Table of Contents

1 Classical_Statistic_Approach

1.1 The statistical characteristics of the time series.

1.1.1 Main statistical characteristics

In the Time Series Analysis we are considering the series analysis as statistical task.

Due to this approach we can introduce the following characteristics of the series:

1. **mean value** (or expected value),

$$ev = \frac{1}{N}\sum_{i=0}^{N-1} y_i,$$

where y is samples of size N;

2. **standard deviation** (or root of the variance),

$$var = \frac{1}{N}\sum_{i=0}^{N-1} (y_i - ev)^2; \; std = \sqrt{var}$$

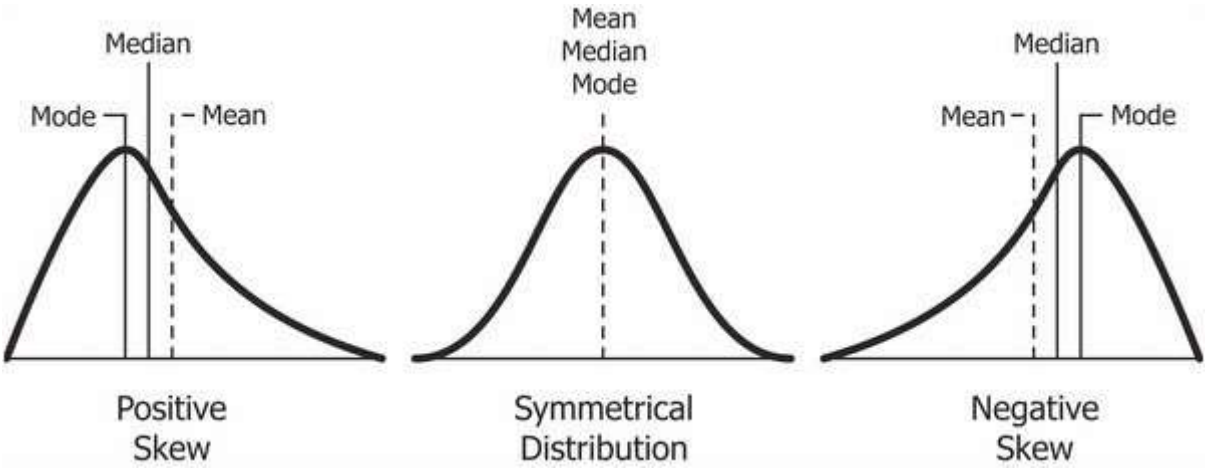
, where var is the variance; std is the standard deviation;

Pleas note beside the mean value we can introduce:

- **Median**: Median is the middle score for a set of data that has been arranged in order of magnitude.

  For example, given an ordered list of student marks, [14 35 45 55 55 56 58 65 87 89 92], median is 56 because it is the middle mark since there are 5 items before it, 5 items after it.

- **Mode**: Mode is the most frequent score in our data set.



Pleas note beside the variance we can introduce:

- **Meand Absolute Deviation** -Difference between mean and other values.
- **Skewness** is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the central point.
- **Kurtosis** is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

3. **autocorrelation** function (**ACF**) is the degree of a time series to be linearity related to a lagged (delayed) version of itself.

$$cor(k) = \frac{1}{N} \frac{\sum_{i=0}^{N-1}(y_k - ev)(y_{i-k} - ev)}{var(y)}$$

Pleas note:

1. cor(k) is called k-th lag.
2. **Correlation** (or cross-correlation) is the degree of two time series to be linearity related one of them to a lagged (delayed) version of another.
3. **Covariance** (also **autocovariance**) is general is simply version of correlation.
4. For **autocorrelation** and **autocovariance**

$$\mathrm{cov}(k) = \frac{1}{N}\sum_{i=0}^{N-1}(y_k - \mathrm{ev})(y_{i-k} - \mathrm{ev})$$

$$\mathrm{cor}(k) = \frac{\mathrm{cov}(k)}{var}$$

   for **cross-correlation** and **cross-covariance**

$$\mathrm{cov}_{xy}(k) = \frac{1}{N}\sum_{i=0}^{N-1}(y_k - \mathrm{ev}(y))(x_{i-k} - \mathrm{ev}(x))$$

$$\mathrm{cor}_{xy}(k) = \frac{\mathrm{cov}_{xy}(k)}{\mathrm{std}(x)\mathrm{std}(y)}$$

5. For some cases we will calculate the correlation without mean value and normalization subtraction as

$$\mathrm{cov}(k) = \sum_{i=0}^{N-1} y_k y_{i-k}.$$

6. Actually covariance can be calculated as for $lags$ as for $-lags$, such as

$$\mathrm{cov}_{full}(k) = \sum_{i=-N-1}^{N-1} y_k y_{i-k}$$

, in this case we have to call the previously mentioned formula as straight correlation, in opposite for the straight here are shown full covariance and also can be introduced same covariance

$$\mathrm{cov}_{same}(k) = \sum_{i=-\frac{N-1}{2}}^{\frac{N-1}{2}} y_k y_{i-k}$$

7. It is shown that mentioned above approach can lead to the bias in prediction, thus in some time the unbiased covariation can be used

$$\mathrm{cov}_{unbiased}(k) = \sum_{i=0}^{N-1} y_k y_{i-k}/(N - |k|)$$

8. Among all cases of covariance behaviors the specific case can be defined, when the covariance value depends only on the difference between lags indexes and nothing more- this case we may call ergodic process. This case corresponds to the constant mean and variance values for each partial sample sizes.

9. The operation or our simplified covariance corresponds to the convolution that are taken for reversed lagged version of samples

$$\mathrm{conv}xy(k) = x * y = \sum_{i=0}^{N-1} y_k x_{k-i}$$

$$\mathrm{cov}xy(k) = x * y = \sum_{i=0}^{N-1} y_k x_{i-k}$$

10. The correlation is the degree in which two samples linearly depend on one to another.

    The autocorrelation is the ability of time series samples to be linearly related to each other

11. The correlation and convolution can be explained geometrically as relative intersection over the area of two figures, while one of them moving into the direction of others.

12. The maximum of the correlation (normalized covariation) is **1**. This mean true linear relation. The opposite case (**-1**) mean This mean true inverse linear relation. The **0** value of correlation mean lack of any linear relation.

13. The zero lag of correlation corresponds to the cosine product of series (angle between them):

$$cos(\alpha_{xy}) = \frac{\sum_{i=0}^{N-1} y_k x_k}{std(x)std(y)}$$

    , where α is the angle between two samples.

    The **1** value its collinearity and **-1** its inverse collinearity. In this case two series x and y can be considered as two vectors, and cosine product as its scalar product.

The **0** value of the angle mean its orthogonality in this case all samples of the series statistically linearly independent.

14. For complex values the covariance (and convolution) includes conjugate product

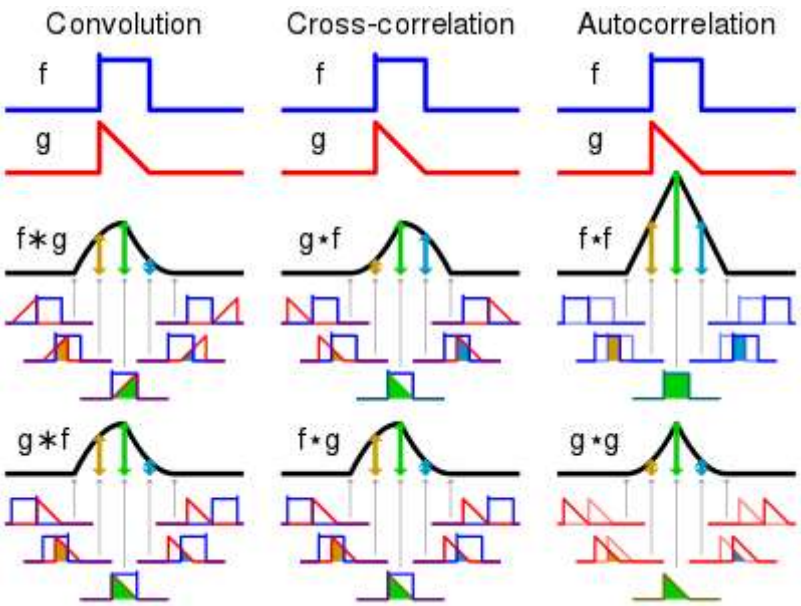
$$\mathrm{cov} = \sum_{i=0}^{N-1} y_k y_{i-k}^*$$

, where sign * is complex conjugation.

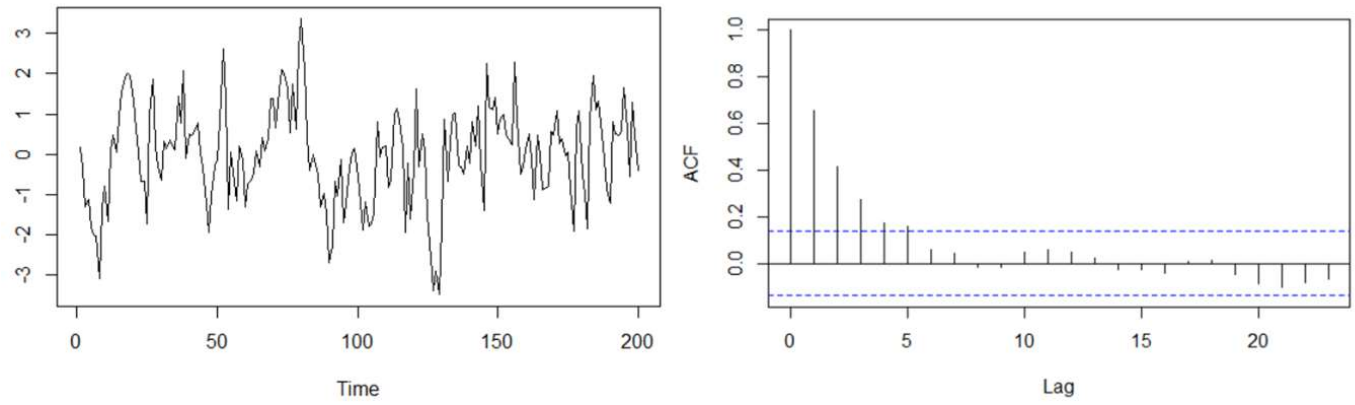
15. There are exist fast method for full-covariation calculation using Fast Fourier Transform

$$\mathrm{cov}_{xy} = \mathrm{ifft}(\mathrm{fft}(x, 2N) * \mathrm{conj}(\mathrm{fft}(y, 2N)), 2N)$$

, where fft and ifft are direct and inverse Fast Fourier Transform respectively, both taken by the series with zero-padding up to the 2N size.



Example of ACF. Blue line show the confidence (equal to the variance).



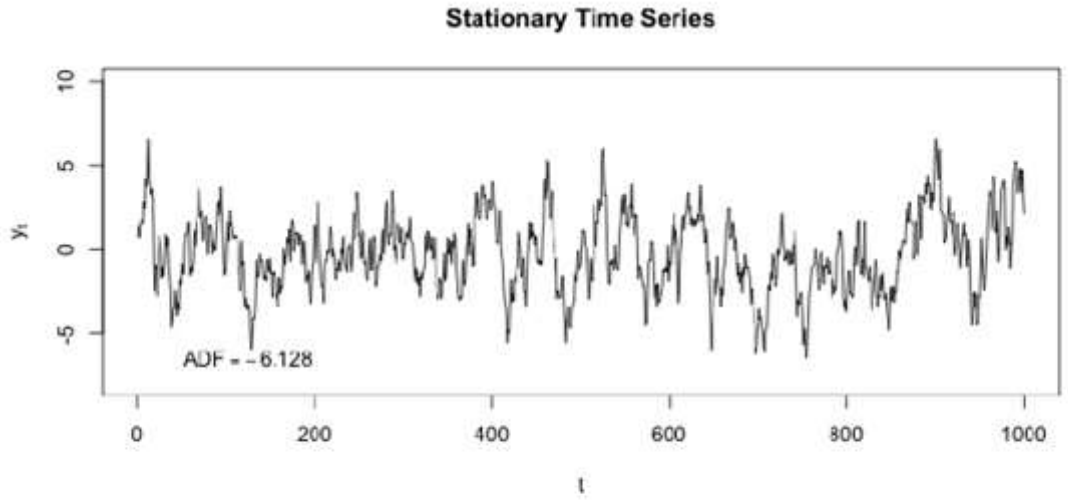
The analysis of mean value and std value behavior of the time series allows to distinguish the following cases:

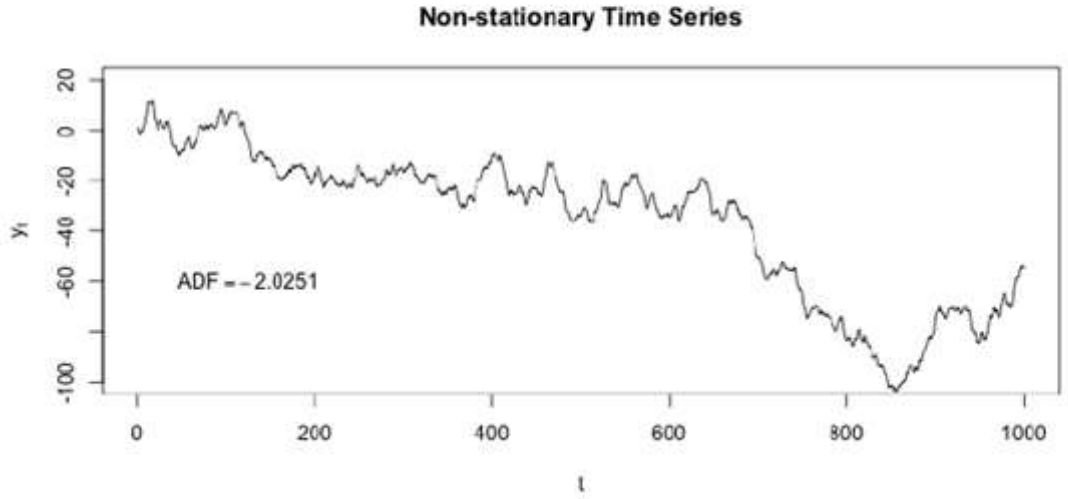
- **Stationary** series - a time series is said to be stationary if its statistical properties do not change over time. In other words, it has constant mean and variance, and covariance depends only on lags index difference.

