Time Series Analysis

Nilotpal Sanyal
(nilotpal.sanyal@gmail.com)

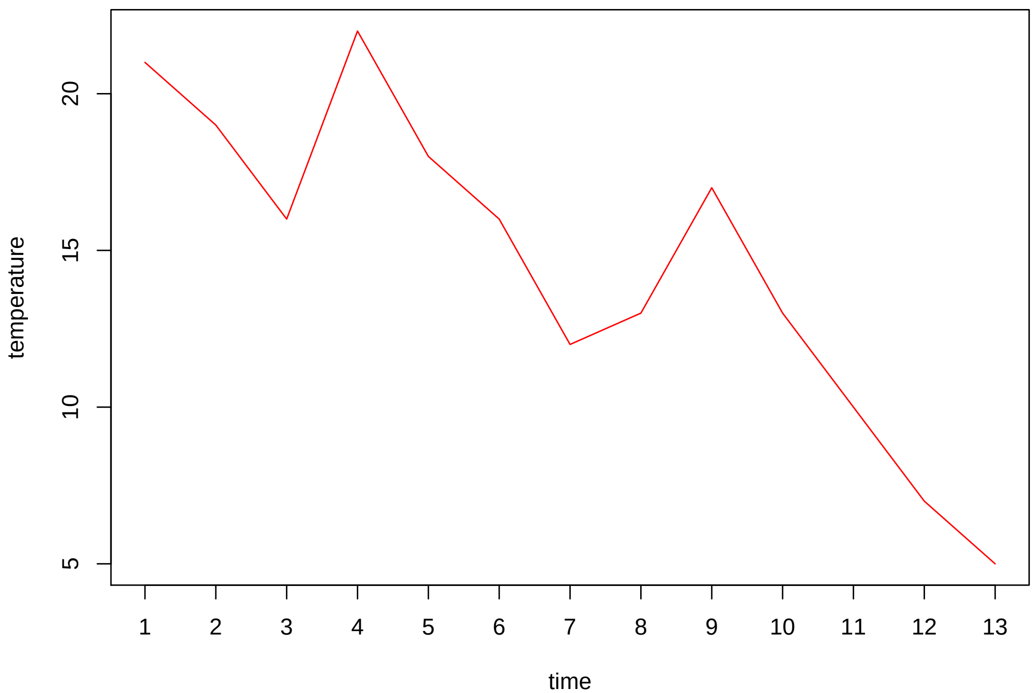
Bayesian and Interdisciplinary Research Unit,
Indian Statistical Institute

What is a time series?

A series of observations on a variable generated sequentially over time, typically at equal intervals.

Ex1: The daily lowest temperature (degree Celcius) in Chicago from September 1 to September 13, 2014:

