Lecture 1

1 Time Series data: Examples and Basic Concepts

Time series analysis deals with analysis of economic data recorded over time. The typical data set will be a macroeconomic time series, e.g. GDP, interest rates, price indexes, although micro time series are becoming more readily available.

In general, in time series observations are generally dependent. Such data series are modelled as random sequences or random (stochastic) processes.

Typical problems in time series analysis are:
a) modeling and testing;
b) forecasting;
c) signal extraction.

Forecasting and signal extraction naturally leads to the use of statistical models/ tools and criterions of success, e.g. estimation of minimum mean square error in case of forecasting.

General observations:

- Data $X_1, X_t, \cdots, X_n$ collected in time often shows dependence. Basic statistical courses assume data to be independent. Hence, it needs to be treated as time series and analyzed accordingly.
- Most of software packages are user friendly and allow to use time series and forecasting.
- Time series: it is best learned by applications similarly as "learning how to swim"
- The joy of time series: discover hidden information in the data during applications.

About this course:

- Understanding comes via intuition and simple examples.
- we keep theory and examples in balance

We discuss:

1. why we observe correlation when data is collected in time?
2. why we seek the simplest model when else is the same?
3. We discuss graphical tools: they are always recommended before any rigorous statistical analysis.
4. We discuss concepts of stationarity and stationary ARMA models.
5. We discuss model building. George Box: "All models are wrong but some are useful". Note: more models than one can fit equally well.
6. We discuss fitting nonstationary models and other problems.

1.1 Examples

In many fields we study data collected over time.
Sampling frequency: monthly, daily, weakly, 10 minutes etc.

The sequence of observations $X_1, X_2, \cdots$, generates times series, e.g.

- the closing prices of the stock market
- country's unemployment rate
- temperature readings.

Time series data are used to understand:
a) dynamics of the system;
b) make sensible forecasts about future behavior.

Why correlation is observed? Most economical processes exhibit inertia, and do not change that quickly.

This combining with sampling frequency makes data correlated. Hence, modeling methods must account for dependence in the data.

Examples of time series can be found in different fields: finance, economics, operational management and so on

Example 1 GNP (Gross national product) of the US.

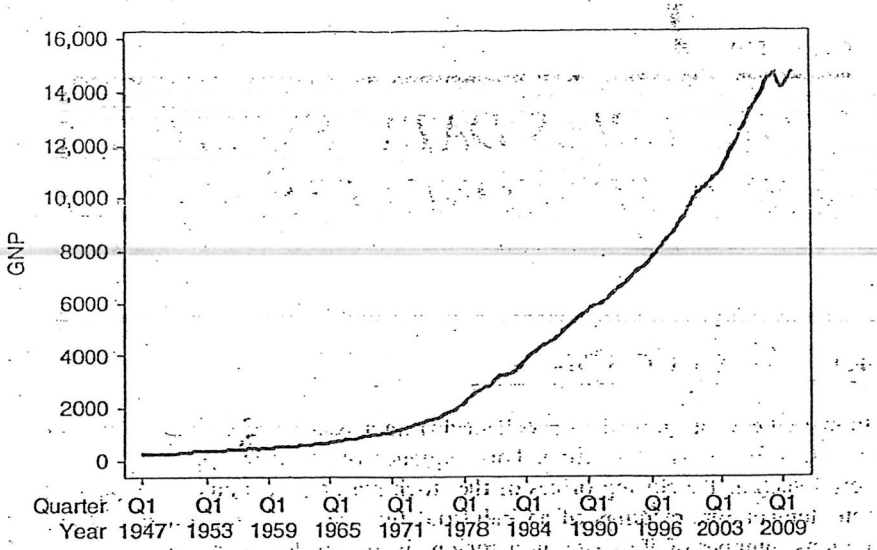

Figure 1.1 GNP (nominal) of the United States from 1947 to 2010 (in billion dollars). *Source:* US Department of Commerce, http://research.stlouisfed.org/fred2/data/GNP.txt.

Observations:

1. Hiccup in the end of period starts 3rd Quarter 2008 (related to financial crises originating from problems in real estate market)
2. Studying such macroeconomic indices is crucial. They help to identify general trends in national economy, impact of public policies, influence of global economy.

Example 2 Real estate market. US median sale prices of houses from 1988 - 2nd Quarter 2010.

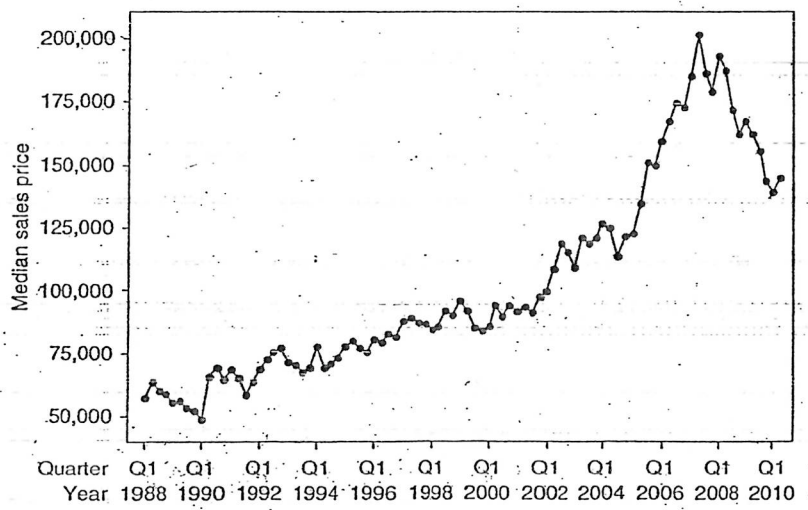

Figure 1.2 Median sales prices of houses in the United States. *Source:* US Bureau of the Census, http://www.census.gov/hhes/www/housing/hvs/historic/index.html.

Observations:

1. Upcoming crises could be noticed as early as 2007.
2. More crucial issue: find what will happen next.

Homeowner: will prices fall further?

Buyers: has the market hit the bottom?

These forecast might be possible using appropriate models.

Example 3 Business is interested in inventory (survey of stock) and sales data. Fig 1.3 shows number of airline passengers in 1949-1960.

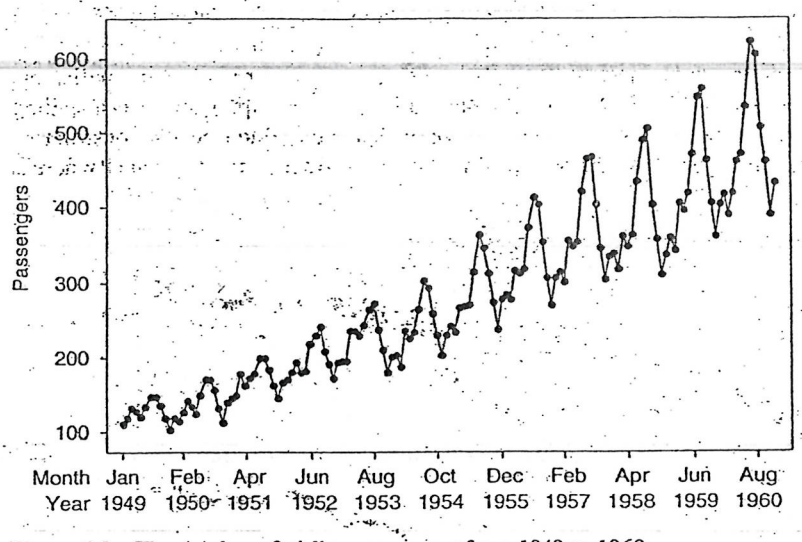

Figure 1.3 The number of airline passengers from 1949 to 1960.

Observations:

1. Cyclical travel pattern, data exhibits seasonal behavior
2. Upward trend: travel becomes more popular
3. Analysis of such data helps resource allocation and investment efforts

Example 4 Quarterly dollar sales in ($1000) of Marshal field company 1960-1975. Fig 1.4

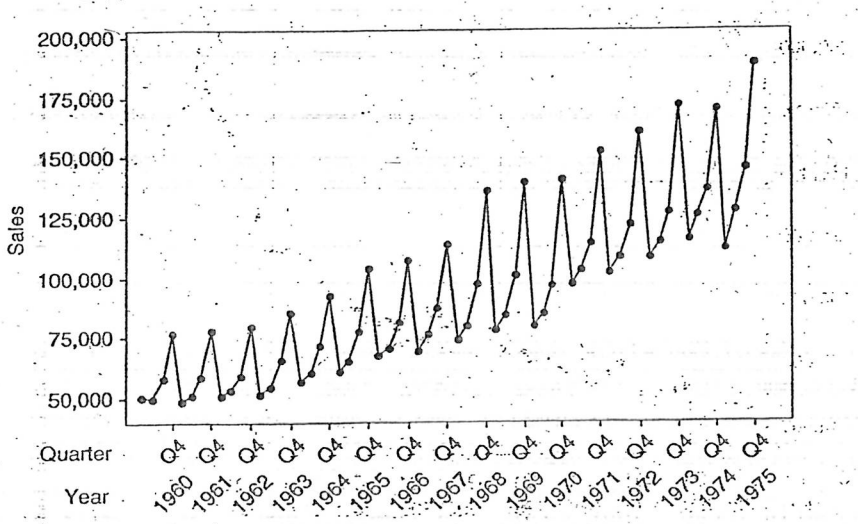

Figure 1.4 Quarterly dollar sales (in $1000) of Marshall Field & Company for the period 1960 through 1975.

Observations:

1. Seasonal pattern, increase of sales in Q4 (Christmas period)
2. For stock inventory this data contains invaluable information.

Example 5 Leading indicator for a variable of interest. Fig 1.5: Building permits is leading indicator for economy influenced by construction activities.

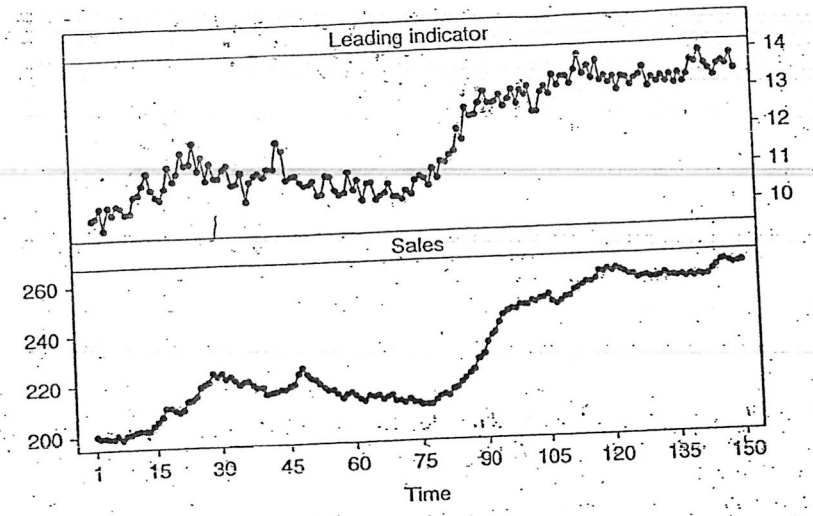

Figure 1.5 Time series plots of sales and a leading indicator.

Observations:

1. Leading indicator and sales show similar patterns.

2. Question: does there exist long terms relationship between these two series? If yes, then from the leading indicator, one can determine sales in the near future.

