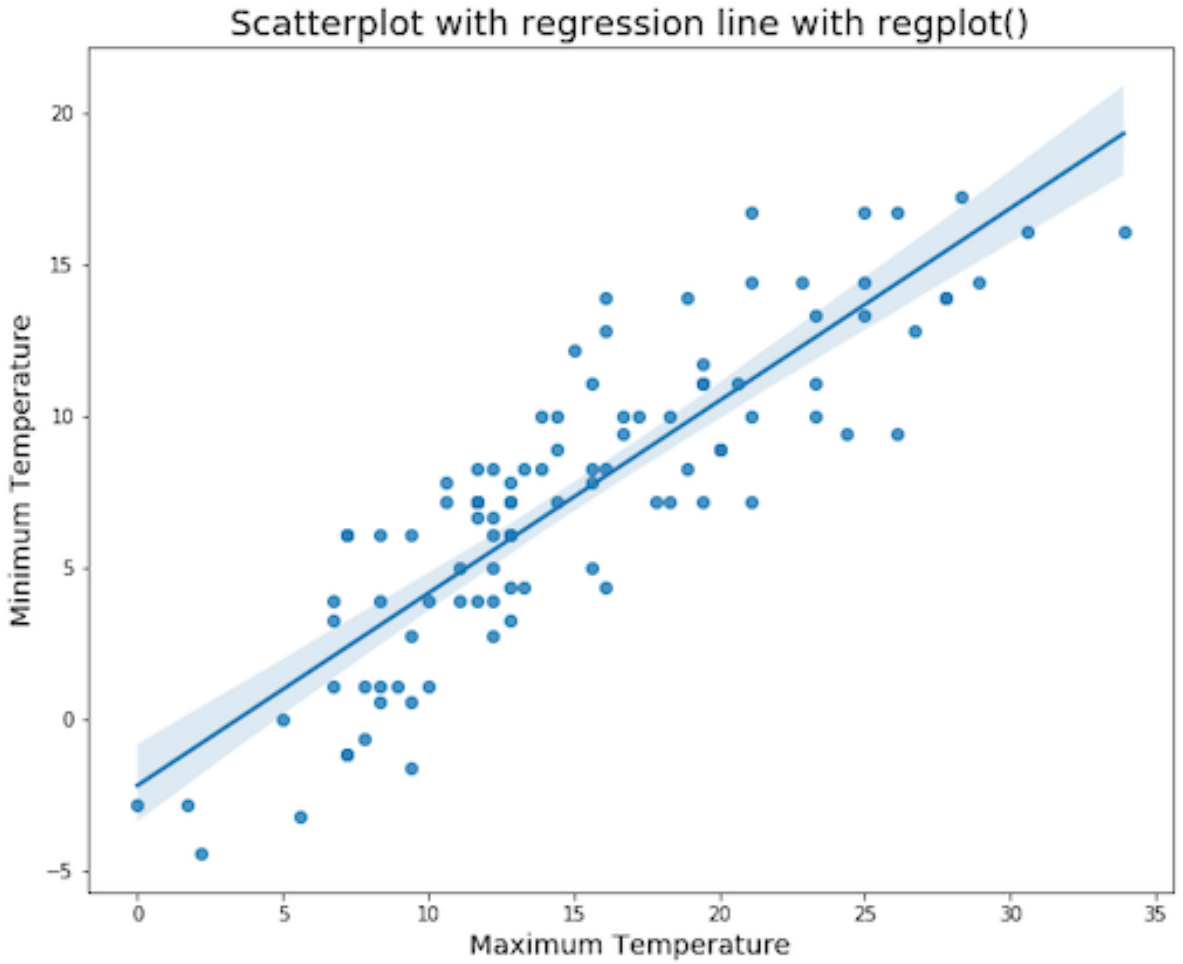
Forecasting focuses on trying to predict future outcomes by analyzing and discovering patterns, trends, and cycles in the historical data. Forecasts may be:

univariate: dependent on one time-series variable

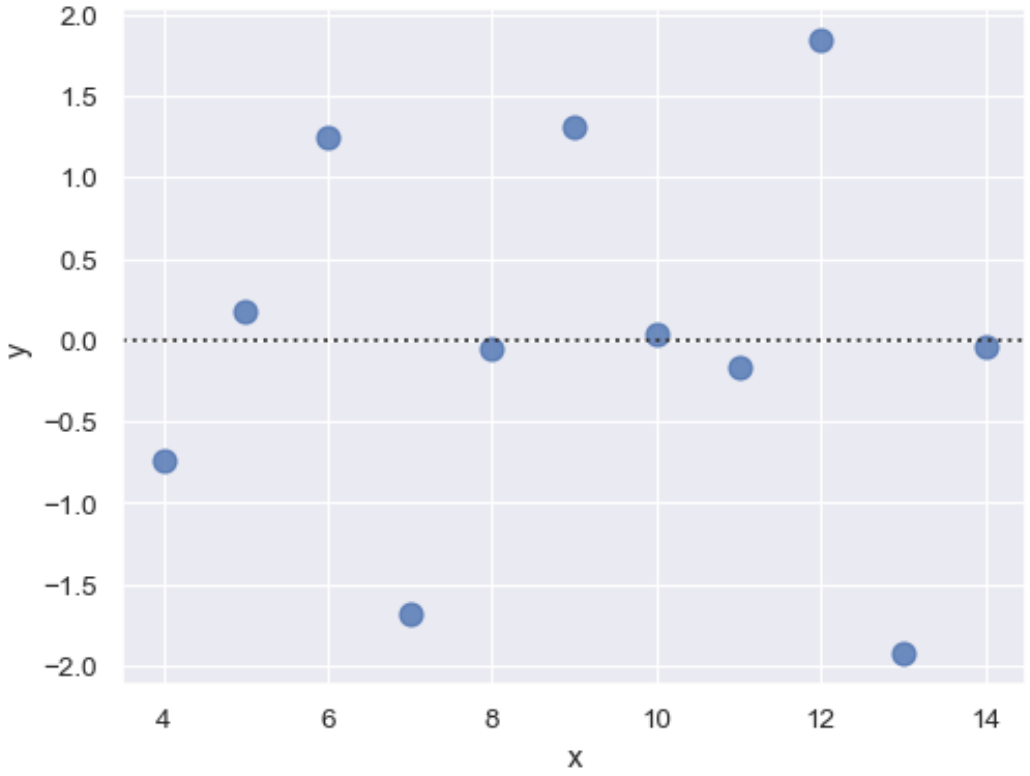
multivariate: dependent on two or more (possibly interrelated) time series variables.



