19 | Time Series

Ivan Corneillet

Data Scientist

GENERAL ASSEMBLY

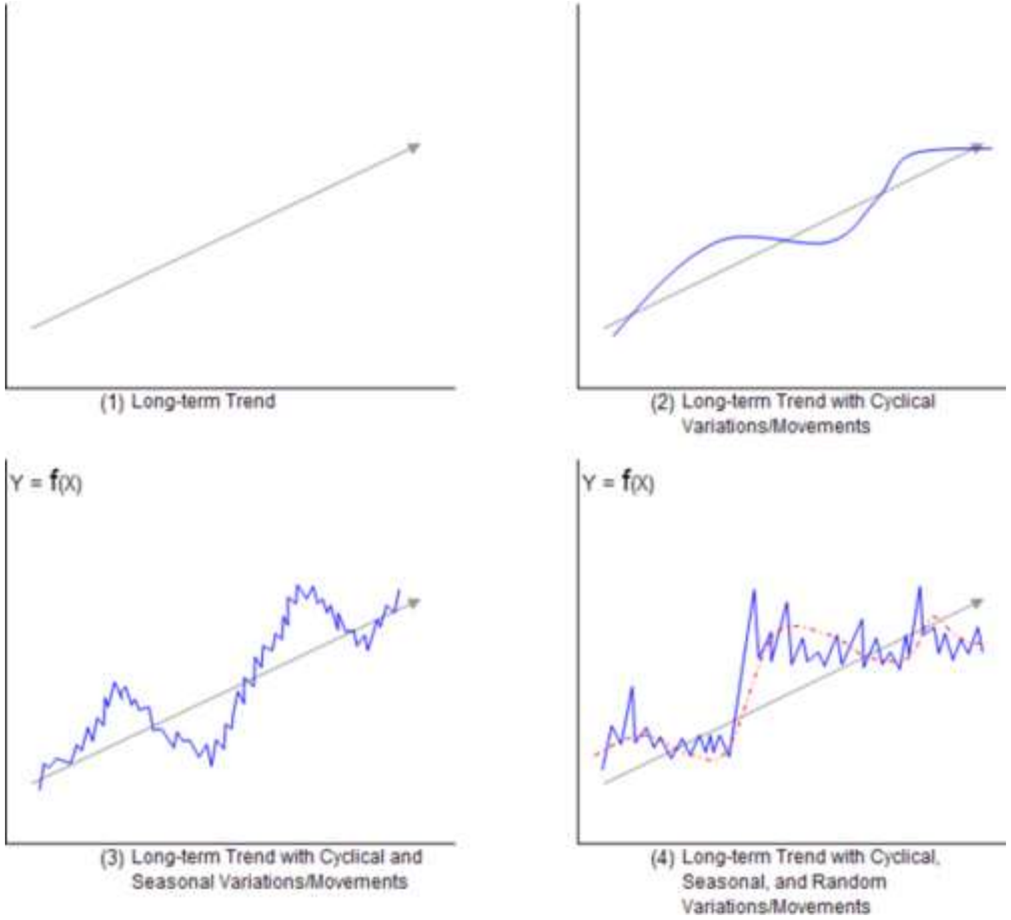
Learning Objectives

After this lesson, you should be able to:

- Understand what time series data is and what is unique about it
- Perform time series analysis in *pandas* including rolling mean/median and autocorrelation
- Model and predict from time series data using AR, MA, ARMA, or ARIMA models
- Specifically, coding these models in *statsmodels*