Example 6 Interrupted time series (by policy changes, strikes, new advertisement, campaigns)

Fig. 1.6: shows market share fight between "Colgate Dental Cream" and "Crest Toothpaste".

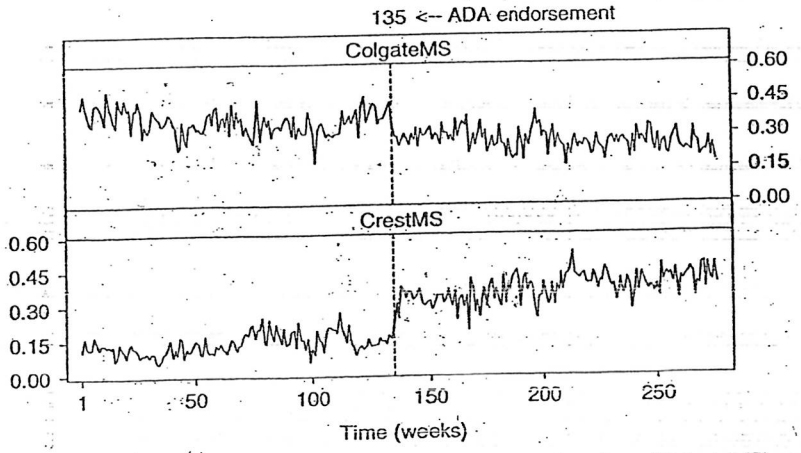

Figure 1.6 Time series plot of the weekly Colgate market share (ColgateMS) and Crest market share (CrestMS).

Observations:

1. Before 1960 Colgate enjoyed 50% market share.

2. In 1960 American Dental Association endorsed Crest as an important aid of dental hygiene.

3. From data one can see impact of endorsement to market share. Was the effect permanent or temporary?

Example 7 (Stationary data). Fig 1.7. Hourly reading from ceramic furnace. Series looks stationary in the sense that the mean and the variance do not seem to vary over time.

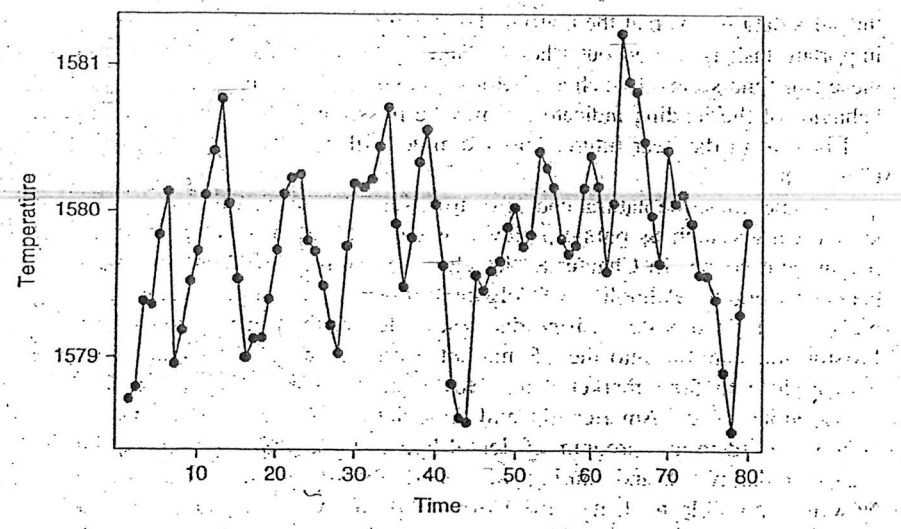

Figure 1.7 A time series plot of 80 consecutive hourly temperature observations from a ceramic furnace.

Example 8 (non-stationary behavior in mean). Concentration and temperature readings of a chemical process, Fig 1.8 and 1.9.

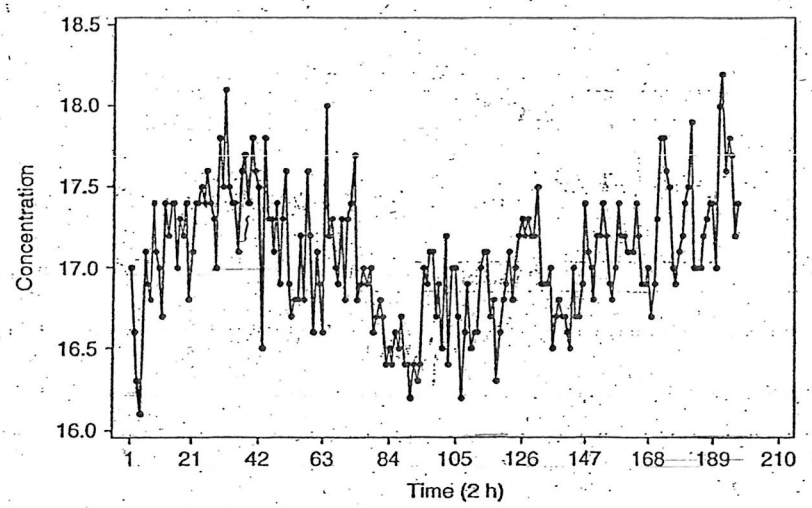

Figure 1.8 Time series plot of chemical process concentration readings sampled every 2 h (BJR series A).

Time series not limited to economics, finance, engineering

Example 9 Internet users over 100 min period. Data wanders around showing signs of "nonstationarity".

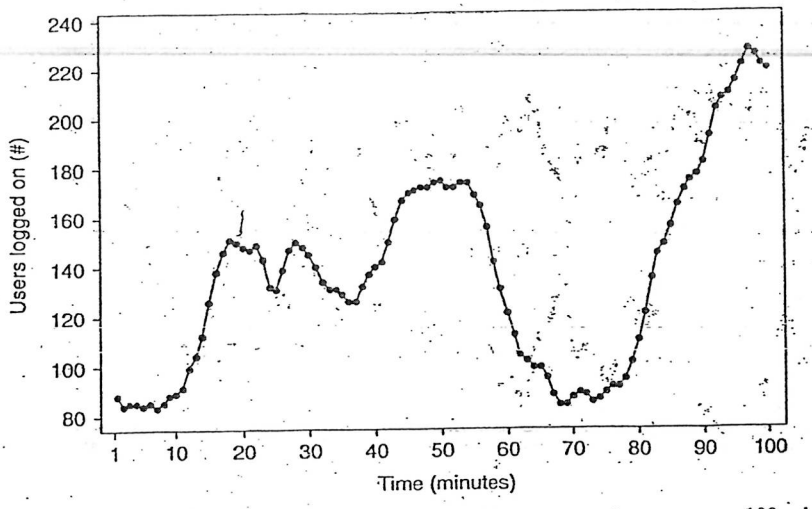

Figure 1.11 Time series plot of the number of internet server users over a 100-min period.

Example 10 Sea level for Copenhagen from 1889-2006.
Observations: Data shows stationary behavior and some subtle increase during the last couple decades.

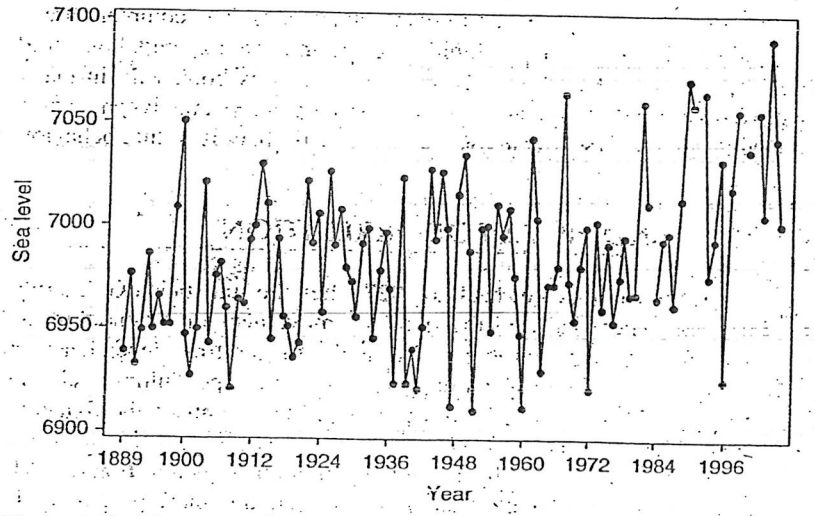

Figure 1.12 The annual sea levels in millimeters for Copenhagen, Denmark. *Source:* www.psmsl.org.

Questions:

1. what can officials expect in the near future?
2. can we make generalizations about sea levels in the other places from this data?

Example 11 Healthcare. Pandemic flu cases in the USA in 2009

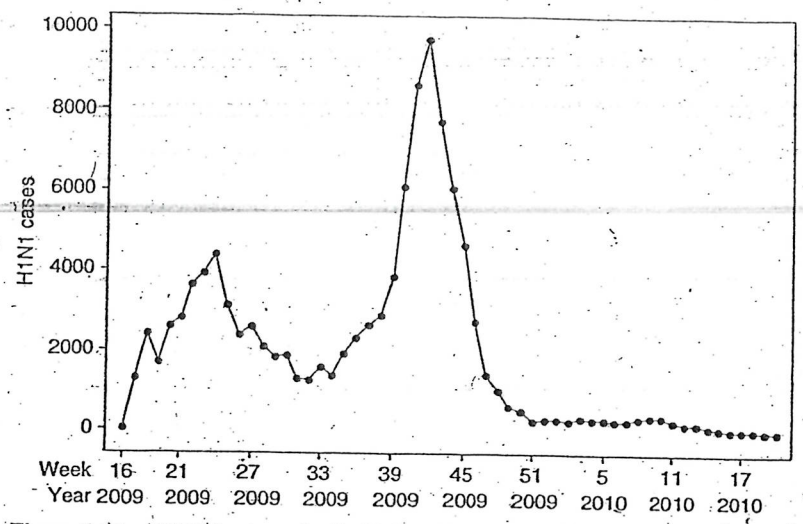

Figure 1.13 H1N1 flu cases in the United States from week 16 of 2009 to week 20 of 2010. *Source:* US Center for Disease Control CDC.

Questions:

1. can we predict the number of flue cases in 2010?
2. what is the reason of decline of the number of cases in the end of 2009? Vaccination campaign? "wash your hands" campaign? People's improved immune system?

Appropriate analysis would help to prepare for the new flue season.

We discuss the tools and methodologies for identifying underling patterns and dynamics, that will allow the analyst to make forecast of future behavior.

Idea of AR dynamics

Idea of autoregressive models for the time series belongs to Yule (1927). Intuition come from pendulum movement under the influence of gravity hit by a single impulse force.

Imagine: Pendulum in equilibrium, but boys throwing pebbles to swing it back and forth. The forces affecting pendulum shown in Fig 1.15.

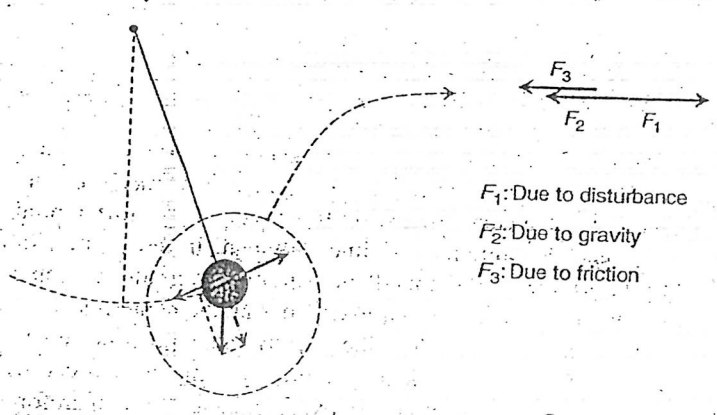

Figure 1.15 A simple pendulum in motion.