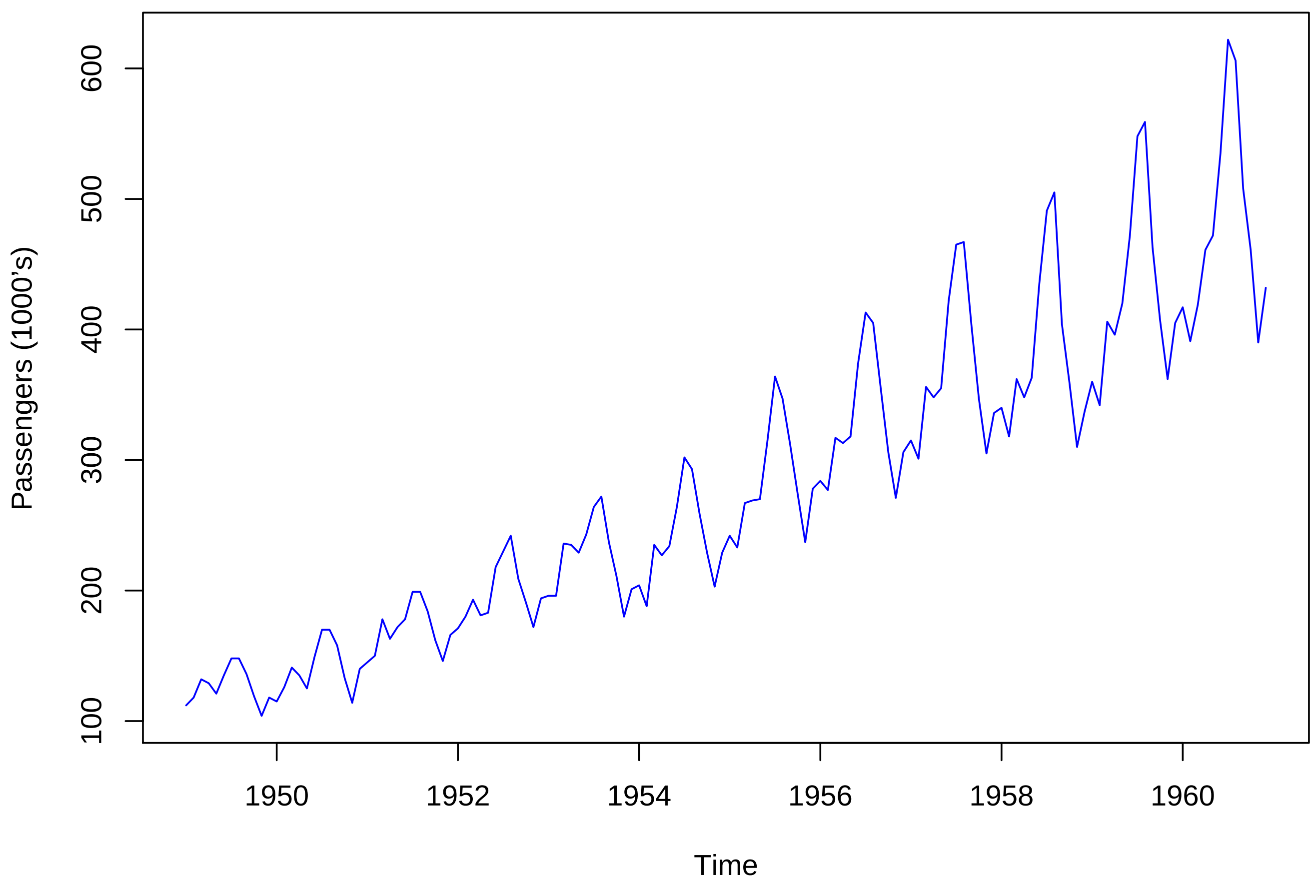
$$21, 19, 16, 22, 18, 16, 12, 13, 17, 13, 10, 7, 5.$$



Ex2: Airline data. Monthly totals (in thousands) of international airline passengers from 1949 to 1960.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1976) Time Series Analysis, Forecasting and Control. Third Edition. Holden-Day. Series G.



Where are time series used?

Time series are used in any domain of applied science and engineering which involves temporal measurements, for example, in statistics, signal processing, pattern recognition, econometrics, mathematical finance, weather forecasting, earthquake prediction, electroencephalography, control engineering, astronomy, communications engineering.

Goal of time series analysis and time series forecasting

The fluctuation of values in a time series is the joint action of a number of forces/components working together. The goal of time series analysis is to isolate and measure the effects of these various components. The goal of time series forecasting is to use a model to predict the future behavior of the time series based on its past behavior.

Time series data, cross-sectional data, longitudinal data, panel data

n = number of subjects
k = number of variables
t = number of time points

Time series data (univariate): n=1, k=1 and large t. Repeated observations on one variable at many time points for one subject.
Ex: Daily lowest temperature of Chicago in a month.

Time series data (multivariate): n=1, small k(>1) and large t. Repeated observations on several variables at many time points for one subject.
Ex: Daily lowest and highest temperatures of Chicago in a month.

Cross-sectional data (univariate): n>1, k=1 and t=1. Observations on one variable at single time point for several/many subjects.
Ex: Lowest temperature of 50 largest cities in a day.

Cross-sectional data (multivariate): n>1, k>1 and t=1. Observations on several variables at single time point for several/many subjects.
Ex: Lowest and highest temperatures of 50 largest cities in a day.