Trends, Cyclical, and Seasonal Variations

Trends, Cyclical, and Seasonal Variations

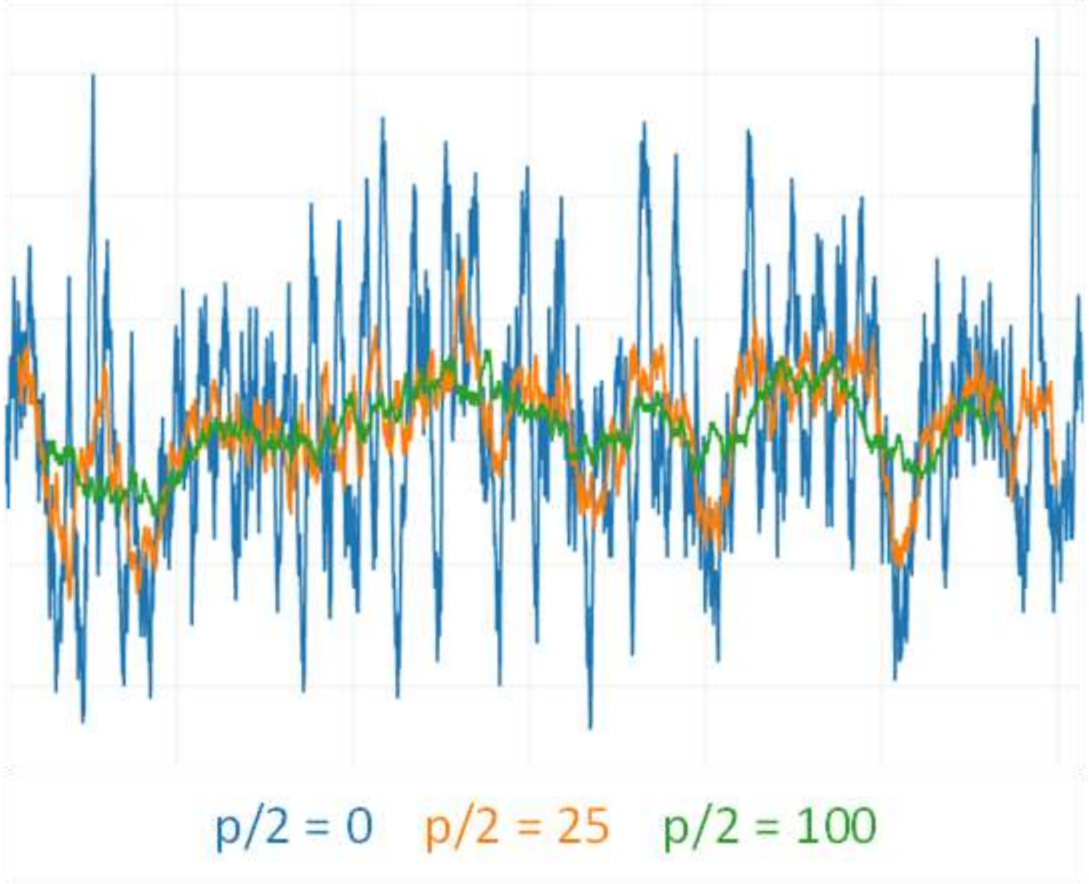


Moving Averages; Rolling Means and Medians

A moving average replaces each data point with an average of p consecutive data points in time

- This could be using the $p/2$ data points prior to and following a given time point; it could also be the p preceding points
- These are often referred to as the "rolling" average
- The measure of average could be mean or median
- The *rolling mean* is

$$F_t = \frac{1}{p} \sum_{k=-\frac{p}{2}}^{\frac{p}{2}} Y_{t+k} \ \ or \ F_t = \frac{1}{p} \sum_{k=0}^{p} Y_{t+k}$$



Rolling means and rolling medians

Rolling mean

- A rolling mean averages all values in its window, but can be skewed by outliers
  - This may be useful if we are looking to identify atypical periods or we want to evaluate these odd periods
  - E.g., this would be useful if we are trying to identify particularly successful or unsuccessful sales days

Rolling median

- The rolling median would provide the 50 percentile value for the period and would possibly be more representative of a “typical” day

Autocorrelation

Autocorrelation

- *Autocorrelation* is how correlated a variable is with itself. Specifically, how related are variables earlier in time with variables later in time