After initial impulse, pendulum will slow down. How fast? Depends is it a short or long pendulum, is it heavy or light, is friction small or large. (Imagine: pendulum= economy).

The harmonic movement $x(t)$ of pendulum is shown in Fig 1.16.

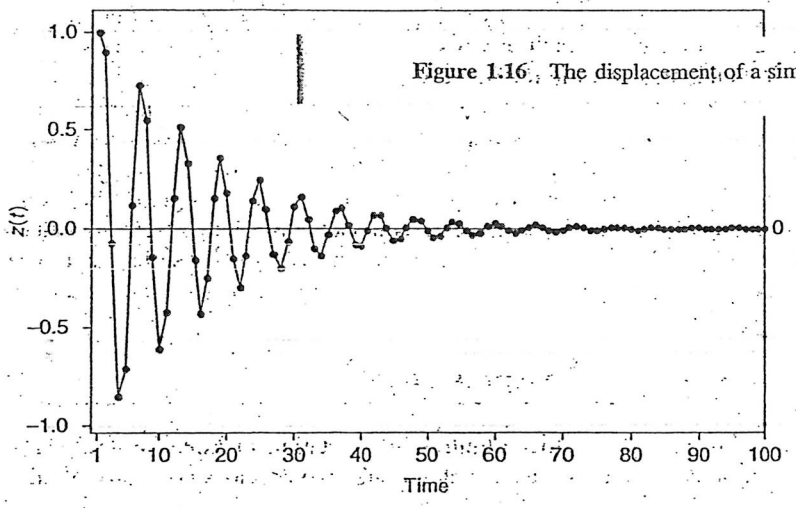

Figure 1.16. The displacement of a simple pendulum in motion.

Note: In continuous time such movements is described by second order differential equation.

In discrete time, replace:

1st derivative by the first difference $\nabla x_t = x_t - x_{t-1}$,

2nd derivative by the second difference

$$\nabla^2 x_t = \nabla(x_t - x_{t-1}) = x_t - 2x_{t-1} + x_{t-2}.$$

After simple rearrangement, we get equation

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \varepsilon_t$$

which is called second order autoregressive time series model.

Here x_t depends on the last two observations x_{t-1} and x_{t-2} and the errors terms ε_t. It follows second order dynamic.

Note: not all time series will be AR(2). We do not have a prior knowledge of such dynamics.

Hint: In time series we will model current observation using previous at appropriate lags which we need to choose.

Impulse response function

In AR(2) model the current observation x_t depends on two previous positions and the current disturbance ε_t.

If equation is valid for x_t it should be valid also for x_{t-1}. Then

$$\begin{aligned} x_{t-1} &= \phi_1 x_{t-2} + \phi_2 x_{t-3} + \varepsilon_{t-1}, \\ x_t &= \phi_1(\phi_1 x_{t-2} + \phi_2 x_{t-3} + \varepsilon_{t-1}) + \phi_2 x_{t-2} + \varepsilon_t \\ &= \phi_1^2 x_{t-2} + \phi_1\phi_2 x_{t-3} + \phi_2 x_{t-2} + \phi_1\varepsilon_{t-1} + \varepsilon_t. \end{aligned}$$

So, x_t depends on disturbances ε_{t-1}, ε_t. By the same argument, it depends on all previous disturbances:

$$x_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots$$

Stationary time series x_t: it does not depend on time shifts: its mean and variance are constant, and autocovariance $Cov(x_t, x_{t+k})$ depends only on the lag k.

Fundamental result A stationary time series x_t can be written as infinite weighted sum of uncorrelated shocks ε_t.

In AR(2) model: some combinations of weights ϕ_1, ϕ_2 will lead to a stable stationary solution, some not.

General idea: one can approximate stationary time series using models with only a few parameters, like ARMA models.

Why should we fit simplest possible model? Because it is easier to understand, interpret and explain.
A. Einstein: "everything should be made as simple as possible, but not simpler"

Impulse response function. A stationary time series x_t (dynamic system, e.g. a pendulum, US economy or something else) can be seen as the impulse response function to random shocks ε_t.

It provides us with information how the impulse propagates through the system and what effect it has over time.

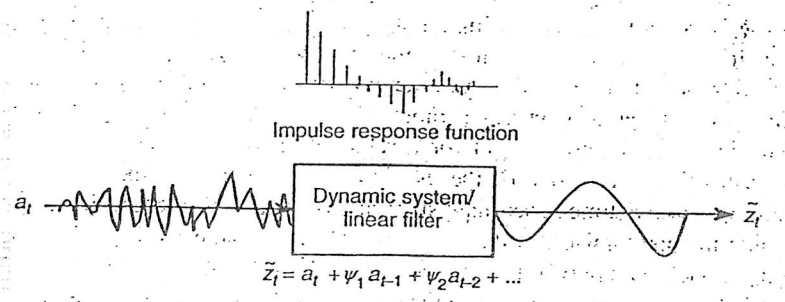

Figure 1.17 A time series model as a linear filter of random shock inputs.

Example. Consider impulse response x_t generated by AR(2) model $x_t = 0.98x_{t-1} - 0.37x_{t-1} + \varepsilon_t$ with a single shock $\varepsilon_0 = 1$, while other shocks are zero: $\varepsilon_j = 0, j \geq 1$. Then the impulse response function will be as in Fig 1.18.

Fig 1.20 shows the response of AR(2) model x_t to 10 shocks.

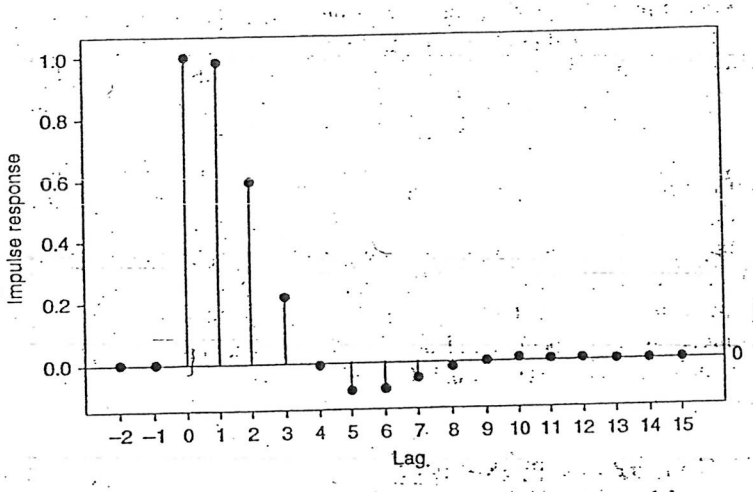

Figure 1.18 Impulse response function for the AR(2) for the pendulum.

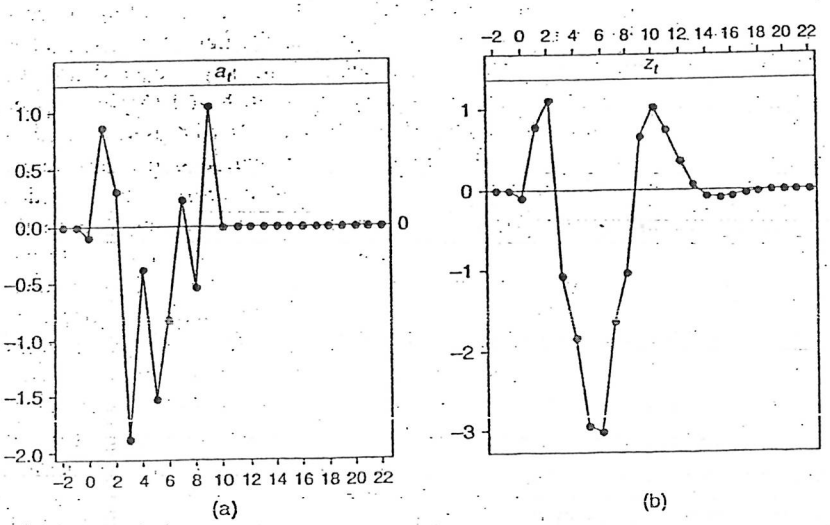

Figure 1.20 (a) Ten independent random white noise shocks $a_t, t = 1, \ldots, 10$ and (b) the superimposed responses of a linear filter generated by the AR(2) model $\tilde{x}_t = 0.9824\tilde{x}_{t-1} - 0.3722\tilde{x}_{t-2} + a_t$.

2 Graphical tools

Properly constructed graphs of times series can:

- dramatically improve statistical analysis
- accelerate discovery of hidden information in the data

First recommendation in data analysis:

1. "plot the data"
2. try to be creative at this

One should avoid mechanical automatic approach to time series analysis using standard software packages
Anscombe (1973): "good statistics is not a purely routine matter, it calls for more than one pass through the computer"

Note: distinguished features of time series lie in its correlation structure. Summary statistics, taken alone, for example, the mean, might be misleading - one needs to take into account dependence and distribution.

Carefully scrutinized plots of time series reveal important features which otherwise could be easily missed.

Graphical analysis of time series

Graphical analysis is essential aspect of time series analysis. It is not less important than the statistical models.

Graphical analysis:

- we learn from it what the data is trying to say.
- In most cases, graphical and mathematical modeling go hand in hand.
- Visualization is important in all steps of analysis and allows to avoid embarrassing mistakes.
- We need to develop a "feel" (experience) for the data which comes from hands-on work with the data.
- From graphical analysis we learn about process behavior and discover relationships.

Graphics is important in data cleaning.

It is not necessarily obvious and trivial and requires skill, techniques and time.

We will discuss

- visualization techniques, and how our eyes and brains process graphical information
- do's and don'ts in graphical data analysis

Note: data always include peculiar patterns. Outliers and strange patterns indicate something important and unusual and need to be carefully examined- they carry information.

Famous saying about baseball: "you can observe a lot by watching" is true also with time series analysis.

Statistical graphing: process involving plotting, thinking, revising, replotting, until graph becomes informative to a reader.

It should display data as accurately and clearly as possible. It highlights important pattern in the data.

There are many ways looking at data, but not all equally good.

When we construct graphic we encode data into some graphical elements and symbols.

Then the reader readers, he/she decodes the elements, see Fig 2.1

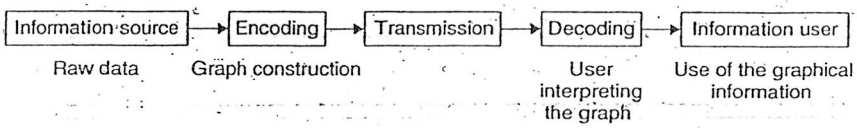

Figure 2.1 Graphical transmission of information.