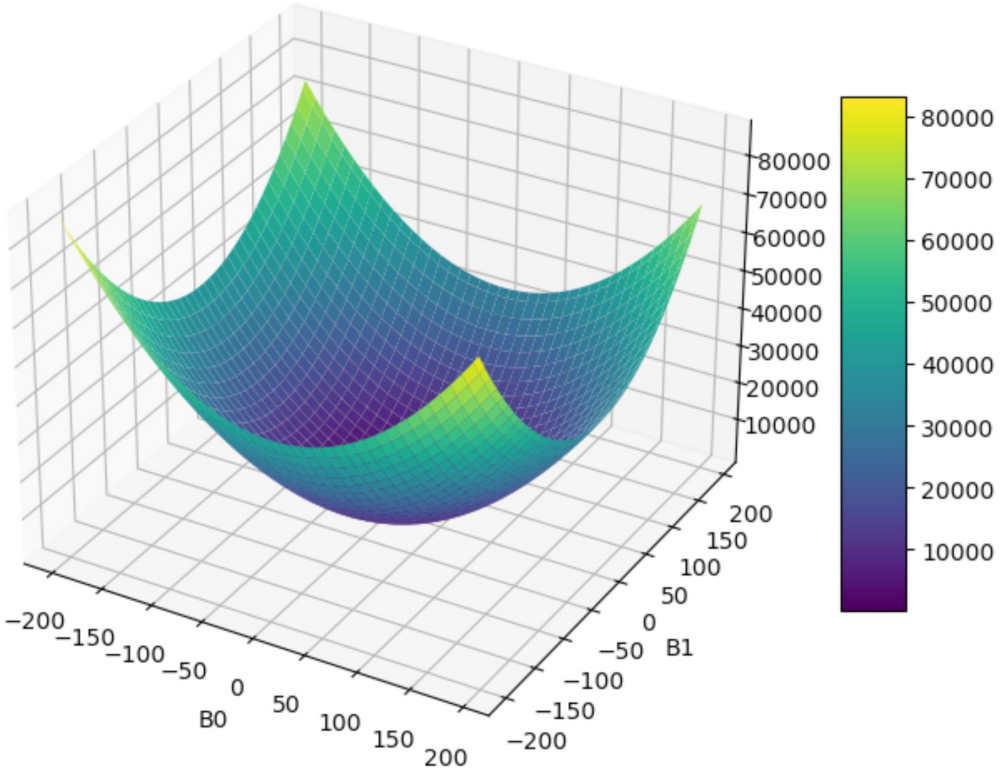
Seaborn regplot



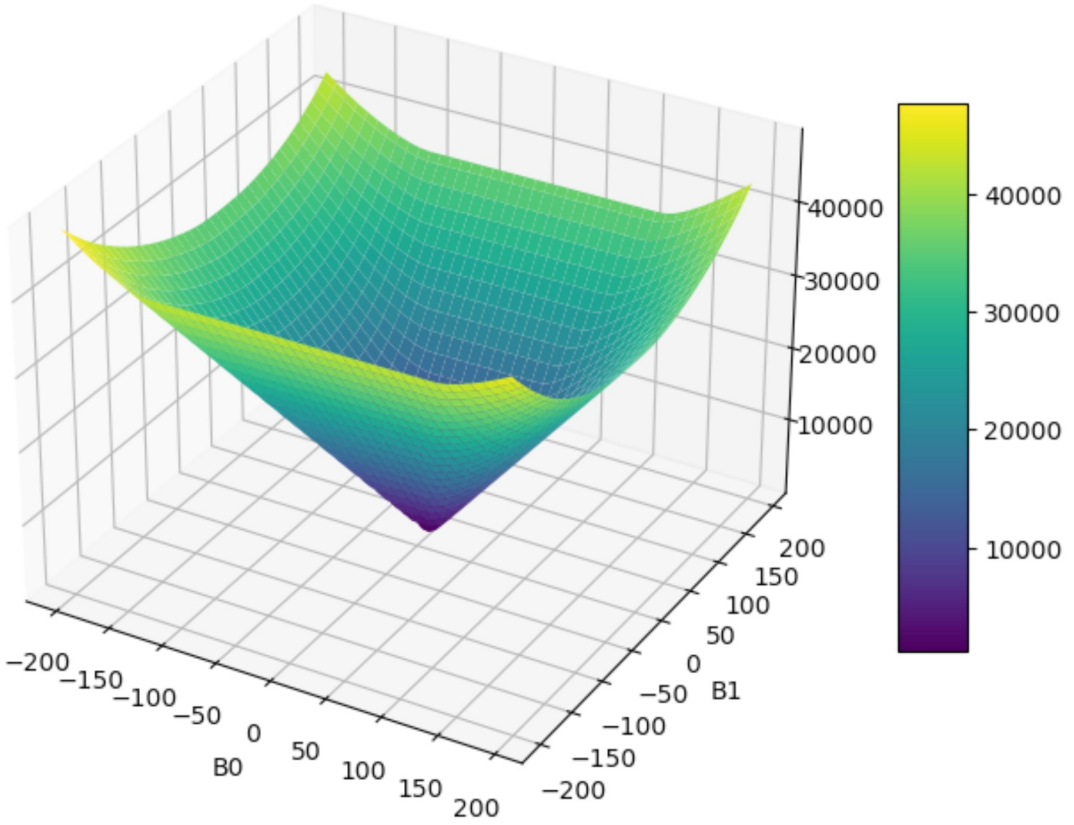
The `residplot()` function can be a useful tool for checking whether the simple regression model is appropriate for a dataset. It fits and removes a simple linear regression and then plots the residual values for each observation. Ideally, these values should be randomly scattered around `y = 0`:



Finding best-fit linear model: $y = b_0 + b_1x + e$



Mean Square Error (MSE)



Mean Absolute Error (MAE)