Longitudinal data or cross-sectional time series data (univariate): large n, k=1, small t. Repeated observations on one variable at several time points for many subjects.
Ex: Weekly lowest temperature of 50 largest cities in a month

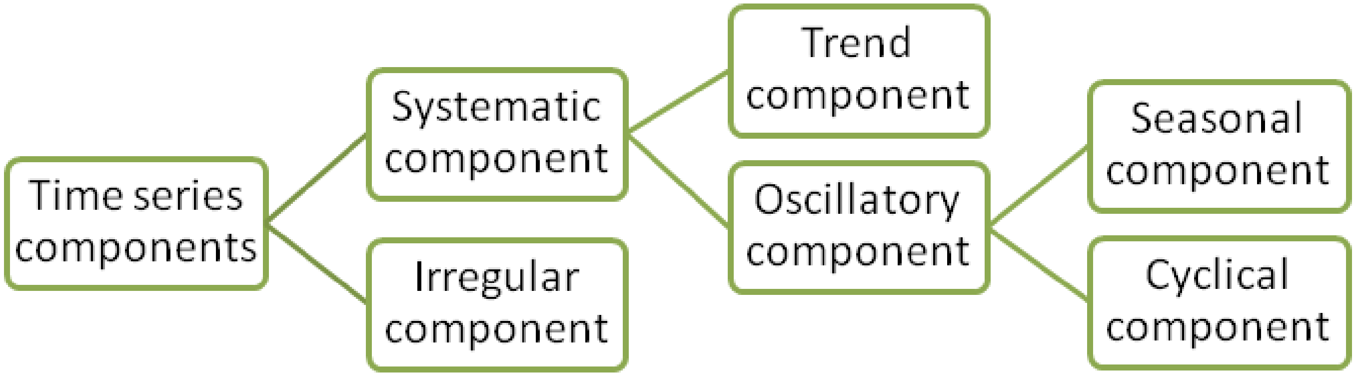
Longitudinal data or cross-sectional time series data (multivariate): large n, k>1, small t. Repeated observations on several variables at several time points for many subjects.
Ex:Weekly lowest and highest temperatures of 50 largest cities in a month

Panel data: Similar to longitudinal data.

How are time series analysis and regression analysis related?

They are different in focus of analysis. In regression analysis, which is a statistical process for estimating the relationships among variables, the focus is on the relationship between one (or more) dependent variable(s) and one or more independent variables. In time series analysis, the focus is on comparing values of a single time series or multiple dependent time series at different points in time. Anyways, one can incorporate time-series analysis tools into regression analysis.

Components of a time series:



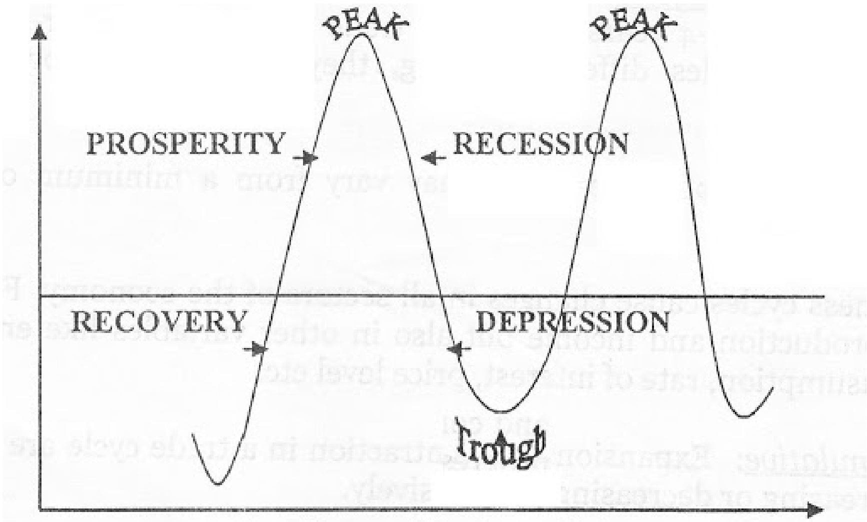
Systematic component: Can be identified, measured and accounted for.

Irregular component: Is random and cannot be accounted for. Ex. change due to flood, famine, war etc.

Trend: Smooth regular long-term tendency of the series to increase or decrease.

Seasonal component: An oscillatory movement which repeats itself at regular intervals of time (periodic), with period of oscillation known and no longer than one year. Ex. sale of ice cream goes up in summer.