  More precisely it have to be added that mentioned statistical characteristics have to be non-infinity values.

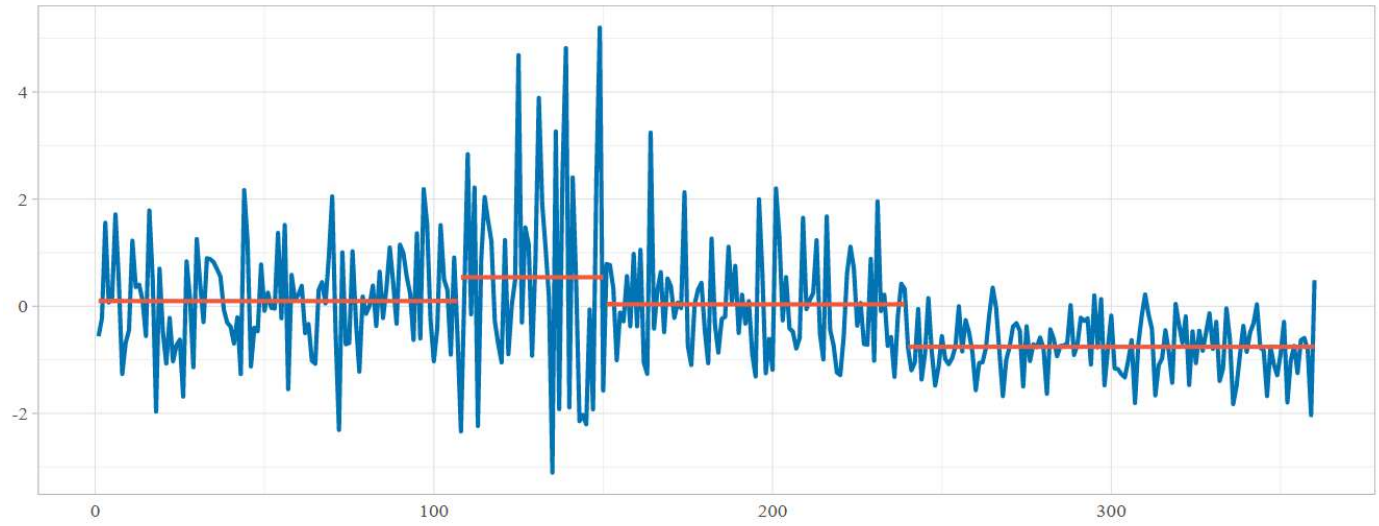
  The previous definition can be called **weak-stationary**. In opposite the **strong-stationary** mean that the series (and its distribution in statistical point of view) do not change over a time.

- **Non-Stationary** series - a time series is said to be non-stationary if some of its statistical properties have changes over time. In other words, it has variable of time mean, variance, or other statistical characteristics.





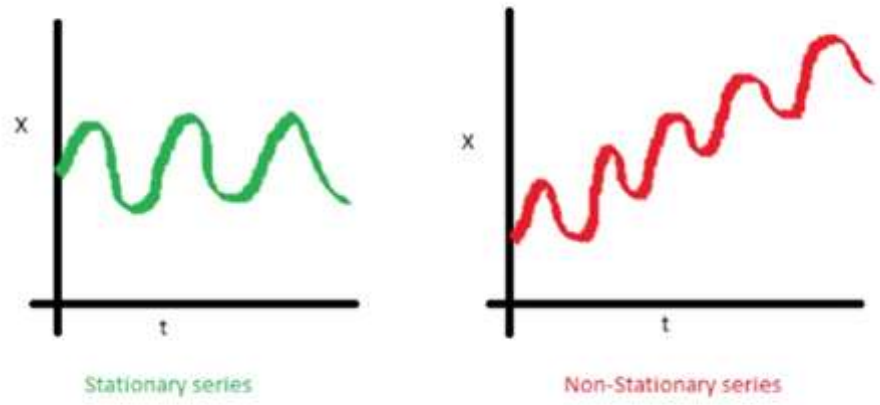
Example of non-stationary series with different mean and variance values



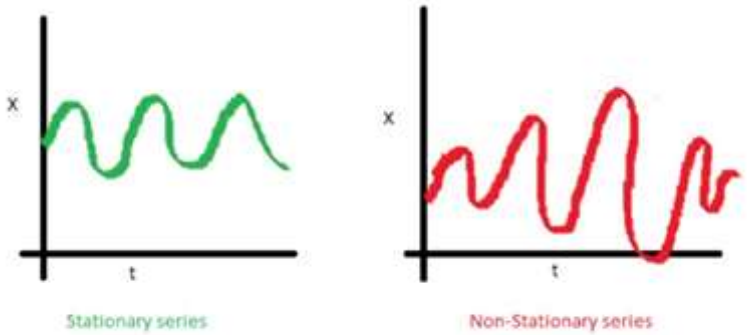
It is important to note:

1. Ideally, we want to have a stationary time series for modeling. Due to the not all of series stationary, we can make different transformations to make them stationary.
2. The stationary time series can be described by only its mean and variance (and covariance) values, for non-stationary time series there are exist a lot of other statistical characteristics.
3. We can consider stationary behavior for each part of time series: trend, cyclicity, seasonality, noises.

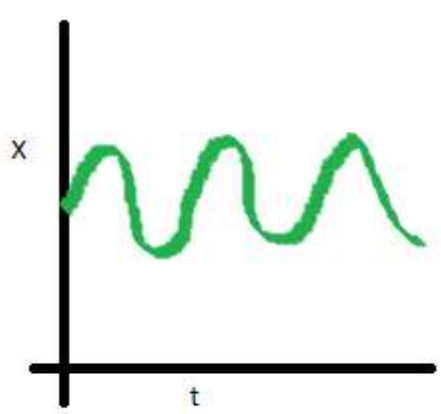
Example of trend line stationarity



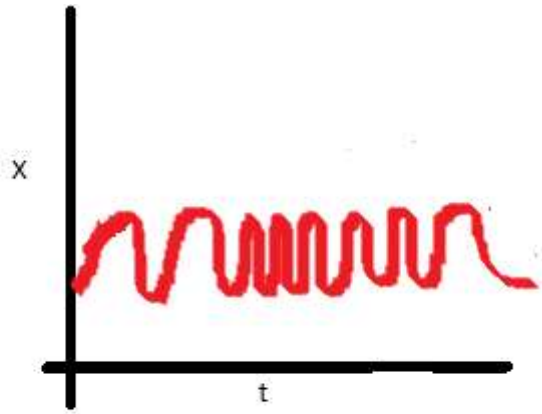
Example of seasonal stationarity (in meaning of the variance)



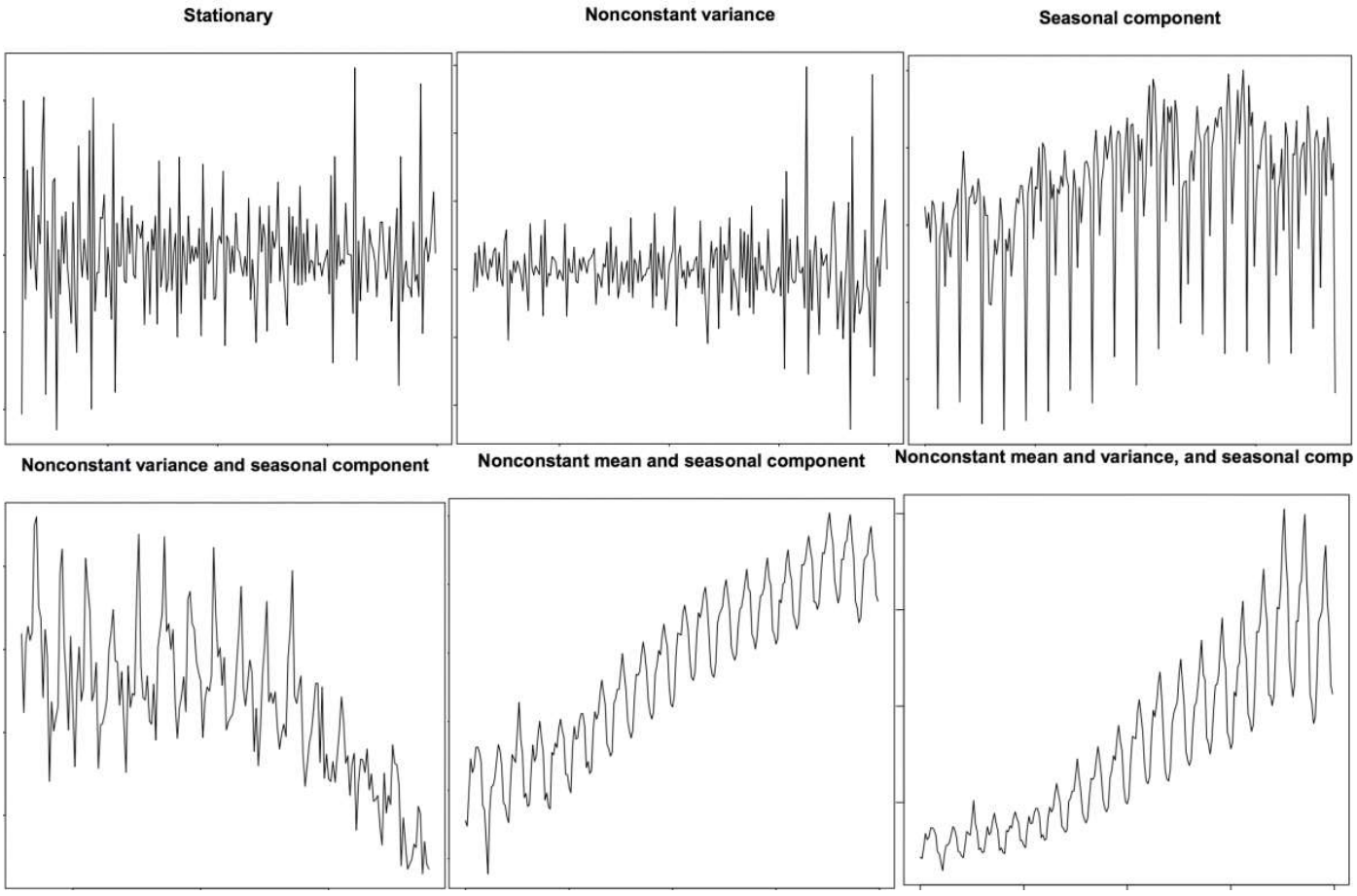
Example of seasonal stationarity (in meaning of the period of seasonality)



Stationary series



Non-Stationary series



One of the most important case of the stationary analysis is the noise stationarity.

The stationarity noise with independent samples over a time (with 0 autocorrelation) are called **White Gauss Noises (WGN)**.

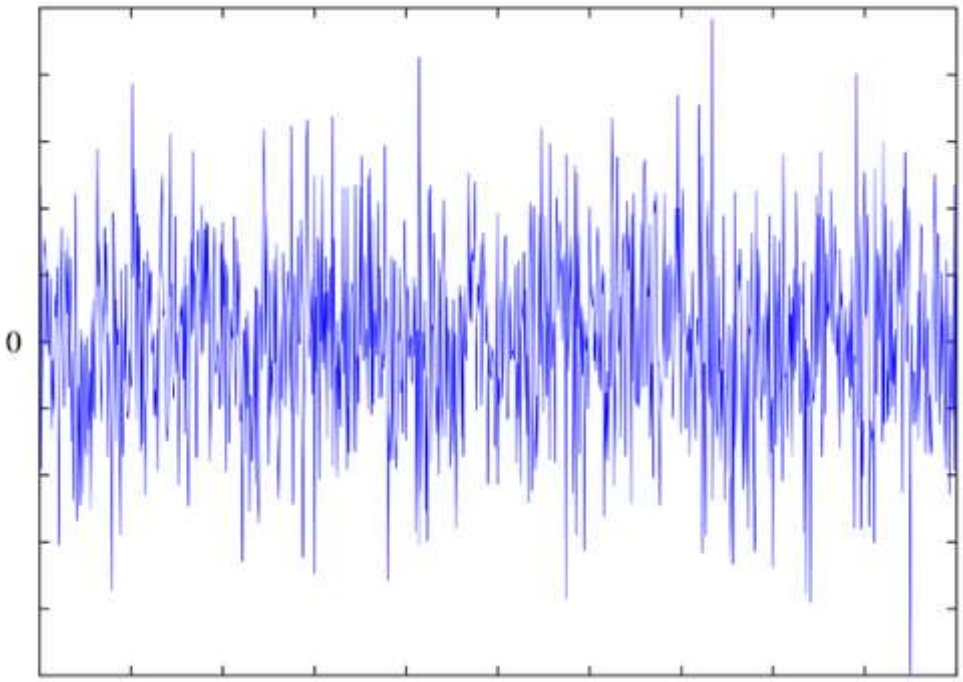
So the WGN is the series with

- $ev = 0; var < \infty;$

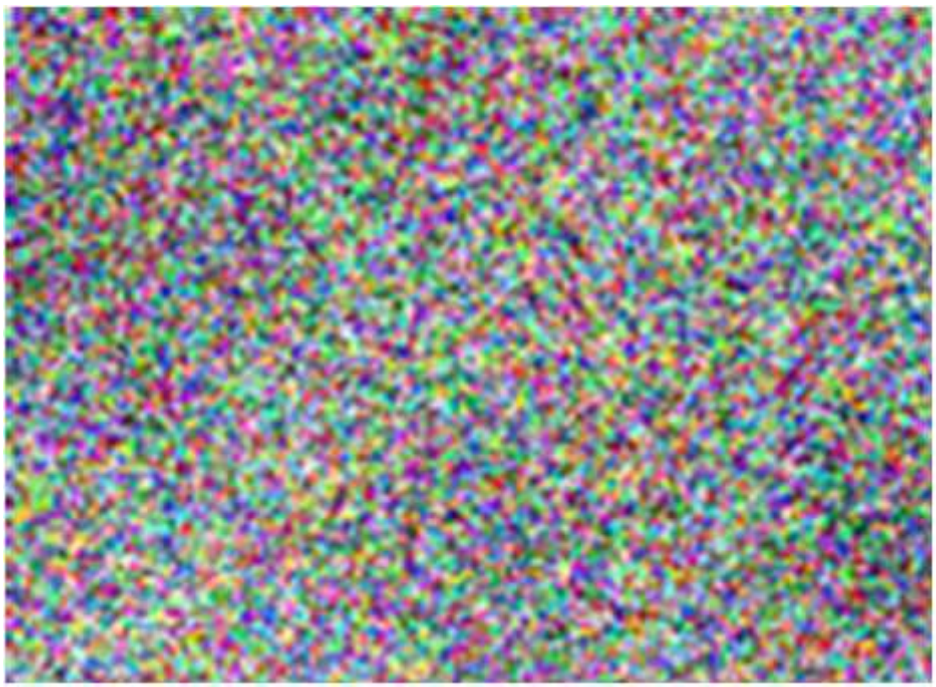
- $cov(n) = std \cdot \delta(k-n)$, where $\delta(k-n) = 1$ if $k = n$ and 0 in other cases.