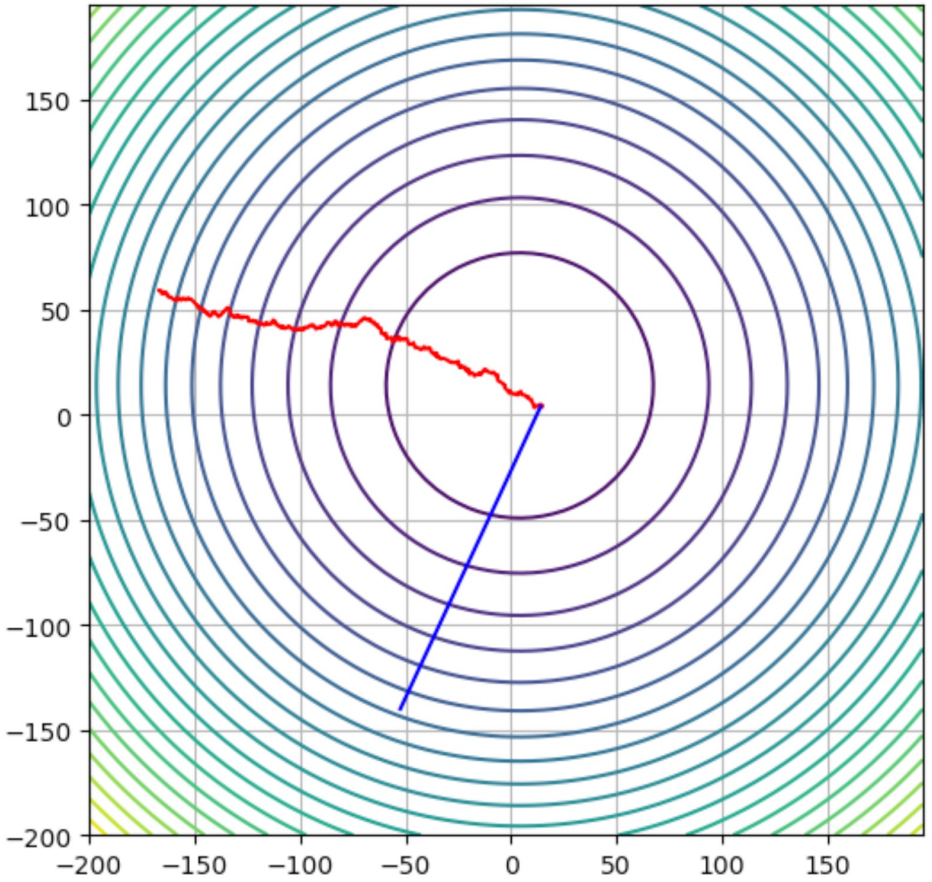
Gradient Descent vs. Random Walk

This graphic shows the error function rendered as a contour plot.

Gradient Descent finds a minimum value by moving in the direction of steepest descent.

But random search might work ok if the error function landscape is complicated.

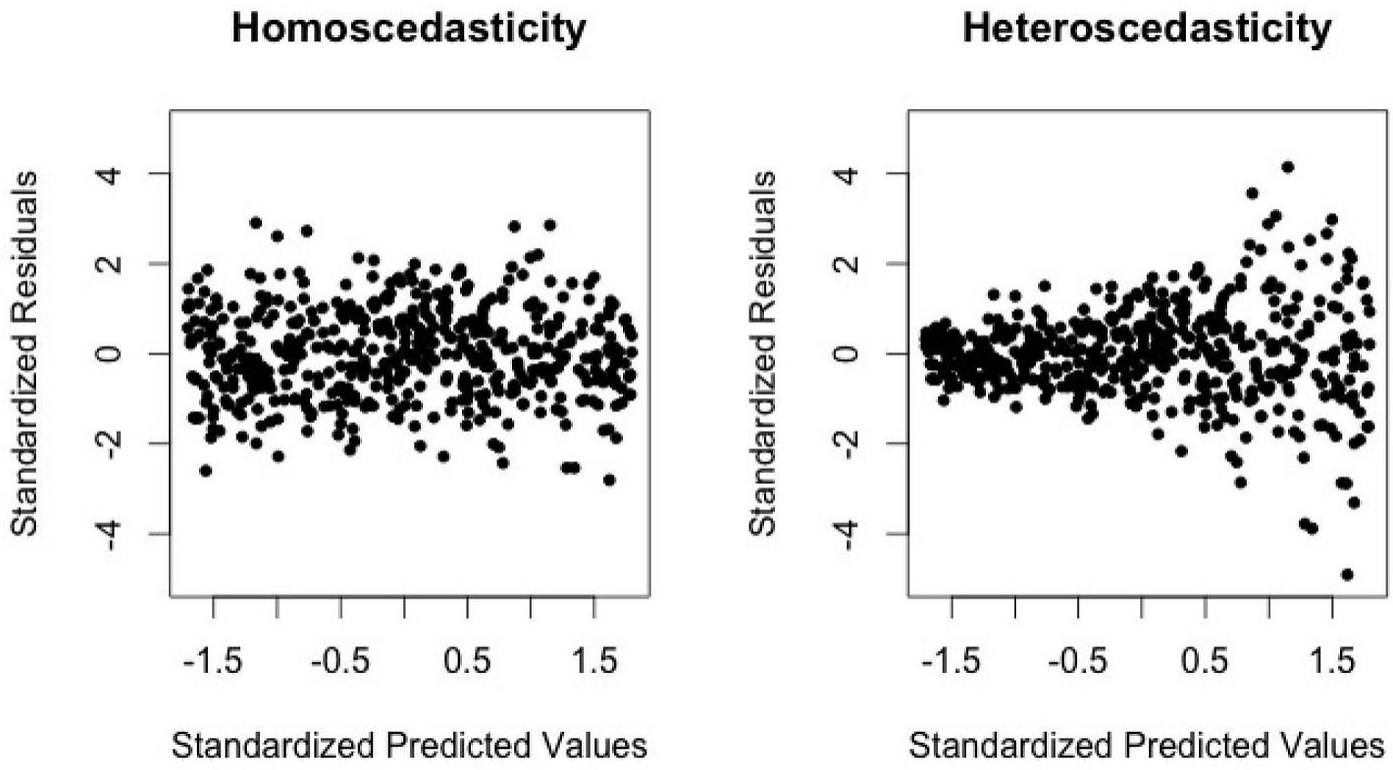
Evolutionary optimization approaches can explore an error landscape with many local minima.