Graph terminology

Terminology provides meaning to words and concepts.

main terms are defined graphically in Fig. 2.2, 2.3 and 2.4

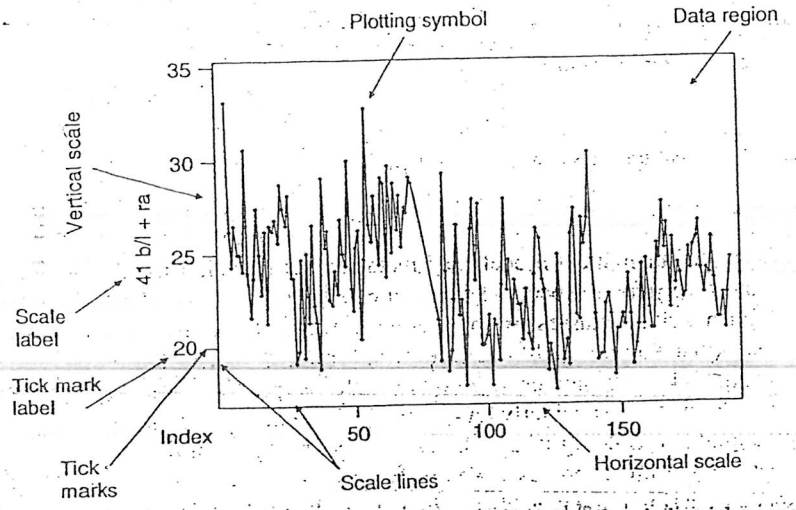

Figure 2.2 A typical time series graph of the temperature of an industrial process. Superimposed on the graph are a number of graph concepts.

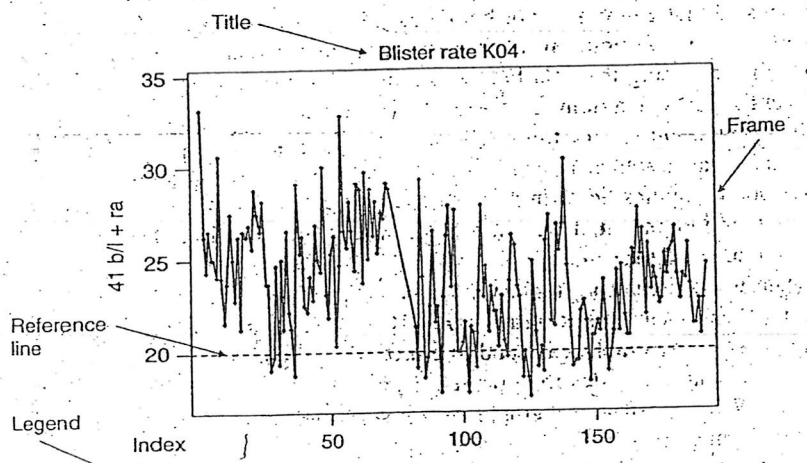

Figure 1. Times series plot of line 41 blister and raised versus time, June and July 2002.

Figure 2.3 The same plot as in Figure 2.2 with additional graphical elements such as legend, title, and reference line.

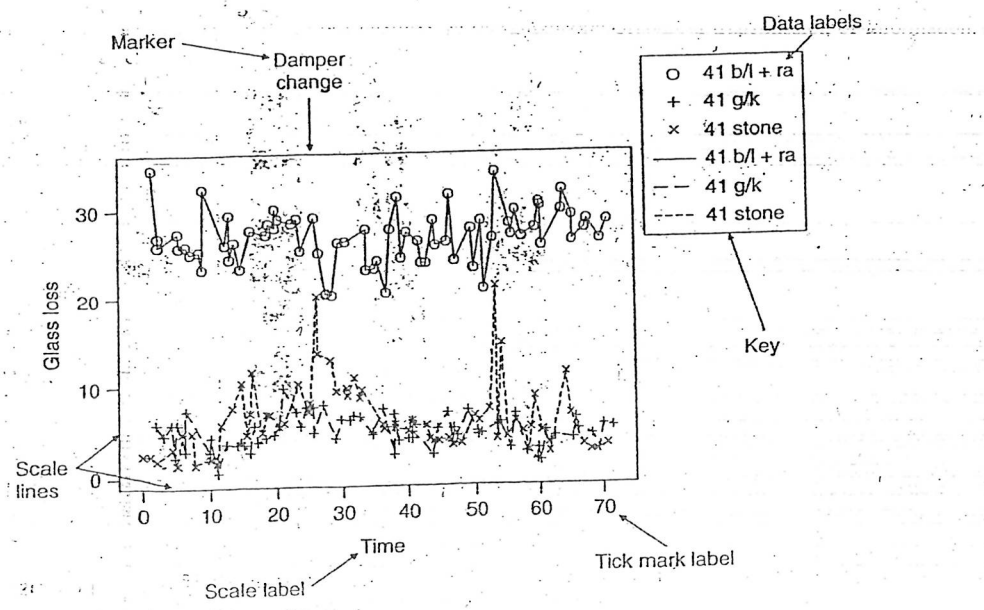

Figure 2.4 Superimposed time series plots.

2.1 Graphical perception

Statisticians and psychologists investigated how human brain processes the data.

Example (Pie chart). This chart very popular in economics and business, when we want to display fraction of the whole.

Fig. 2.5: five categories A, B, C, D, E make up the whole. Their fractions are 23, 21, 20, 19, and 17%.
Drawback : looking at the graph impossible to determine differences between categories.

Pie chart is good when objective is to show that one category is dominating.

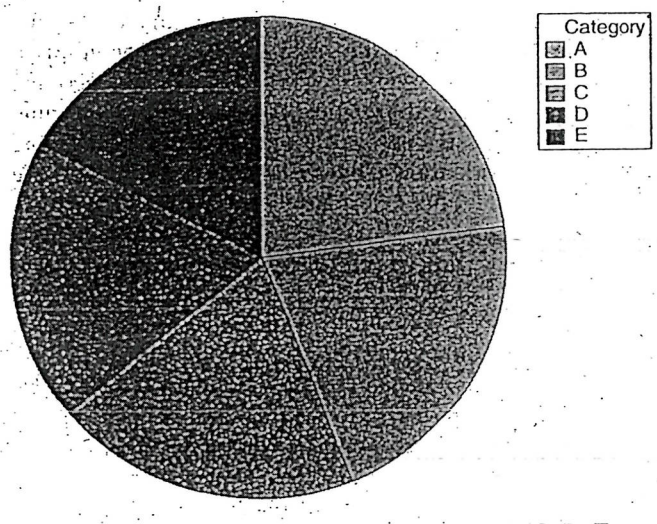

Figure 2.5 A pie chart of the five categories, A, B, C, D, E.

Bar chart: it privides information about size of the categories. We see that A is larger than B.

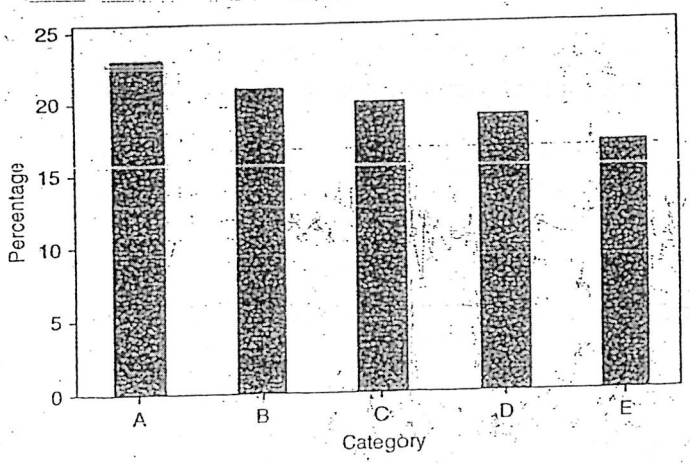

Figure 2.6 A bar chart of the same five categories.

General issues: making a graph, data is encoded into graph by a number of mechanisms:

- shape of the graph
- selection of plotting symbols
- scale

That allows to convey us information the way we prefer, accurately, so another person can visually decode the graph. Graphs can also hide information and be used for propaganda purposes.

Example (Common baseline issue).

Fig 2.7: Bars A1 and B1 have common base, we can compare them: they are of unequal length.

Bars A2 and B2 are the same as A1 and B1, but do not have common baseline, so it is difficult to see that they are different.

Bars A3 and B3 are of the same length and position as A2 and B2. Adding a reference box, which provides a scale, we see that A3 and B3 are unequal.

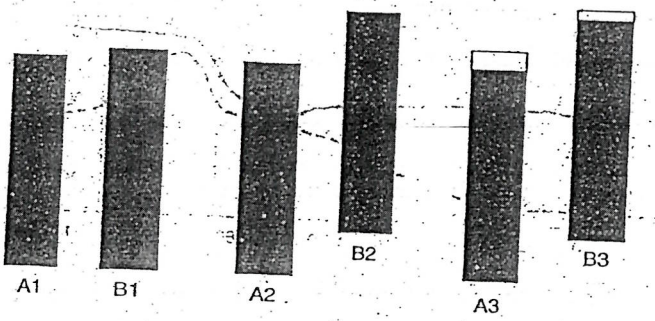

Figure 2.7 Bar charts on common and different baselines.

Example (Differences between superimposed graphs).

Often problem is not detecting difference between curves, but our perception of that difference.

Fig 2.8 shows early imports between England and East India.

Observations: difference is small between 1755-1770. Problem: we tend to compare horizontal distance rather than vertical.

Fig 2.9 provides difference between imports and exports.

We see: trading deficit is not constant 1775-1770, and has sudden increase around 1765.

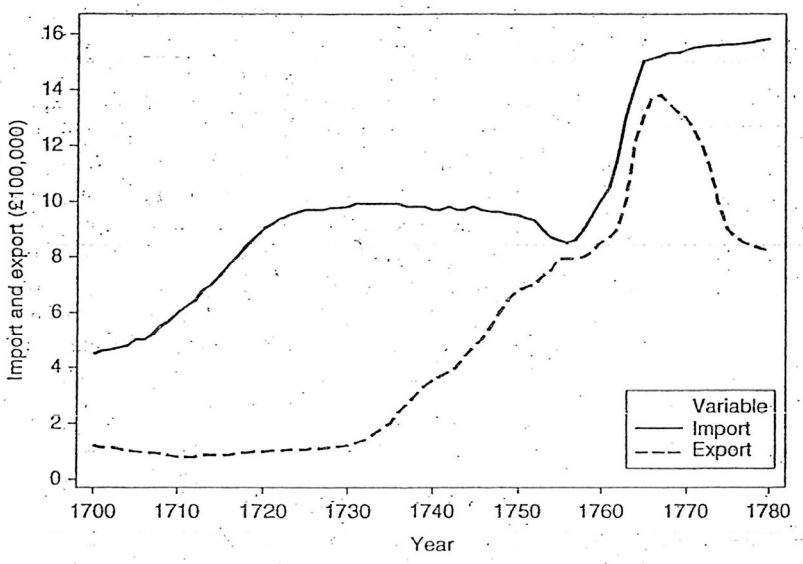

Figure 2.8 Import and export between England and East Indies from 1700 to 1780.