  Pleas note that variance is also can be called the noise power.

  If WGN has const mean and const var it also can be called **independent and identically distributed (i.i.d)**.

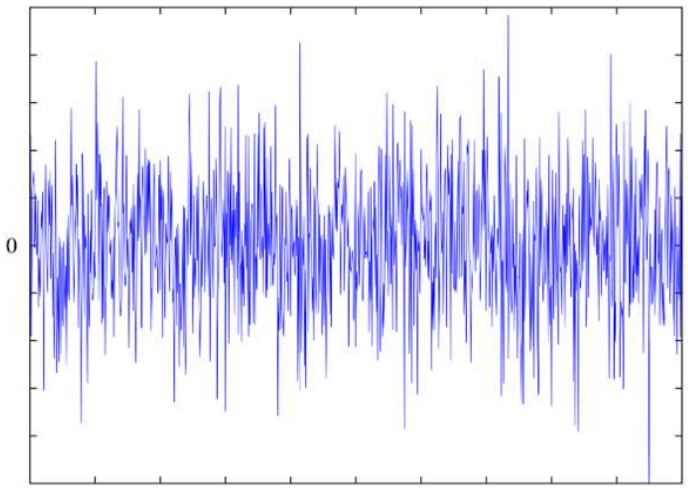


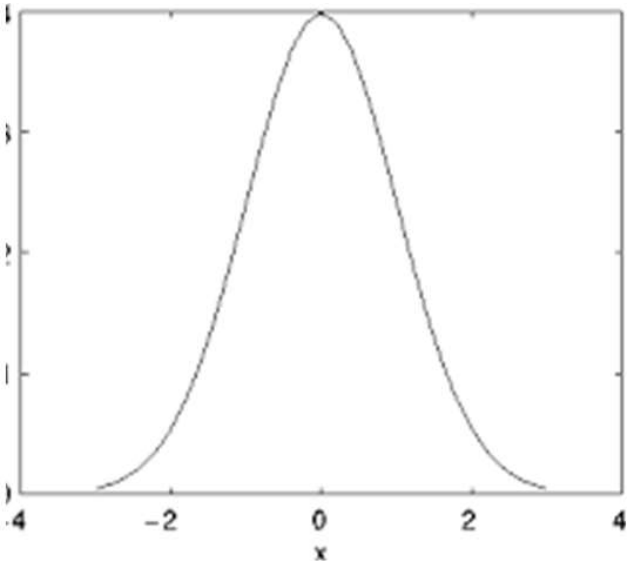
Example of wgn - is when you turn out antenna from tv and see so-called "snow"



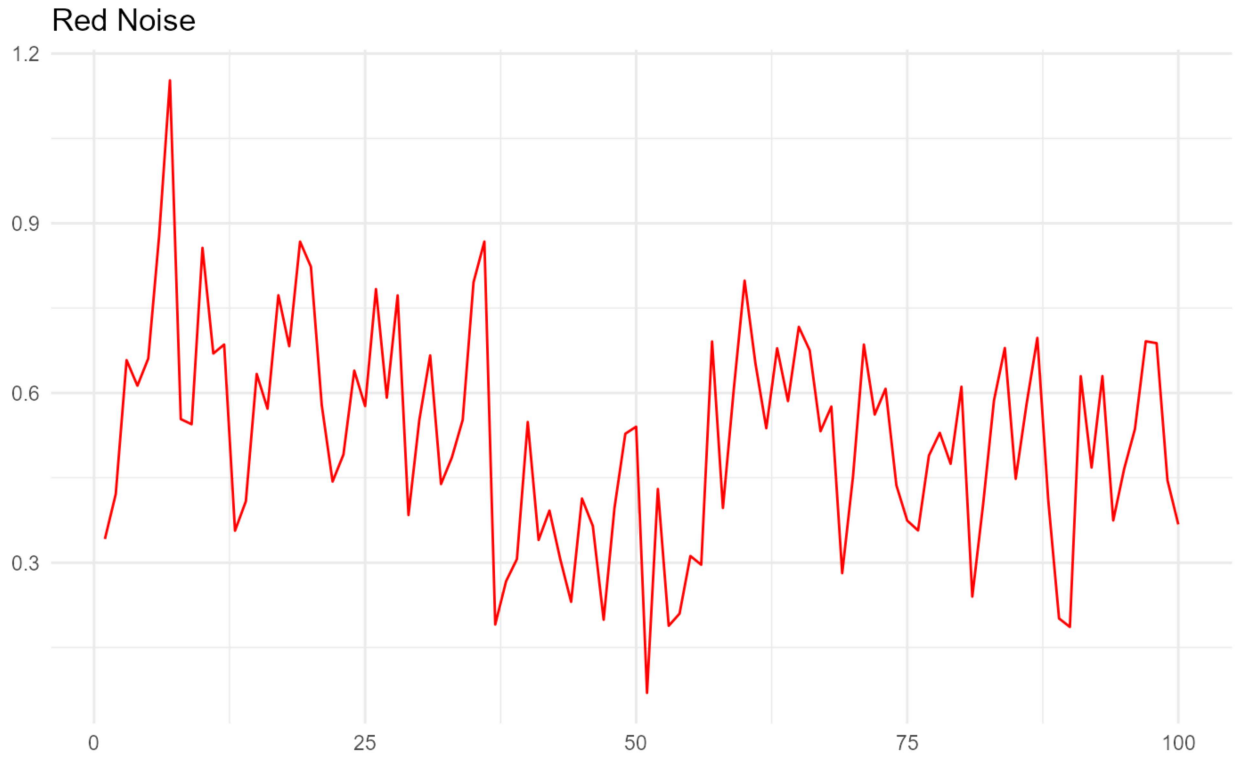
Many stationary processes (tasks) in time series are simulated as deterministic on the back-ground of white Gaussian noises.

> *Pleas note* the WGN has so-called **normal distribution** with zero-mean
>
> $$f(wgn) = \frac{1}{std(wgn)\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{wgn}{std(wgn)})^2},$$
>
> where $f(wgn)$ is the distribution.





Pleas note: in opposite to wgn we can introduce the noises that are differ from wgn as colored noises, for instance, it can be simulated as some random (stochastic) seasonality.



By the mentioned above we can say that on of the main task of time series analysis is to eliminate (or reduce) the noise influence on the other series components.
In the stationary case the main task is to reduce wgn influence.

Actually, almost all time series are non-stationary (in some degree). However, depends on the requirement accuracy in some cases we can approximate series (make supposition) that it is stationary for our task.

pleas note:

1. Not all methods of time series analysis are require stationary of time series.
2. In other cases it can be make some transforms (preprocessing) of series to make it stationary.

The one of the simplest way to transform series is to take its numerical derivative, obtaining the series

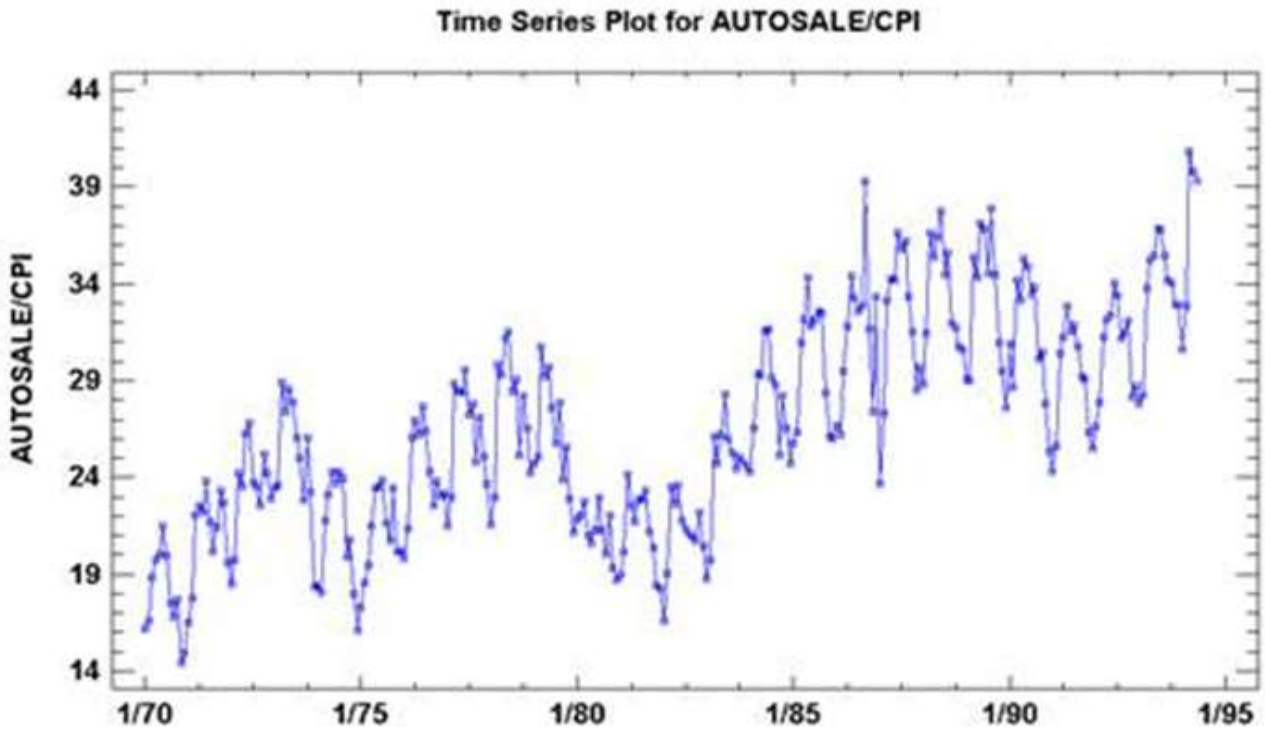
$$\Delta y_n = y'_n = y_{n+1} - y_n$$
$$\Delta^2 y_n = y''_n = y'_{n+1} - y'_n$$
$$and\ so\ on..$$

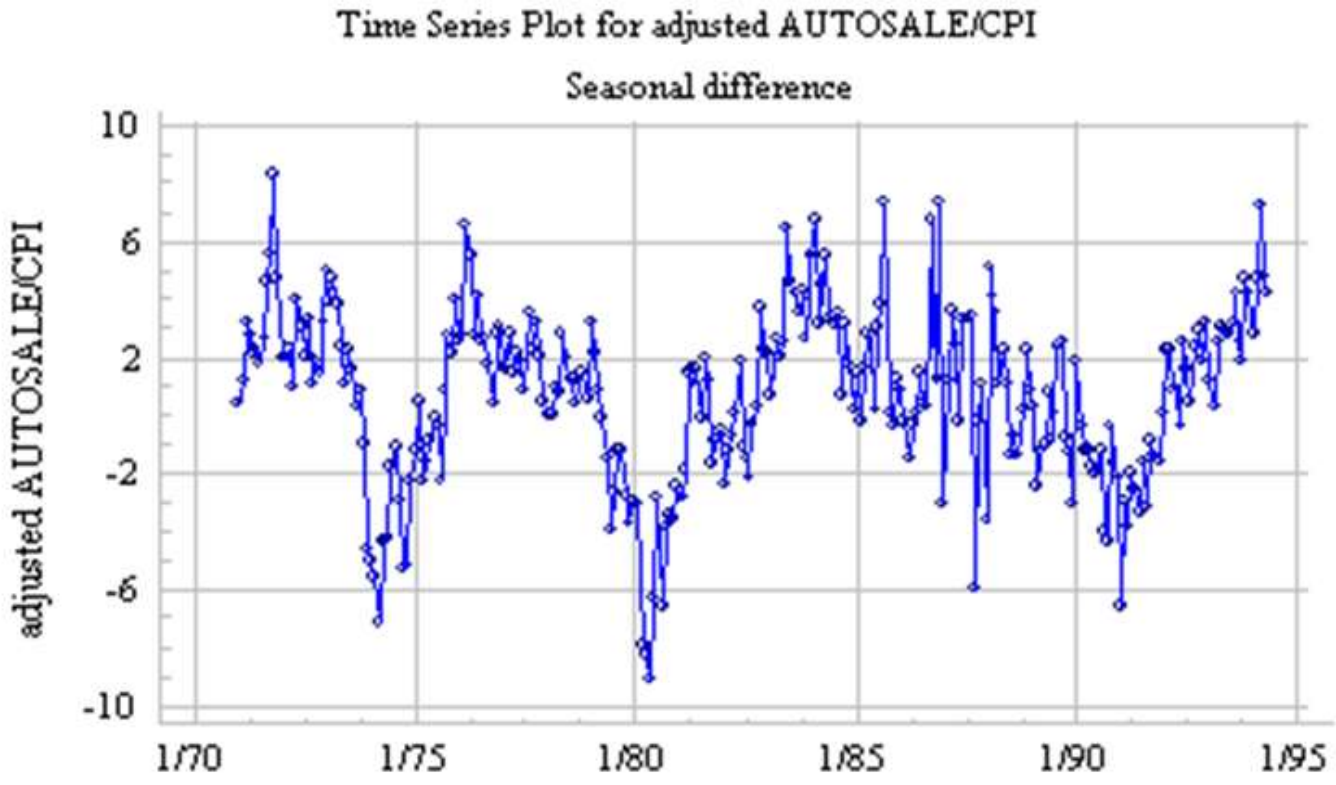
The degree of derivative when the stationary is obtaining are called integration parameter.