- Typically, for a high quality model, we require some autocorrelation in our data
- We can compute autocorrelation at various lag values to determine how far back in time we need to go

Autocorrelation

- To compute autocorrelation, we fix a "lag" k denoting how many time points earlier we should use to compute the correlation
- A lag of $k = 1$ computes how correlated a value is with the prior one. A lag of $k = 10$ computes how correlated a value is with one 10 time points earlier

$$r_k = \frac{\sum_{i=1}^{n-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

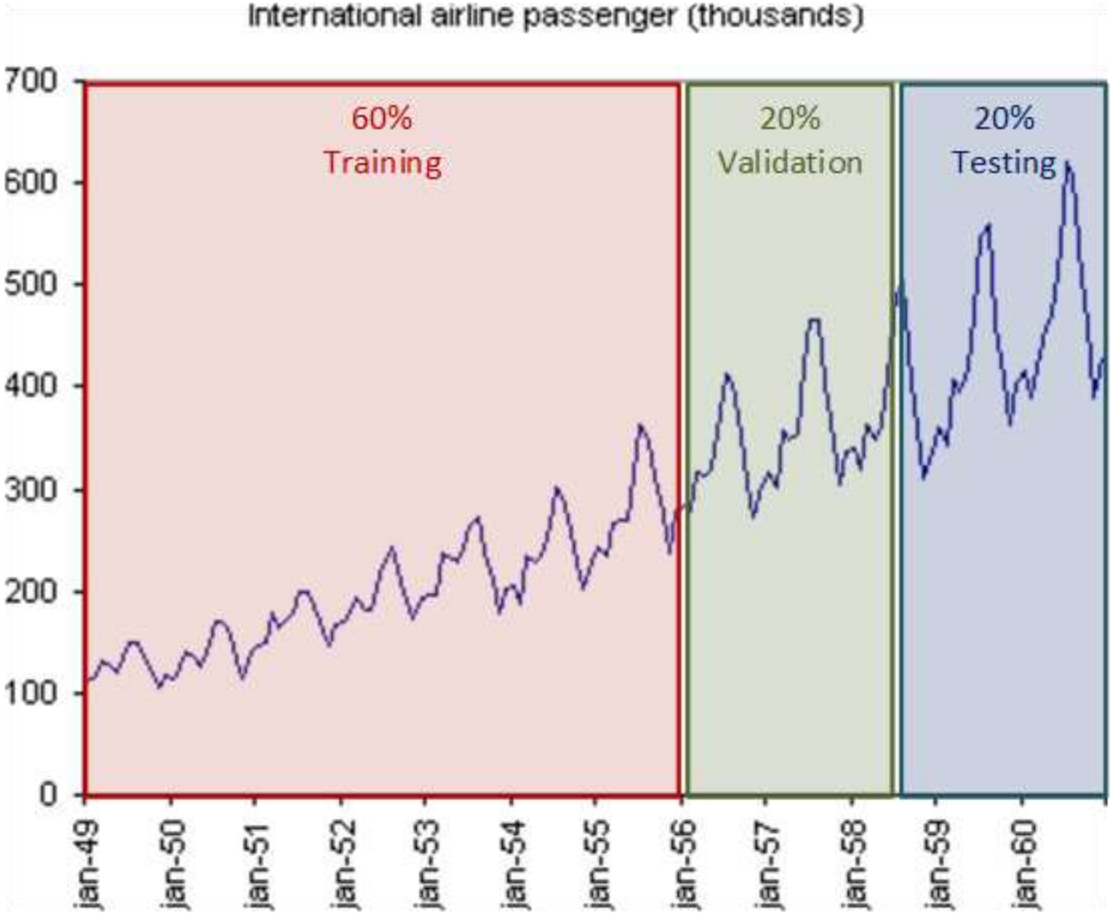
(n observations; $\bar{x}$, overall mean)