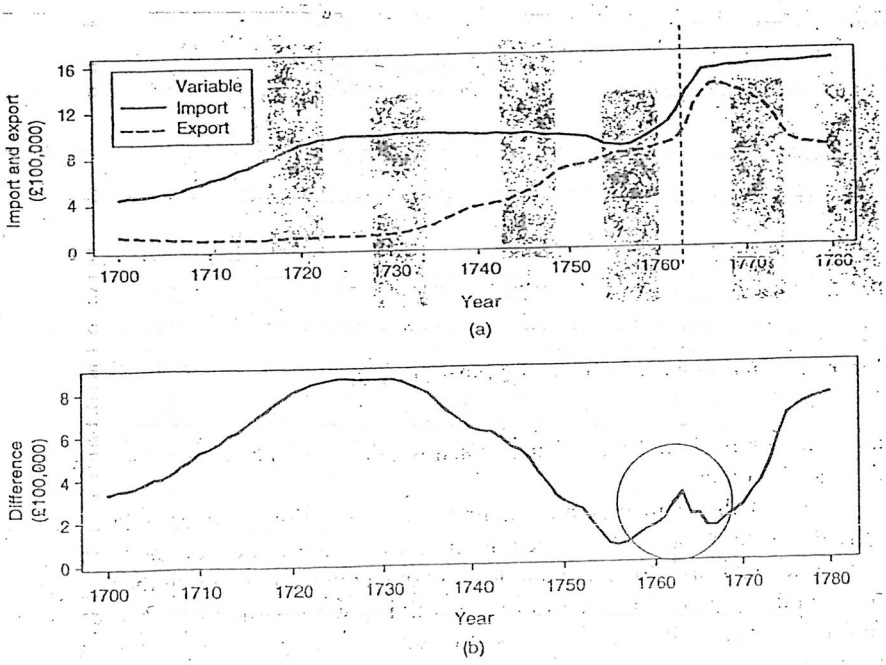

Figure 2.9 (a) Import and export between England and East Indies from 1700 to 1780 and (b) the difference between the import and the export.

Example (optical illusion).

Fig 2.9: imagine solid line shows expenditures of start-up company, and dashed the revenues. Both increase exponentially.

Impression: difference between the two is getting smaller. In fact, it remains the same.

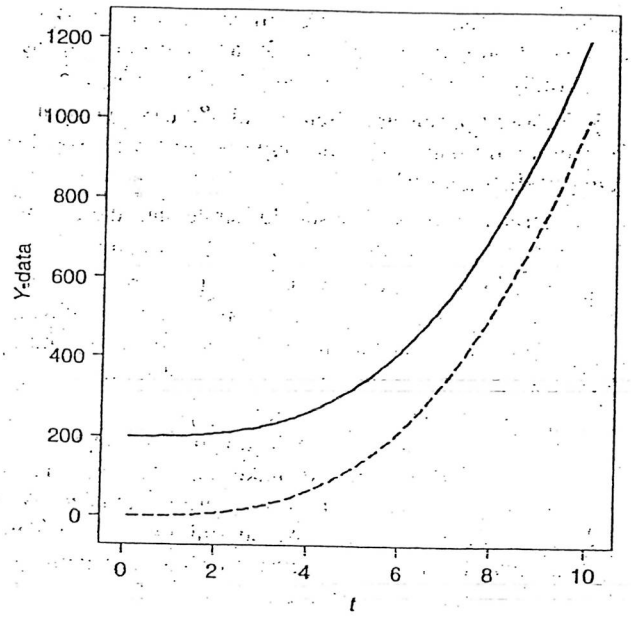

Figure 2.10 Two exponentially increasing curves. Note that the difference between the numerical values and hence the vertical distance is the same for all values of t.

2.2 Principles of graph construction

Two purposes of statistical graphics:
(i) data scrutiny, discovery, analysis.
(ii) communication of findings in the data

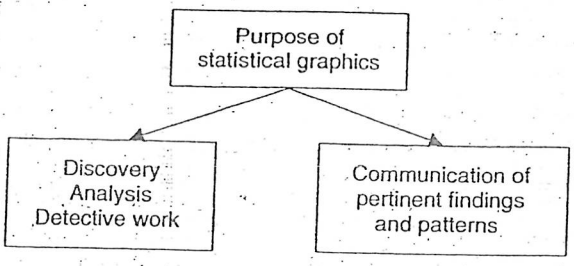

Figure 2.11 Purposes of statistical graphics.

Rules/guidlines:

- communicate key points
- use minimum amount of notes in the data region.
  Keys and markers should be left out of the data region.
- compact graphs are often useful.
- graphs construction is an iterative process. They should tell the story.
- several graphs may be needed to show different aspect of the data.
- scale used in the graph is crucial. We should carefully choose axis scales of the variables.

If different graphs compared, they should have the same scales. Zero does not have to be included.

Use logarithmic scale when data spans several orders.

2.3 Aspect ratio

Aspect ratio $\alpha = h/w$, where h is height and w is width of the data region.

The aspect ratio has impact on our ability to visually decode slopes.

Fig. 1.12a,b,c. Which of the plots provides best discrimination between convex and the linear part of the curve?
The difference is best apparent on 2.12 a. Why?

Answer: because our perception of the curve is based on orientation of the the segments: eye has best discrimination power when orientation of what we would like to compare is close to a 45% angle.

Idea: We need to rearrange the aspect ratio that the angle is around 45%

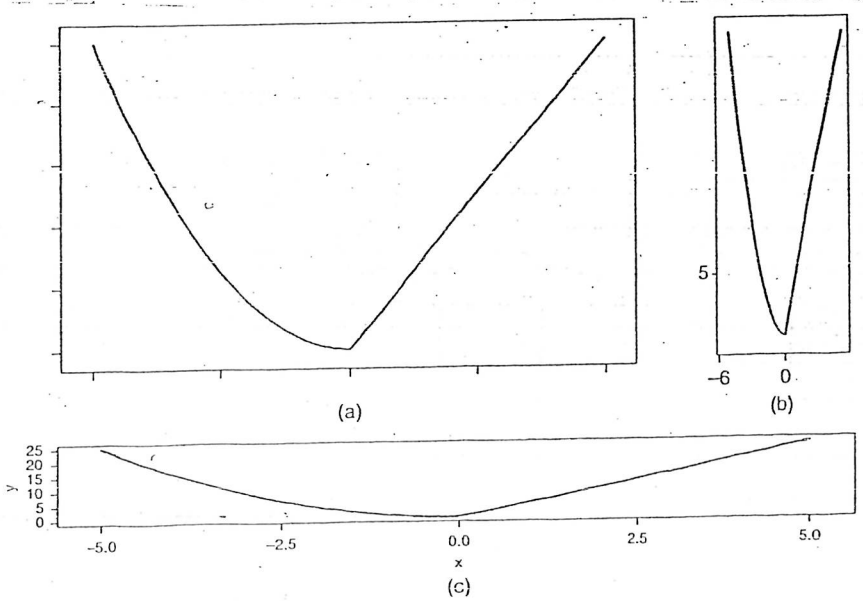

Figure 2.12 The same graph with different aspect ratios.

Example (Annual sun spots numbers) from 1770-1869.

Fig 2.13 uses aspect ration $\alpha = 4/6$ which typical for many graphs. It shows the cycle, but does not show that the curve rises faster than falls, which is feature of this data (lack of symmetry).

Fig 2.14: $\alpha = 6/6$ This feature even more difficult to see.

Fig 2.15: $\alpha = 1/6$: we see cycle where curve rises more rapidly than falls. This is possible because slopes are almost at 45% angle. We would miss this if we use the default setting in software packages.

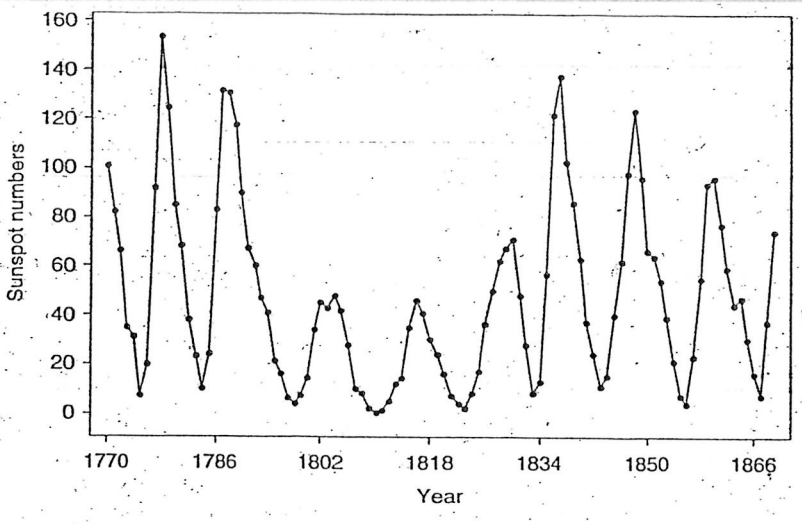

Figure 2.13 The annual sunspot numbers from 1770 to 1869. The aspect ratio is 4/6.

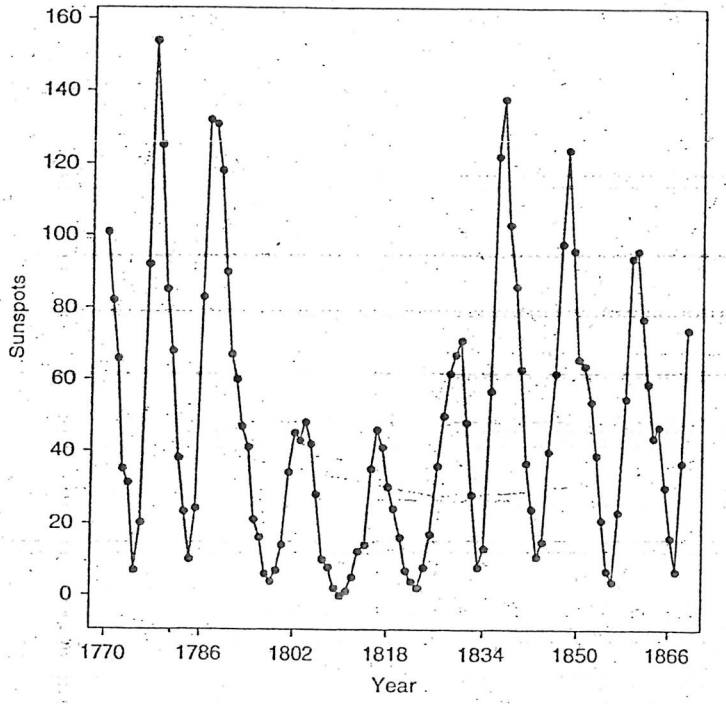

Figure 2.14 The annual sunspot numbers from 1770 to 1869. The aspect ratio is 1.

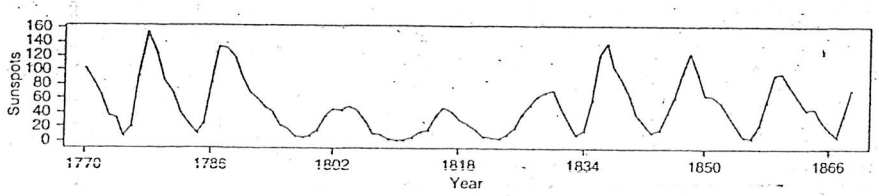

Figure 2.15 The annual sunspot numbers from 1770 to 1869. The aspect ratio is 1/6.

Example (Airline Passenger number) from 1949-1960.

Fig 2.16, aspect ration 4/3. When we look at general trend and ignore seasonality, eye connects peaks. Because with this aspect ratio, trend is at 45%, we can see a bit of non-linear exponential trend.

Fig 2.17, aspect ration 5/16. Now trend looks more or less linear: they eye looks at and connects the peaks, while the angle is far from 45%. However, now we see clearly asymmetry around the peaks.

Conclusion: the best aspect ration depends on what we are looking for, and what message we are trying to convey.

It is beneficial to try a few different aspect ratios.