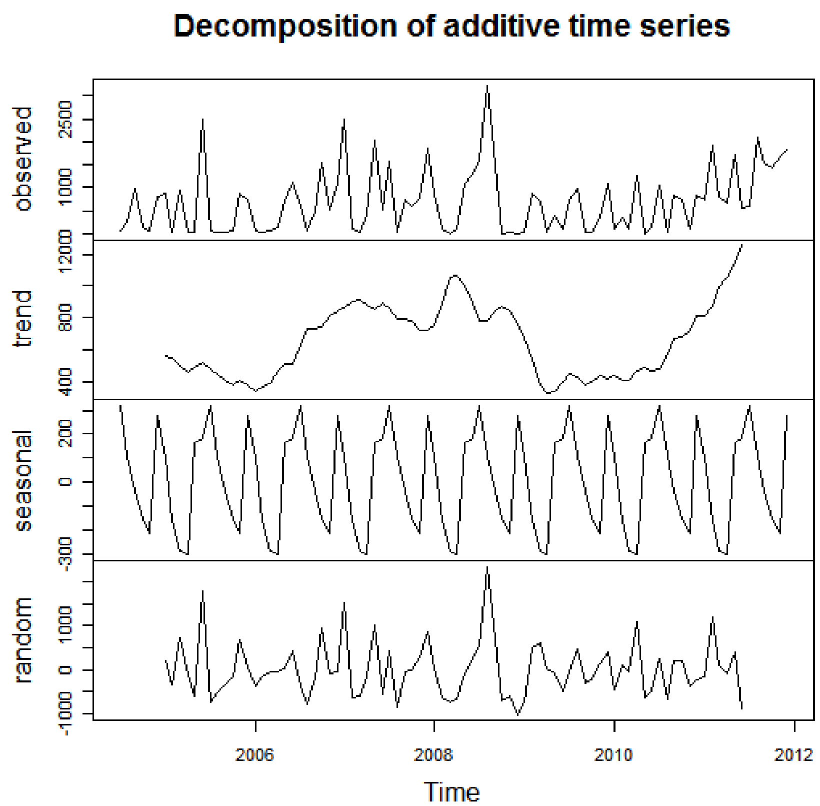
Cyclical component: An oscillatory movement which repeats itself at differing intervals of time (not periodic), with period of oscillation more than one year. Ex. business cycles with four phases prosperiry, recession, depression and recovery.



Four Phases of Business Cycle

Source:http://kalyan-city.blogspot.com/2011/06/4-phases-of-business-cycle-in-economics.html

In time series analysis, we go on measuring and accounting for each systematic component and any other pattern until only the irregular component is left.



Source:http://www.simafore.com/Portals/64283/images/time-series-forecasting-using-r-to-decompose-resized-600.png

Notation:

Let us consider univariate time series and use the following notations.

For time points $t = 1, 2, 3, \ldots,$

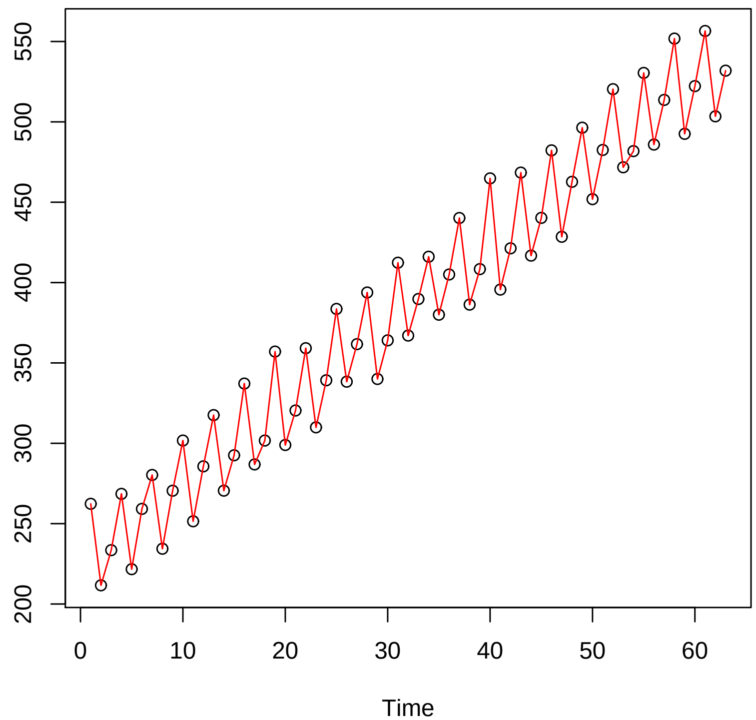
y_t = time series data
T_t = trend
S_t = seasonal component
C_t = cyclical component
I_t = irregular component

What models are used for time series?