Time Series Modeling

Time series models predicts future values in the time series

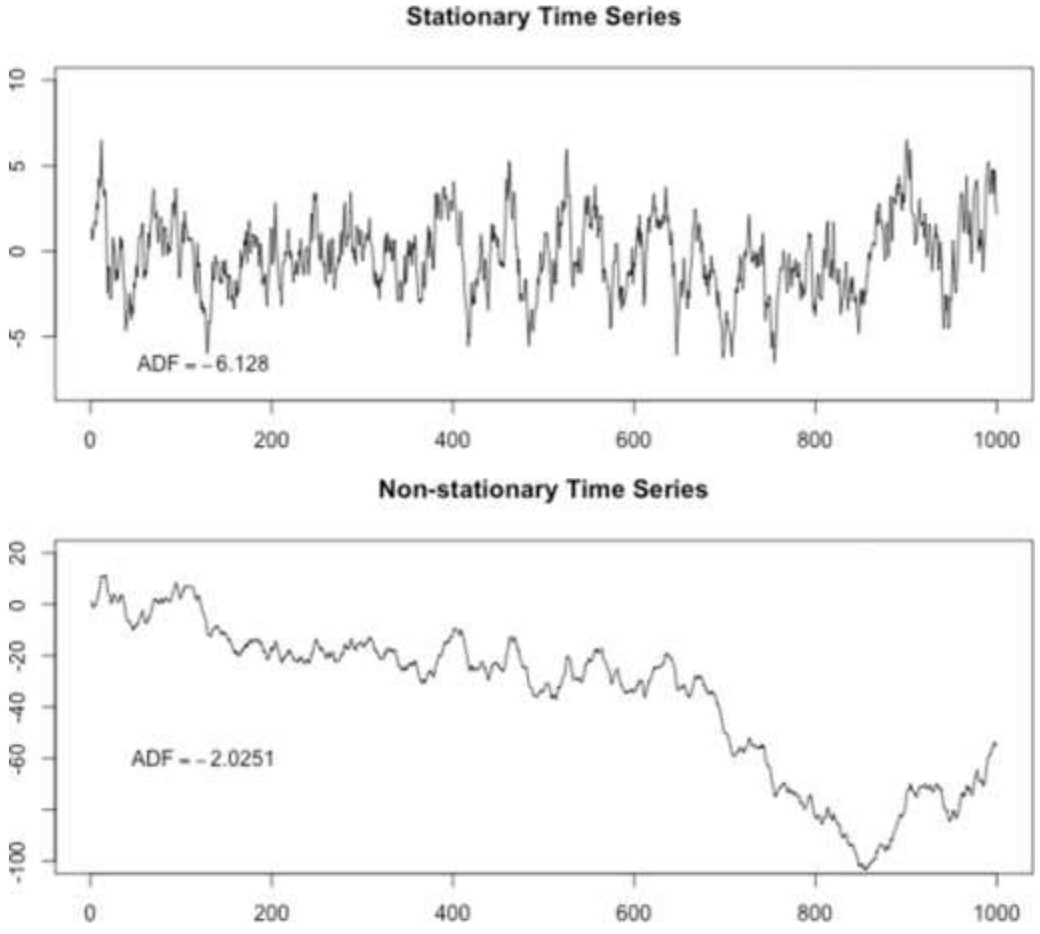
- **Like** other predictive models, we will use prior history to predict the future
- **Like** previous modeling exercises, we will have to evaluate the different types of models to ensure we have chosen the best one
  - We will want to evaluate on a held-out set or test data to ensure our model performs well on unseen data
- **Unlike** previous models, we will use the earlier in time outcome variables as inputs for predictions
- **Unlike** previous modeling exercises, we won't be able to use standard cross-validation for evaluation
  - Since there is a time component to our data, we cannot choose training and test examples at random

Instead, we will exclusively train on values earlier (in time) in our data and test our model on values at the end of the data period



Stationarity

Many models (e.g., AR, MA, $ARMA$) assume that time series are *stationary*, i.e., that their mean and variance is the *same* throughout (no trend)



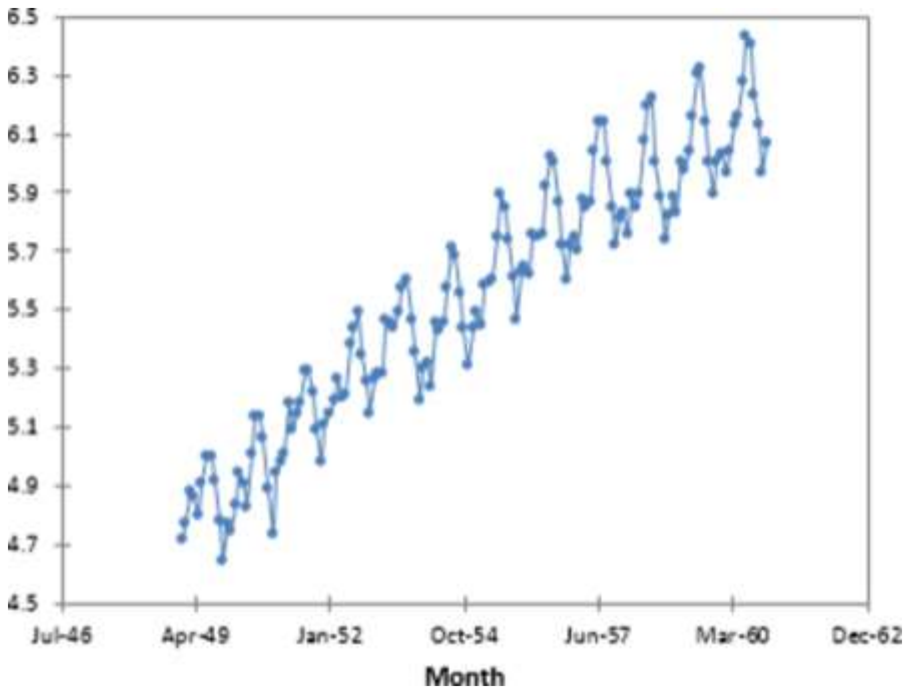
- E.g., while sales may shift up or down over time, the mean and variance of sales is constant; i.e., there aren't many dramatic swings up or down

Many time series data aren't stationary (e.g., stock market performance); e.g., the S&P 500 mean performance since 1993 is increasing over time

A simple method to get a stationary time series from a non-stationary time series is *differencing*; instead of predicting the series, we predict the difference between two consecutive values (e.g., $ARIMA$)

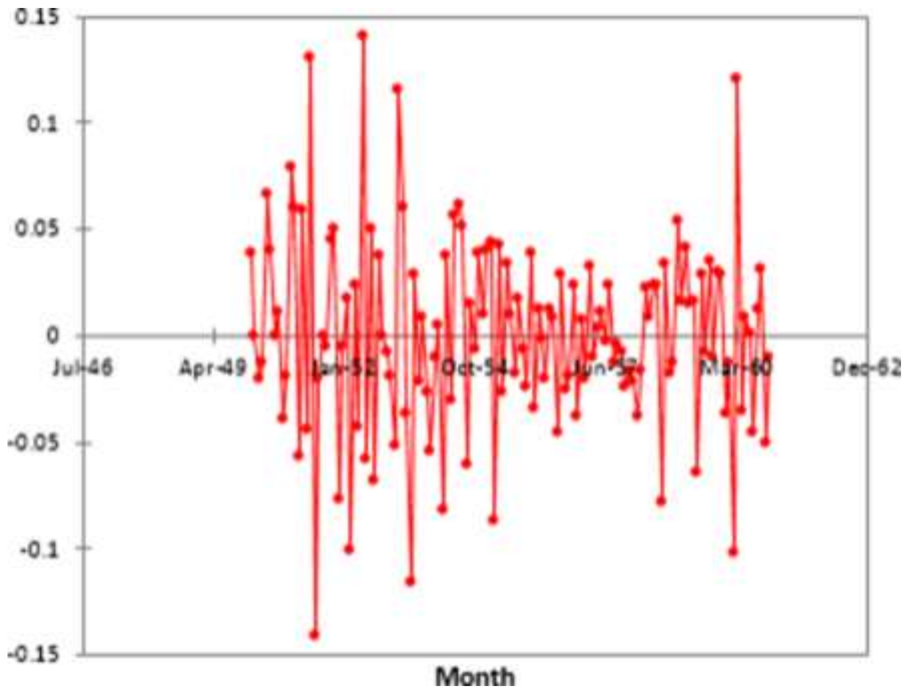
Before differencing

(non-stationary series)



After differencing

(stationary series)



Auto-regressive (AR) Models

In an auto-regressive $AR(p)$ model, we are learning regression coefficients for each of the p previous values

$$y_t = \beta_0 + \beta_1 y_{t-1} + \cdots + \beta_p y_{t-p} + \varepsilon_t$$

- A model with high autocorrelation implies that the data is highly dependent on previous values and that an auto-regressive model would perform well

Auto-regressive $AR(p)$ models

- Auto-regressive models are useful for learning falls or rises in our series
  - This will weight together the last few values to make a future prediction

- Typically, this model type is useful for small-scale trends such as an increase in demand or change in tastes that will gradually increase or decrease the series

Moving Average (MA) Models

In a moving-average $MA(q)$ model, we are learning regression coefficients for each of the q lag error terms

$$y_t = \mu + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$$

(with μ as the mean of the time series)

- Moving-average models attempt to predict the next value based on the overall average and how off our previous predictions were

Moving-average $MA(q)$ models

- Auto-regressive models slowly incorporate changes in the system by combining previous values
- MA models use prior errors to quickly incorporate changes

- This model is useful for handling specific or abrupt changes in a system, e.g., something going out of stock or a sudden rise in popularity affecting sales

ARMA Models

$ARMA$ (auto-regressive moving-average) models combine the auto-regressive AR and moving-average MA models

$$ARMA(p, q) = AR(p) + MA(q)$$

Incorporating both models allows us to mix two types of effects

- AR models slowly incorporate changes
  - E.g., in preferences and tastes
- MA models base their prediction on the prior error, allowing to correct sudden changes based on random events
  - E.g., supply and popularity spikes

AR, MA, and $ARMA$ models

$$AR(p) = ARMA(p, 0)$$

$$MA(q) = ARMA(0, q)$$

ARIMA Models

$ARIMA(p, d, q)$ (auto-regressive integrated moving-average model) models predict the differences of the series (as opposed to their value)

$$y_t - y_{t-1} = ARIMA(p, 1, q)$$

- $ARIMA(p, d, q)$ handles the stationarity assumption we wanted for our data. We don't need to *detrend* or *differentiate* manually, the model does this for us

d is the degree of differencing

$$y_t = ARIMA(p, 0, q) = ARMA(p, q)$$

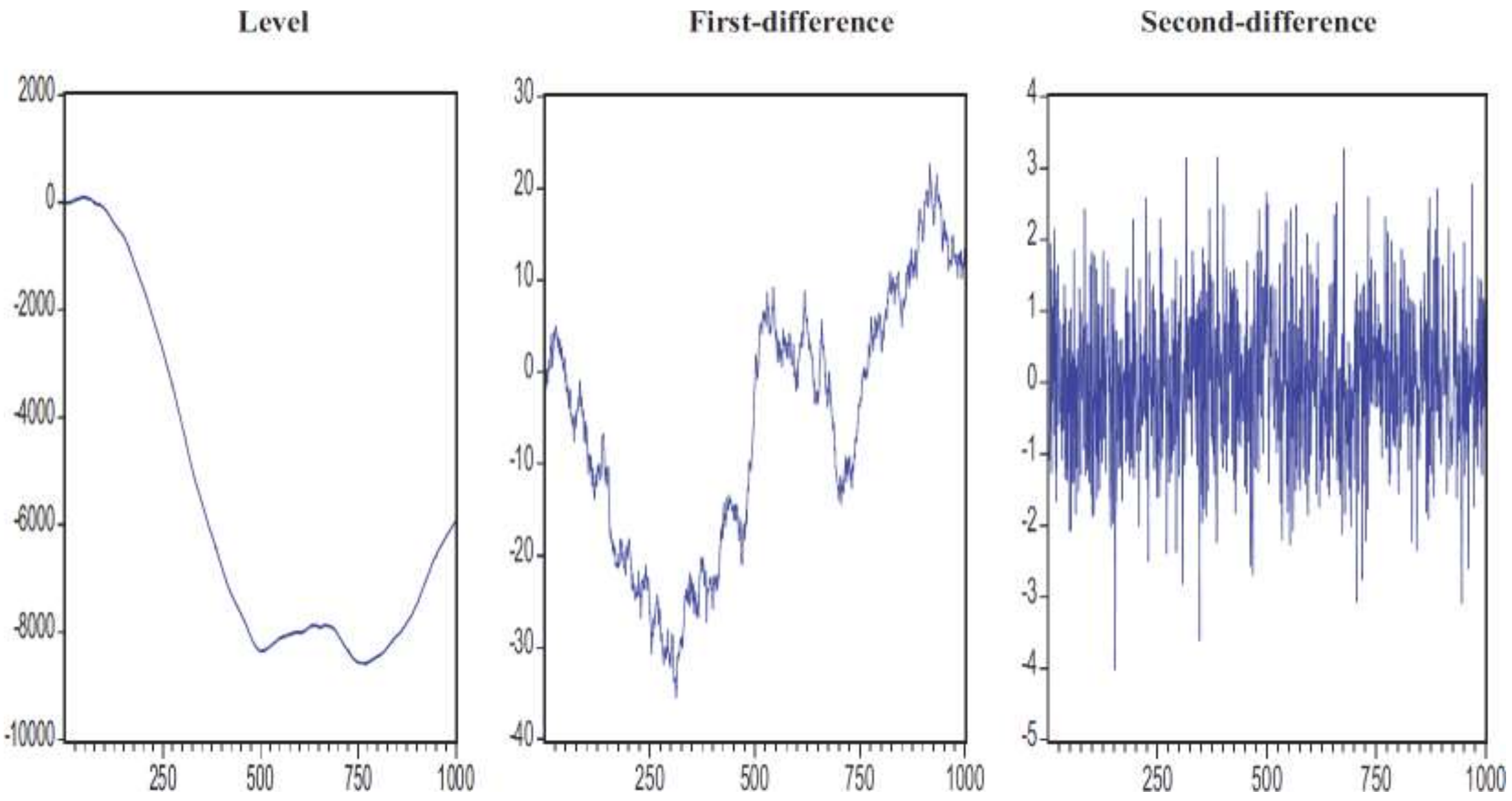
$$y_t - y_{t-1} = ARIMA(p, 1, q)$$

(a.k.a, the *first-difference*)

$$(y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) = y_t - 2y_{t-1} + y_{t-2} = ARIMA(p, 2, q)$$

(this is not the difference from two periods ago; rather, the *second-difference* is the first-difference of the-first difference, a discrete analog of a second derivative, i.e., the local acceleration of the series rather than its local trend)

d is the degree of differencing (cont.)



AR, MA, $ARMA$, and $ARIMA$ models

$$AR(p) = ARIMA(p, 0, 0)$$

$$MA(q) = ARIMA(0, 0, q)$$

$$ARMA(p, q) = ARIMA(p, 0, q)$$