Linear Regression confidence intervals assume Homoscedasticity

Homoscedasticity means:

- residual terms are normally distributed
- residuals have constant variance
- Magnitude of residual term is independent of $\mathcal{X}$



Time Series (Additive) Decomposition: $y_t = T_t + S_t + C_t + I_t$

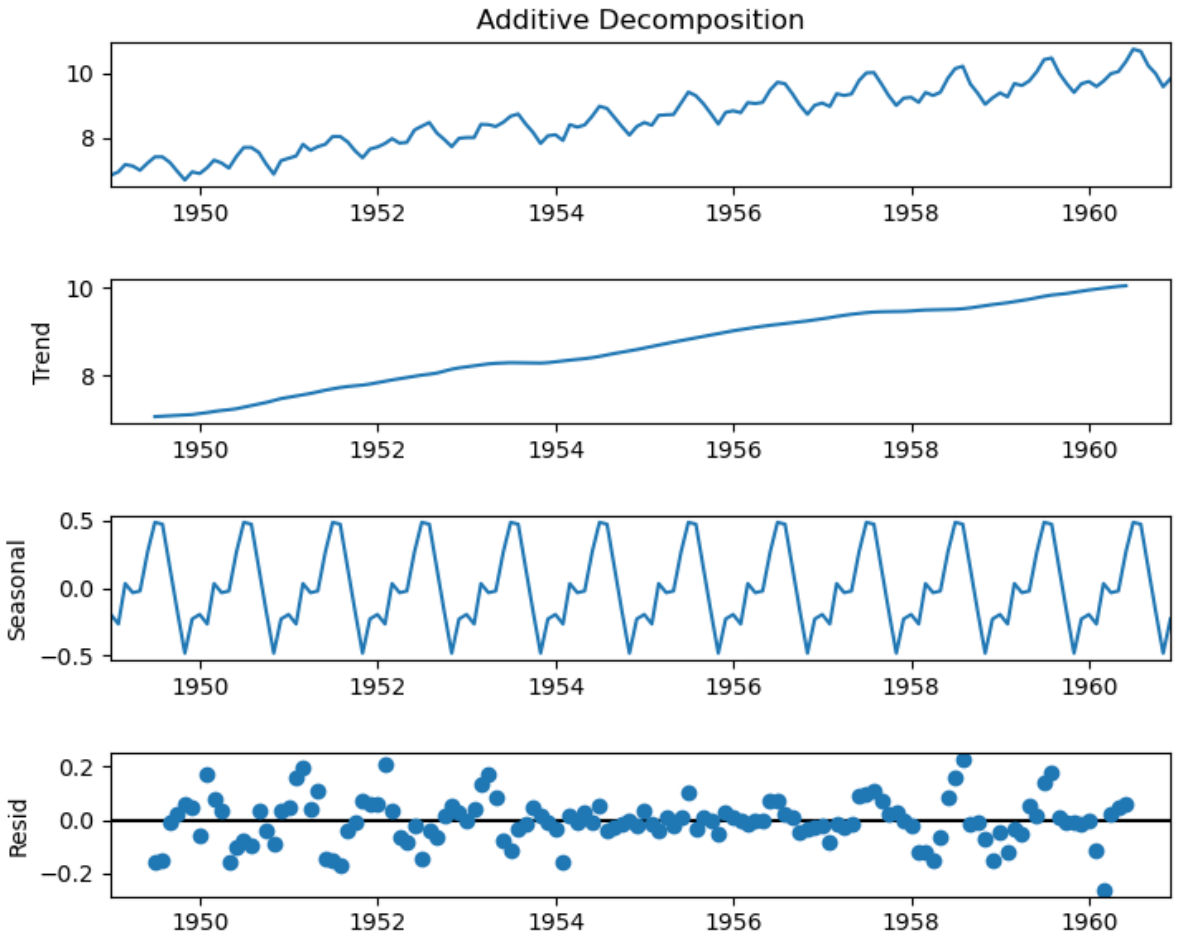
Time series decomposition is a statistical method that deconstructs a time series into several distinct components, each representing an underlying pattern category.