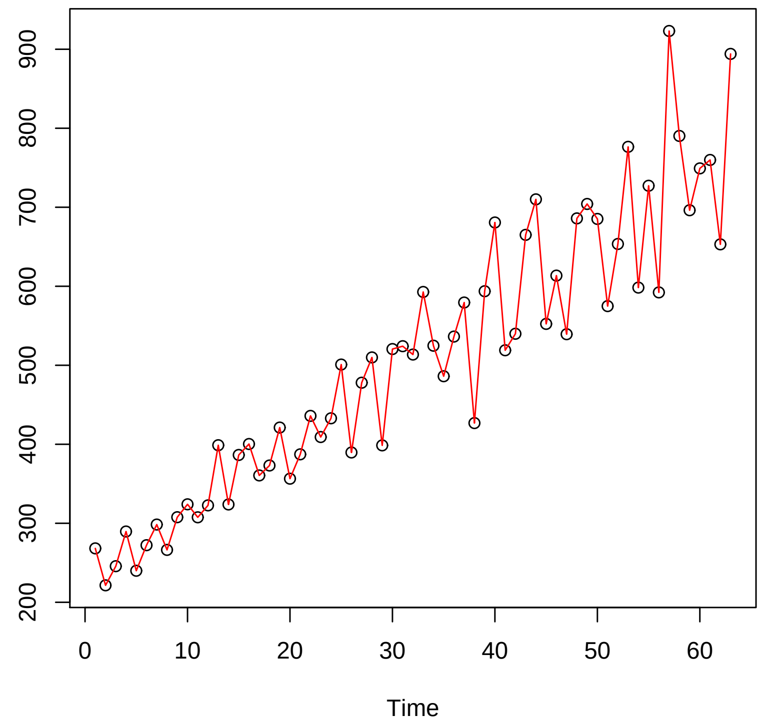
- Additive model: $y_t = T_t + S_t + C_t + I_t$.

- Multiplicative model: $y_t = T_t \times S_t \times C_t \times I_t$.

Choose the additive model when the magnitude of the seasonal pattern in the data does not depend on the magnitude of the data, that is, does not change as the series goes up or down. Consider the following time series plot. The seasonal variation looks to be about the same magnitude across time, so an additive decomposition might be good.



Choose the multiplicative model when the magnitude of the seasonal pattern in the data depends on the magnitude of the data, that is, increases as the data values increase, and decreases as the data values decrease. Consider the following time series plot. The seasonal variation increases as we move across time. A multiplicative decomposition could be useful.