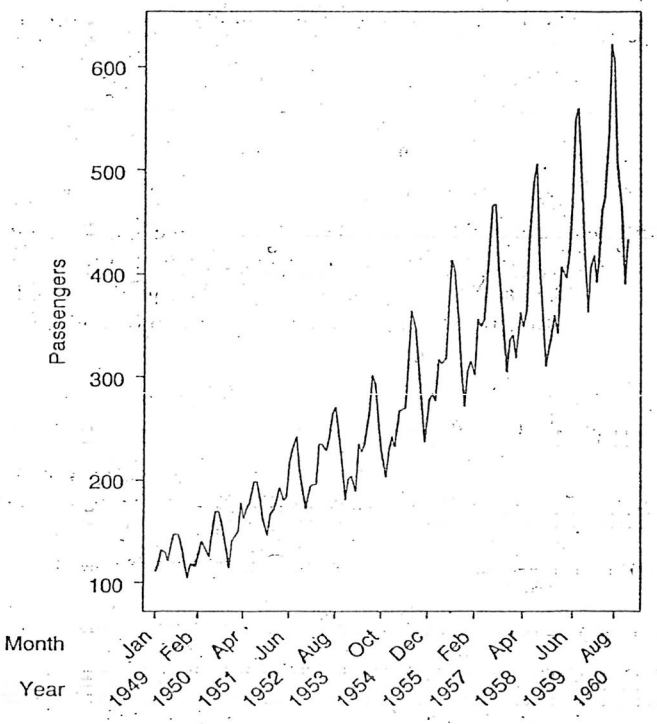

Figure 2.16 The number of airline passengers from 1949 to 1960. The aspect ratio is 4/3.

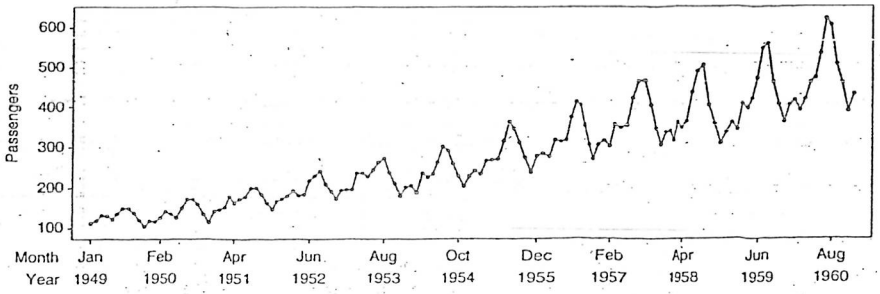

Figure 2.17 The number of airline passengers from 1949 to 1960. The aspect ratio is 5/16.

2.4 Time Series Plots

In time series, the dependent variable x_t is plotted against time variable t. Time values are usually equally spaced.

There are various ways of plotting time series data.

Connected symbols graph.

It is most common time series graph.

Advantage: one can see clearly each individual data points and ordering of the points.

Fig 2.18: annual cycle is clearly visible.

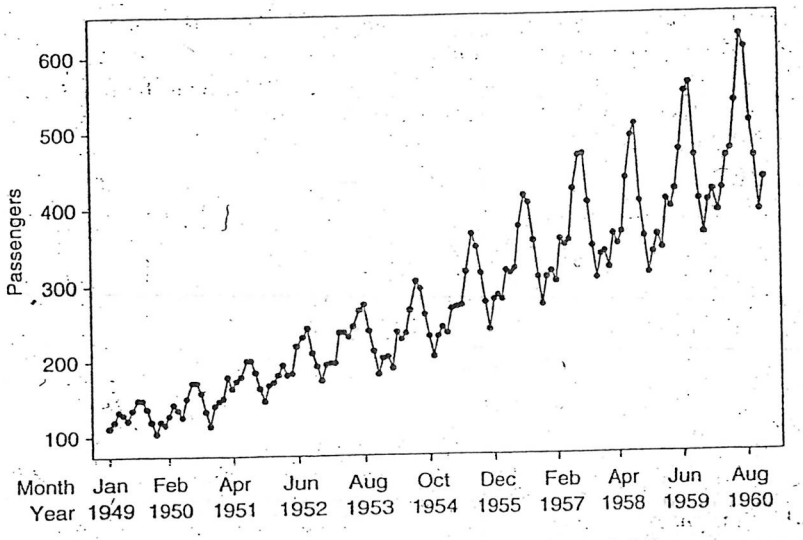

Figure 2.18 The connected symbols graph for the number of airline passengers data.

Connected lines graph.

It provides clarity about flow of the data and cyclic pattern, but it does not show how many individual values are involved and where they are. Not clear how may points we have around the peak.

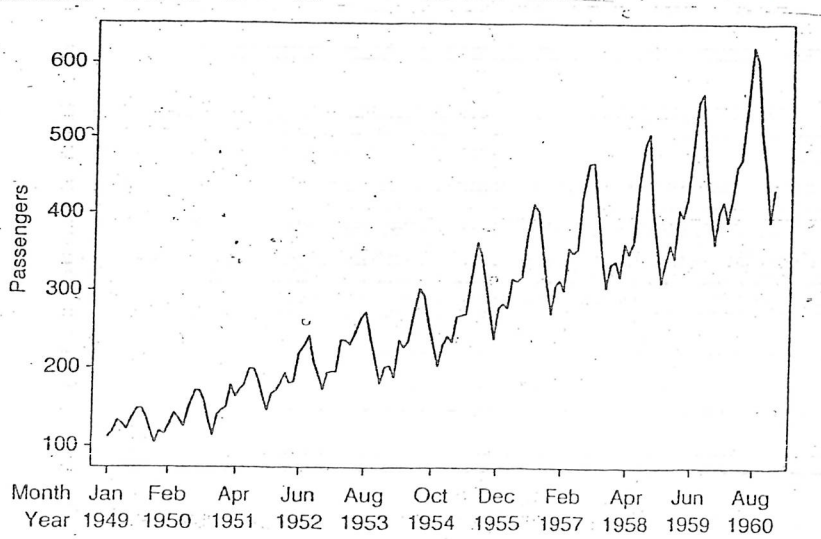

Figure 2.19 The connected lines graph for the number of airline passengers data.

Graph with symbols only

Fig 2.20: In this graph symbols are not connected. It may show general trend in time series, but ordering is no longer obvious. It is hard to see short and mid-term patterns.

It is also harder to see cycles.

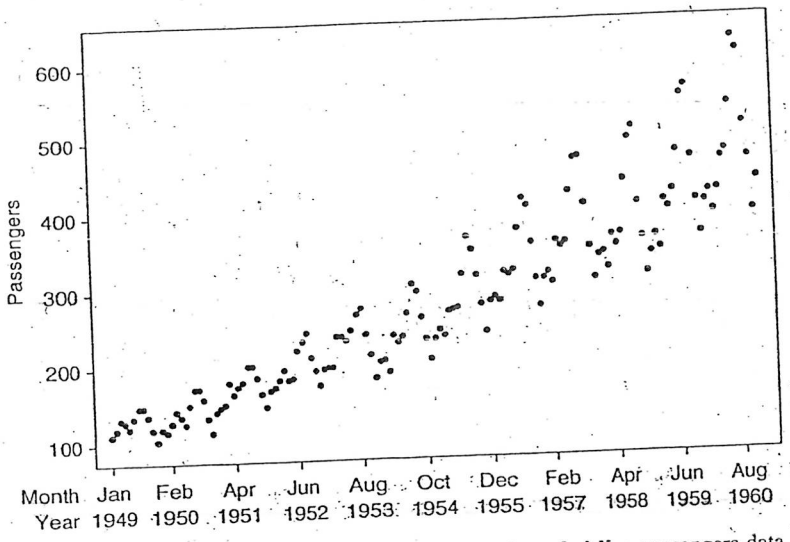

Figure 2.20 The symbols only graph for the number of airline passengers data.

Graph with projected bars and symbols

Projected bar charts are popular in economics, and popular in newspapers. Symbols make them too crowded, one can see peaks but low values are hard to see.

They can be useful if bars are projected to a reference horizontal line, in middle of the graph. For example, in autocorrelation plots.

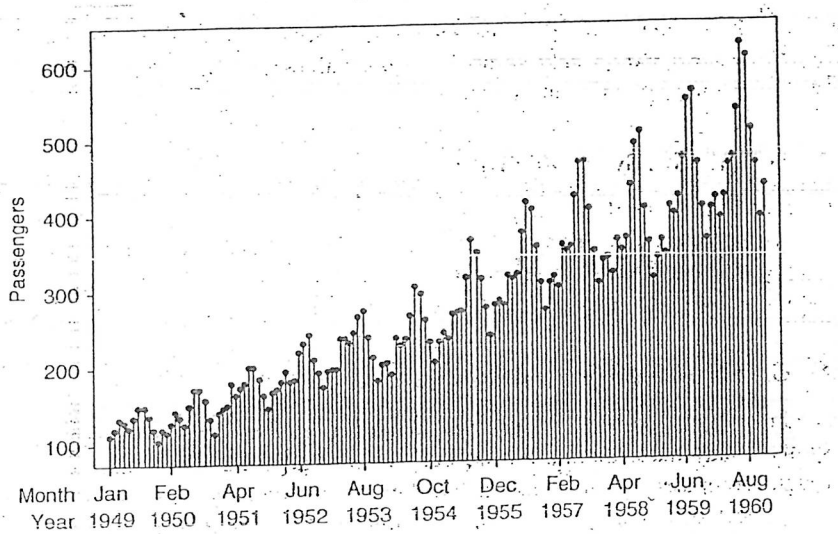

Figure 2.21 The projected bars and symbols graph for the number of airline passengers data.

Graph with projected bars (Vertical Line plot)

Fig 2.22 does not contain symbols and looks less crowded.

Drawback: creates asymmetry, peaks peaks stand out more than low points. In time series hardly used.

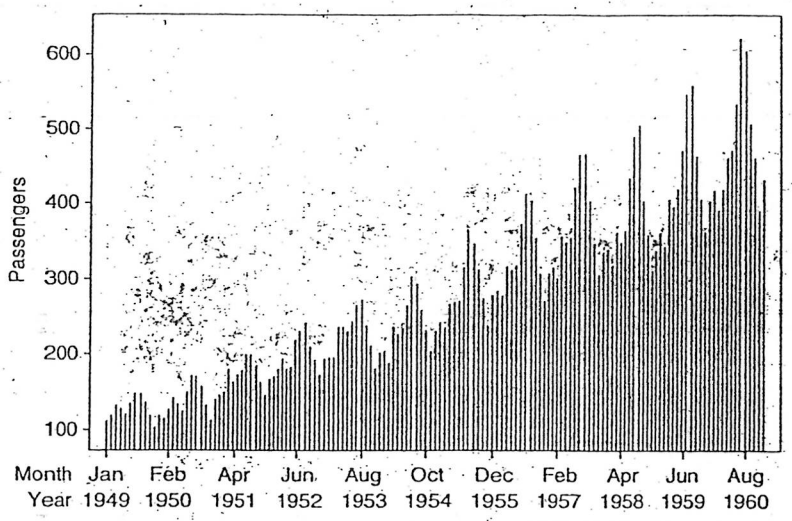

Figure 2.22 The projected bars and symbols graph for the number of airline passengers data.

Area graph

Often used with superimposed time series

Fig 2.23 is even worse than Fig 2.22.: asymmetry not seen well, peaks peaks stand out more than low points. In time series hardly used.