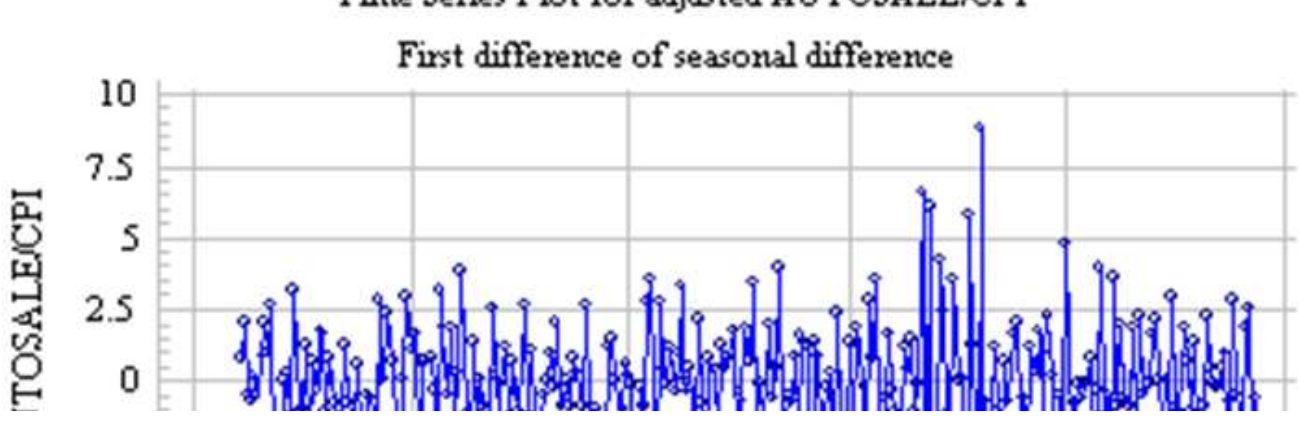
In the same manner it can be done seasonality compensation using such differences, when samples are expected to have the same values in the case of without trend

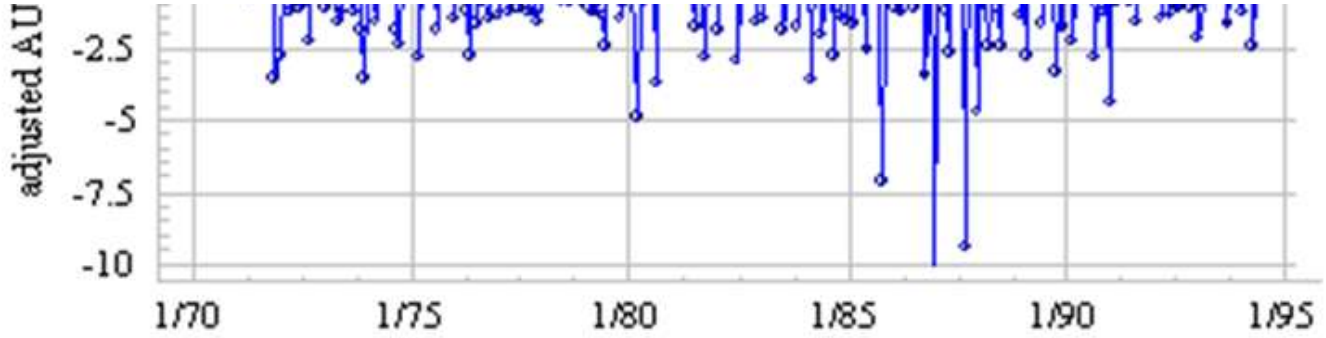
$$\Delta_s y_n = y^s_n = y_{n+k_s} - y_n$$
$$\Delta^2_s y_n = y^{2s}_n = y_{n+k_{s_2}} - y_n$$
$$and\ so\ on..$$

where Δ_s is the seasonal derivative; k_s is season period; k_{s_2} is the period of the the second-order seasonality.

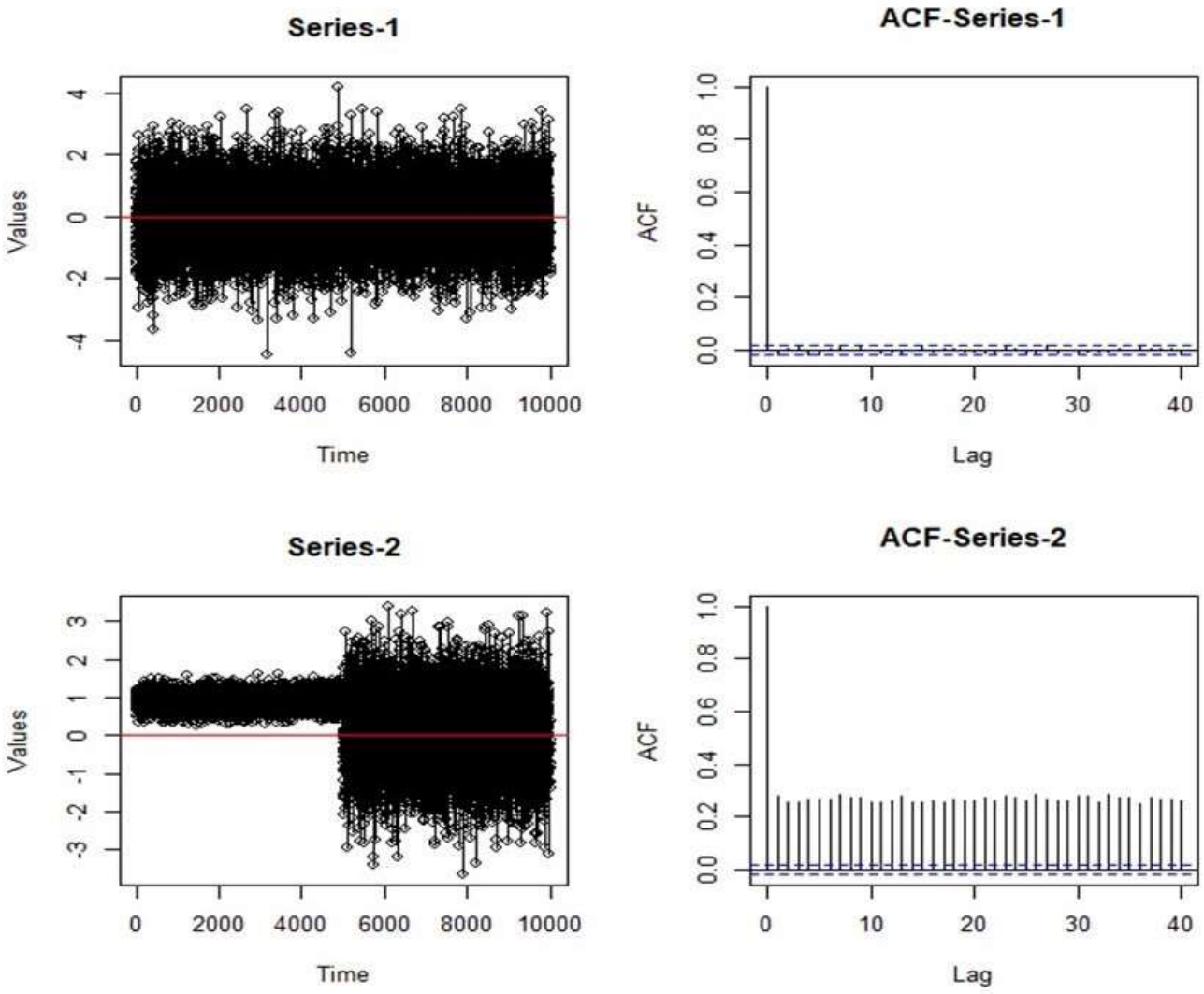

Time Series Plot for AUTOSALE/CPI
AUTOSALE/CPI
44
39
34
29
24
19
14
1/70
1/75
1/80
1/85
1/90
1/95


Time Series Plot for adjusted AUTOSALE/CPI
Seasonal difference
adjusted AUTOSALE/CPI
10
6
2
-2
-6
-10
1/70
1/75
1/80
1/85
1/90
1/95


Time Series Plot for adjusted AUTOSALE/CPI
First difference of seasonal difference
TOSALE/CPI
10
7.5
5
2.5
0



c (https://uk.mathworks.com/matlabcentral/fileexchange/43172-auto-correlation-partial-auto-correlation-cross-correlation-and-partial-cross-correlation-function?s_tid=srchtitle)



1.1.2 Residual analysis

After make time series decomposition or forecast it is important to check the error behavior.

As usual we will calculate the error of prediction as

$$e_n = y_n - \hat{y}_n$$

where e_n is error value for the sample n; $\hat{y}_n$ is the predicted value of y_n.

We expect the residual errors to be random white Gaussian noise-like - it means that the model (i.e. the model predictions $\hat{y}_n$) has captured all of the structure

and the only error left is the random fluctuations in the time series that cannot be modeled (or explained).

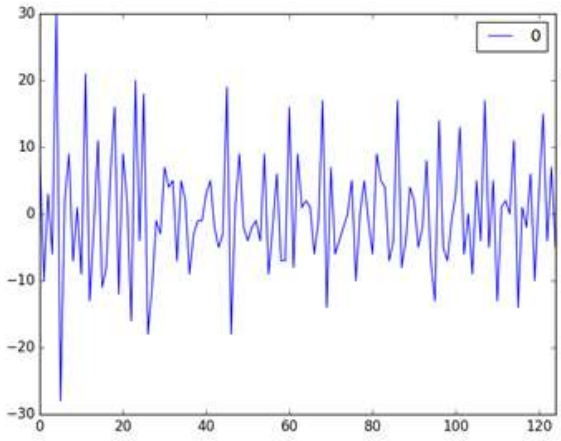
In other cases (if the residuals contain some structure or patterns) it means that the model does not include all the possible information.

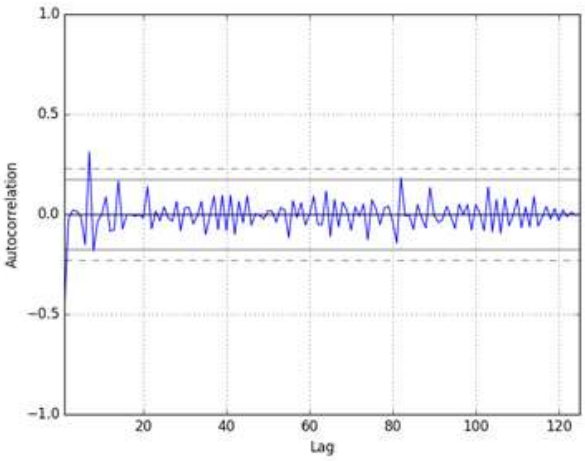
There are a several **methods for the residual analysis**.

1. **Visual analysis of residuals** (in time or frequency domain, and ACF of residuals, so-called partial ACF (**PACF**), the least accurate method).
2. **Summary statistic analysis** (mean value, std value it is behavior, values spreading).
3. **Histogram plot** (and its approximation).
4. **Q-Q plot** (probability plot, graphical comparison of probability distribution).

5. **Non-parametric statistical-hypothesis testing** (most popular is ch-square (χ^2) test, F-test, t-test, ADF-test, Ljung–Box test).
6. **Distance between distributions measurement** (parametric statistical-hypothesis testing).

The example of the time series forecast residuals. Here can be checking the presence of the trend, cyclic or seasonal component.





Here is the example of summary statistic for the previously shown graphic

```
count    125

mean     0.064000

std      9.187776

min    -28.000000

25%     -6.000000

50%     -1.000000

75%      5.000000

max     30.000000
```

```
<br>
```

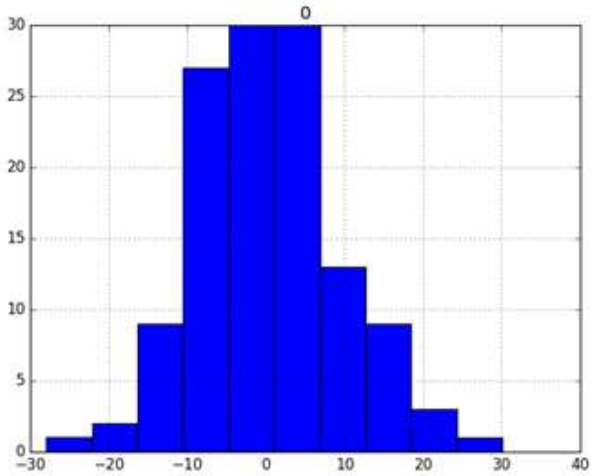
There are the following conclusions are possible.

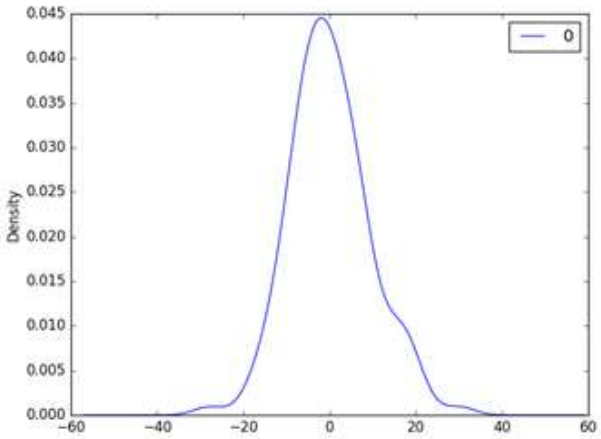
1. Mean error value close to zero, but perhaps not close enough (actually we can not say weather it is due to small sample size or by the bias of series or presence of valuable component).
2. The non-symmetric min and max values also approve some abnormal behavior.
3. The std in almost 3 times smaller then min value it is normal.

Here are the histogram and its normalized approximation are shown.

We can see some bias of mean value and presence of skewness (3-rd order statistical moment, some distribution asymmetry).

- *The small skewness* allow to suppose that some model improvement is required.
- *The large skewness* allow to suppose that the other model is need to be checked or some pre-processing (log, square root of data) is necessary before forecast (or decomposition).





