STAT302: Time Series Analysis

Chapter 1. Introduction

Sangbum Choi, Ph.D

Department of Statistics, Korea University

Outline

Introduction to Time Series Data

Time Series Models

General Approach to TS Modeling

Time series (TS) data

- A **(univariate) time series** is a sequence of measurements of the same variable collected over time. Most often, the measurements are made at regular time intervals.
- One difference from standard linear regression is that the data are not necessarily independent and not necessarily identically distributed.
- One defining characteristic of time series is that this is a list of observations where the ordering matters. Ordering is very important because there is dependency and changing the order could change the meaning of the data.

Basic objectives of the analysis

The basic objective usually is to determine a model that describes the pattern of the time series. Uses for such a model are:

- To **describe** the important features of the time series pattern.
- To **explain** how the past affects the future or how two time series can interact.
- To **forecast** future values of the series.
- To possibly serve as a **control standard** for a variable that measures the quality of product in some manufacturing situations.

Important characteristics to consider first

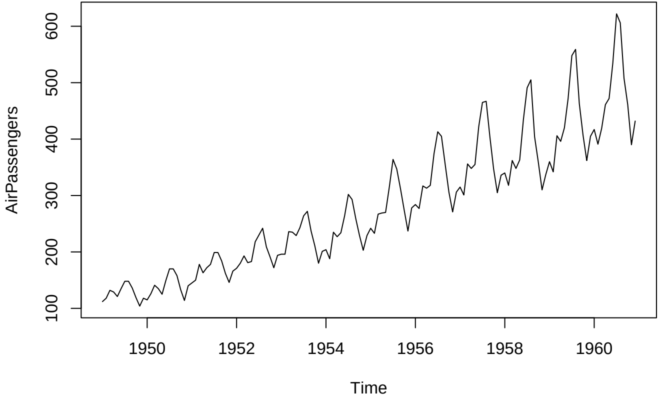
Some important questions to first consider when first looking at a time series are:

- Is there a **trend**, meaning that, on average, the measurements tend to increase (or decrease) over time?
- Is there **seasonality**, meaning that there is a regularly repeating pattern of highs and lows related to calendar time such as seasons, quarters, months, days of the week, and so on?
- Are their **outliers**? In regression, outliers are far away from your line. With time series data, your outliers are far away from your other data.
- Is there a **long-run cycle** or period unrelated to seasonality factors?
- Is there **constant variance** over time, or is the variance non-constant?

Example: Air passenger bookings

The number of international passenger bookings (in thousands) per month on an airline (Pan Am) in the United States for the period 1949–1960.

```r
ts.plot(AirPassengers)
```



Example: Air passenger bookings

There are a number of features:

- The number of passengers travelling on the airline is increasing with time. In general, a systematic change in a time series that does not appear to be periodic is known as a **trend**.
- A repeating pattern within each year is known as **seasonal variation**. The seasonal variation coincides with vacation periods.
- Sometimes we may claim there are **cycles** in a time series that do not correspond to some fixed natural period; examples may include business cycles or climatic oscillations such as El Nino.

Time series patterns

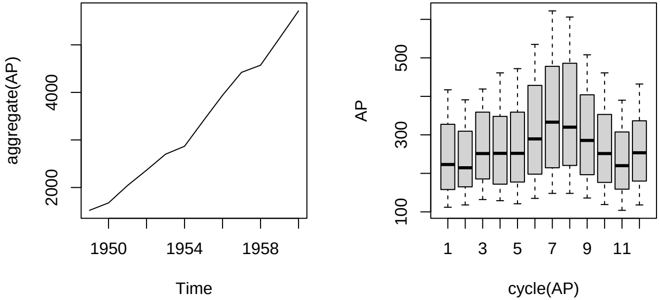
In describing these time series, we have used words such as **trend** and **seasonal**, which need to be defined more carefully.

- **Trend**: A trend exists when there is a long-term increase or decrease in the data.
- **Seasonal**: A seasonal pattern occurs when a time series is affected by seasonal factors such as the time of the year or the day of the week.
- **Cyclic**: A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency.