1. **Trend**: The trend shows a general direction that the data is moving towards over a long period of time.

2. **Seasonality (or Seasonal Variation):** These are periodic fluctuations that occur regularly based on a season. They are predictable movements that can be attributed to the time of year, month, or even a particular day.

3. **Cyclic Variation**: These are oscillations around the trend line that are not of a fixed period. They occur due to economic cycles or other broad systemic factors

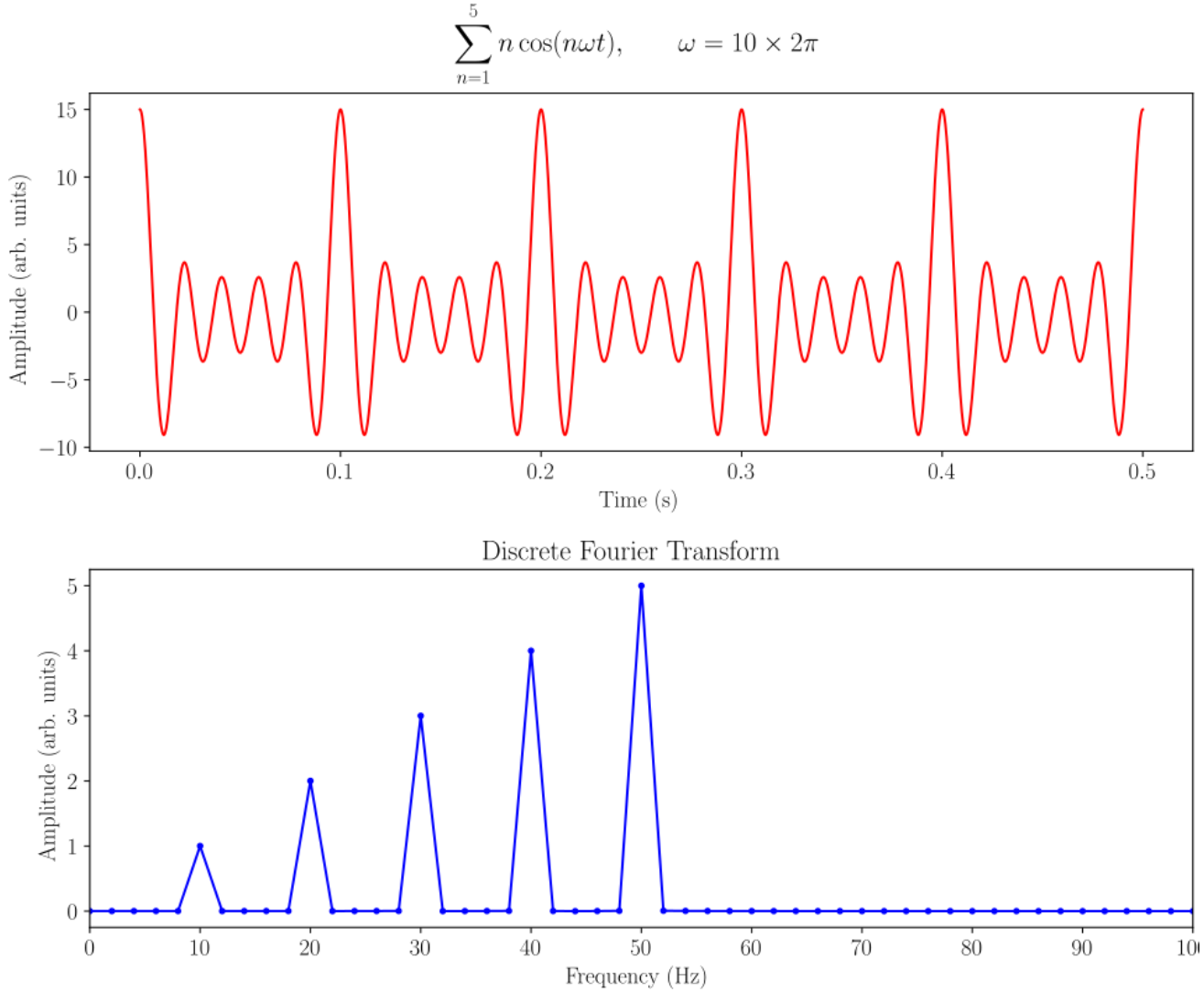
4. **Residual (or Irregular Component):** Residuals ideally should resemble white noise, meaning they are random and do not contain any pattern.