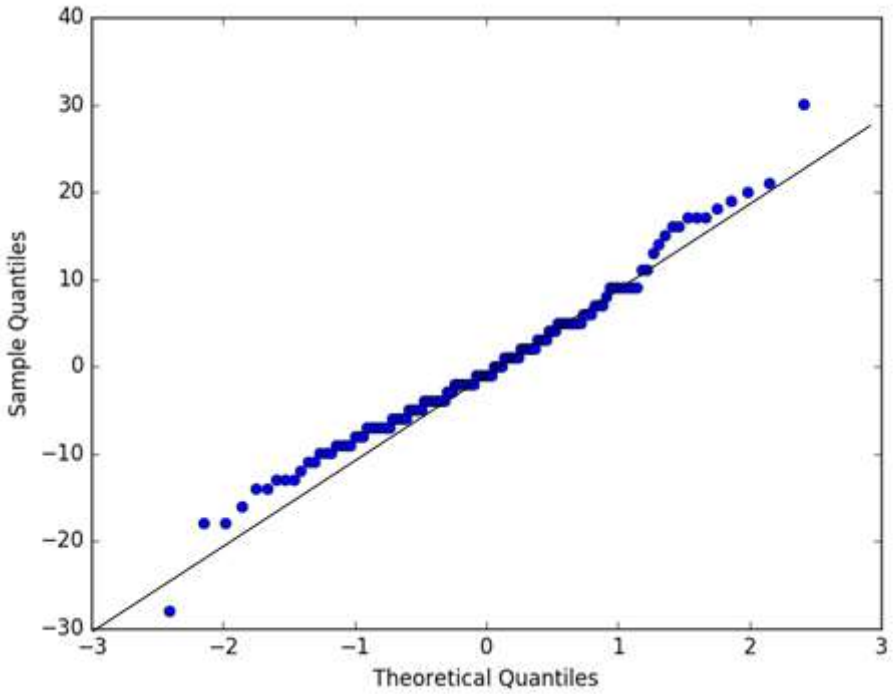
A Q-Q plot, or quantile plot is the method for compares two distributions graphically.

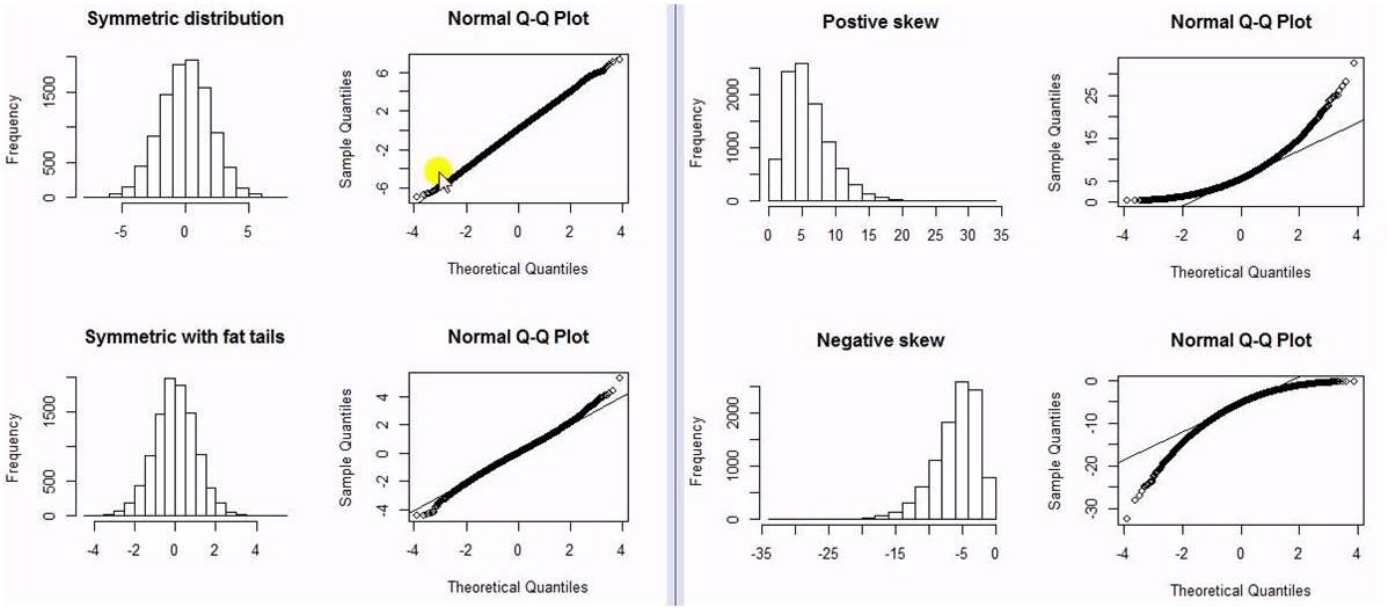
The method can be use for visual estimation how similar or different distribution happen to be.

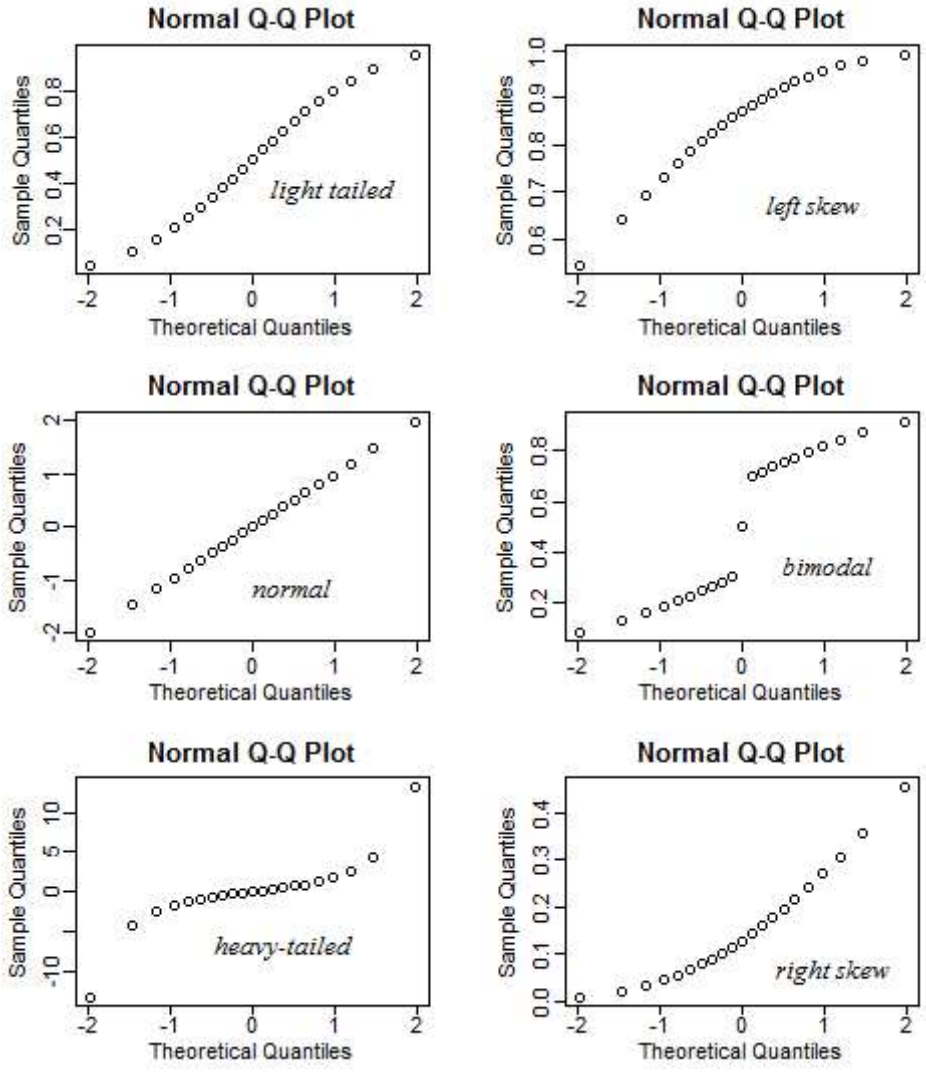
In the residual analysis the obtained distribution values are ordered and compared to an idealized Gaussian distribution.

The comparison is shown as a scatter plot (theoretical on the x-axis and observed on the y-axis) where a match between the two distributions is shown as a diagonal line from the bottom left to the top-right of the plot (with angle 45 degree $y = x$).

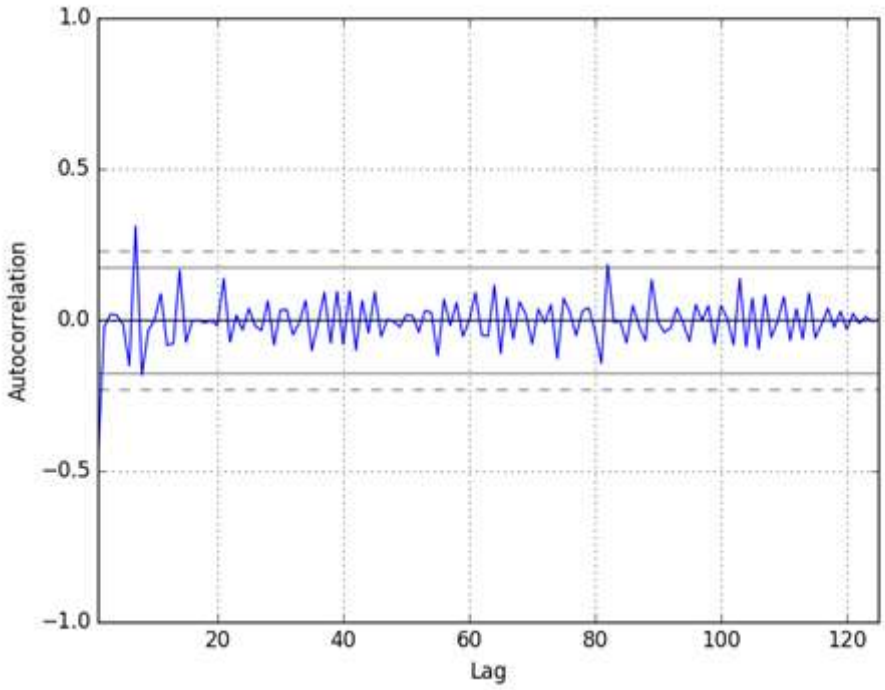
The Q-Q plot can help to rapid show the departures from this expectation.







Example of the non-parametric statistic



On the picture the continuous lines and dashed lines. These lines show the levels below which (or between +level and -level) with the confidence 95% for the continuous and 99% for the dashed line-types the samples belong to the white noise.

These lines values are obtained by the Ljung – Box test - the test checking whether the m lags of the series ACF belong to the White Gaussian Noise with the set confidence level.
The calculated value calculated by the Ljung – Box test is so-called **p-value by the Ljung – Box test**.

> *Pleas note the confident value 95% corresponds to the* $1.96 \cdot std$ *and 99% corresponds to the*$2.6 \cdot std$ *for normal distribution which can be interpreted as the law of* 3σ *(the law which says that the values out of the range* $\pm 3\sigma$ *can be considered as outliers) but here the values out of the range* $\pm 3\sigma$ *can be considered as valuable.*

There are a several techniques to check stationarity.

One of them can be done using so-called

Partial Autocorrelation Functio (PACF):

$$PACF(k,p) = \frac{\sum_{i=0}^{N-1}(y_k - \hat{y}_{k|k-p+1}) \cdot (y_{i-k} - \hat{y}_{i-k|i-k-p+1})}{std(y_k - \hat{y}_{k|k-p+1}) \cdot std(y_{i-k} - \hat{y}_{i-k|i-k-p+1})} = \frac{PA(}{std(y_k - \hat{y}_{k|k-p+1})}$$

where

$\hat{y}_{k|k-p+1}$ is the $\hat{y}_k$ predicted by $y_{k-1}, \ldots, y_{k-p+1}$ samples;

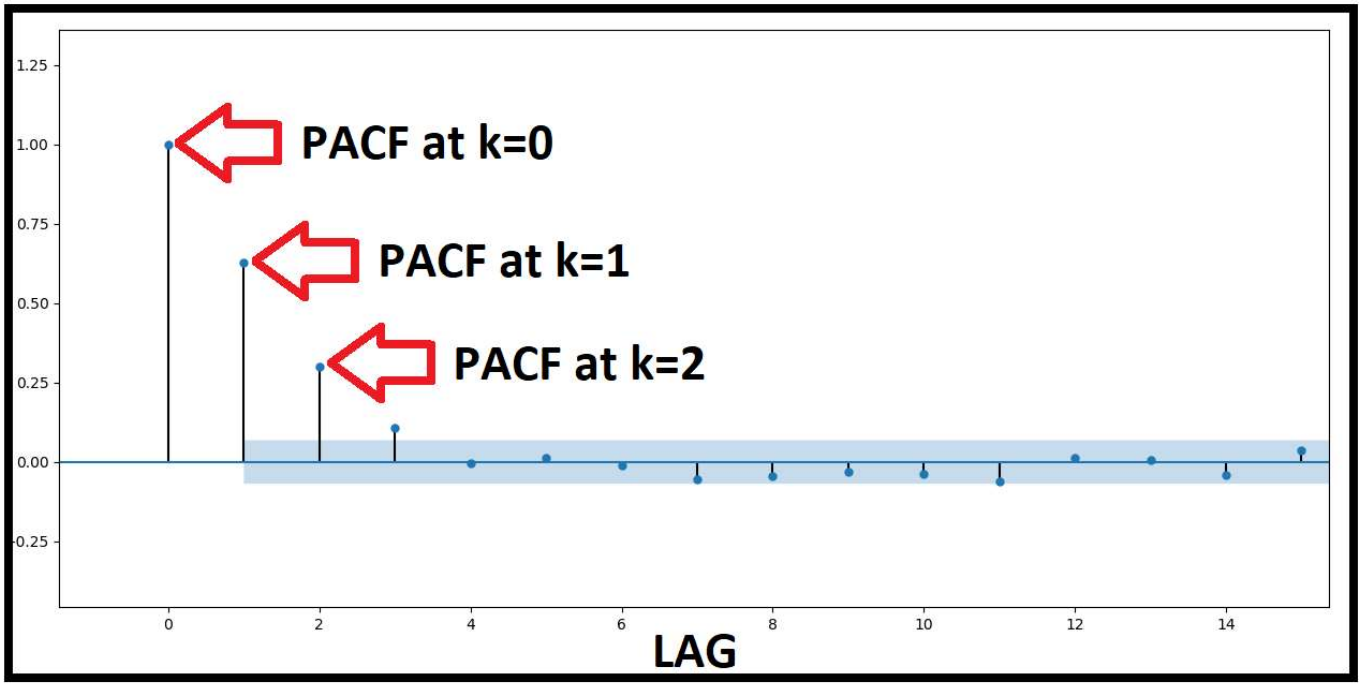
$\hat{y}_{i-k|i-k-p+1}$ is the $\hat{y}_{i-k}$ predicted by $y_{i-k-1}, \ldots, y_{i-k-p+1}$ samples;