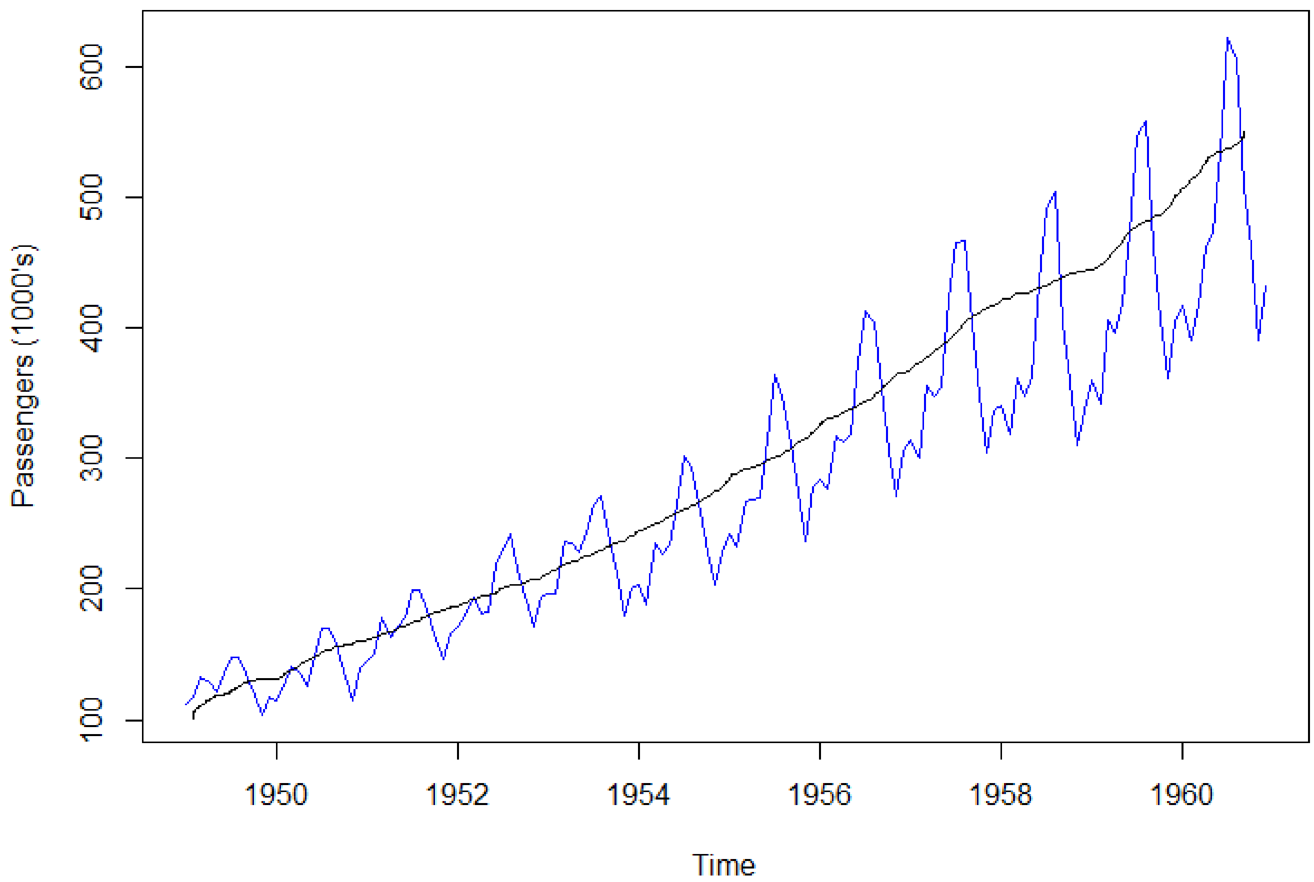
How to measure trend (T_t)?

- To measure T_t, we have to eliminate S_t, C_t and I_t from y_t.
- S_t can be eliminated by taking average/total over the period.
- C_t and I_t can be eliminated by the following methods.

1) Free-hand curve method.
2) Semi average method.
3) Moving average method.
4) Least square curve fitting method.