Example: Air passenger bookings

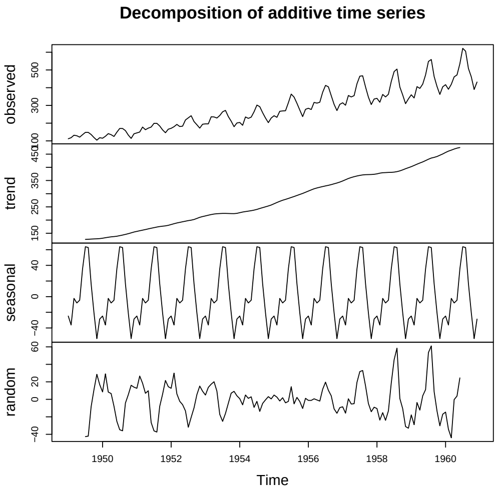
- To get a clearer view of the trend, the seasonal effect can be removed by **aggregating** the data to the annual level.
- You can see an increasing trend in the annual series and the seasonal effects in the boxplot.

```
AP=AirPassengers
par(mfrow=c(1,2))
plot(aggregate(AP))
boxplot(AP ~ cycle(AP))
```



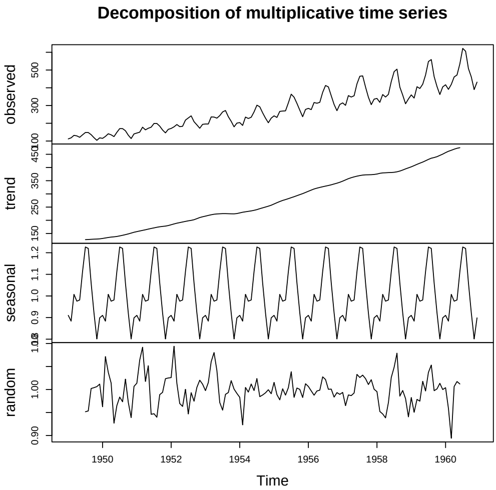
Time series decomposition

```r
plot( decompose(AirPassengers, type="additive") )
```



Time series decomposition

```
plot( decompose(AirPassengers, type="multiplicative") )
```



Example: Dow Jones Industrial (DJI) average

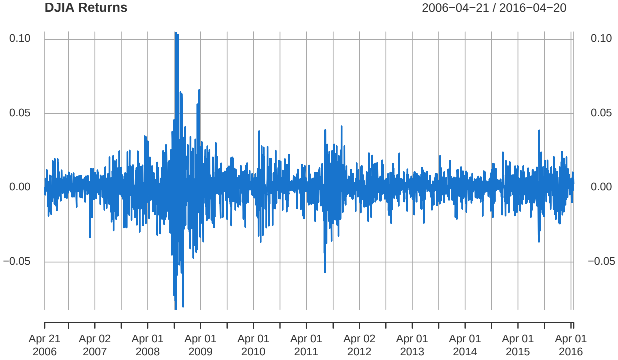
As an example of financial time series data, the following shows the daily prices and returns (or percent change) of the Dow Jones Industrial Average (DJIA) from April 20, 2006 to April 20, 2016.

```
library(astsa); library(xts)
plot(djia$Close[-1] , col=4, main="DJIA Prices")
```

Example: Dow Jones Industrial (DJI) average

Financial time series data have some special features, shown in this example; the volatility is generally clustered and persistent.

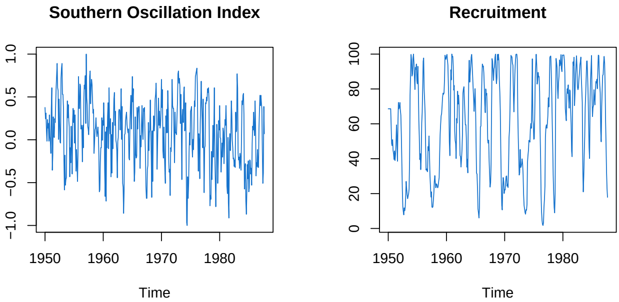
```
ldjiar = diff(log(djia$Close))[-1]
plot(ldjiar, col=4, main="DJIA Returns")
```