$PACVF(k,p)$ is partial auto-covariance,

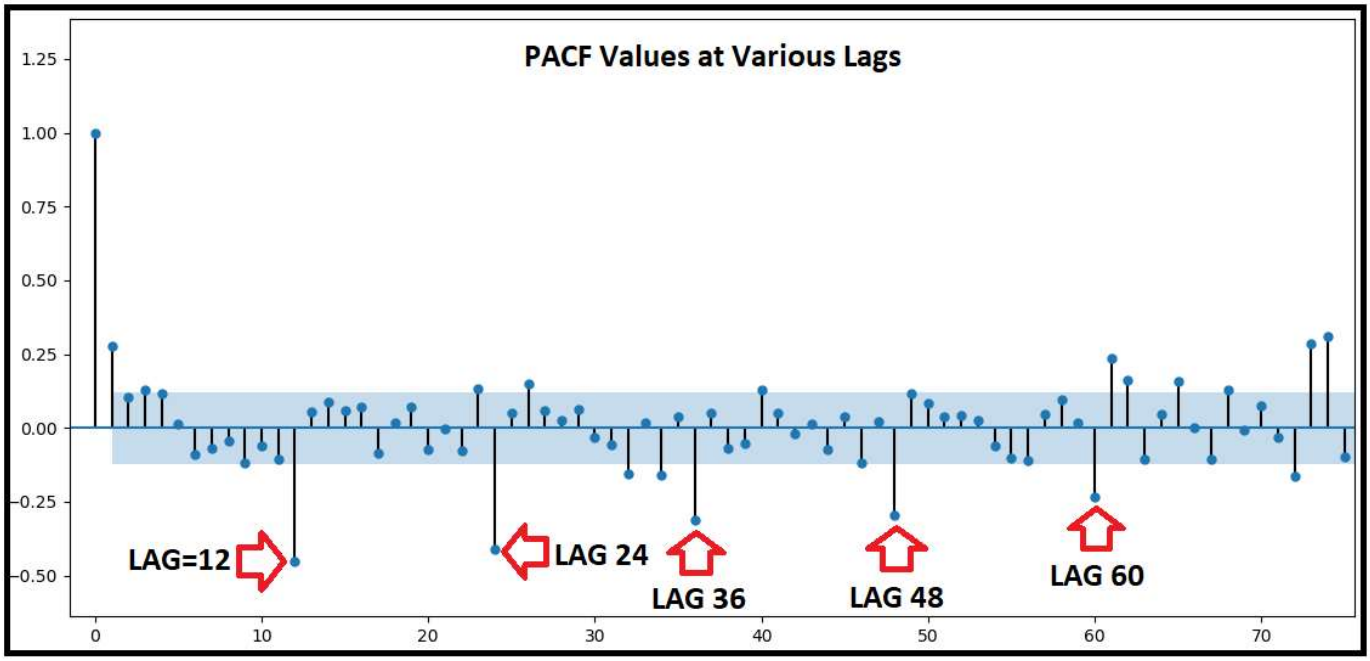
$PACVF(k,p) = \sum_{i=0}^{N-1}(y_k - \hat{y}_{k|k-p+1}) \cdot (y_{i-k} - \hat{y}_{i-k|i-k-p+1})/N$

> The "partial" correlation between two variables is the amount of correlation between them which is not explained by their mutual correlations with a specified set of other variables.
>
> For example, if we are regressing a variable Y on other variables X1, X2, and X3, the partial correlation between Y and X3 is the amount of correlation between Y and X3 that is not explained by their common correlations with X1 and X2. Here X1 X2 and X3 can be lagged versions of Y.

Example of stationary series PACF



Example of stationary series PACF with seasonality



You can put PACF to very effective use for the following things:

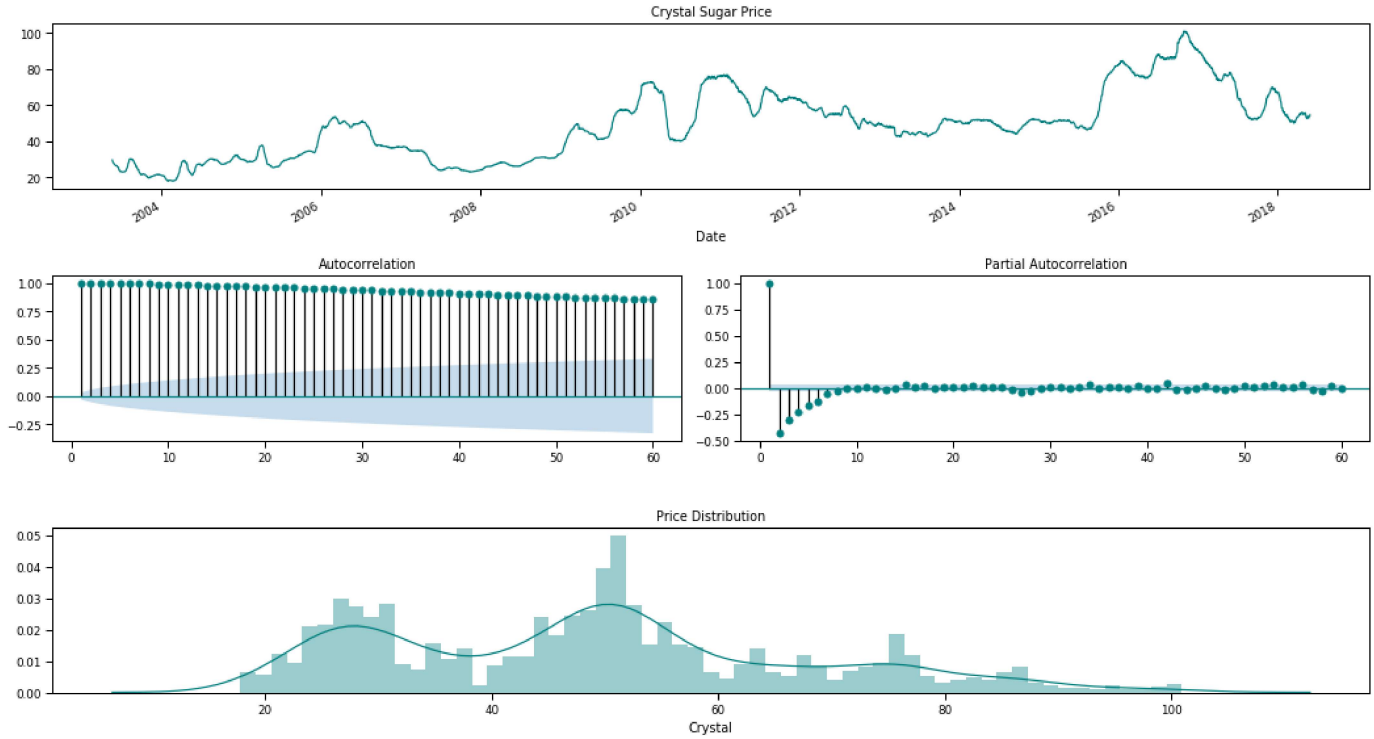
- To determine how many past lags to include in the forecasting equation of an auto-regressive model. This is known as the Auto-Regression (AR) order of the model.

- To determine, or to validate, how many seasonal lags to include in the forecasting equation of a moving average based forecast model for a seasonal time series.

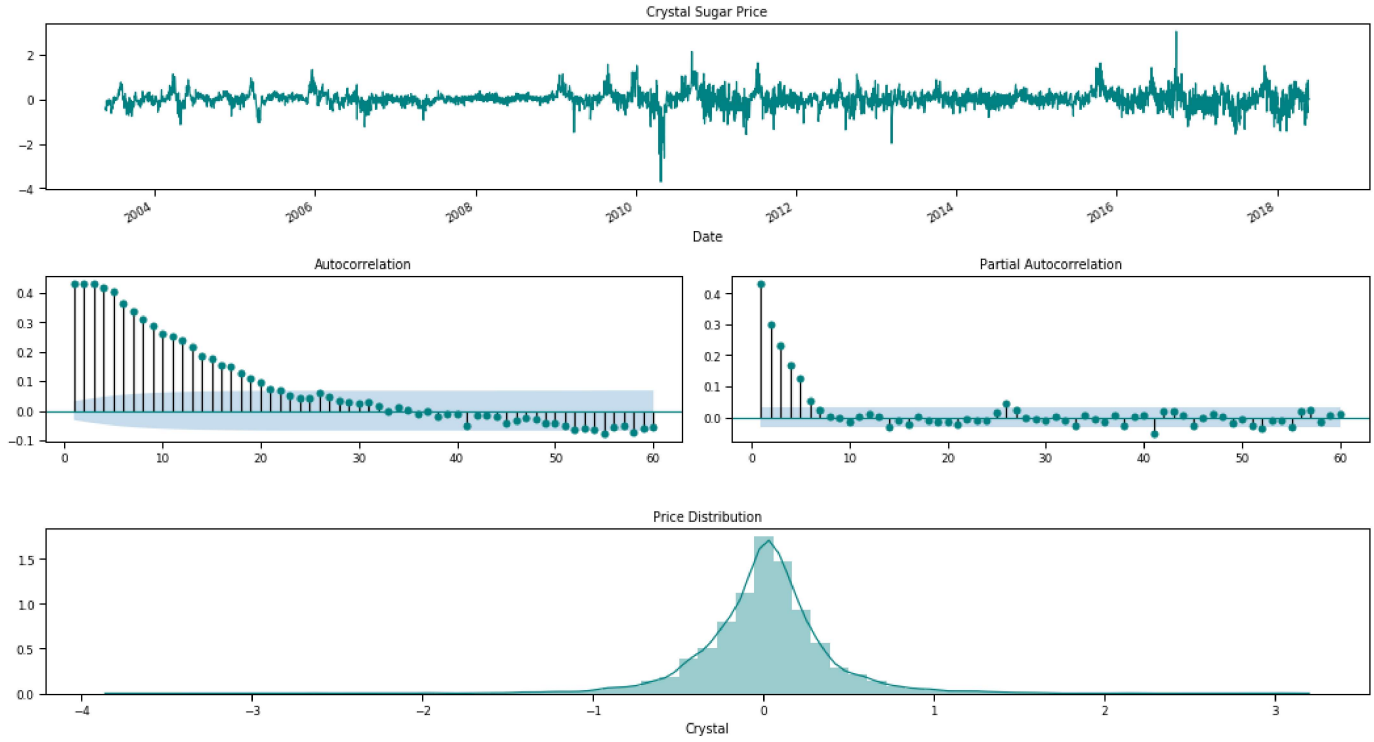
  This is known as the Seasonal Moving Average (SMA) order of the process.

The PACF can be used within ACF for stationary analysis and also will be applied further for other aims.

Example of Non-stationary case ACF and PACF



Example of Non-stationary case ACF and PACF



1.1.3 Accuracy metrics

After its prediction or its parameters estimation the accuracy have to be estimated.

There are a several techniques for this, the main one is:

- **Root of square sum (RSS)**:

$$RSS = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i))^2$$

RSS can be also called **Sum of Squared Errors (SSE)**.

The other techniques are:

- R^2 **score**:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - ev(y))^2} = 1 - \frac{RSS}{var(y)}$$

- **Mean absolute error (MAE)**:

$$MAE = \frac{1}{n}\sum_{n=0}^{N}|\hat{y_n} - y_n|,$$

- **Median Absolute Error (MedAE)**:

$$MedAE = median(|y_1 - \hat{y_1}|, \ldots, |y_n - \hat{y_n}|)$$

- **Mean absolute percentage error (MAPE)**:

$$MAPE = \frac{1}{N}\sum_{n=0}^{N}\frac{|\hat{y_n} - y_n|}{|y_n|},$$

- **Symmetric Mean absolute percentage error (sMAPE)**:

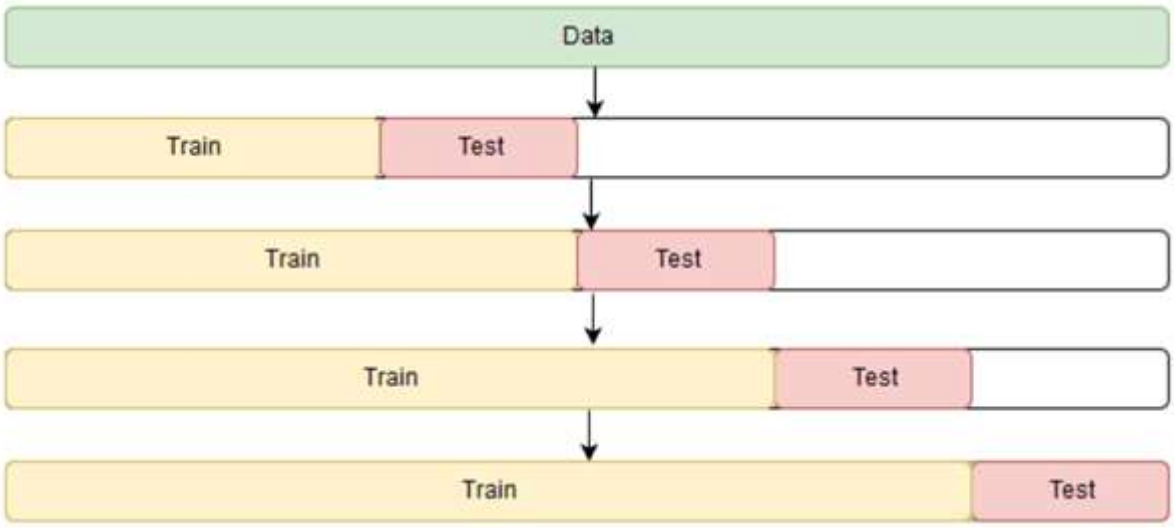
$$sMAPE = \frac{1}{N}\sum_{n=0}^{N}\frac{|\hat{y_n} - y_n|}{|y_n| + |\hat{y_n}|},$$

And many other that can be proposed for some complex cases.