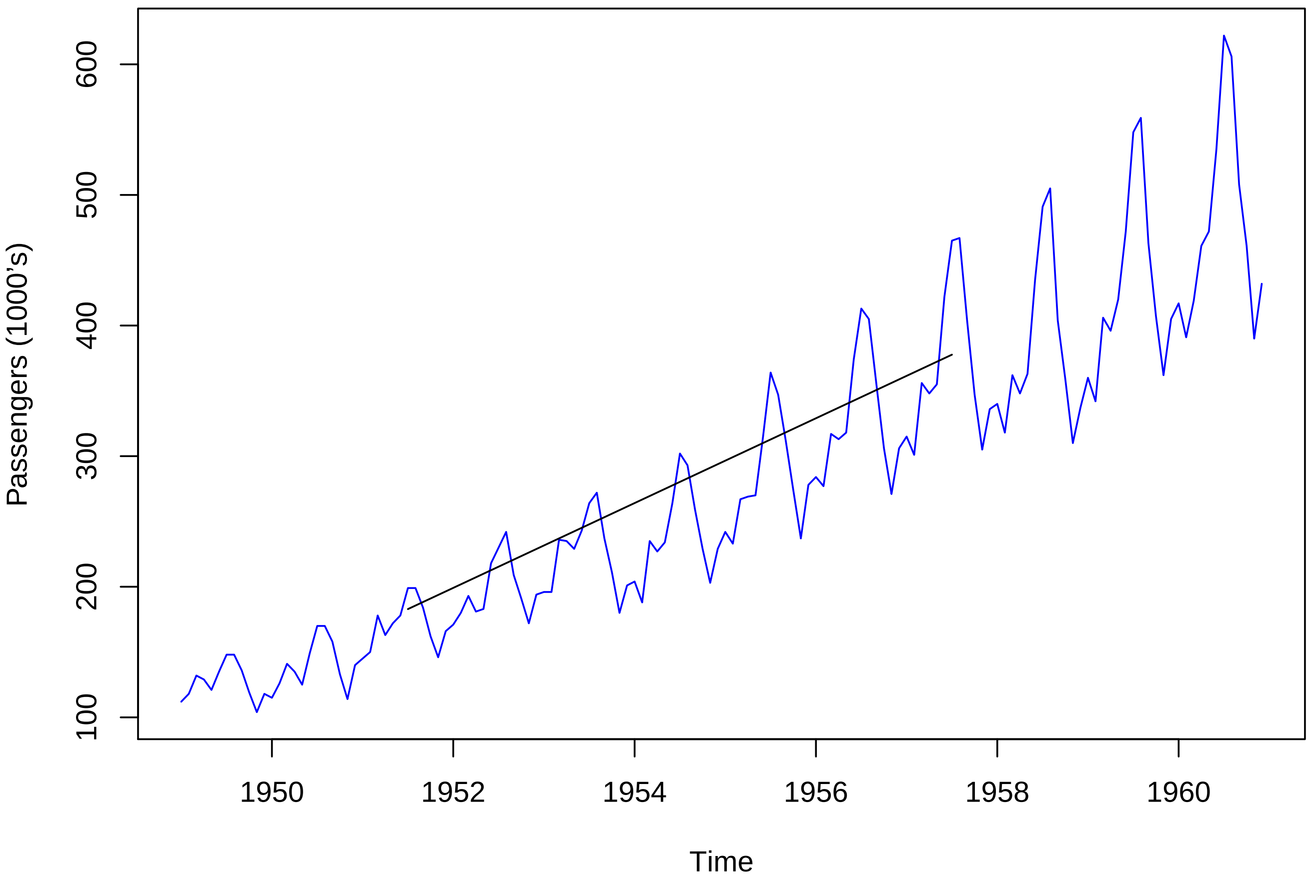
1) Free-hand curve method

Draw line diagram of data with time t in X-axis and y_t in Y-axis.
Draw a smooth free-hand curve through the plotted points.



2) Semi average method

Divide the data into two halves (if odd number of observations, omit the middlemost one). Compute mean of each halves. Plot those means against the midpoints of the two halves. Join the plotted points and extend both sides.



3) Moving average (MA) method

Popular method. For MA values with period k, compute the means of each successive set of k consecutive observations. Plot the moving averages and join the points.

		3 month MA
Jan	112	
Feb	118	120.667
Mar	132	126.333
Apr	129	127.333
May	121	128.333
Jun	135	134.667
Jul	148	143.667
Aug	148	144.000
Sep	136	134.333
Oct	119	119.667
Nov	104	113.667
Dec	118	.
.	.	.

		4 month MA	2 item MA
Jan	112		
Feb	118		
		122.750	
Mar	132		123.875
		125.000	
Apr	129		127.125
		129.250	
May	121		131.250
		133.250	
Jun	135		135.625
		138.000	
Jul	148		139.875
		141.750	
Aug	148		139.750
		137.750	
Sep	136		132.250
		126.750	
Oct	119		123.000
		119.250	
Nov	104		.
		.	
Dec	118		.
		.	
.	.		.

4) Least square curve fitting method

Mostly used method. Determine an appropriate trend equation by looking at plots or otherwise. Then estimate the parameters of the trend equation mathematically by applying least square method.