Example: El Nino and fish population

- We may also be interested in analyzing several time series at once $\Rightarrow$ **multivariate time series** data.
- Monthly values of the Southern Oscillation Index (SOI) and associated Recruitment (number of new fish)

```
par(mfrow = c(1,2))  # set up the graphics
plot(soi, col=4, ylab="", main="Southern Oscillation Index")
plot(rec, col=4, ylab="", main="Recruitment")
```



Manipulating time series data with xts & zoo

- The xts and zoo are two R packages that provide tools and functions for manipulating time series data.

```
library(xts)
ts_time = rnorm(100)
ts_index = seq(as.Date("2009-01-01"),
               length=100,
               by="month")
ts_data = xts(x = ts_time, order.by = ts_index)
ts_data["2017"]
##                   [,1]
## 2017-01-01 -0.62194416
## 2017-02-01 -0.38024168
## 2017-03-01 -0.10625383
## 2017-04-01 -0.09180913
```

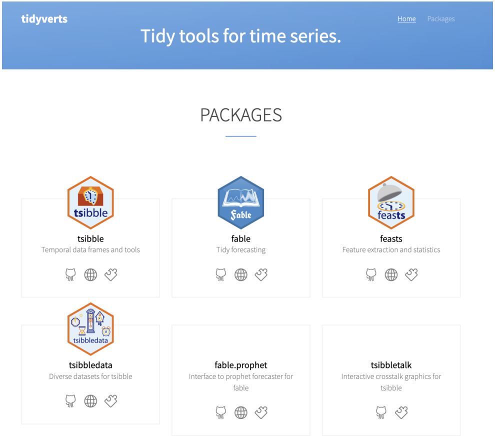
Manipulating time series data with tsibble

The tsibble package also provides a data infrastructure for tidy temporal data with wrangling tools.

```
library(tsibble)
ts_index = yearmonth("2009 Jan") + 0:99
ts_data = tsibble(ts_time, ts_index)
head(ts_data)
## # A tsibble: 6 x 2 [1M]
##   `<dbl>`  `<mth>`
##     <dbl>    <mth>
## 1 -1.20    2009 Jan
## 2  0.385   2009 Feb
## 3  0.0984 2009 Mar
## 4  3.18    2009 Apr
## 5  0.616   2009 May
## 6 -0.687   2009 Jun
```

Learn more about tsibble

An ecosystem, the `tidyverts`, is built around the `tsibble` object for tidy time series analysis (https://tidyverts.org).



Outline

Introduction to Time Series Data

Time Series Models

General Approach to TS Modeling

Time series models

- The primary objective of time series analysis is to develop mathematical models that provide plausible descriptions for sample data.
- A time series model is a probabilistic model that describes the different ways that the series data $\{X_t\}$ could have been generated.
- Not just a model for one observation at one time point. Rather a model for **the entire set of observations in time**. That is, find the joint distribution of observations.

Time series models

- More formally, a time series model is usually a probability model for $\{X_t, t \in T_0\}$, a collection of random variables (RVs) indexed in time (this is the population).
- The fundamental TS approach is to use the supposition that adjacent points in time are correlated, i.e.,

$$X_t \sim X_{t-1}, X_{t-2}, X_{t-3}, \ldots$$

White noise (pure noise) processes

- Perhaps the simplest model for a time series is one in which there is no trend or seasonal component.
- Such observations are simply independent and identically distributed (iid) random variables with zero mean.
- White noise (WN) or iid noise

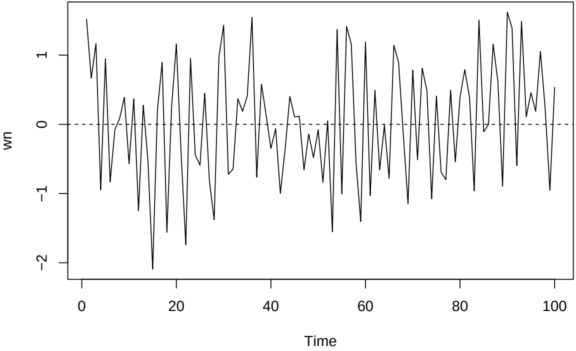
$$\varepsilon_t \sim \text{iid}(0, \sigma_w^2)$$