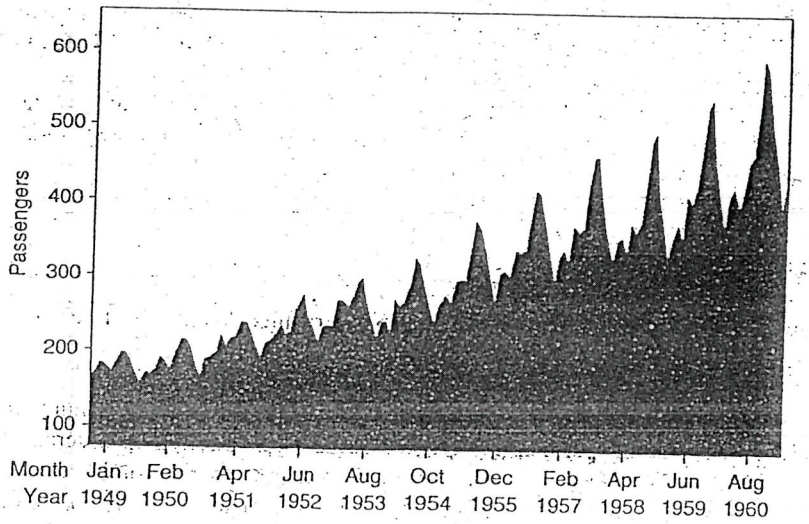

Figure 2.23 The area graph for the number of airline passengers data.

Cut and stack graph

Used when time series has large number of observations

Fig 2.24: shows hourly temperatures of a cooling tower. Graph is crowed. It is hard to identify short term oscillations.

Solution: Instead of using very low aspect ration, we cut time series in segments, plot them and stack them.

Fig 2.25 shows 11 panels stacked on the same graph.

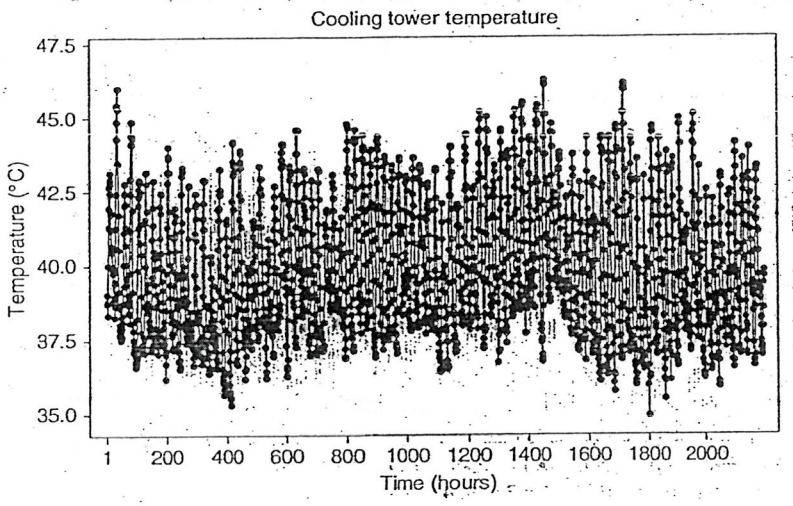

Figure 2.24 The hourly temperature observations from a cooling tower.

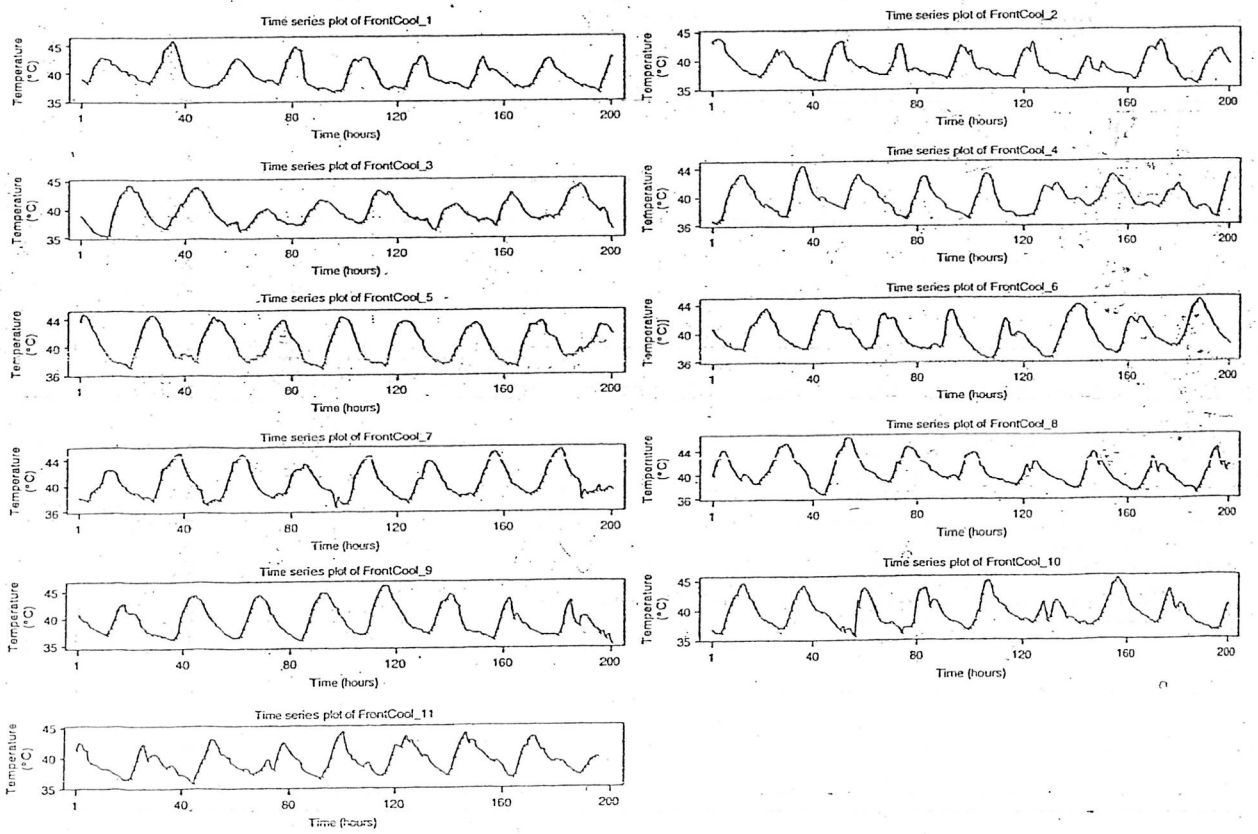

Figure 2.25 Eleven segments of the hourly temperature observations from a cooling tower.

2.5 Bad graphics

Examples of what not to do.

Defect report to upper management

Fig 2.26 (good graph) shows daily defect counts for 6 months of production. It shows some variation form month to month.

Fig 2.27 Box plot (good graph). It summarizes the data and sometimes preferred in report as summary to upper management.

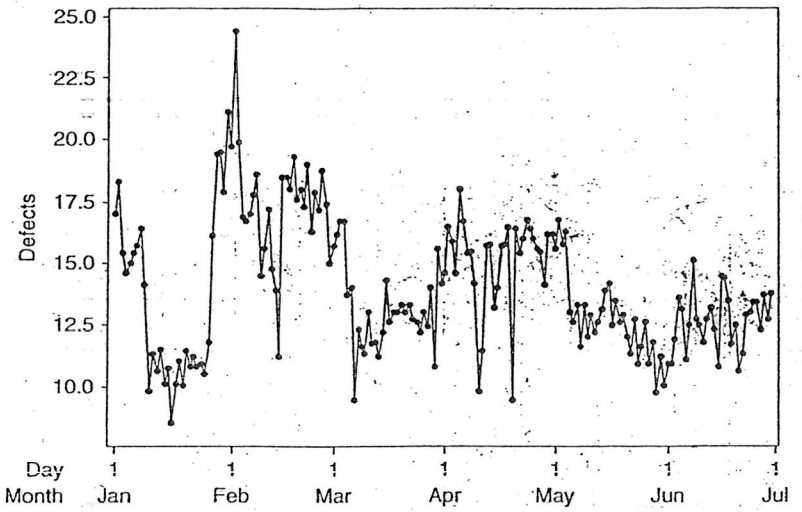

Figure 2.26 Time series plot of the daily defect count in percentage for 6 months of production in a ceramic production process.

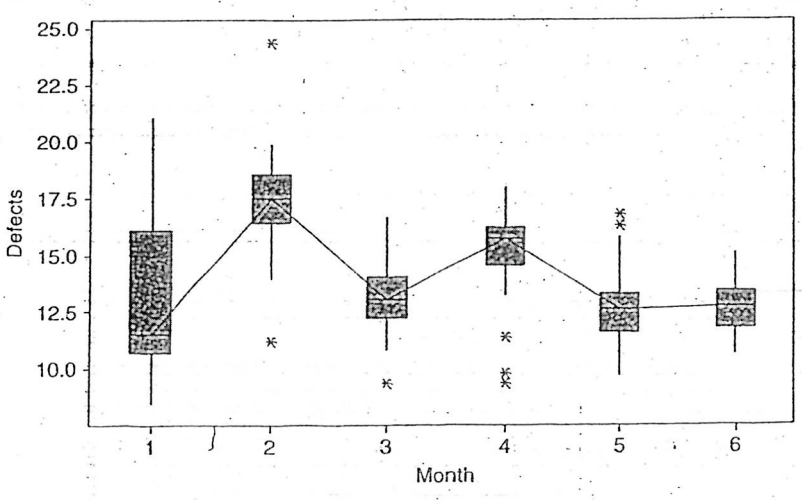

Figure 2.27 Time series connected box plot showing monthly defect counts in percentage for 6 months of production in a ceramic production process.

Fig 2.28 (Bad graph): we adjust graph to entire year, and y axis includes 0. Consequently, scale gets compressed, some features are no longer observed.

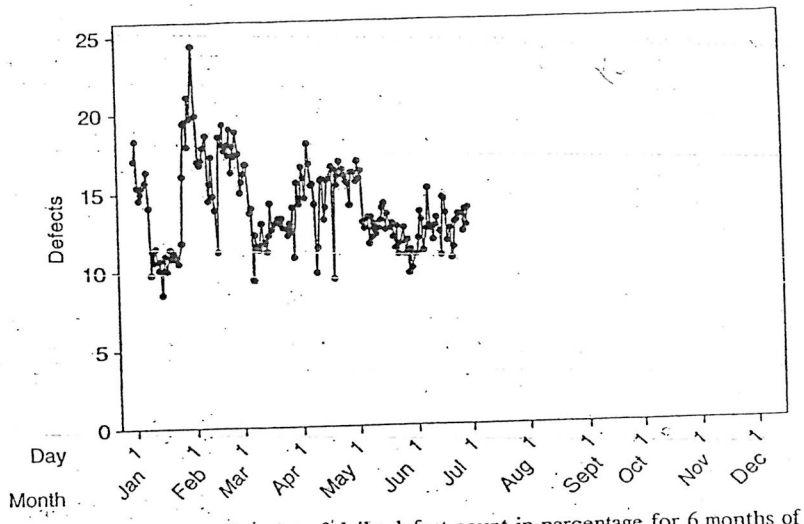

Figure 2.28 Time series plot of daily defect count in percentage for 6 months of production in a ceramic production process.

Example. Temperature of Industrial Furnace. [Bad scaling]

Temperature in industrial furnace can get quit high.

Fig 2.29 shows 200 observations. From the figure, it is hard to conclude if the temperature was stable during that period.