Fast-Fourier Transform (FFT) Decomposition

Goal: Break down an original time series into a set of sinusoidal waves with different frequencies and amplitudes.

The resulting frequency *spectrum* identifies the contribution of each wave to the final output series.



Smoothing time series data

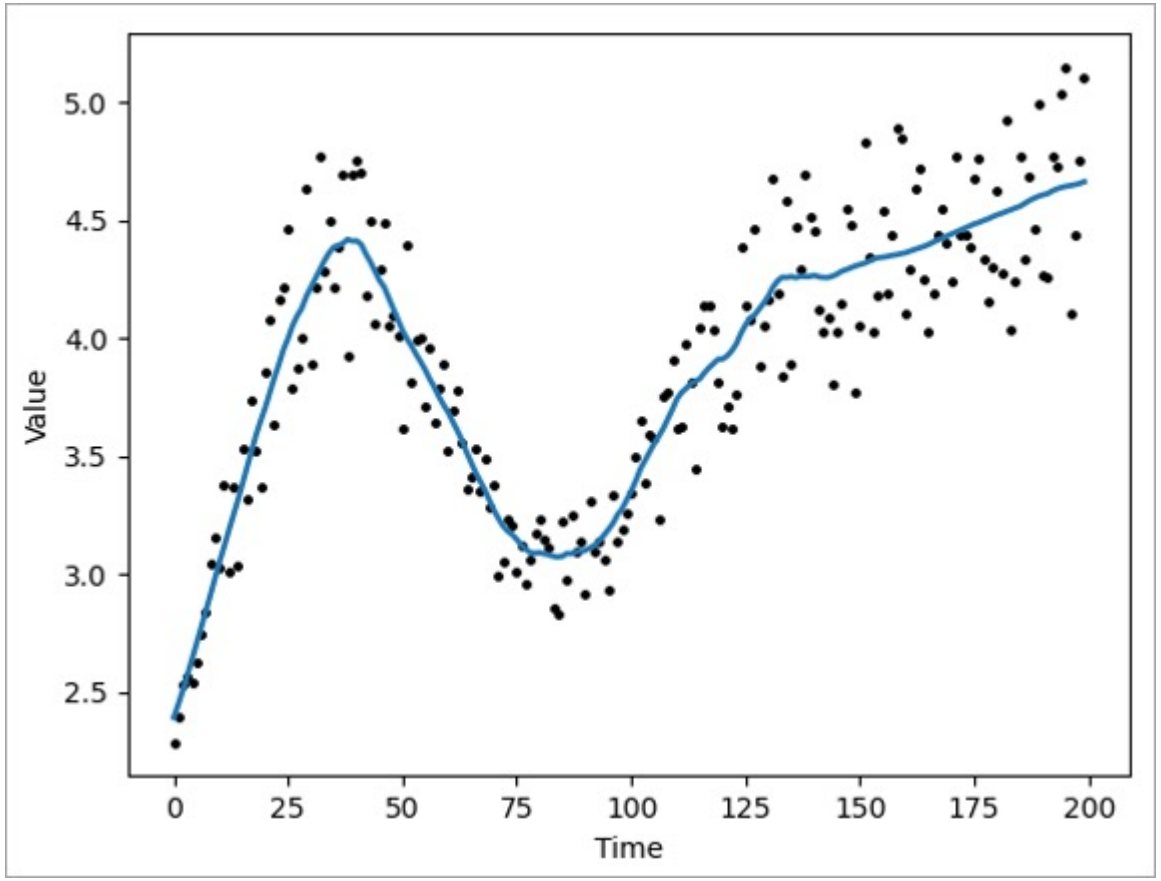
Smoothing methods are used in time series analysis to remove noise and better expose the signal of the underlying causal processes. They can be particularly helpful for data visualization, and they also form the basis of many methods for time series forecasting.

Moving Average: Calculate means over a sliding window.

Exponential Smoothing: Weights older points exponentially lower so that more recent observations have more influence.