Examples of trend equation are
- $T_t = a + bt$, linear trend
- $T_t = a + bt + ct^2$, parabolic trend
- $T_t = a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k$, k'th degree polynomial trend
- $T_t = ab^t$, exponential trend
- $T_t = a + bc^t$, modified exponential trend
- $T_t = abc^t$, Gompertz trend
- $T_t = \frac{K}{1+e^{a+bt}}, b < 0$, logistic trend

Least square method: For any trend equation T_t, mathematically minimize

$$S = \sum_t (y_t - T_t)^2$$

by taking partial derivative of S w.r.t the constants (a,b etc.) and setting them to zero, which leads to the set of equations

$$\frac{\partial S}{\partial const} = 0, const = a, b, \ldots$$

Solving the above set of equations, we get the estimates of the constants ($\widehat{a}$, $\widehat{b}$ etc.) and thus T_t is determined.

How to measure seasonal component (S_t)

Seasonal component can be measured by the following methods.

1) Seasonal average method.
2) Ratio to moving average method.
3) Ratio to trend method.
4) Link relative method.

1) Seasonal average method

Simplest measure of S_t. Assumes that the time series contains no trend (T_t) and no cyclical component (C_t). Not of much practical utility.

- Seasonal avergaes = Total of seasonal values/Number of years (eliminates I_t)
- General average = Total of seasonal averages/Number of seasons
- Seasonal indices = Seasonal averages/General average

2) Ratio to moving average method

- Calculate centered 12-month moving averages Y_t (will represent T_t and C_t).
- Moving average ratio $u_t = y_t/Y_t \times 100$ (will represent S_t and I_t).
- Preliminary seasonal indices $\bar{y}_i$ = average of u_t over the years - Adjusted seasonal indices $S_i = \bar{y}_i/\bar{y} \times 100$, where $\bar{y} = \sum_{i=1}^{12} \bar{y}_i/12$.

2) Ratio to trend method

- Using an appropriate trend equation, obtain the monthly trend values Y_t for each year.
- Trend ratio $u_t = y_t/Y_t \times 100$ (will represent S_t, C_t and I_t).
- Preliminary seasonal indices $\bar{y}_i$ = average of u_t over the years - Adjusted seasonal indices $S_i = \bar{y}_i/\bar{y} \times 100$, where $\bar{y} = \sum_{i=1}^{12} \bar{y}_i/12$.

4) Link relative method

Not much used now.

How to measure cyclical component (C_t)?

1) Residual method.
2) Harmonic analysis method.

1) Residual method - First determine T_t and S_t. - Then obtain $u_t = y_t/(T_t S_t)$. C_t can be obtained by removing I_t from u_t. - Moving average with suitable period can remove I_t. But difficult to find suitable period. - So, less practical utility.

2) Harmonic analysis method - Periodogram analysis.

Pattern in residuals: Autocorrelation

After accounting for trend, seasonality and cyclicality, if the residuals look random then there is nothing left to exploit. But if the residuals still show a pattern, then that might indicate the presence of 'autocorrelation' or 'serial correlation' that can be accounted for by taking into model.

At this point, we need to know the stochastic (probabilistic) representation of time series.

Stochastic representation of time series

- Stochastically, the observations in a time series are treated as realizations of random variables. - A sequence of random variables defined at fixed sampling intervals is sometimes referred to as a discrete-time stochastic process; shorter name time series model is often preferred. - Time series is denoted by $\{y_t : t \in T\}$.

Autocorrelation

$\rho(t, t+h) = \frac{cov(y_t, y_{t+h})}{\sqrt{V(y_t)}\sqrt{V(y_{t+h})}}$, autocorrelation function. h is called lag.

Stationary time series

- Mean doesn't change in time – no trend, no seasons, no cycles.
- Variance doesn't change in time.
- Correlations don't change in time.

Up to here, weakly stationary.

- Joint Distributions don't change in time (t).

That makes it strongly stationary.

A time series $\{y_t : t \in T\}$ is stationary if

- $E(y_t) = constant$, for all $t \in T$
- $V(y_t) = \sigma^2(constant)$, for all $t \in T$
- $\gamma(h) = cov(y_t, y_{t+h})$ is a function of lag h only, but not of t, for all $t, t+h \in T$.

$\gamma(h)$ is called auto-covariance function. Here, autocorrelation is $\rho(h) = \frac{\gamma(h)}{\sigma^2} = \frac{\gamma(h)}{\gamma(0)}$.