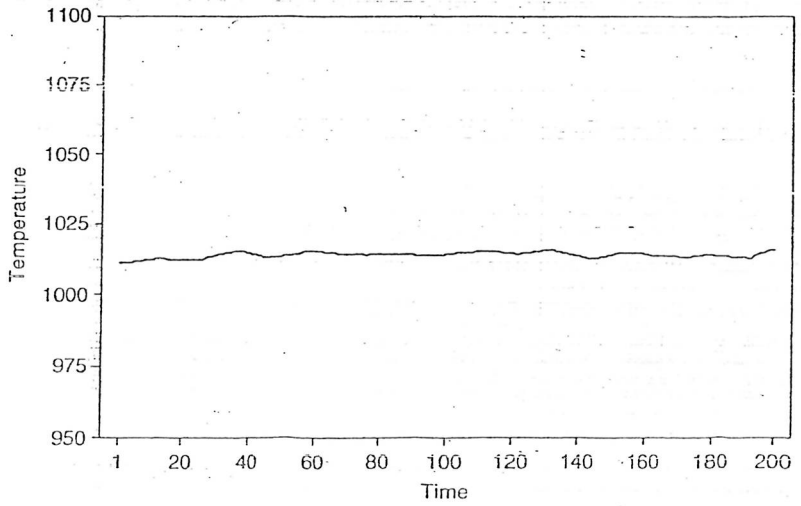

Figure 2.29 Temperature measurements for an industrial furnace.

Fig 2.30: we further "hide" the data including "0" on y axis.
Fig 2.31: we make things worse adding horizontal gridlines.

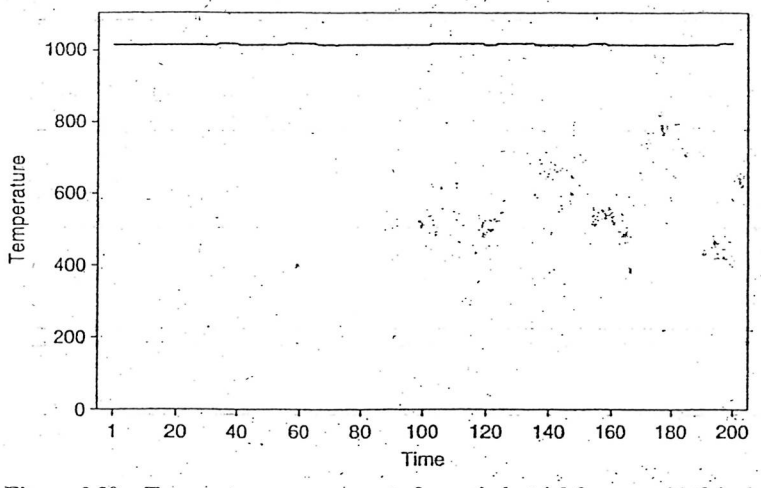

Figure 2.30 Temperature measurements for an industrial furnace with 0 included in the y-axis.

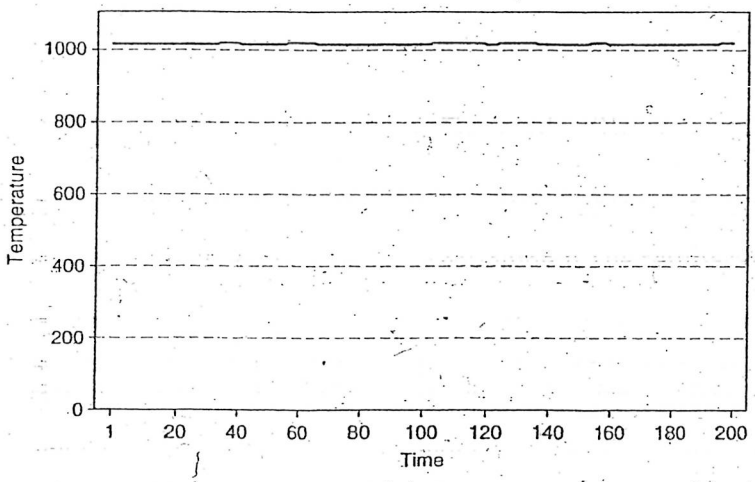

Figure 2.31 Temperature measurements for an industrial furnace with 0 included in the y-axis and horizontal gridlines.

Solution: Fig 2.32: use more appropriate limits for y axis. It shows the range of variation of 5^oC degrees compared to average values 1014^o.

Note: Small changes in temperature can cause defects, and studying the behavior of temperature is important controlling quality.

So, variation should not be hidden in the scale.

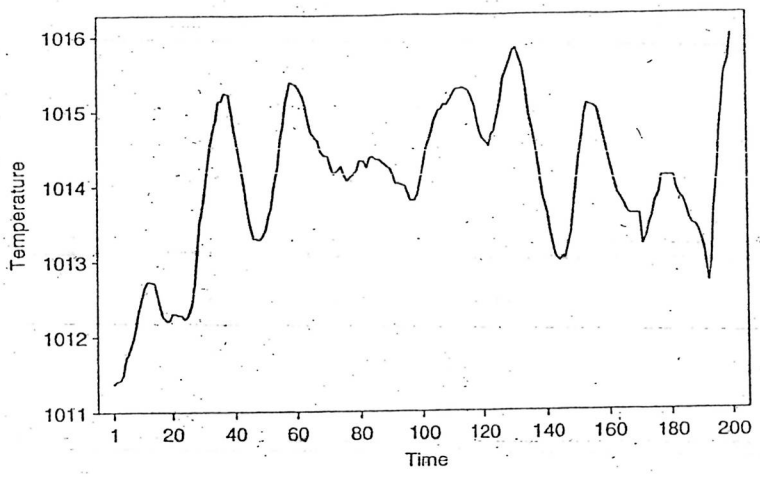

Figure 2.32 Temperature measurements for an industrial furnace with an appropriate scale for y-axis.

Example. Auto and Truck Production in US. Stacked area graphs can be uninformative.

Fig 2.35 show production of autos and trucks in US 1986-2007. Stacked graph makes hard to compare between production of truck and autos.

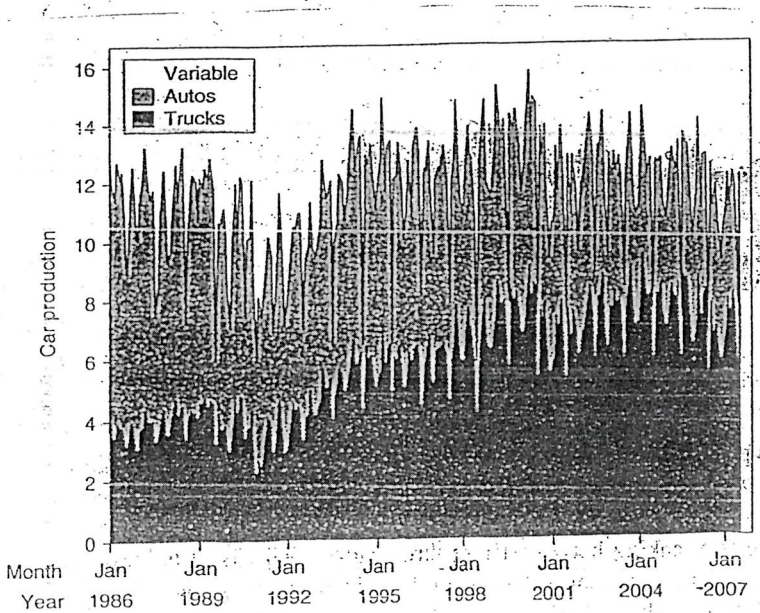

Figure 2.35 Monthly US motor vehicle production from January 1986 to September 2007.

Fig 2.36 shows graphs for autos and trucks using the same baseline. It indicates: auto production is declining, trucks production is increasing. It is still difficult to compare the plots because of large number of observations

Fig 2.37 plots these time series using two separate panels. Both panels have the same y-axis, to easy comparison.

Conclusion: 1. Plot the data before any statistical analysis
2. Be creative

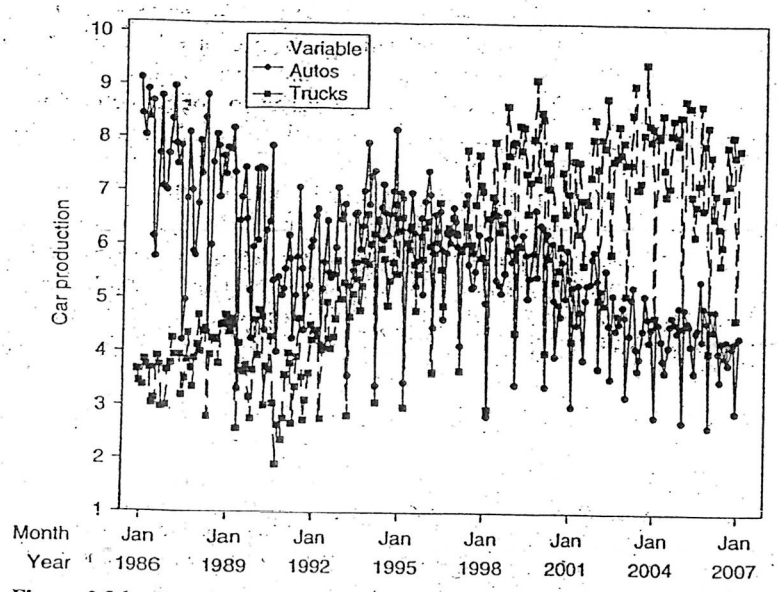

Figure 2.36 Superimposed time series plots of auto and truck production in the United States.

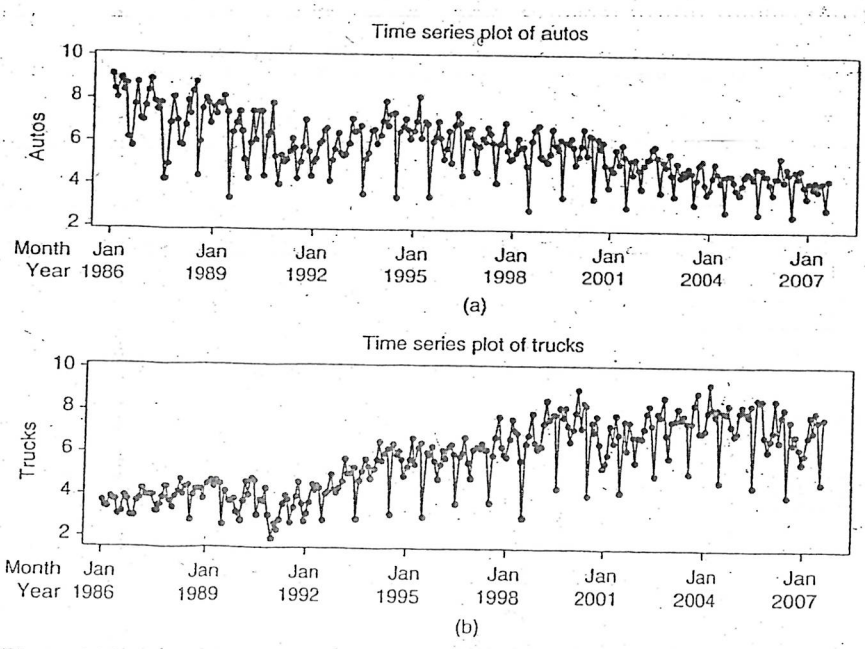

Figure 2.37 Time series plots of auto and truck production in the United States in separate panels.