It is have to be noted that the error value depends on the parameters (i.e. features) of a chosen algorithm.

These parameters as a rule can be optimized using some part of a series as testing (or validation samples) in opposite to other which is taken as training one. This routine is known as **cross-validation**.

In analogue to **k-fold** it can be introduced **cross-validation on a rolling basis** technique.



1.2 Smoothing

1.2.1 Moving average

For noise reduction, a lot of methods exists.

Simplest group of techniques are based on the smoothing idea.

The first which is frequently coming out with smoothing is the simple **moving average (MA)**.

MA can be defined as

$$y_{ma}(n) = \frac{1}{m} \sum_{i=n-m}^{n} y_i,$$

or in backward direction as

$$y_{ma}(n) = \frac{1}{m} \sum_{i=n}^{n+m} y_i,$$

where m is the order of averaging.

The moving average work due to supposition that time series components are change with rate too slowly than m samples (which is usually true for trend).

In the case of wgn, due to the strong non-stationarity moving average reduce its $variance$ in m times (std in $\sqrt{m}$ times).

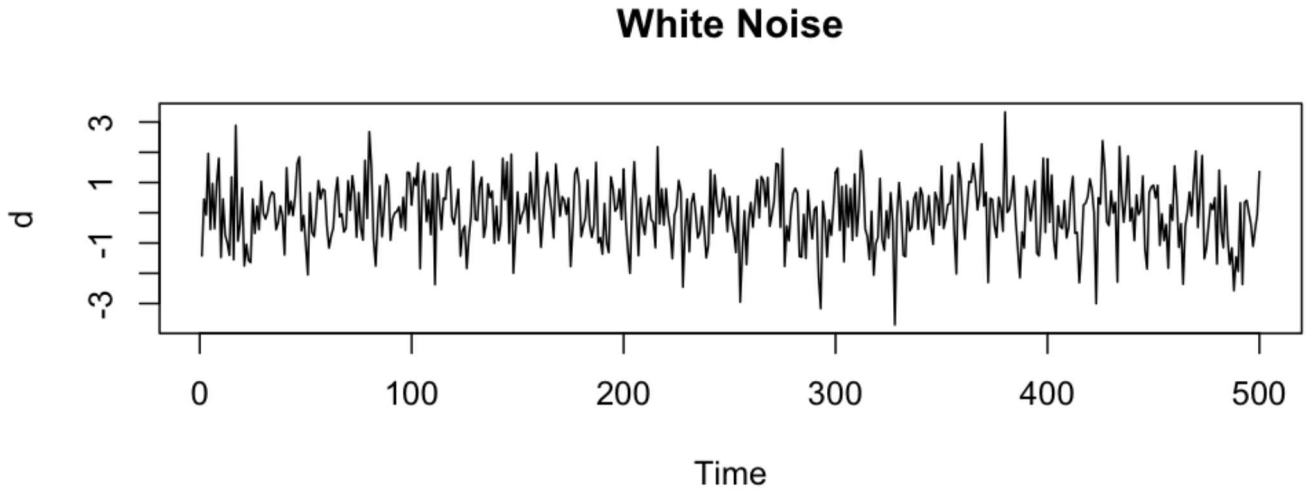
If by some reasons we do not want to take uniformly all the samples in the averaging we may take a so-called **weighted moving average (WMA)** as:

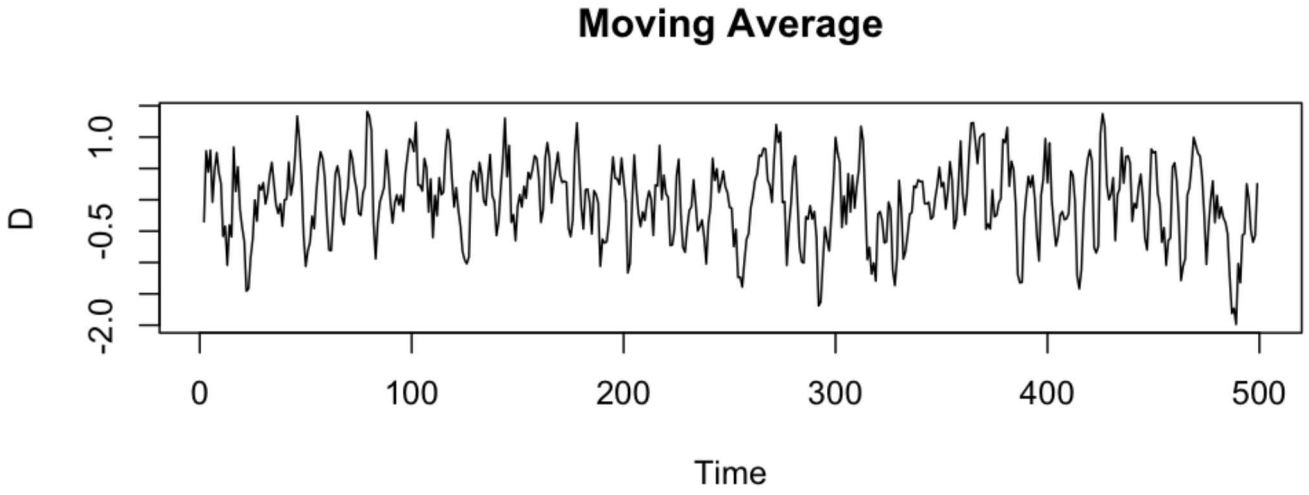
$$y_{ma}(n) = \frac{1}{m} \sum_{i=n-m}^{n} w_i y_i,$$

where w_i is the set of weights such that $\sum_{i=0}^{m-1} w_i = 1$.

Note

> Too big order of average can lead to the error in the trend or cyclic part prediction.





1.2.2 Exponential Smoothing

The next idea, to reduce the ma problem is to using the exponential smoothing.

Single Exponential Smoothing, SES (or Single Exponential Moving Average, **SEMA**):

$$\hat{y}_0 = y_0;$$

$$\hat{y}_n = \alpha y_n + (1 - \alpha)\hat{y}_{n-1},$$

where α is the smoothing parameter; $\hat{y}$ is forecast value.

> *Pleas note*, that SEMA work well for data with zero level, however for complex series it is better to choose other methods.

Double Exponential Smoothing, DES (or Double Exponential Moving Average, **DEMA**, **Holt Model**):

$$b_0 = y_1 - y_0; \rightarrow trend$$

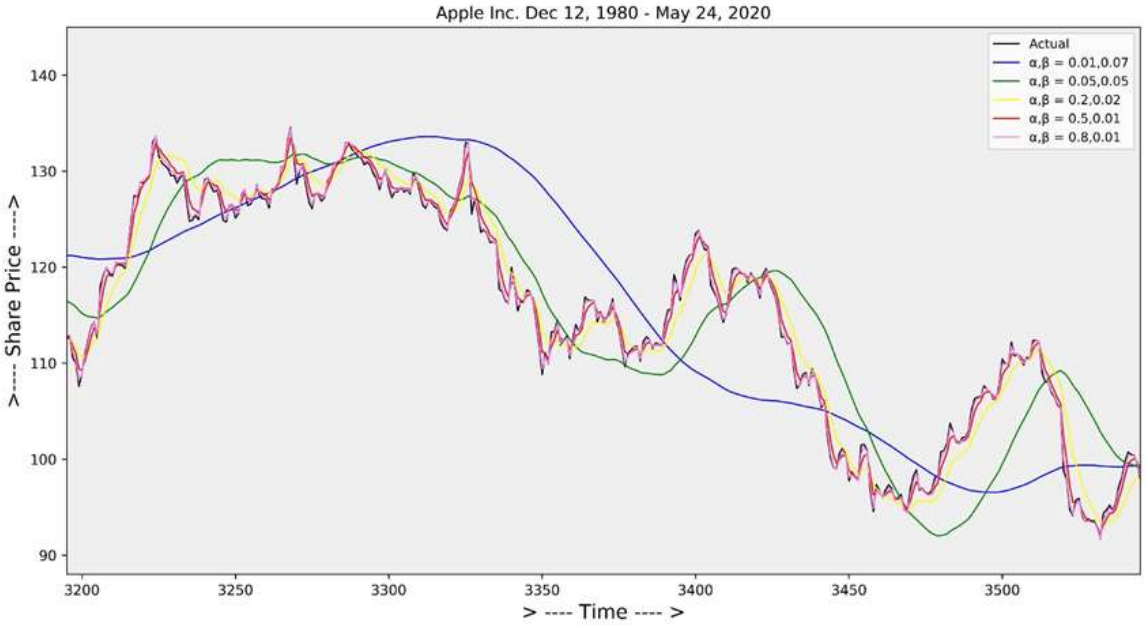
$$l_0 = y_0; \rightarrow level$$

$$l_n = \alpha y_n + (1-\alpha)(l_{n-1} + b_{n-1});$$

$$b_n = \beta(l_n - l_{n-1}) + (1-\beta)b_{n-1};$$

$$\hat{y}_{n+1} = l_n + b_n$$

where β is the additional smoothing parameter.



Triple Exponential Smoothing (or Triple Exponential Moving Average, **TEMA,**

Holt-Winters exponential smoothing, HW):

$$b_0 = y_1 - y_0; \rightarrow trend$$

$$l_0 = y_0; \rightarrow level$$

$$s_0 = \sum_{i=0}^{L-1} (y_{L+i} - y_i)/L^2; \rightarrow seasonality$$

$$l_n = \alpha(y_n - s_{n-L} + (1-\alpha)(l_{n-1} + b_{n-1});$$

$$b_n = \beta(l_n - l_{n-1}) + (1-\beta)b_{n-1};$$

$$s_n = \gamma(y_n - l_n) + (1-\gamma)s_{n-L};$$

$$\hat{y}_{n+m} = l_n + mb_n + s_{n-L+1+(m-1)modL}$$

where γ is the triple smoothing parameter; L is the estimation of season length (in samples); m is where m can be any integer meaning we can forecast any number of points into the future.

The index $n - L + 1 + (m - 1) mod L$ in the forecast equation for TEMA is the offset into the seasonal components from the last full season from observed data (i.e. if we are forecasting the 3rd point into the 45 season into the future, we cannot use seasonal components from the 44th season in the future since that season is also forecasted - we can use only the points form the observed data).

For TEMA an addition equations for estimation of deviation values can be add

$$\hat{y}_{max_x} = l_{n-1} + b_{n-1} + s_{n-L} + md_{k-L},$$

$$\hat{y}_{min_x} = l_{n-1} + b_{n-1} + s_{n-L} - md_{k-L},$$

$$d_k = \gamma \mid y_k - \hat{y}_k \mid + (1 - \gamma) d_{k-L},$$

where d is expected deviation.

Beside the shown additive Holt-Winters exponential smoothing there are exist a several Holt-Winters models for instance, for the multiplicative series case and for dumped series case.

In the generalized form the smoothing techniques can be joint in the so-called **Error-Trend-Seasonality (ETS)** model.

The model can be described as **ETS(Error,Trend,Seasonal)s = ETS(X,X,X)s**, where **X can be N-None, A-additive, M-multiplicative, Ad-additive dumped, s-sesonal period if S is not None**.

With this notations:

- Simple Exponential smoothing corresponds to the ETS(A,N,N).

- Triple Exponential smoothing corresponds to the ETS(A,A,A). where ε_t is the error)For Additive Error model depends on the empty or presence of the trend and seasonality the following cases are possible (where ε_t is the error).

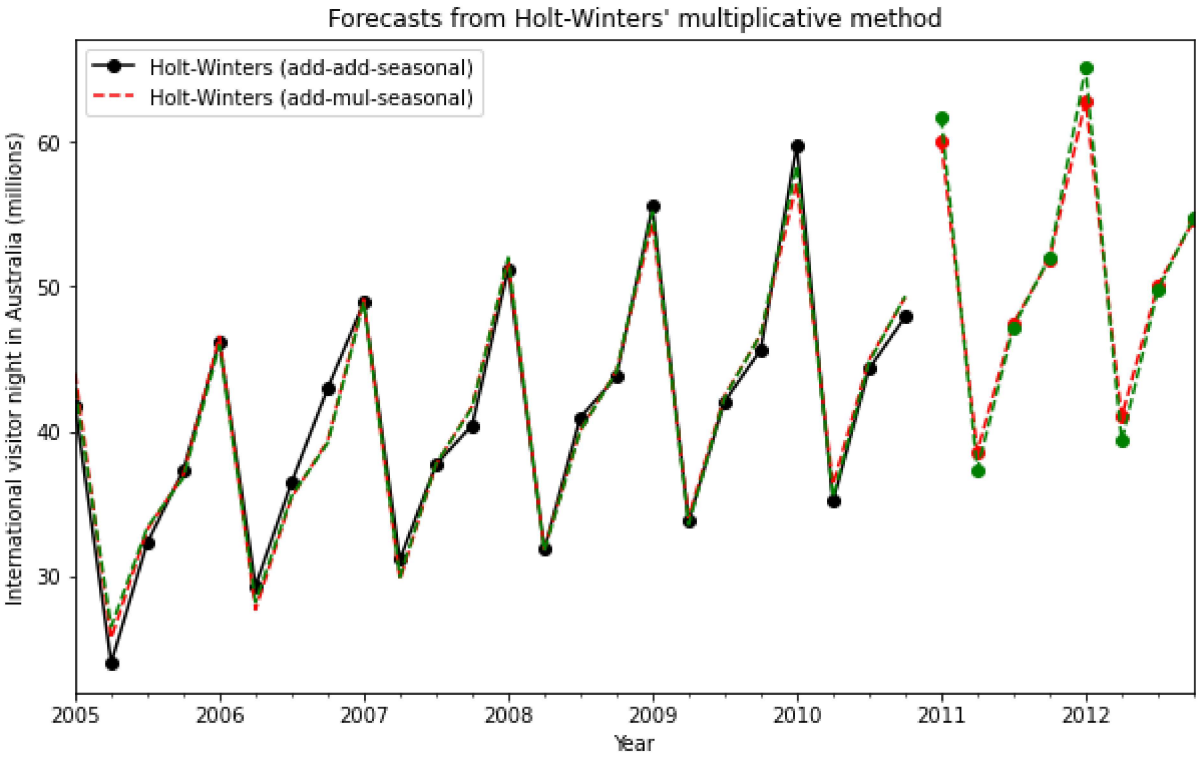
  For Additive/Multiplicative Error model depends on the empty or presence of the trend and seasonality the following cases are possible

Short hand	Method
(N,N)	Simple exponential smoothing
(A,N)	Holt's linear method
(A_d,N)	Additive damped trend method
(A,A)	Additive Holt-Winters' method
(A,M)	Multiplicative Holt-Winters' method
(A_d,M)	Holt-Winters' damped method

where notation are

Trend Component	Seasonal Component		
	N	A	M
	(None)	(Additive)	(Multiplicative)
N (None)	(N,N)	(N,A)	(N,M)
A (Additive)	(A,N)	(A,A)	(A,M)
A_d (Additive damped)	(A_d,N)	(A_d,A)	(A_d,M)

the named above is fairly both for multiplicative and additive error models. [here are the formal description for the models (https://otexts.com/fpp2/statespacemodels.png)](https://otexts.com/fpp2/statespacemodels.png)



1.3 Regression Analysis

1.3.1 Linear regression

The simplest method of regression is the **linear regression**.