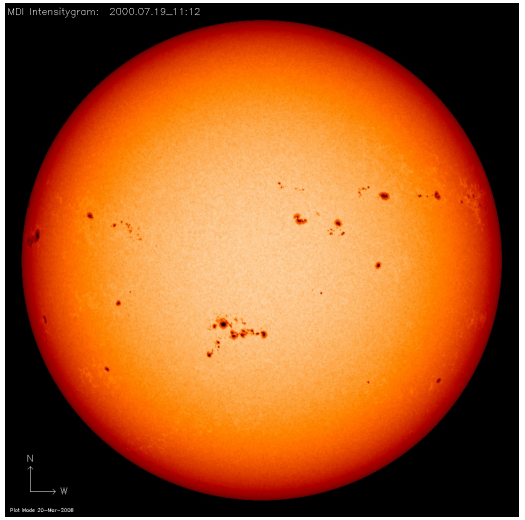
LOESS: Locally Estimated Scatterplot Smoothing using local regression.

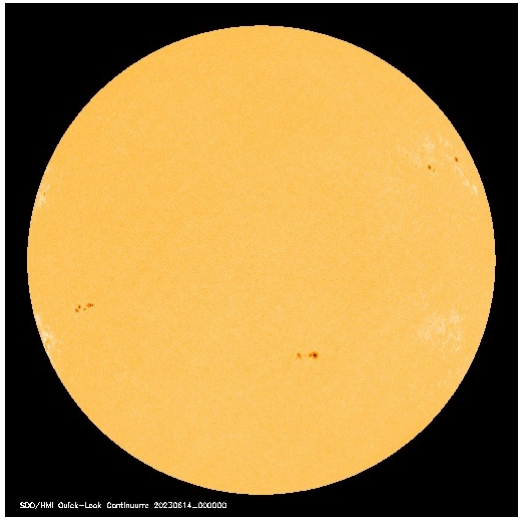
Kalman Filtering: Effective for identifying statistical noise. More widely used in econometrics.

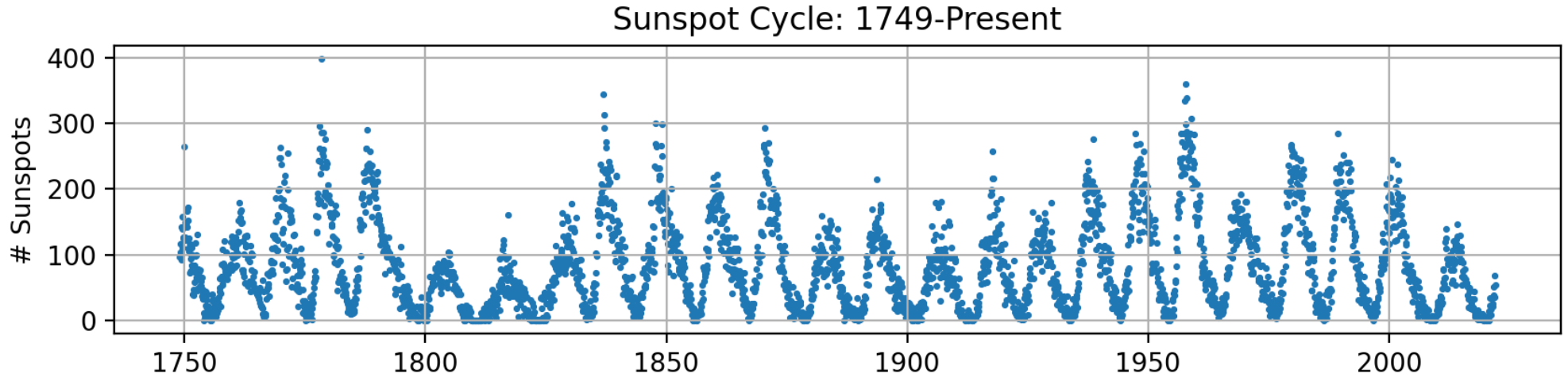


Sunspot Data

An 11-year cycle driven by the changing magnetic field which itself is generated by a dynamo caused by the differential rotation of the Sun's interior. Highly tangled magnetic field lines → increased sunspot activity.







Cyclogram / Phase Plotting of Sunspot Data

Fold the time series so that points in the same phase of each cycle are aligned vertically by plotting:
observation vs. (time MOD period).

By plotting data points corresponding to the same phase or point in the cycle on top of each other, you can examine the consistency of the cycle and observe any variability or irregularity within the cycle.