- Gaussian white noise (GWN)

$$\varepsilon_t \sim \text{iid } N(0, \sigma_w^2)$$

White noise (pure noise) processes

```
wn=rnorm(100)
ts.plot(wn)
abline(h=0,lty=2)
```



Moving averages (MA) and filtering

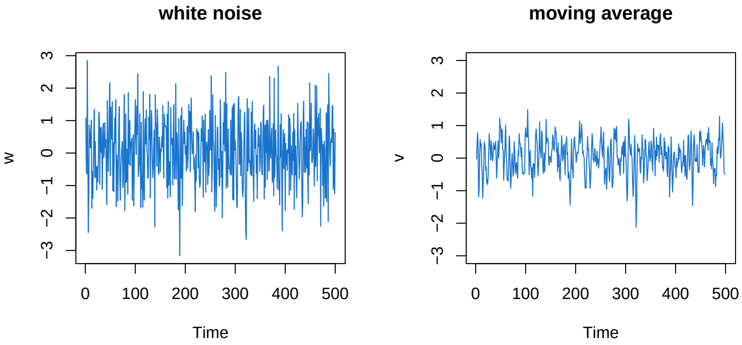
- We might replace the white noise series ε_t by a moving average (MA) that smooths the series.
- For example, consider replacing ε_t by an average of its current value and its immediate neighbors in the past and future. That is, let

$$V_t = \frac{1}{3}(\varepsilon_{t-1} + \varepsilon_t + \varepsilon_{t+1}).$$

- A linear combination of values in a time series is referred to, generically, as a **filtered** series:

Moving averages (MA) and filtering

```
w = rnorm(500,0,1)                       # 500 N(0,1) variates
v = filter(w, sides=2, rep(1/3,3))  # moving average
par(mfrow=c(1,2))
ts.plot(w, col=4, main="white noise")
ts.plot(v, col=4, ylim=c(-3,3), main="moving average")
```



Autoregressions (AR)

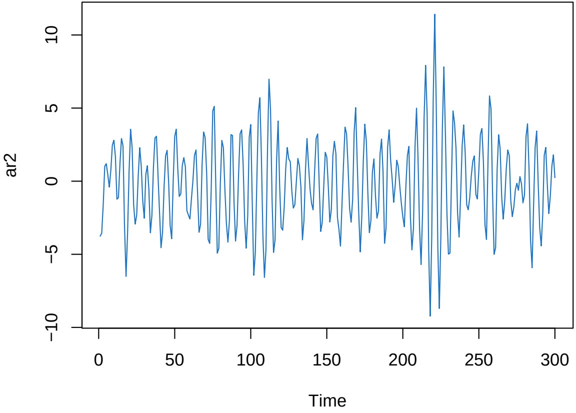
- Consider the second-order equation:

$$X_t = X_{t-1} - 0.9X_{t-2} + \varepsilon_t$$

- It represents a regression or prediction of the current value X_t of a time series as a function of the past two values of the series, and, hence, the term **autoregression** is suggested for this model.
- Later, we will call this model as an AR(2) process, which can be generated by `arima.sim` function in R.

Autoregressions (AR)

```
ar2 = arima.sim(n=300,model=list(order=c(2,0,0),ar=c(1,-0.9)))
ts.plot(ar2, col=4)
```



Random walk processes

- The random walk $\{X_t, t = 0, 1, 2, ...\}$ (starting at zero) is obtained by cumulatively summing (or "integrating") iid random variables.
- Thus a random walk with zero mean is obtained by defining $X_0 = 0$ and

$$X_t = \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_t, \;\; t = 1, 2, ...,$$

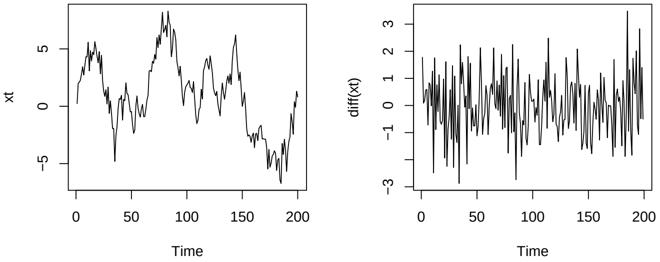
  where $\{\varepsilon_t\}$ is iid white noise.
- Equivalently, a random walk has the form

$$X_t = X_{t-1} + \varepsilon_t$$

Random walk processes

- A realization of length 200 of a simple symmetric random walk is shown.
- Notice that the outcomes can be recovered from $\{\varepsilon_t, t = 0, 1, \ldots\}$ by differencing.

```
xt=cumsum(rnorm(200))
layout(matrix(c(1,2),1,2))
ts.plot(xt); ts.plot(diff(xt))
```



Random walk with drift

- A model for analyzing a trend is the random walk with drift model given by

$$X_t = \delta + X_{t-1} + \varepsilon_t$$

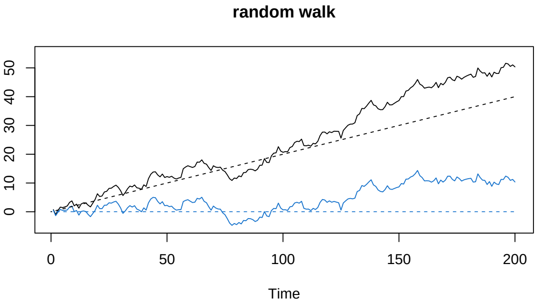
for $t = 1, 2, \ldots$, with initial condition $X_0 = 0$, and where ε_t is white noise.

- Note that we may rewrite it as a cumulative sum of white noise variates:

$$X_t = \delta t + \sum_{j=1}^{t} \varepsilon_t$$

Random walk with drift

```r
set.seed(154) # so you can reproduce the results
w = rnorm(200); x = cumsum(w) # two commands in one line
wd = w +.2; xd = cumsum(wd)
ts.plot(xd, ylim=c(-5,55), main="random walk", ylab='')
lines(x, col=4)
clip(0,200,0,50)
abline(h=0, col=4, lty=2)
clip(0,200,0,50)
abline(a=0, b=.2, lty=2)
```



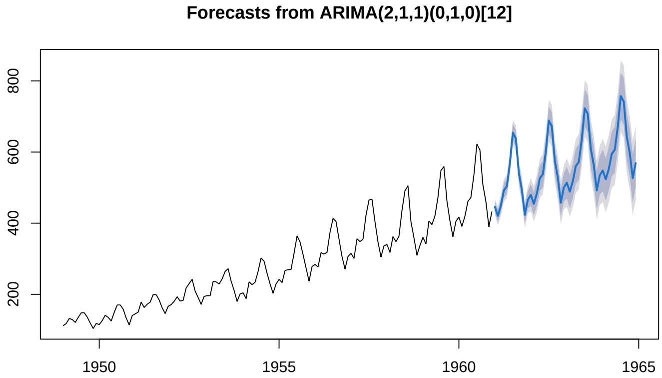
Outline

The forecasting process

- A process is a series of connected activities that transform one or more inputs into one or more outputs.
- The activities in the forecasting process are:
  1. Problem definition
  2. Data collection
  3. Data analysis
  4. Model selection and fitting 5. Model validation
  5. Forecasting model deployment
  6. Monitoring forecasting model performance

The forecasting processes

```
library(forecast); library(tidyverse)
AirPassengers %>% auto.arima() %>% forecast(h=48) %>% plot()
```



General approach to TS modeling

The examples of the previous section illustrate a general approach to TS analysis

1. Plot the series and examine the main features of the graph, checking in particular whether there is
   (a) a trend,
   (b) a seasonal component,
   (c) any apparent sharp changes in behavior,
   (d) any outlying observations.

General approach to TS modeling

2. Remove the trend and seasonal components to get stationary residuals ("detrending" and "deseasonalizing"):
    (a) regression approach;
    (b) smoothing;
    (c) differencing.
3. Choose a model to fit the residuals, making use of various sample statistics. Forecasting will be achieved by forecasting the residuals and then inverting the transformations to arrive at forecasts of the original series X_t.