This method is based on the supposition that the trend has the following model:

$$\hat{y_n} = a \cdot x_n + b + \eta_n$$

where $\hat{y_n}$ is the estimation of y_n value; a and b are the slope coefficient and the bias; η_n is the noise (random factor) variable for sample n.

This model is valid for any linear time series, for instance, for a linear trend.

Using this model the previous and future values of samples could be found.

Thus our task here is to find the coefficients which approximate series the best in correspondence with some criteria. The last is mean that we have to introduce some metric for estimating the accuracy relation on our model parameters selection.

Intuitively we may suppose that the minimum of average by distances between each sample and approximation curve could be selected as such metric. This metric was mentioned above as RSS.

The approximation series in the writen above form by RSS minimization is called **Least-Square Method (LSM)**.

The simplest casr of LSM is **Ordinary LSM (OLS)**.

OLSM task can be given as:

$$\sum_{n=0}^{N-1}(y_n - \hat{y_n})^2 \to 0$$

In our case it lead to the following:

$$\sum_{n=0}^{N-1}(y_n - (a \cdot x_n + b))^2 \to 0$$

This problem can be solved analitically

$$a = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y - a\sum x}{N}$$

In more general case, when the series can be expressed as some linear combination of terms, for instance polynomial regression $\hat{y}_n = \sum_{i=0}^{p} w_i x_n^i + \eta_n = w_0 + w_1 x_n + w_2 x_n^2 + \ldots + \eta_n$ the series can be given for each sample y_n estimation as

$$\hat{y}_n = \sum_{i=0}^{p} w_i x_i(n) = W^T X(n),$$

where $x_i(n)$ is map of i-th term, and the zero-term is bias; $W = (w_0, w_1, \ldots w_p)^T$; $X(n) = (1, x_1(n), \ldots x_p(n))^T$. By introduction vector $Y = (y_0, y_1, \ldots y_{N-1})$ and matrix $X = (X(0), X(1), \ldots, X(N-1))$ we can write linear regression in the following form

$$Y = W^T X + \eta,$$

where $\eta = (\eta_0, \eta_1, \ldots)$; Y has dimension $N \times 1$ (raw); W has dimension $1 \times N$ (column); X has dimension $(p+1) \times N$ (matrix).

For the discussed model it could be said that or series has $p+1$ features or the order of the model is p.

The OLS problem solution in the our generalized case

$$W = (X^T X)^{-1} X^T Y = X^+ Y$$

where X^+ is the pseudo inverse matrix.

Pleas note:

1. The OLS solution work for for any symmetric distribution of noises η,with zero mean value and restricted variance

   in particular for Gaussian Noise distribution, such that:

   $$\eta \sim N(0, \sigma),$$

   $$\text{ev}(\eta) = 0,$$

   $$\text{var}(\eta) = \sigma^2 < \infty,$$

   $\text{cov}(\eta_i, \eta_j) = 0$ if $i \neq j$ i.e. stationary i.i.d. noises

2. **The Gauss-Markov Theorem** state that for Gaussian Noise distribution OLS provide an statistically effective beast linear unbiased estimation (**BLUE**) with the variance of this estimation $var(W) = var(\eta)(X^T X)^{-1}$.

3. The more general case of least-square solution for slightly non-stationary case is the weighted LS (WLS) solution

   $$W = (X^T A X)^{-1} X^T A Y,$$

   where A is some weight matrix. The particular case (Generalized LS, GLS),when $A^{-1} = cov(X_i, X_j)$ appropriate for complex noise with $cov(\eta_i, \eta_j) \neq 0$.

Ill-conditioned problem

One of the main constrain to the using of ordinary regression is the Ill condition number of the processed data.

Ill-conditioned problem – the relatively high changing of the estimation results, caused by the small perturbation (changing) in the data.

Pleas note

- Ill or well condition are depends on the inverse condition number.
- Physically condition number number depends on the noise influence on the series. Thus then more relation of deterministic part to the noise one then better the condition of series (i.e. then smaller the perturbation influence of the data of the estimation result).
- The problem of Ill-conditions can implicitly appears in the growing of variance and outliers in the regression prediction.

The problem of Ill-conditions or high noise influence leads to the possibility of overfitting the approximated series.

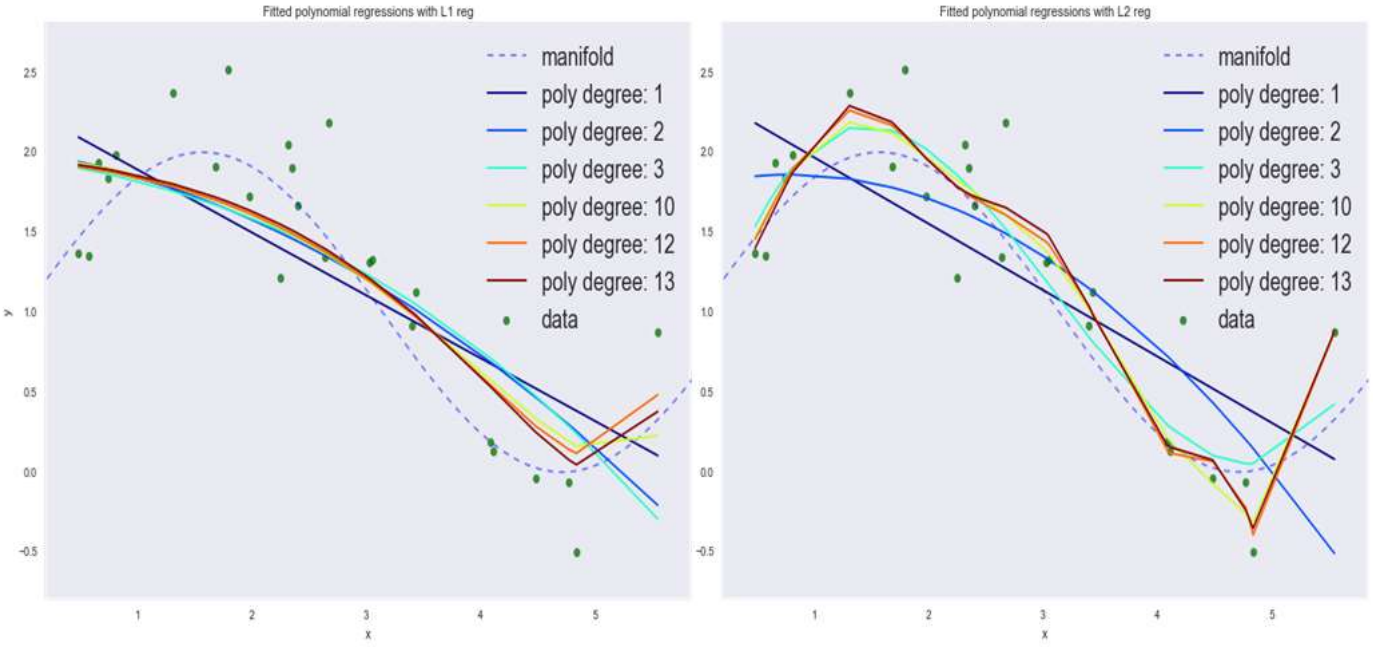
The problem of Ill-conditions or high noise influence leads to the possibility of overfitting the approximated series.

The solutions of ill-condition problem.

1. **Robust regression** is the group of heuristically proposed methods for traditional statistical problem solutions. This methods can be based on change the criteria (RSS) on the other which provide statistically uneffective but more robust to noise results (for instance median regression, MAE-regression, or so-called M-estimators).
2. **Features selection (Dimension reduction)** - select only restricted (reduced) number of features which contribution number to small (then it smaller - then better).
3. **Feature transformation** to the form with better condition (for instance PCA, and other decomposition technique).

4. **Matrix regularization** and normalization (L1 lasso, L2 ridge (Tichonov) and e.t.c).

Example of previously shown problem solution with using L1 and L2 regularization



In addition to mentioned above in some cases series could be re-scale (**normalization and standardization**):

1. **Standardization** - make distribution with zero meaning and variance 1 by transform

$$y' = (y - ev(y))/\sqrt{var(y)},$$

where y' is transformed value y.

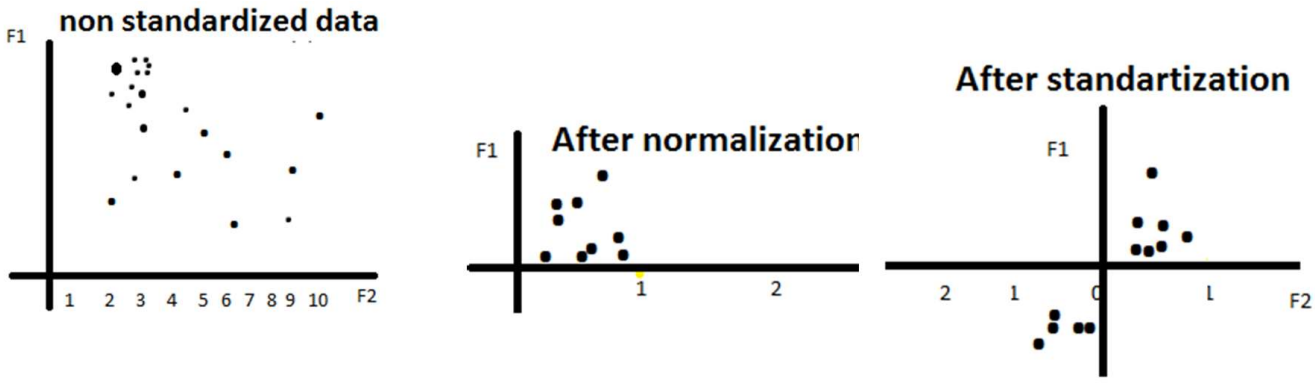
2. **Min-max normalization**:

$$y' = (y - min(y))/(max(y) - min(y)).$$

3. After normalization data can be **scaling**

$$y' = ay + b,$$

for instance b may be instrumental error, and a – graduation coefficient.



1.3.2 Non-Linear regression

The generalized regression problem can be written as

$$L(Y - f(W, X)) \to min$$

where L is some criteria (or better to call it here loss function); $f(W, X)$ is some functional relation between weight vector and feature matrix (in particular $f(W, X) = f(W^T X)$).

The generalized regression problem (in particular non-linear regression) can appear if:

1. data can't be reduced to the linear form.

2. frequently the relation between data and answers can't be formulated analytically, it is only supposed that the relation exist, and we want to approximated.
3. We want to classify some data into known classes.

The non-linear regression can be solved using the gradient descent method, in which iteratively weights vector renewed as

$$W^t = W^{t-1} - \mu \nabla L(W^{t-1}, X),$$

where μ is the learning rate.

Pleas note

1. If $L = \sum(Y - f(W^T X))^2$, than

   $W^t = W^{t-1} - 2\mu \nabla f^T((W^{t-1})^T X)X.$

2. The gradient descent procedure require initialization of weight vector

   $W^0 = w_0^0, w_1^0, \ldots, w_N^0$ and μ_0 - first approximation.

3. In general the result can be depended on the initial values If the loss function supposed no to be smooth everywhere.

4. In general μ_t might be changed! The loss minimum could be skipped depends on the learning rate!

5. The procedure of gradient descent should be continuous until reach the condition:

   $t \leq t_{\text{limited}}$;
   $W^t - W^{t-1} < \xi$, where ξ is some small number;
   $L(W^t, X) - L(W^{t-1}, X) < \xi$;
   Cross-validation techniques

6. The procedure do not guaranty lack of under/over fitting problem.

7. In general modified gradient descent can be used (such as stochastic, adaptive, second-order, and e.t.c.).

The Neural-network can be considered as non-linear regression.

The Cibenko-Hronik Theorem (1989) state that the each restricted function $f(x)$ can be approximated by the following solution:

$$\hat{f}(x) = \sum_{i=1}^{M} \alpha_i \sigma_i(W^T X)$$

such that

$$L(f(x), \hat{f}(x)) < \xi(N, P)$$

Where

- $\xi(N, M)$ is sufficiently small error;
- $\hat{f}(x)$ is the approximation result;
- $\sigma_i(W^T X)$ is the nonlinear regression of input layer (represent a hidden layer in the neural network);
- α_i is the weight for each output of hidden layer;
- M is the number of synapses between the hidden and output layer;
- P is the number of input feature.

Uitson Lema: for data with p linear independent features the $2p + 1$ hidden layer dimension enough.

In [ ]: