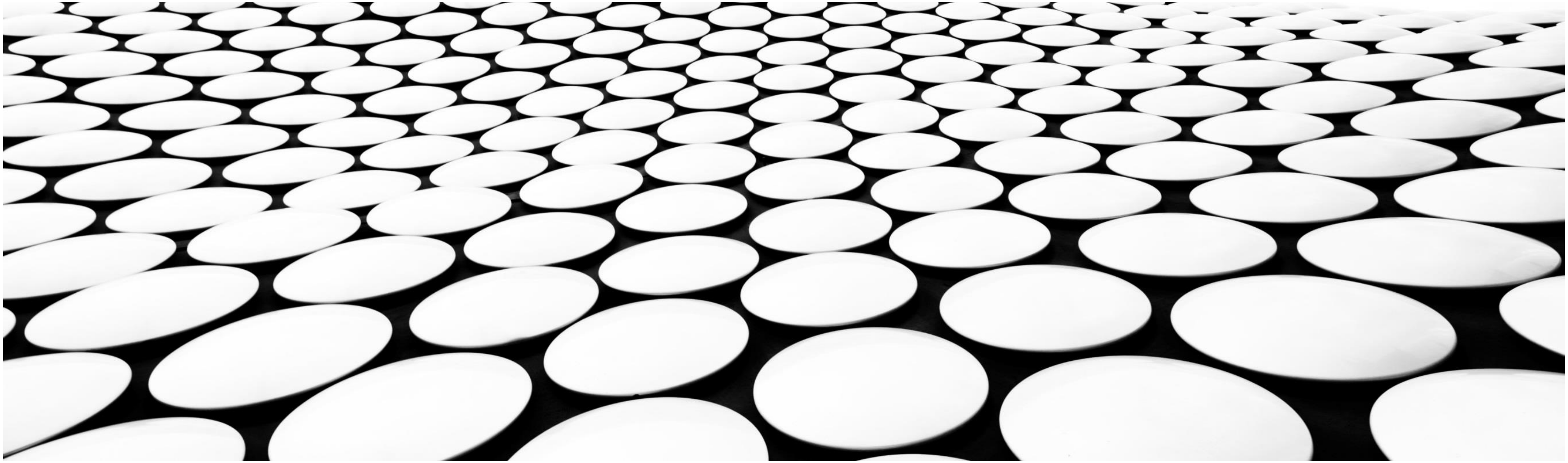
TIME SERIES

DESCRIPTIVE STATISTICS AND DATA PROCESSING

Prof. Dr. Klaus Fabian Côco
Prof. Dr. Patrick Marques Ciarelli

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO



DESCRIPTIVE STATISTICS OF TIME SERIES

DESCRIPTIVE STATISTICS OF TIME SERIES

There are various statistical measures applicable when describing a time series. These measures allow capturing information such as:

- Central tendency or location,
- Dispersion or spreading, and
- Distribution or format.

The measures presented are most effective when the time series are stationary.

For the next slides consider:

- Time series are discrete in time, and they can be univariate, when there is only one variable, or multivariate, when there are several variables over time;
- A discrete variable X in a time series is composed of a series of samples (values) equally spaced in time by an interval ΔT, so that there are values of X in the instants $X[0]$, $X[\Delta T]$, $X[2\Delta T]$, $X[3\Delta T]$, ..., $X[i\Delta T]$, ..., $X[(n-1)\Delta T]$, where n is the total number of samples;
- For simplicity, the value of X at instant i, $X[i\Delta T]$, can be represented by $X[i]$ or x_i;
- <u>Note that the first sample of the variable X is $X[0]$ (x_0) instead of $X[1]$ (x_1).</u>

DESCRIPTIVE STATISTICS OF TIME SERIES

CENTRAL TENDENCY

A measure of central tendency (also referred to as measures of center or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or center of its distribution. Some main measures of central tendency are:

- **Mean** (μ): This is the mean value of an entire or a windowed time series, where n is the number of data points in the variable T and t_i is the data value at index i.

$$\mu(T) = \frac{1}{n}\sum_{i=0}^{n-1} t_i$$

- **Median** (*median*): The median is the "middle" of n numbers when those numbers are arranged from smallest to greatest. This measure is more robust to outliers than the mean.

$$median(T) = \begin{cases} \dfrac{t_r + t_{r+1}}{2}, \text{if } n \text{ is even } (r = n/2) \\ t_{r+1}, \text{if } n \text{ is odd } (r = (n-1)/2) \end{cases}$$

The values of the variable T must be arranged from smallest to largest, such that $r = 1$ is the smallest value and $r = n$ is the largest value.

DESCRIPTIVE STATISTICS OF TIME SERIES

DISPERSION

Measures of <u>dispersion</u> describe how similar or varied the set of observed values are for a particular variable. Some main measures are:

- **Standard Deviation** (σ): This measure indicates how much variation exists around the mean.

$$\sigma(T) = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}[t_i - \mu(T)]^2}$$

- **Variance** (σ^2): It is the square of the standard deviation.

- **Range** ($range$): This measures returns the interval value between minimum and maximum values.

$$range(T) = maximum(T) - minimum(T)$$

$$minimum(T) = \min_{i=0,\ldots,n-1}(t_i) \quad maximum(T) = \max_{i=0,\ldots,n-1}(t_i)$$

- **Coefficient of Variation** (cv): It is defined as the ratio of the standard deviation (σ) to the mean (μ) (or its absolute value, $|\mu|$). It is not indicated when the mean value is close to zero, because the coefficient of variation will approach infinity.

$$cv(T) = \frac{\sigma(T)}{|\mu(T)|}$$

DESCRIPTIVE STATISTICS OF TIME SERIES

DISPERSION

Some more robust estimates of dispersion are:

- **Absolute Average Deviation** (AAD):

$$AAD(T) = \frac{1}{n}\sum_{i=0}^{n-1}|t_i - \mu(T)|$$

- **Median Absolute Deviation** (MAD):

$$MAD(T) = \underset{i=0,\dots,n-1}{median}(|t_i - median(T)|)$$

- **Interquartile Range** (IQR):

$$IQR(T) = P_{75\%}(T) - P_{25\%}(T) = Q_3(T) - Q_1(T)$$

where Q_1 is the first quartile and Q_3 is the third quartile. With n numbers arranged from smallest to greatest, compute:

$$P_k(T) = \begin{cases} \dfrac{t_r + t_{r+1}}{2}, \text{if } n \times k \text{ is an integer } (r = n \times k) \\ t_r, \quad \text{otherwise } (r = \lceil n \times k \rceil) \end{cases}$$

with $0 \leq k \leq 1$

The values of the variable T must be arranged from smallest to largest, such that $r = 1$ is the smallest value and $r = n$ is the largest value.

DESCRIPTIVE STATISTICS OF TIME SERIES

EXAMPLE

Example: For the values of the variable T, calculate the statistical measures presented above.
T = [1, 2, 13, 10, 9, 16, 18, 20, 24, 15].

Solution:

$$\mu(T) = \frac{1}{n}\sum_{i=0}^{n-1} t_i = 12.8 \qquad median(T) = \frac{13+15}{2} = 14$$

$$\sigma(T) = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}[t_i - \mu(T)]^2} = 7.4356 \qquad \sigma^2(T) = \frac{1}{n-1}\sum_{i=0}^{n-1}[t_i - \mu(T)]^2 = 55.2889$$

$$range(T) = maximum(T) - minimum(T) = 24 - 1 = 23 \qquad cv(T) = \frac{\sigma(T)}{|\mu(T)|} = \frac{7.4356}{12.8} = 0.5809$$

$$AAD(T) = \frac{1}{n}\sum_{i=0}^{n-1}|t_i - \mu(T)| = 5.84 \qquad MAD(T) = 4.5 \qquad IQR(T) = P_{75\%}(T) - P_{25\%}(T) = 18 - 9 = 9$$

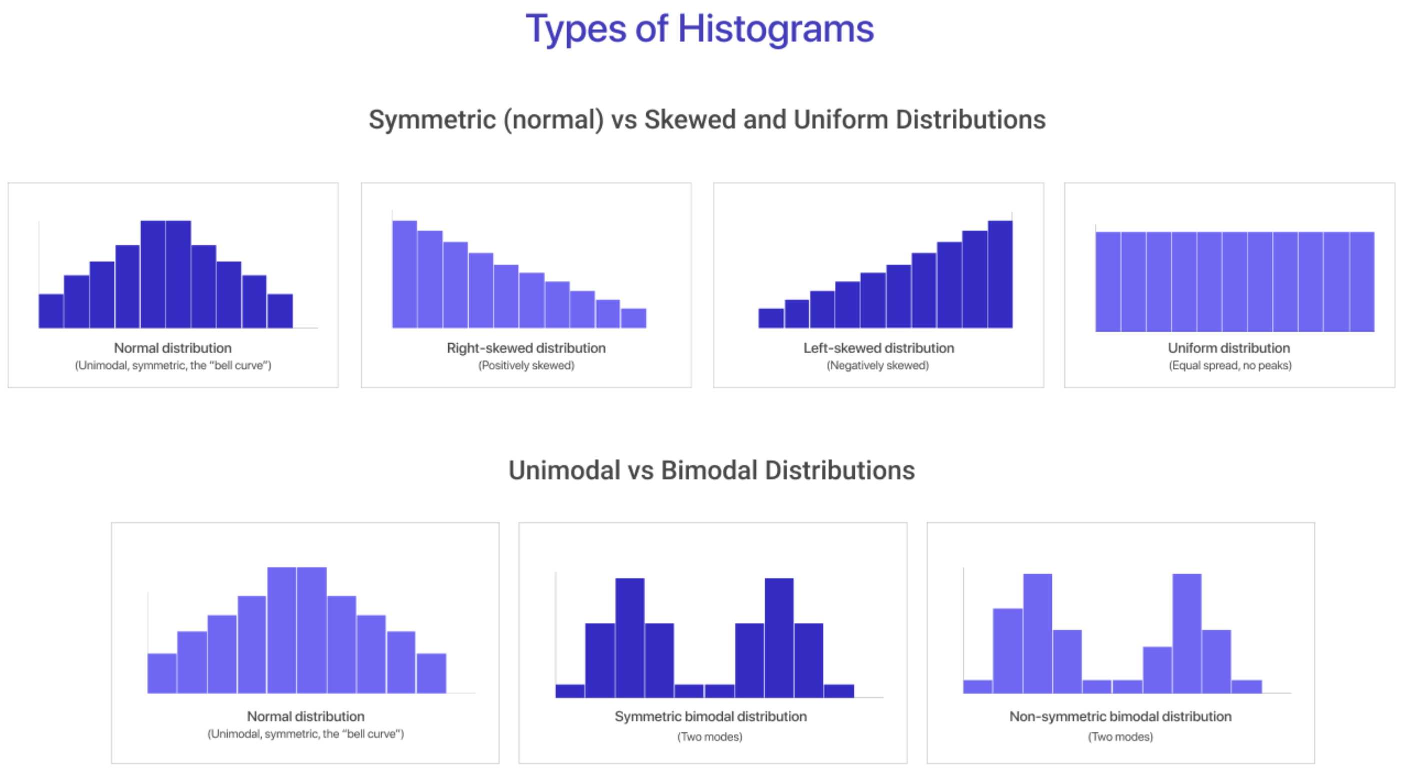
Note that *IQR* represents the entire range of values, while σ, *AAD*, and *MAD* represent only half of the range around the central tendency, for example $\mu \pm \sigma$

DESCRIPTIVE STATISTICS OF TIME SERIES

DISTRIBUTION

A statistical distribution, or probability distribution, describes how values are distributed for a variable. In other words, the statistical distribution shows which values are common and uncommon. Some measures and ways to analyze a distribution are:

- **Histogram:** histogram is a visual representation of the distribution of quantitative data. Histogram can give an approximate idea of the probability distribution function (PDF) of a variable.



DESCRIPTIVE STATISTICS OF TIME SERIES

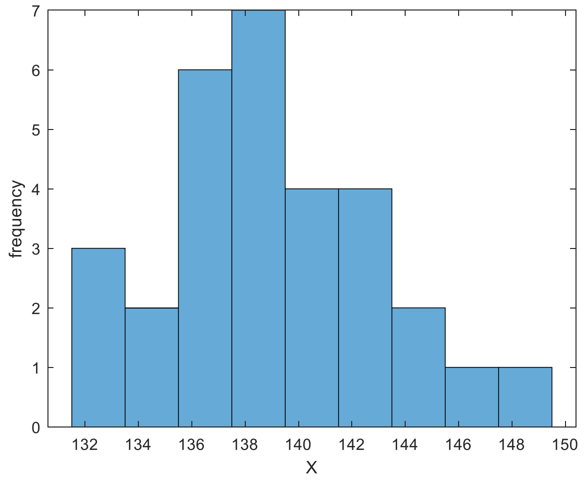
DISTRIBUTION

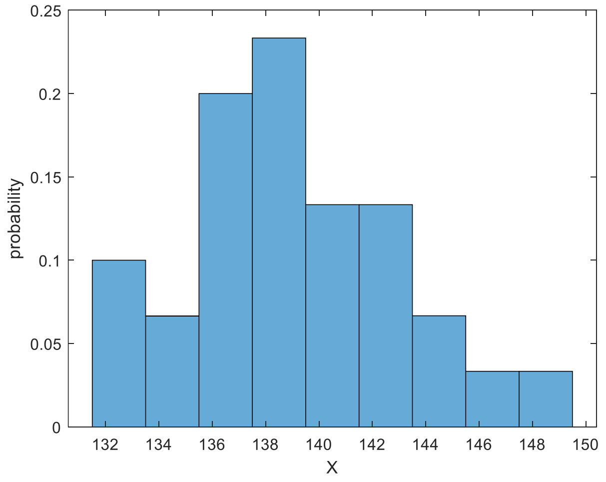
How to create a histogram:

To create a histogram, the data need to be grouped into intervals (bins) of equal length.
Then, compute the frequency (or relative frequency) of the data into each interval.
The relative frequency is the frequency of the data in a interval divided by the total number of data.
The bars are as wide as the interval and as tall as the frequency (or relative frequency).

Example: Compute the histogram of a time series with the values [135, 137, 136, 137, 138, 139, 140, 139, 137, 140, 142, 146, 148, 145, 139, 140, 142, 143, 144, 143, 141, 139, 137, 138, 139, 136, 133, 134, 132, 132].
For histograms, we usually want to have from 5 to 20 intervals. Since the data range is from 132 to 148, it is convenient to have a interval of length 2 since that will give us 9 intervals:

131.5 - 133.5 -> 3 -> 1/10
133.5 - 135.5 -> 2 -> 1/15
135.5 - 137.5 -> 6 -> 1/5
137.5 - 139.5 -> 7 -> 7/30
139.5 - 141.5 -> 4 -> 2/15
141.5 - 143.5 -> 4 -> 2/15
143.5 - 145.5 -> 2 -> 1/15
145.5 - 147.5 -> 1 -> 1/30
147.5 - 149.5 -> 1 -> 1/30





DESCRIPTIVE STATISTICS OF TIME SERIES

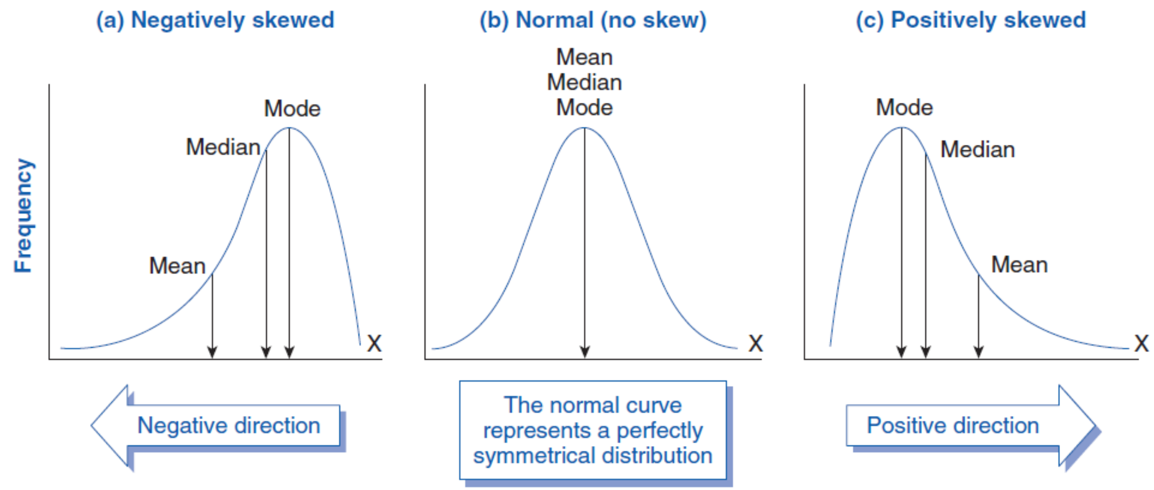
DISTRIBUTION

The skewness and kurtosis are higher-order statistical attributes and relate to the distribution format.

- **Skewness** ($skewness$): this measure indicates the symmetry of the probability density function (PDF) of the amplitude of a time series.

$$skewness(T) = \frac{\sum_{i=0}^{n-1}[t_i - \mu(T)]^3}{(n-1)\sigma^3(T)}$$

A time series with an equal number of large and small amplitude values has a skewness of zero.
When a time series has many large values and few small values (left tail), the skewness value is negative.
When a time series has many small values and few large values (right tail), the skewness value is positive.



DESCRIPTIVE STATISTICS OF TIME SERIES

DISTRIBUTION

The skewness and kurtosis are higher-order statistical attributes and relate to the distribution format.

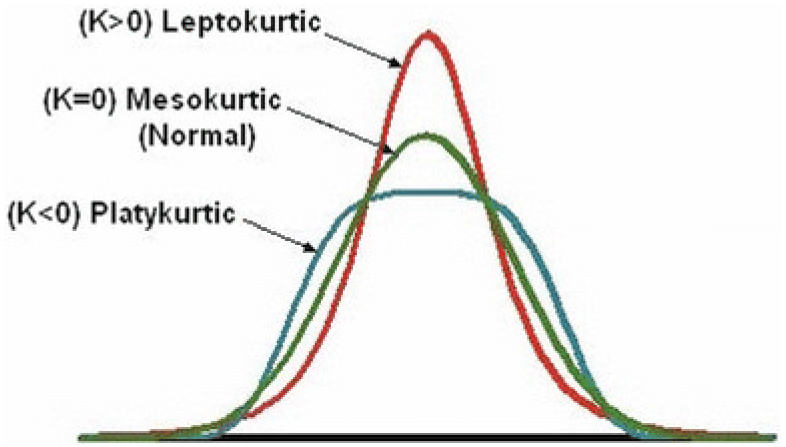
- **Kurtosis** ($kurtosis$): it measures the peakedness of the PDF of a time series.

$$kurtosis(T) = \frac{\sum_{i=0}^{n-1}[t_i - \mu(T)]^4}{(n-1)\sigma^4(T)}$$

The kurtosis value of a normal distribution (Gaussian distribution) equals 3. Normally, this distribution is used as a reference, and a correction is made in the equation:

$$kurtosis(T) = \frac{\sum_{i=0}^{n-1}[t_i - \mu(T)]^4}{(n-1)\sigma^4(T)} - 3$$

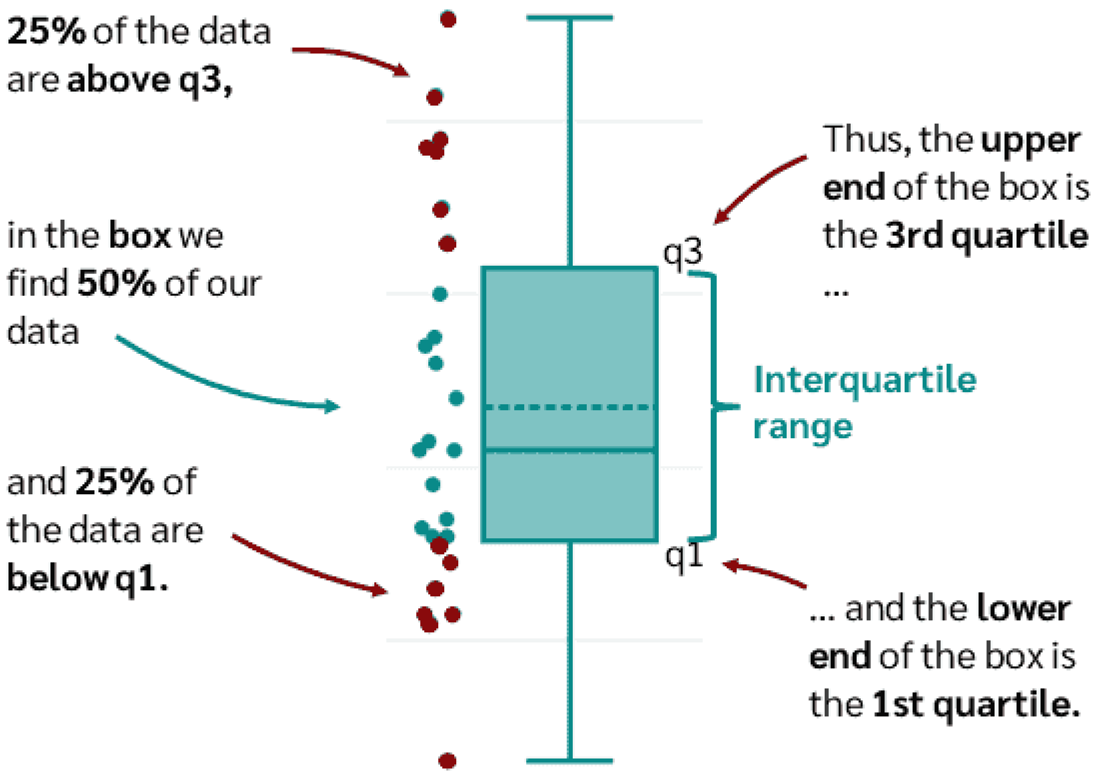
Therefore, distributions close to zero indicates a Gaussian-like peakedness.
PDFs with relatively sharp peaks have kurtosis greater than zero.
PDFs with relatively flat peaks have kurtosis less than zero.

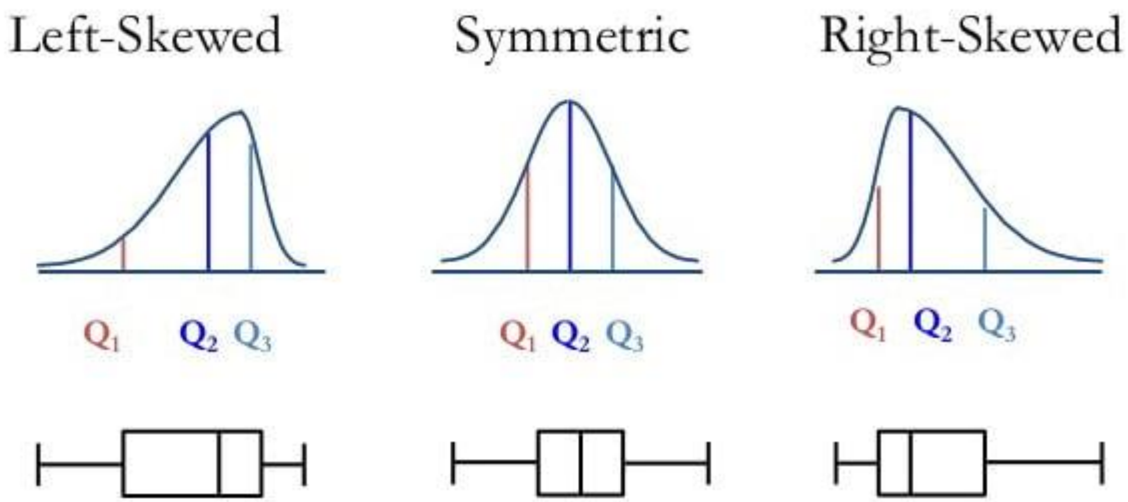


DESCRIPTIVE STATISTICS OF TIME SERIES

DISTRIBUTION

- **Boxplot:** It is a method for demonstrating graphically the locality, spread and skewness groups of numerical data through their quartiles. In addition, outliers, which are points that differ significantly from the rest of the data, may be plotted as individual points beyond the whiskers on the boxplot.





DESCRIPTIVE STATISTICS OF TIME SERIES

DISTRIBUTION

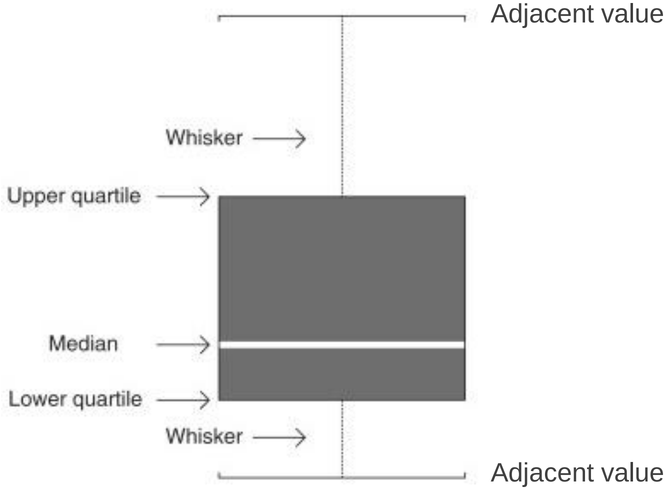
How to create a boxplot:

To create this plot we need 3 numbers: Q_1 (lower quartile) ($P_{25\%}$), Q_2 (median), Q_3 (upper quartile) ($P_{75\%}$);
Mark the numbers above the horizontal axis with vertical lines (or vertical axis with horizontal lines).
Connect Q_1, Q_2 and Q_3 to form a box.
Compute the lower and upper limits:

$$lower\ limit = Q_1 - 1.5 \times IQR$$
$$upper\ limit = Q_3 + 1.5 \times IQR$$

where $IQR = Q_3 - Q_1$

Potential outliers are observations that lie outside the lower and upper limits.
Identify the adjacent values, that are the most extreme values that are not potential outliers. Then, connect them to the box to form the whiskers.



DESCRIPTIVE STATISTICS OF TIME SERIES

DISTRIBUTION

How to create a boxplot:

Example: Let X be a time series with the values [24, 58, 61, 67, 71, 73, 76, 79, 82, 83, 85, 87, 88, 88, 92, 93, 94, 97]. Create a boxplot with these values.

Solution:

$Q_1 = 71$ $Q_2 = 82.5$ $Q_3 = 88$ IQR = 88 – 71 = 17

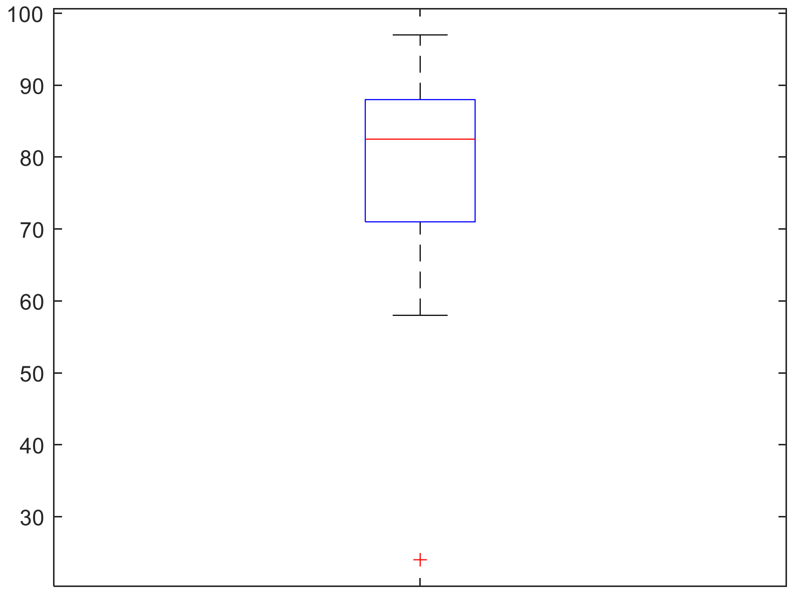
lower limit = 71 – 1.5x17 = 45.5

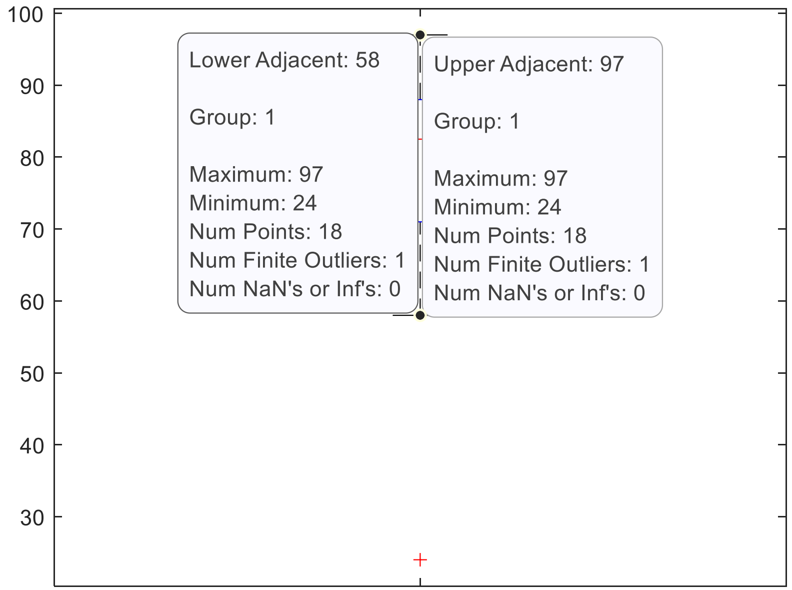
upper limit = 88 + 1.5x17 = 113.5

24 is smaller than 45.5: potential outlier

Lowest valid value: 58

Highest valid value: 97





DESCRIPTIVE STATISTICS OF THE TIME SERIES

AUTOCORRELATION

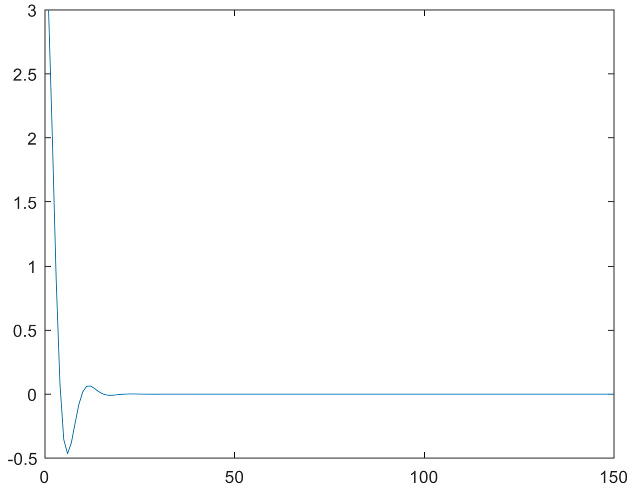
Autocorrelation measures the degree of similarity between a given time series and a lagged version of itself over successive time intervals. it measures the linear relationship between a variable's current value and its past values. It is conceptually similar to the correlation between two different time series, but autocorrelation uses the same time series twice: once in its original form and once lagged one or more time periods.

There are several autocorrelation coefficients, corresponding to each lag in the time. For example, r_1 measures the relationship between t_i and t_{i-1}, r_2 measures the relationship between t_i and t_{i-2}, and so on. The value of r_k can be written as:

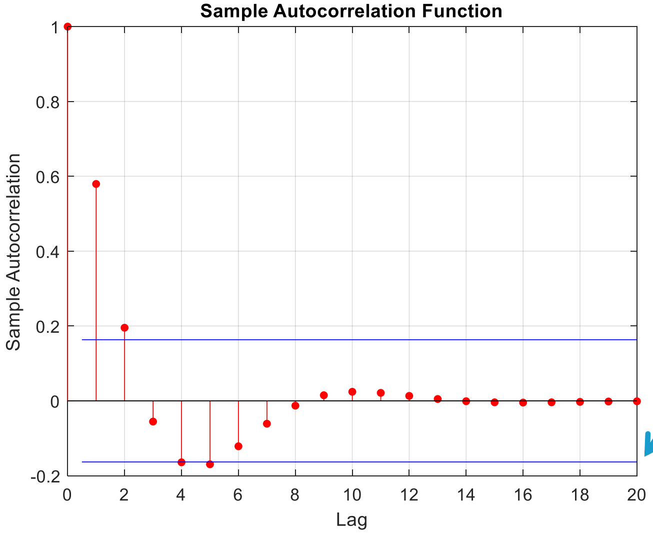
$$r_k(T) = \frac{\sum_{i=k}^{n-1}[t_i - \mu(T)][t_{i-k} - \mu(T)]}{\sum_{i=0}^{n-1}[t_i - \mu(T)]^2}$$

k is the number of lags.

$Y[0] = 3$
$Y[1] = 2$
$Y[i] = 1.2Y[i-1] - 0.5Y[i-2]$



autocorrelation



statistical thresholds

DESCRIPTIVE STATISTICS OF THE TIME SERIES

AUTOCORRELATION

Example: Let T be a time series with the values [1, -1, 1, -1, 1, -1, 1, -1]. Compute the autocorrelation of T.

$$\mu = 0 \qquad \sum_{i=0}^{n-1} [t_i - \mu(T)]^2 = 8 \qquad r_k(T) = \frac{\sum_{i=k}^{n-1} [t_i - \mu(T)][t_{i-k} - \mu(T)]}{\sum_{i=0}^{n-1} [t_i - \mu(T)]^2} = \frac{\sum_{i=k}^{n-1} t_i t_{i-k}}{8}$$

k = 0

t_i	t_i
1	1
-1	-1
1	1
-1	-1
1	1
-1	-1
1	1
-1	-1

$r_0 = 1$

k = 1

t_{i+1}	t_i
-1	1
1	-1
-1	1
1	-1
-1	1
1	-1
-1	1
-	-1

$r_1 = -7/8$

k = 2

t_{i+2}	t_i
1	1
-1	-1
1	1
-1	-1
1	1
-1	-1
-	1
-	-1

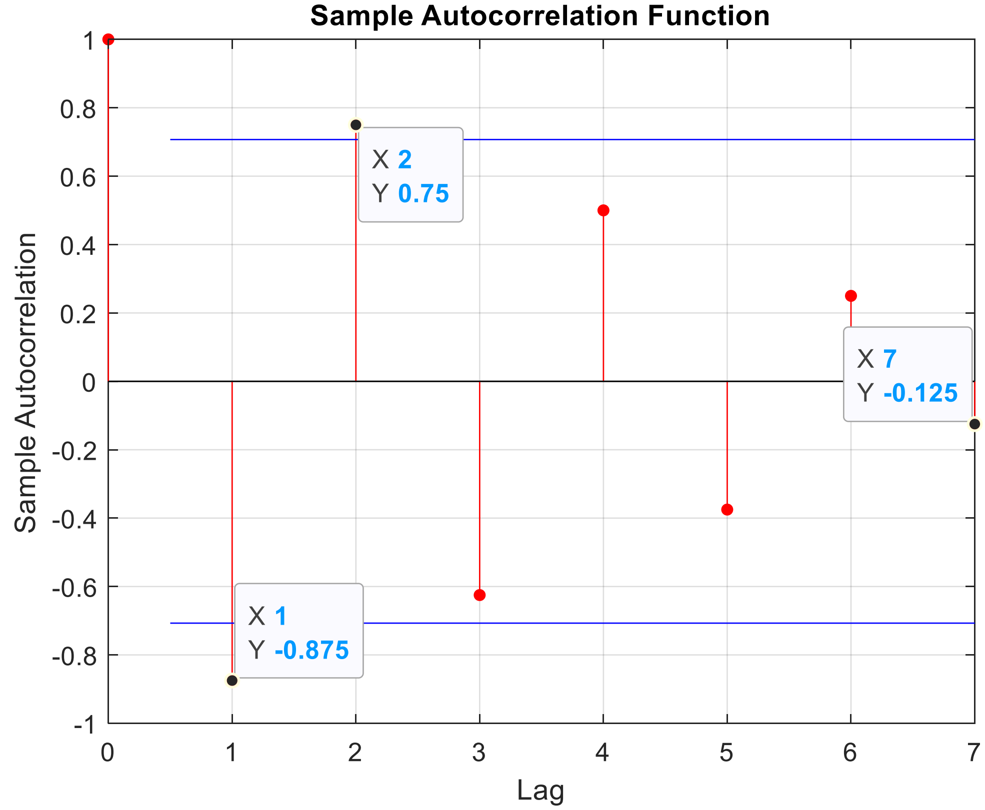
$r_2 = 3/4$

…

k = 7

t_{i+7}	t_i
-1	1
-	-1
-	1
-	-1
-	1
-	-1
-	1
-	-1

$r_7 = -1/8$



DESCRIPTIVE STATISTICS OF THE TIME SERIES

AUTOCORRELATION

The **partial autocorrelation function** (PACF) gives the partial correlation of a stationary time series with its own lagged values, removing the effect of any correlations due to the terms at shorter lags.

Given a time series T, the partial autocorrelation of lag k, denoted θ_k, is the autocorrelation between t_i and t_{i+k} with the linear dependence of t_i on t_{i+1} through t_{i+k-1} removed.

The partial autocorrelation function is similar to the ACF except that it shows only the correlation between two observations that the shorter lags between those observations do not explain. For example, the partial autocorrelation for lag 3 is only the correlation that lags 1 and 2 do not explain.

Partial autocorrelation is a commonly used tool for identifying the order of an autoregressive (AR) model. The partial autocorrelation of an AR(p) process is zero at lags greater than p.

DESCRIPTIVE STATISTICS OF THE TIME SERIES

AUTOCORRELATION

Example: Let T be a time series with the values [3, 2, 0.9, 0.1, -0.4, -0.5, -0.4, -0.2, -0.1, 0]. Compute the partial autocorrelation of T.

For $k = 0$: $\rightarrow \theta_0 = 1$ (lag zero)

$k = 1$: minimize the mean squared error between t_{i+1} and $\widehat{t_{i+1}}$ (using ordinary least squares (OLS)), where $\widehat{t_{i+1}} = \beta_0 + \theta_1 t_i \rightarrow \theta_1 = \boxed{0.6238}$

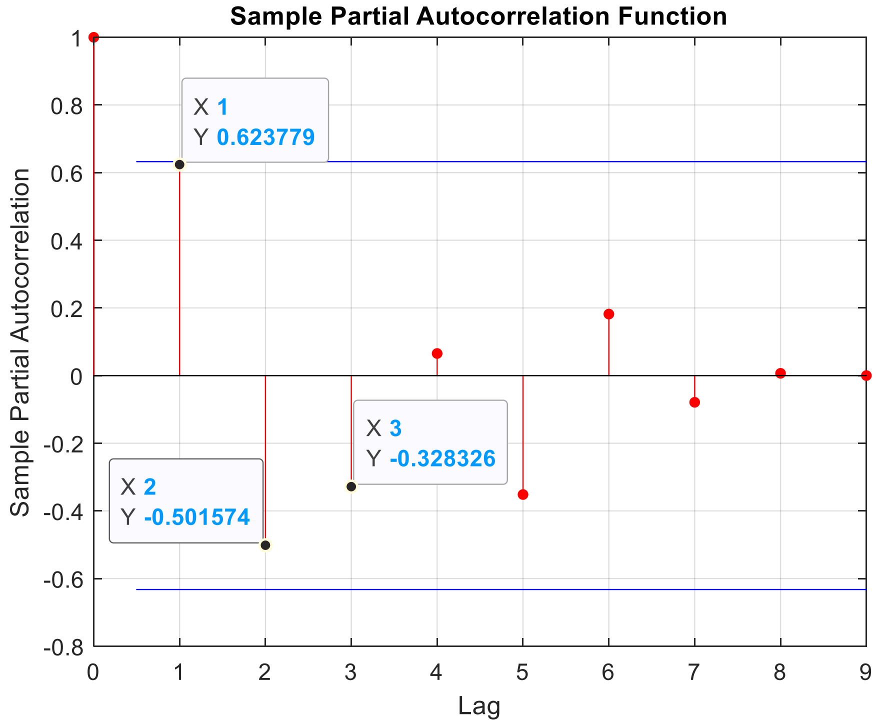
$k = 2$: minimize the mean squared error between t_{i+2} and $\widehat{t_{i+2}}$, where $\widehat{t_{i+2}} = \beta_0 + \beta_1 t_{i+1} + \theta_2 t_i \rightarrow \theta_2 = \boxed{-0.5016}$

$k = 3$: minimize the mean squared error between t_{i+3} and $\widehat{t_{i+3}}$, where $\widehat{t_{i+3}} = \beta_0 + \beta_1 t_{i+2} + \beta_2 t_{i+1} + \theta_3 t_i \rightarrow \theta_3 = \boxed{-0.3283}$

$\vdots$

$$\widehat{t_{i+k}} = [1 \quad t_{i+k-1} \quad \cdots \quad t_i] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \theta_k \end{bmatrix}$$

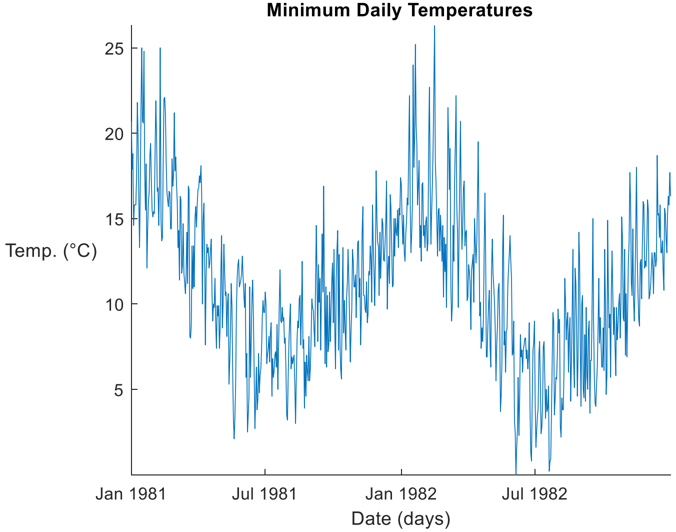
$$\varepsilon = \min[(t_{i+k} - \widehat{t_{i+k}})^2]$$

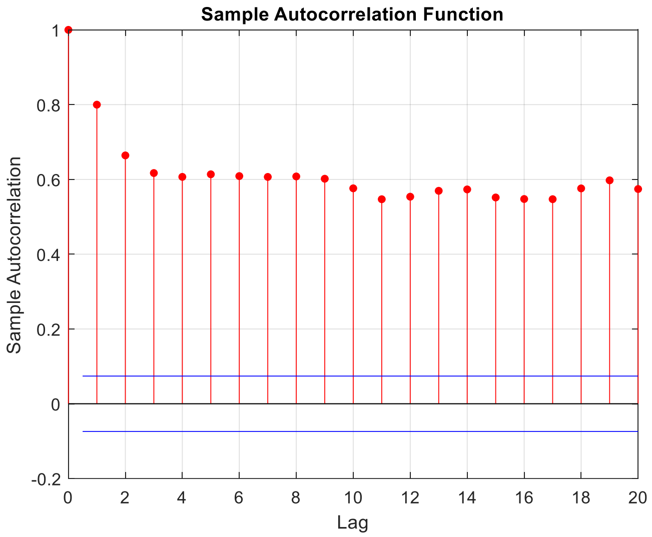


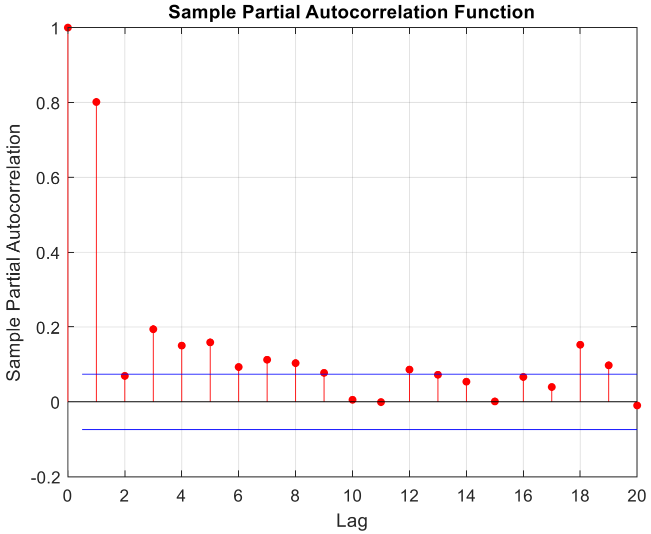
DESCRIPTIVE STATISTICS OF THE TIME SERIES

AUTOCORRELATION

Autocorrelation Function (ACF) vs Partial Autocorrelation Function (PACF):







Both functions play an important role in data analysis aimed at identifying the extent of the lag in time series models.

For moving average MA(q) models, the ACF will be zero for lags greater than q. Thus, the ACF provides a considerable amount of information about the order of the dependence when the process is a moving average process.

If the process, however, is Autoregressive Moving Average (ARMA) or Autoregressive (AR), the ACF alone tells us little about the orders of dependence. PACF can help to determine the appropriate lags p in an AR(p) model or in an extended ARMA(p,q) model.

DESCRIPTIVE STATISTICS OF THE TIME SERIES

CROSS-CORRELATIONS - MULTIVARIATE

For a univariate time series, the autocorrelations summarize the linear time dependence in the data. With a multivariate time series each component has autocorrelations but there are also cross lead-lag correlations between all possible pairs of components. The cross lag k correlations between the variables X and Y are defined as:

$$p_{XY}(k) = corr(x_i, y_{i-k}) = \frac{\sum_{i=k}^{n-1}[x_i - \mu(X)][y_{i-k} - \mu(Y)]}{\sqrt{\sum_{i=0}^{n-1}[x_i - \mu(X)]^2}\sqrt{\sum_{i=0}^{n-1}[y_i - \mu(Y)]^2}}$$

They are not necessarily symmetric in k, in general:

$$p_{XY}(k) = corr(x_i, y_{i-k}) \neq corr(y_i, x_{i-k}) = p_{YX}(k)$$

If $p_{XY}(k) \neq 0$ for some $k > 0$ then Y is said to lead X. This implies that past values of Y are useful for predicting future values of X.

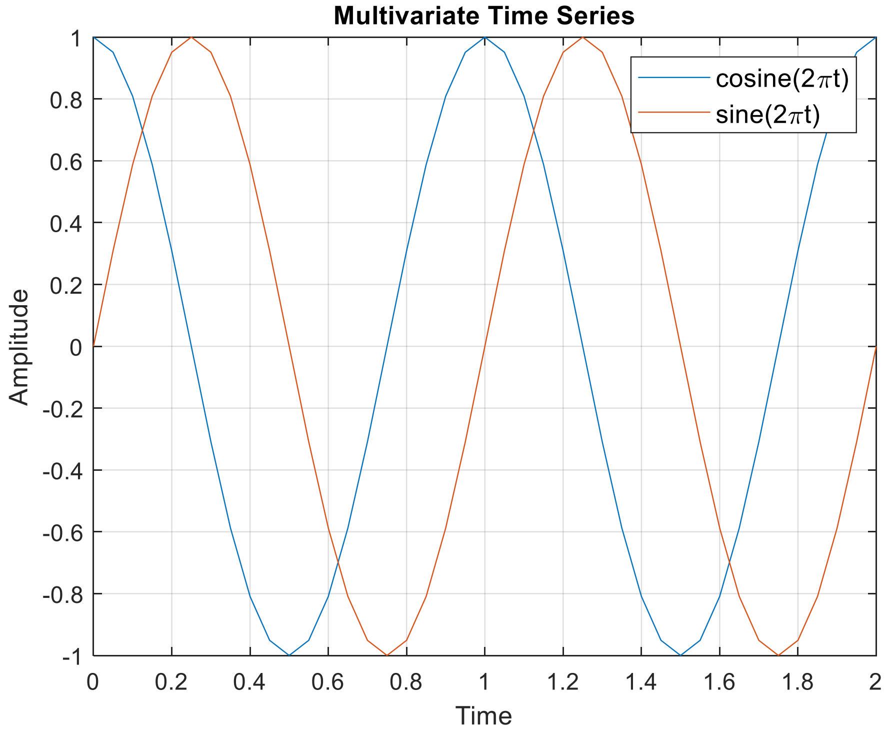
Similarly, if $p_{YX}(k) \neq 0$ for some $k > 0$ then X is said to lead Y. It is possible that X leads Y and vice-versa. In this case, there is said to be dynamic feedback between the two series.

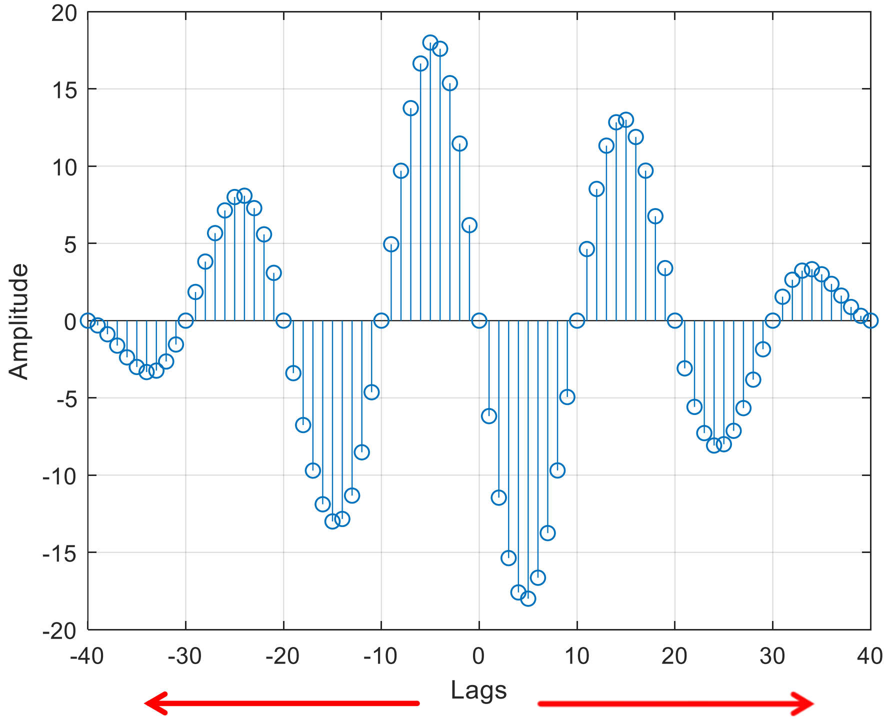
DESCRIPTIVE STATISTICS OF THE TIME SERIES

CROSS-CORRELATIONS - MULTIVARIATE

Example: Calculate the cross-correlation between variables $X = cos(2\pi t)$ and $Y = \sin(2\pi t)$.

$$p_{XY}(k) = corr(x_i, y_{i-k})$$





$k < 0$: past values of X and current values of Y

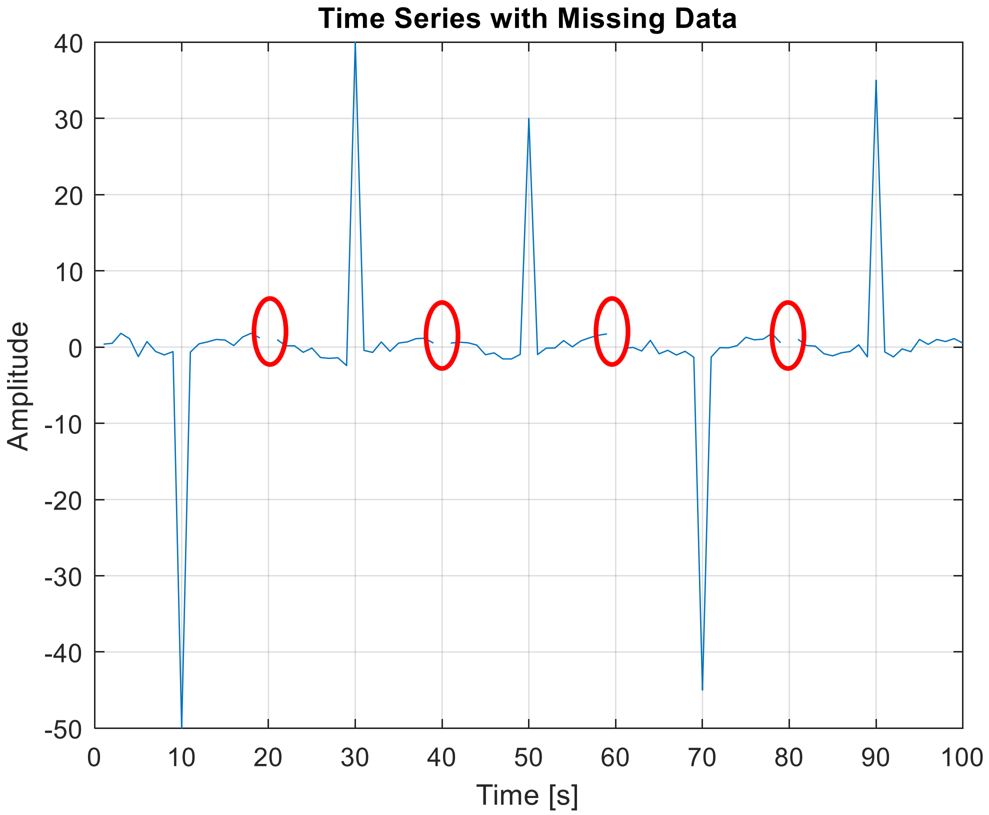
$k > 0$: past values of Y and current values of X

TYPES OF ANOMALIES IN TIME SERIES

TYPES OF ANOMALIES IN TIME SERIES

MISSING DATA

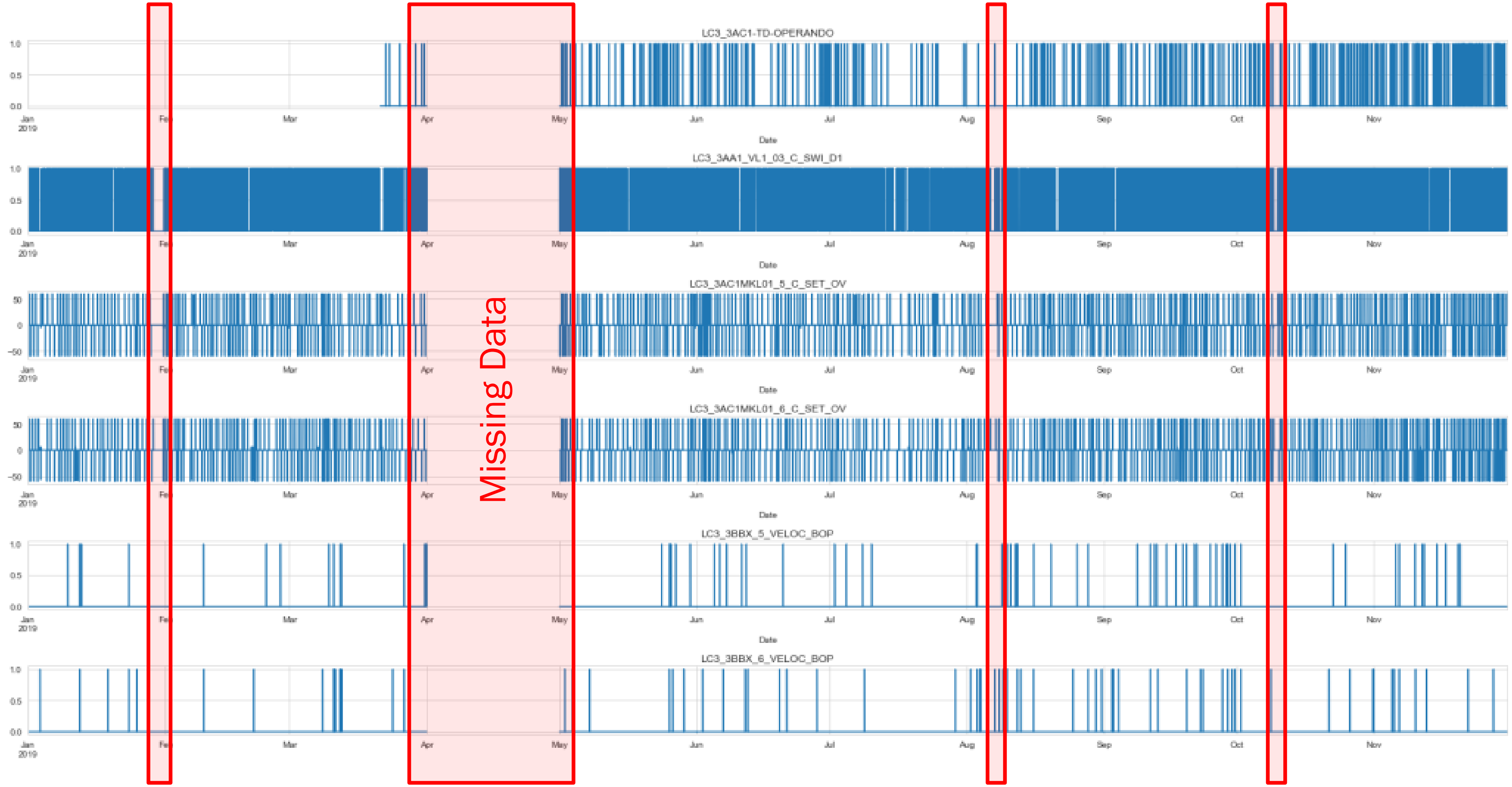
Missing data, or missing values, occur when no data value is available for the variable in an observation. Although sometimes missing values signify a meaningful event in the data, they often represent unreliable or unusable data points.



TYPES OF ANOMALIES IN TIME SERIES

MISSING DATA

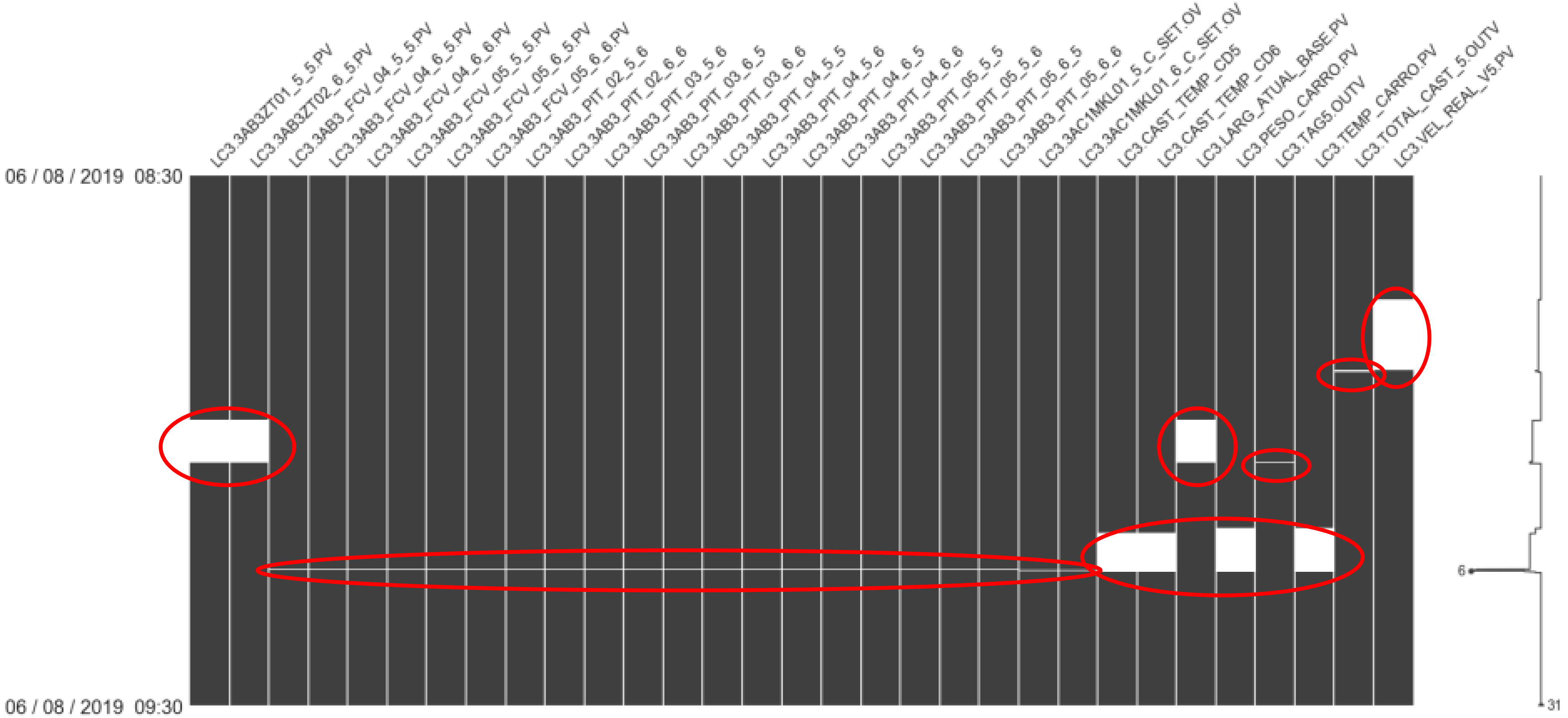
Missing data is a problem that occurs frequently in real datasets. They are typically represented by NaN (not a number) or values outside the scope of the variable.



TYPES OF ANOMALIES IN TIME SERIES

MISSING DATA

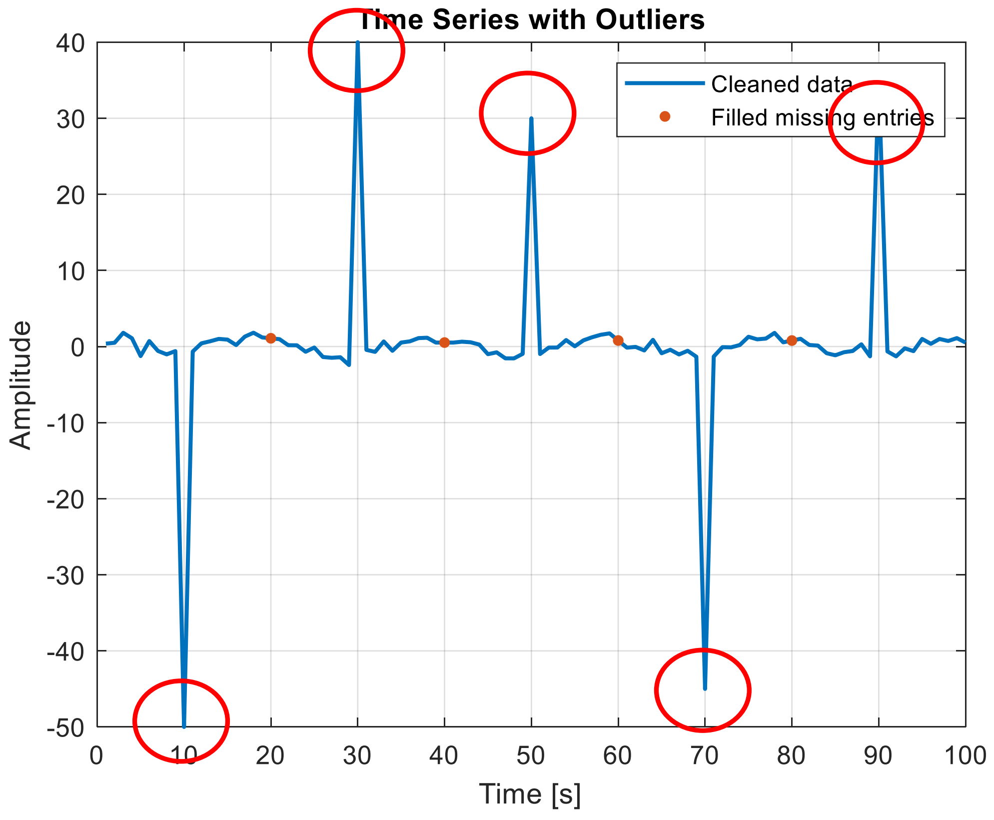
Understanding the reasons why data are missing is important for handling the remaining data correctly. If values are missing completely at random, the data sample is likely still representative of the population. But if the values are missing systematically, analysis may be biased. For example: a sensor can have problems.

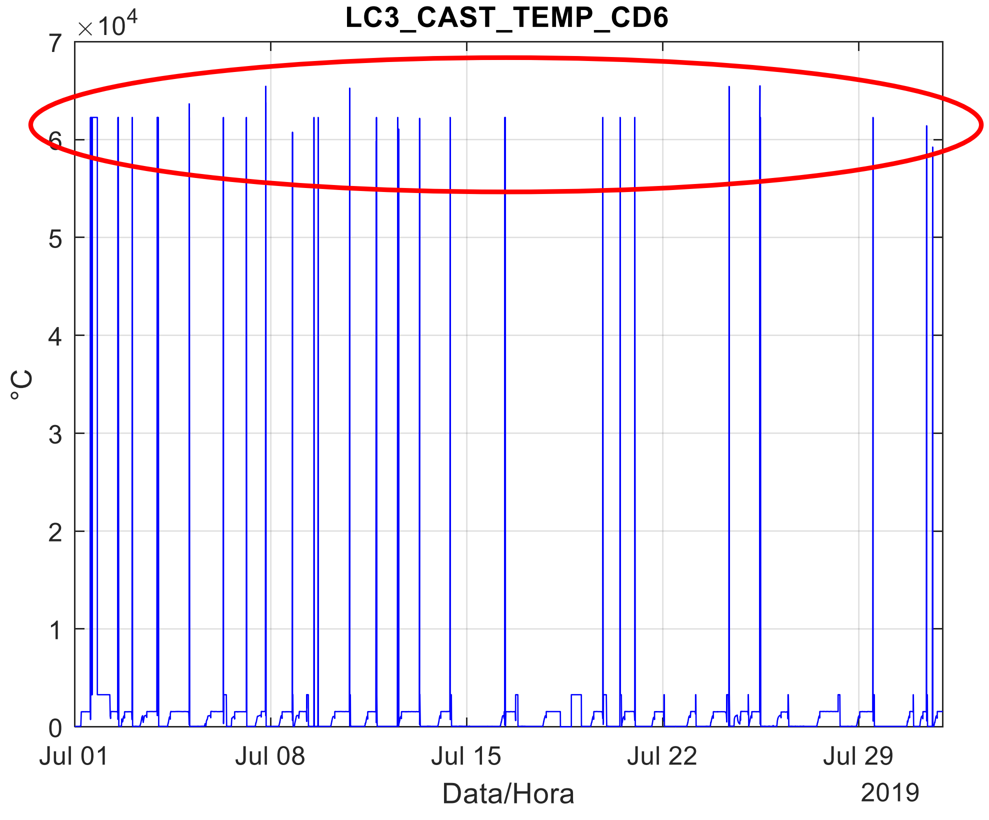


TYPES OF ANOMALIES IN TIME SERIES

OUTLIER

Outlier: is an observation that deviates greatly from the others in the time series or is inconsistent.

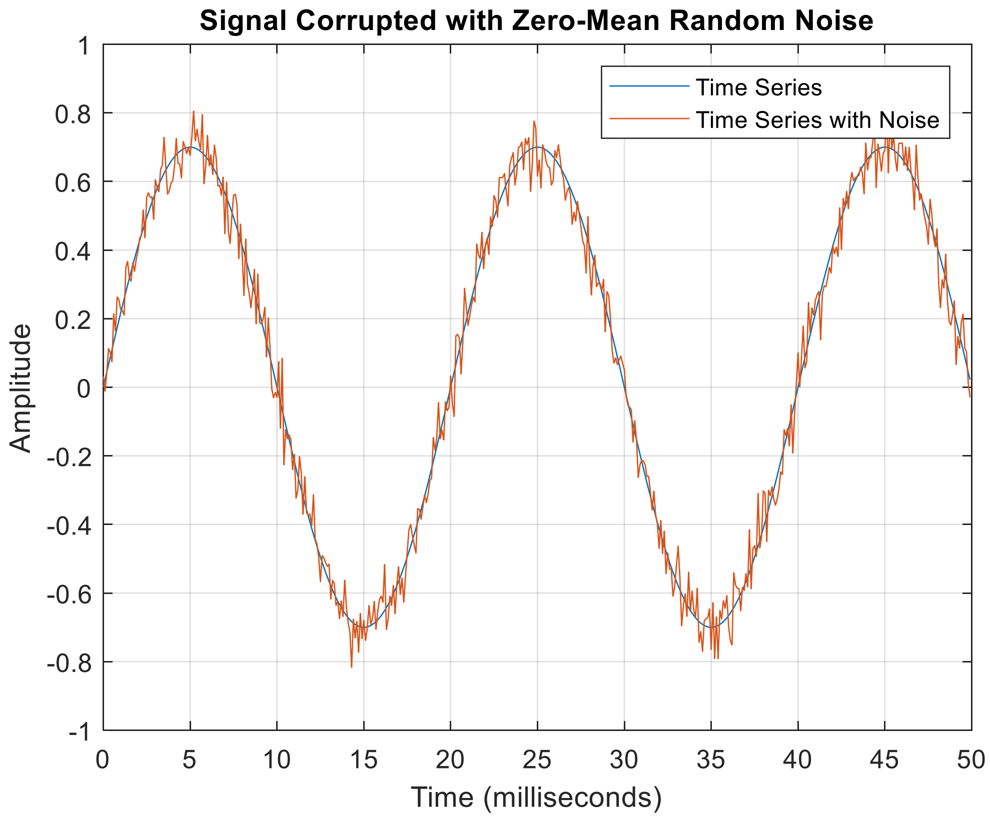


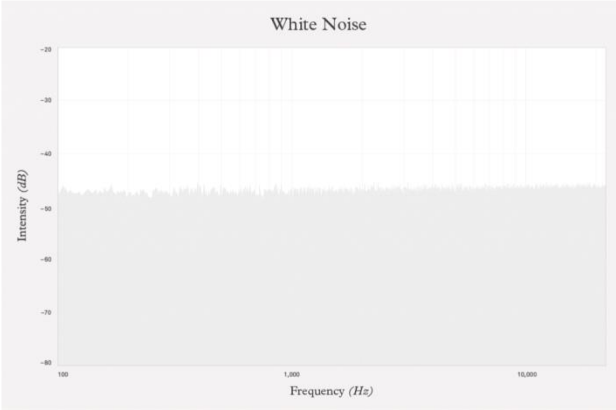


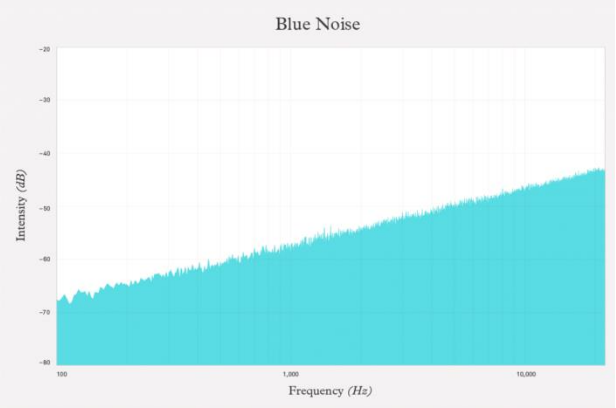
TYPES OF ANOMALIES IN TIME SERIES

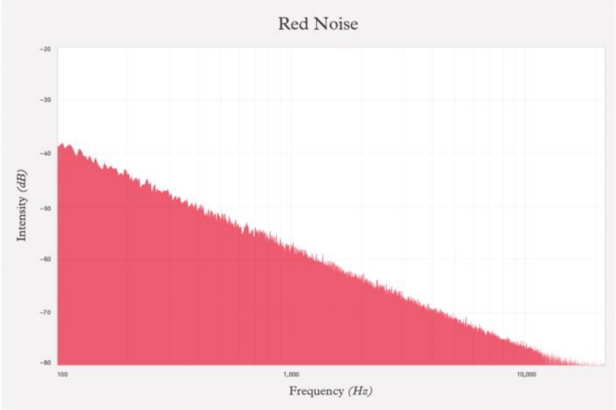
NOISE

Noise is any deviation from the true value of samples in a time series. Almost all datasets will contain a certain amount of unwanted noise. White noise is a random signal having equal intensity at different frequencies, giving it a constant power spectral density. Colored noise is a random signal that is more prominent at specific frequencies. Noise can be characterized by its probability density function.









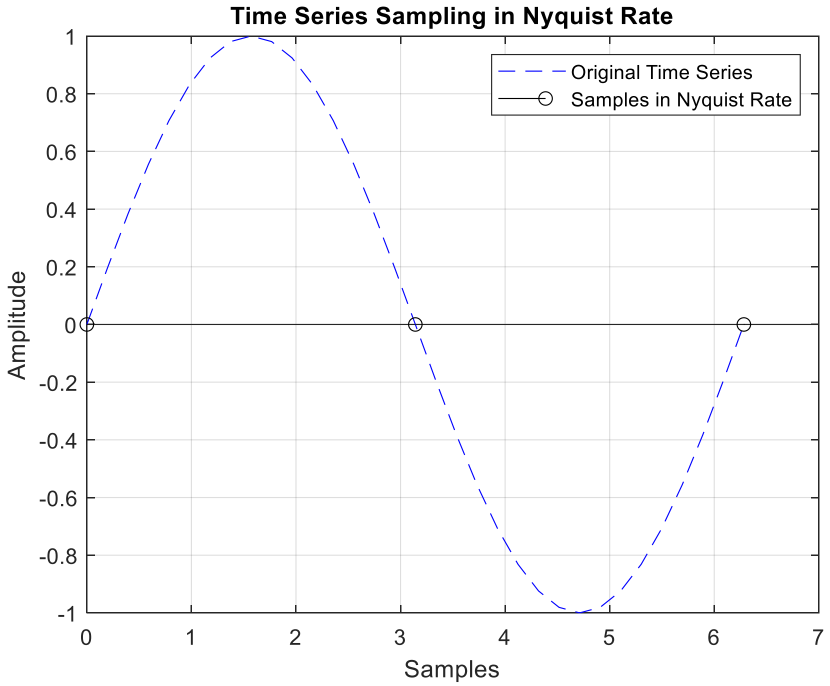
TYPES OF ANOMALIES IN TIME SERIES

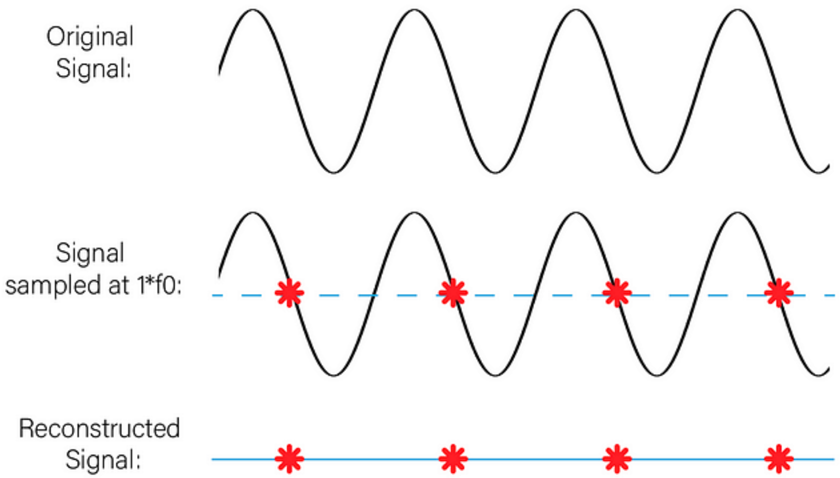
WRONG SAMPLING RATE

Wrong Sampling Rate: A common anomaly in time series is the acquisition of samples with insufficient sampling rate.

The sample rate (f_s) must be greater than twice the highest frequency component (f_0) of interest in the measured signal, i.e., $f_s > 2f_0$. This frequency, f_0, is often referred to as the Nyquist frequency.

Even if the sampling rate satisfies Nyquist frequency ($f_s = 2f_0$) are not guarantees for a good understanding of the dependence between samples of a time series.

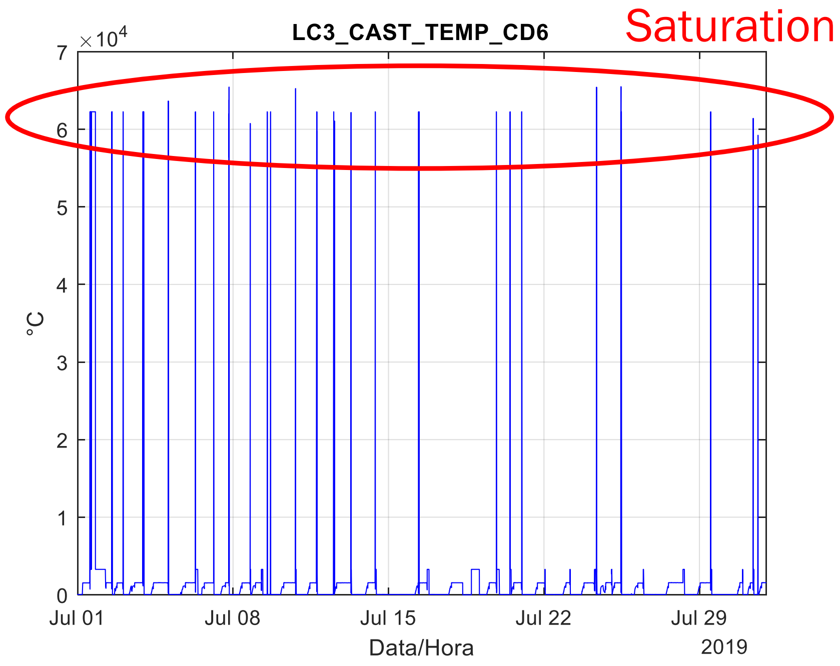


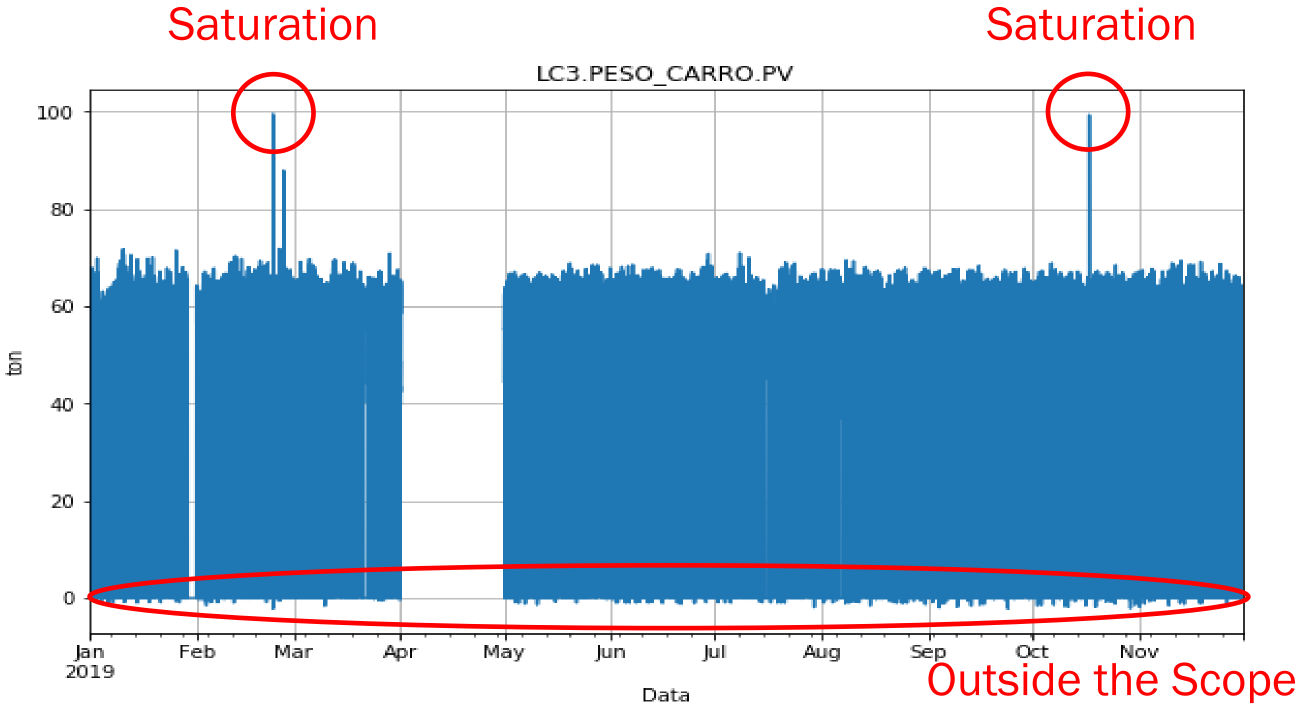


TYPES OF ANOMALIES IN TIME SERIES

SENSOR ERROR

Sensor Error: Sensor errors can result in incorrect values. The values can be noise or can be scaled outside the scope of the variable, they can be null (no reading) or extremely high (saturation), or they can be wrong values (calibration) difficult to detect when compared to the time series data itself.





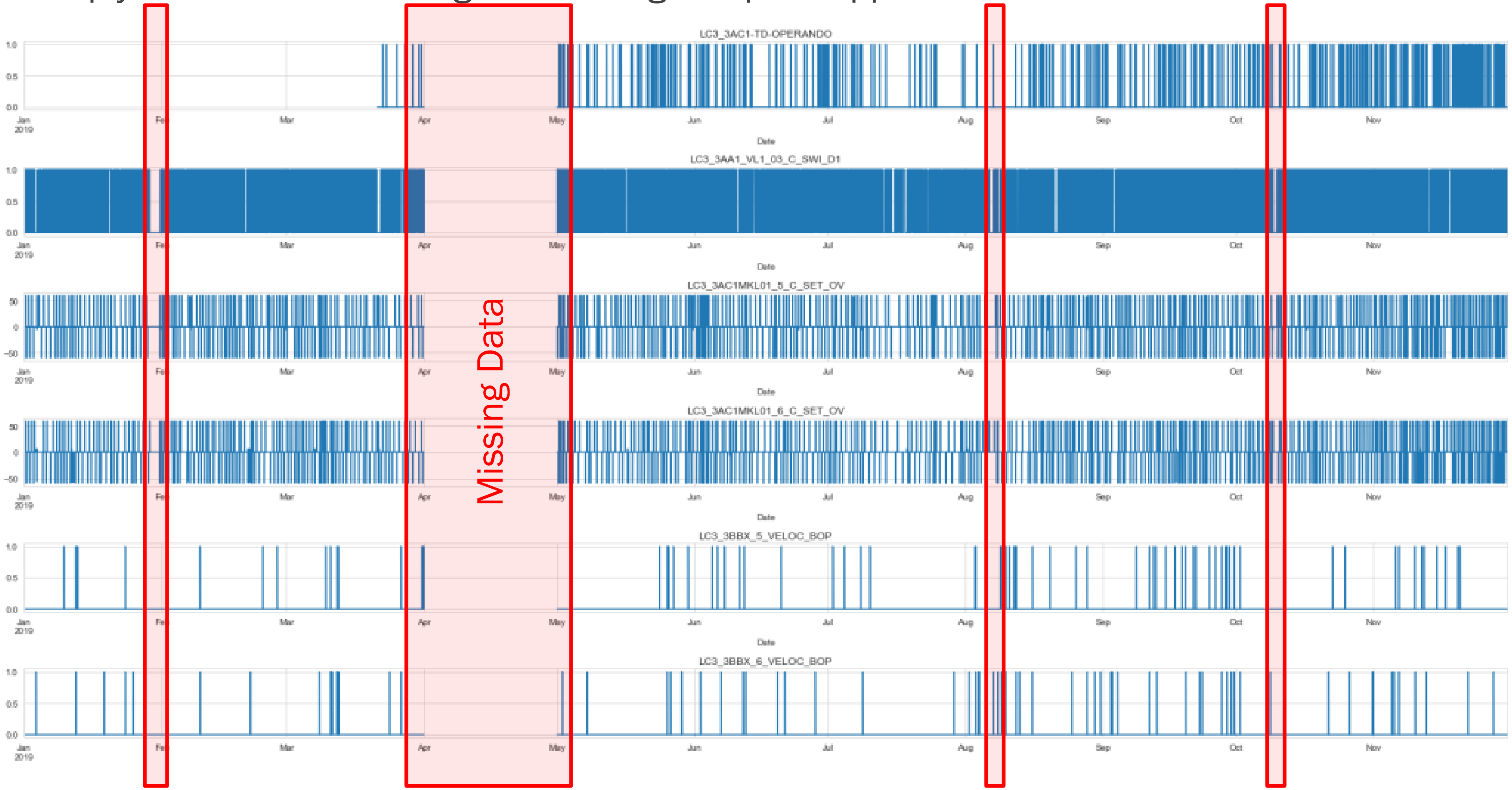
MATHEMATICAL TREATMENTS IN TIME SERIES

MATHEMATICAL TREATMENTS IN TIME SERIES

MISSING DATA

There are different ways of handling missing data from a time series:

Removal: it simply consists of removing the missing sample snippets.

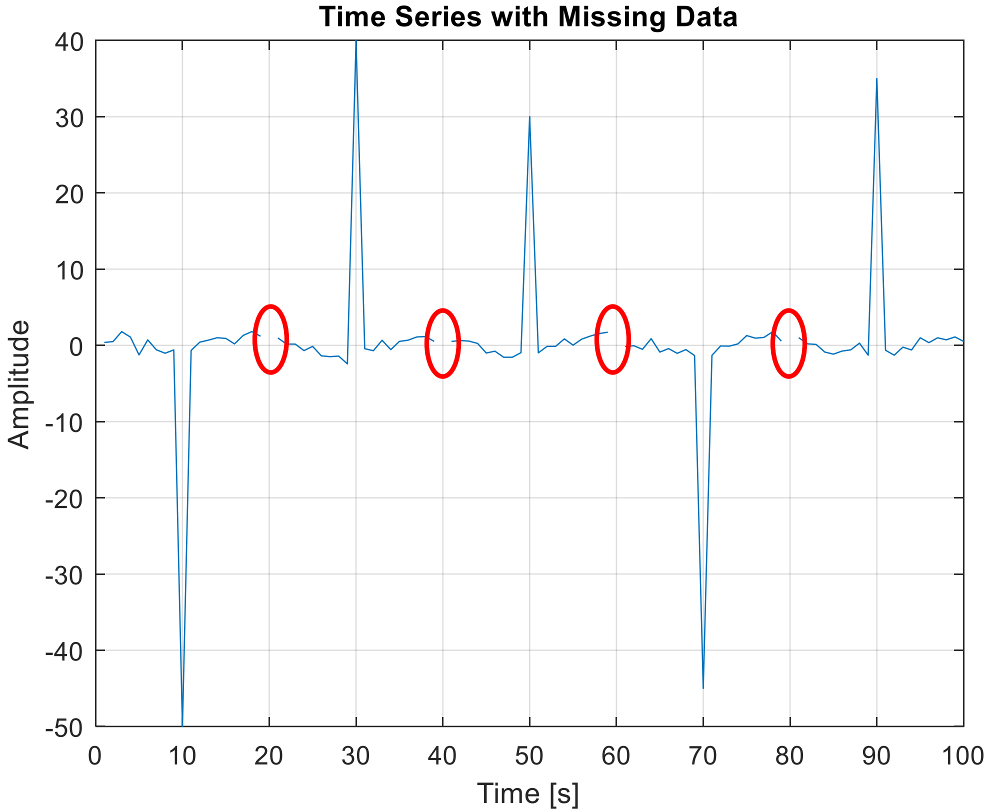


MATHEMATICAL TREATMENTS IN TIME SERIES

MISSING DATA

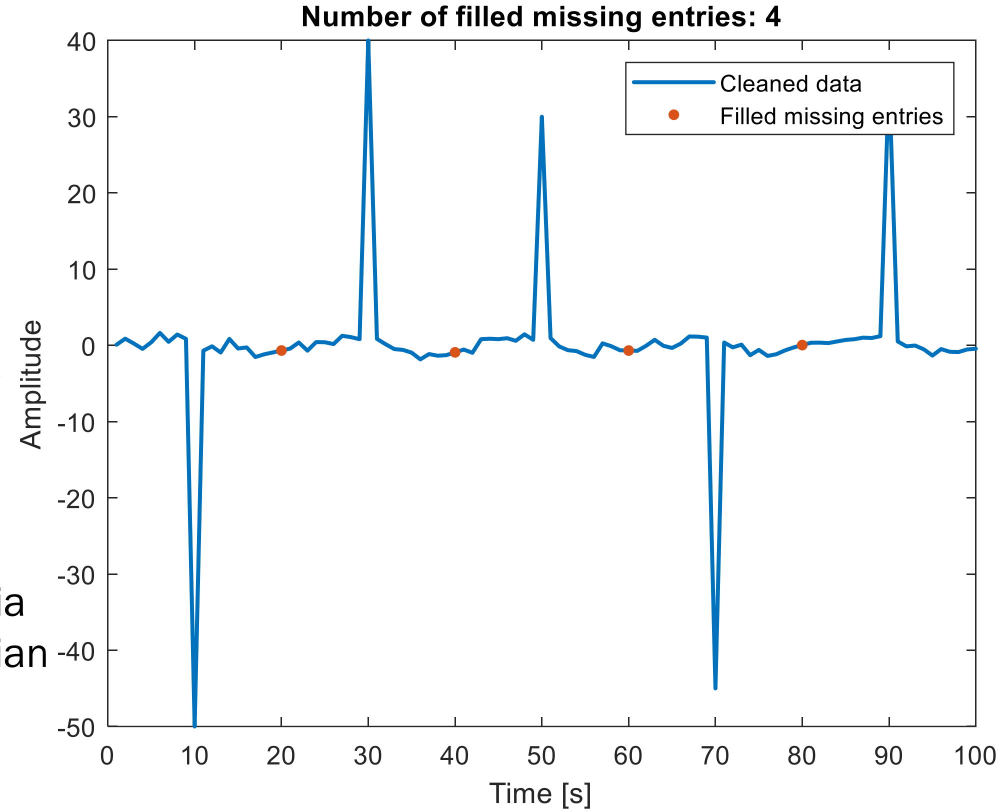
There are different ways of handling missing data from a time series:

Interpolation: replaces missing data with samples from an interpolation function.



Linear
Spline
Cubic
Akima
Pchip

Nearest
Previous
Next
MovMedia
MovMedian



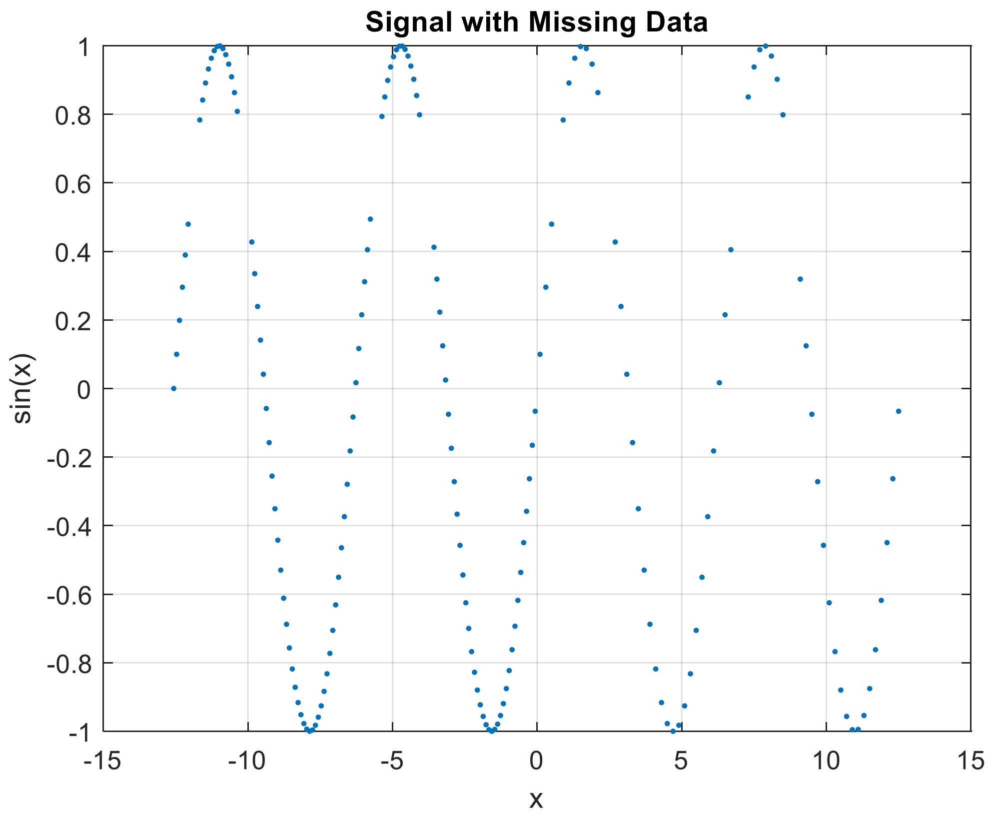
MATHEMATICAL TREATMENTS IN TIME SERIES

MISSING DATA

Matlab Ex1Cap2

There are different ways of handling missing data from a time series:

Interpolation: replaces missing data with samples from an interpolation function.



Linear
Spline
Cubic
Akima
Pchip

Nearest
Previous
Next
MovMedia
MovMedian

MATHEMATICAL TREATMENTS IN TIME SERIES

MISSING DATA

Matlab Ex2Cap2

There are different ways of handling missing data from a time series:

PCA with ALS: Finds principal components using the Alternating Least Squares (ALS) algorithm when there are missing values in the data. The result is used to estimate missing data.

ALS can work well for data sets with a small percentage of missing data at random but might not perform well on sparse data sets.

```
ingredients =
```

7	26	6	60
1	29	15	52
11	56	8	20
11	31	8	47
7	52	6	33
11	55	9	22
3	71	17	6
1	31	22	44
2	54	18	22
21	47	4	26
1	40	23	34
11	66	9	12
10	68	8	12

Including Missing Data ≈30%

7	26	6	NaN
1	29	15	52
NaN	NaN	8	20
11	31	NaN	47
7	52	6	33
NaN	55	NaN	NaN
NaN	71	NaN	6
1	31	NaN	44
2	NaN	NaN	22
21	47	4	26
NaN	40	23	34
11	66	9	NaN
10	68	8	12

Restoring Missing Data with PCA

7.0000	26.0000	6.0000	51.5250
1.0000	29.0000	15.0000	52.0000
10.7819	53.0230	8.0000	20.0000
11.0000	31.0000	13.5500	47.0000
7.0000	52.0000	6.0000	33.0000
10.4818	55.0000	7.8328	17.9362
3.0982	71.0000	11.9491	6.0000
1.0000	31.0000	-0.5161	44.0000
2.0000	53.7914	5.7710	22.0000
21.0000	47.0000	4.0000	26.0000
21.5809	40.0000	23.0000	34.0000
11.0000	66.0000	9.0000	5.7078
10.0000	68.0000	8.0000	12.0000

MATHEMATICAL TREATMENTS IN TIME SERIES

OUTLIER

A common practice for detecting outliers is identifying elements whose values are outside an interval spanning over the mean (μ) plus/minus three standard deviations (σ). For a random variable vector, A, with n scalar observations, a_i, $i = 0,1, \dots, n-1$, is an outlier if:

$$\left|\frac{a_i - \mu}{\sigma}\right| > 3$$

Unfortunately, three problems can be identified when using this approach: 1) it assumes that the distribution is normal; 2) the mean and standard deviation are strongly impacted by outliers; 3) this method is very unlikely to detect outliers in small samples (when n is small).

An alternative is to use the median absolute deviation (MAD) [1]: an outlier is a value that is more than three scaled Median Absolute Deviation (MAD) away from the median. The scaled MAD (sMAD) is defined as:

$$sMAD(A) = b \times \underset{i=0,\dots,n-1}{median}(|a_i - median(A)|)$$

where $b = 1.4826$, when it is assumed normality of the data. If another underlying distribution is assumed, this value changes to $b = 1/Q(0.75)$, where $Q(0.75)$ is the 0.75 quantile of that underlying distribution. Three (3) sMAD away from the median is very conservative, while two (2) is poorly conservative.

[1] Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766, 2013.

MATHEMATICAL TREATMENTS IN TIME SERIES

OUTLIER

Example: consider a small set of $n = 8$ observations with values 1, 3, 3, 6, 8, 10, 10, and 1000. Let's identify the outliers in this set.

First method: mean and standard deviation:

$$\mu(A) = \frac{1}{n}\sum_{i=0}^{n-1} a_i = 130.125 \qquad \sigma(A) = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n-1}(a_i - \mu)^2} = 351.4986$$

$$\mu - 3\sigma \leq \text{inliers} \leq \mu + 3\sigma$$
$$-924.3708 \leq \text{inliers} \leq 1184.6208$$

Thus, no observation is an outlier.

Second method: median and scaled Median Absolute Deviation:

To calculate the median, observations have to be sorted in ascending order to identify the mean rank and to determine the value associated with that rank.

$$mean\_rank = \frac{n+1}{2} = 4.5$$

The median is the mean of 4th and 5th observations: $median = \frac{6+8}{2} = 7$

$$sMAD(A) = 1.4826 \times \underset{i=0,\ldots,n-1}{median}(|a_i - median(A)|) = 5.1891$$

$$median - 3sMAD \leq \text{inliers} \leq median + 3sMAD$$
$$-8.5672 \leq \text{inliers} \leq 22.5672$$

Therefore, 1000 is considered as an outlier.

[1] Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766, 2013.

MATHEMATICAL TREATMENTS IN TIME SERIES

OUTLIER

Other methods for detecting outliers:

- **Moving methods**: These methods are based on a sliding window on the data. They can be combined with the two previous approaches, but, in this case, the measurements are computed locally.

- **Grubbs' Test**: This method assumes that the data are normally distributed and is based on statistical test. The statistic test corresponds to a p-value that represents the likelihood of existing outliers, assuming the data have a Gaussian distribution. Typical significance levels (α) are 0.05 and 0.01. If the p-value is smaller than α then at least one outlier is present in the data. This method gives a general answer to the question "Is there at least one outlier in this data?".

- **ROUT method**: This method was developed to identify outliers from nonlinear regression. Basically, it has three steps: 1) Fit a curve using a robust nonlinear regression method; 2) Analyze the residuals of the robust fit, and determine whether one or more values are outliers; 3) Remove the outliers and obtain again a regression model.

MATHEMATICAL TREATMENTS IN TIME SERIES

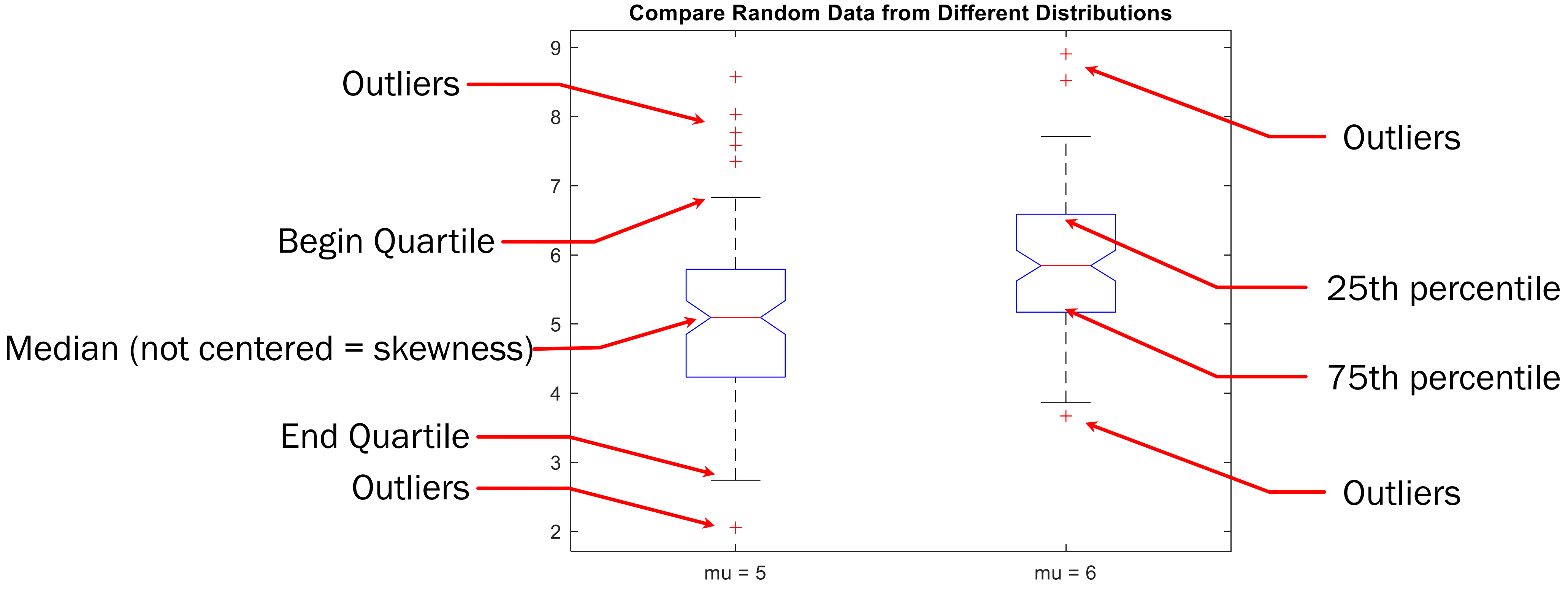
OUTLIER

Matlab Ex3Cap2

Other methods for detecting outliers:

- **Quartiles:** Returns true for elements more than 1.5 interquartile ranges above the upper quartile or below the lower quartile. This method is useful when the data in A is not normally distributed.

BoxPlot



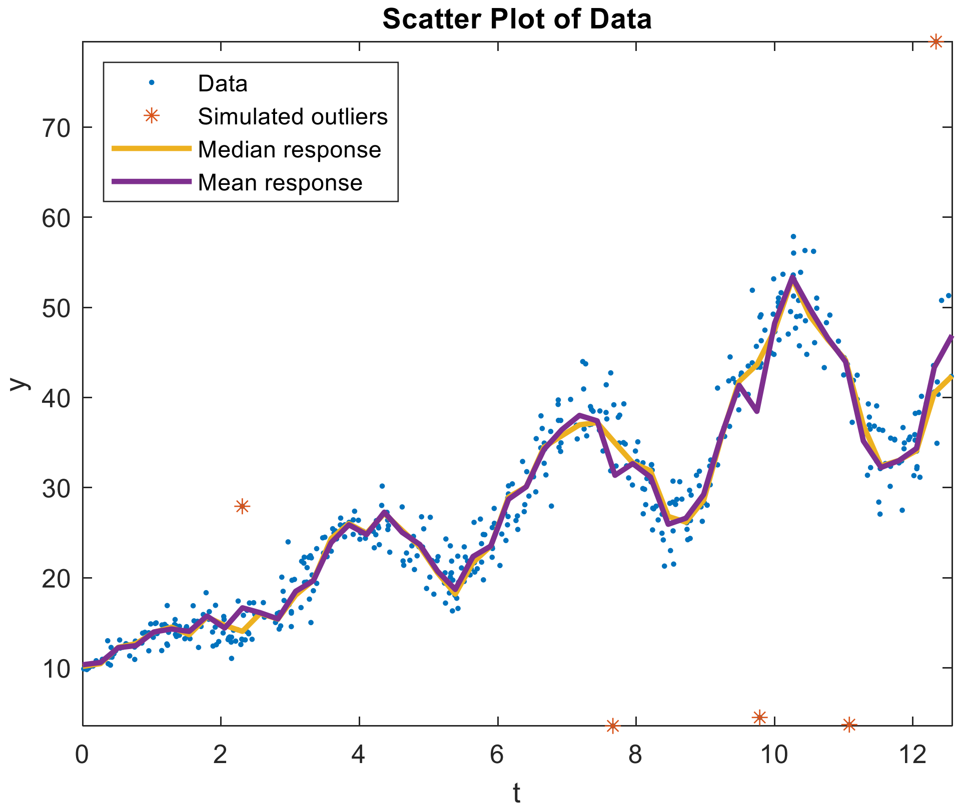
MATHEMATICAL TREATMENTS IN TIME SERIES

OUTLIER

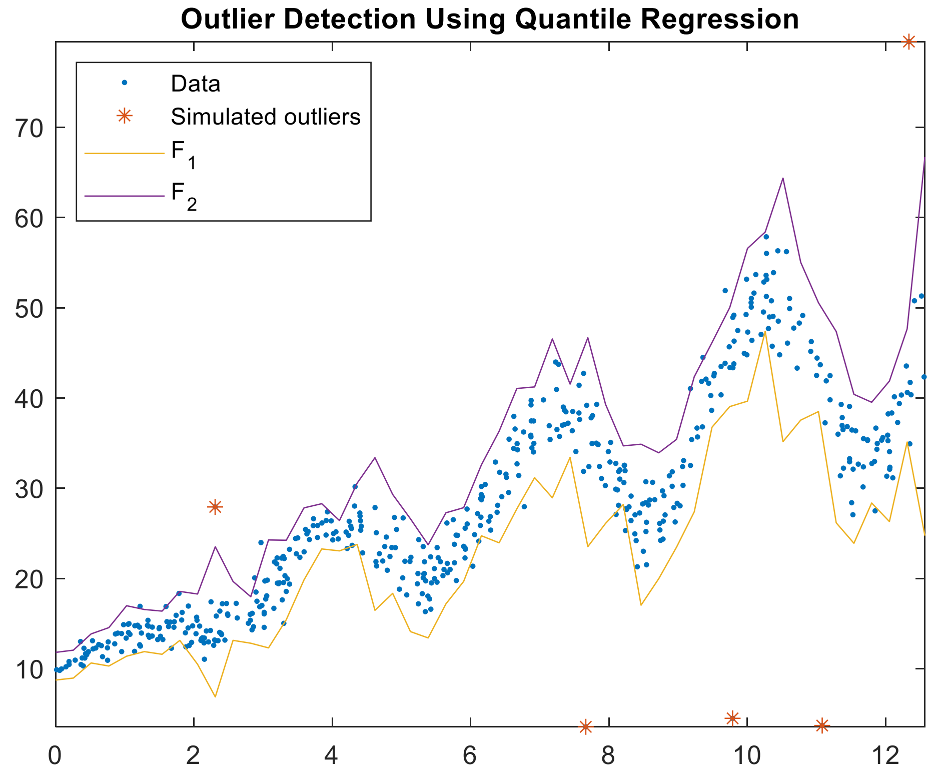
Matlab Ex4Cap2

Other methods for detecting outliers:

- Quartile Regression



$$iqr = 3Q - 1Q$$
$$\mathrm{F}_1 = 1Q - 1.5iqr$$
$$\mathrm{F}_2 = 3Q + 1.5iqr$$



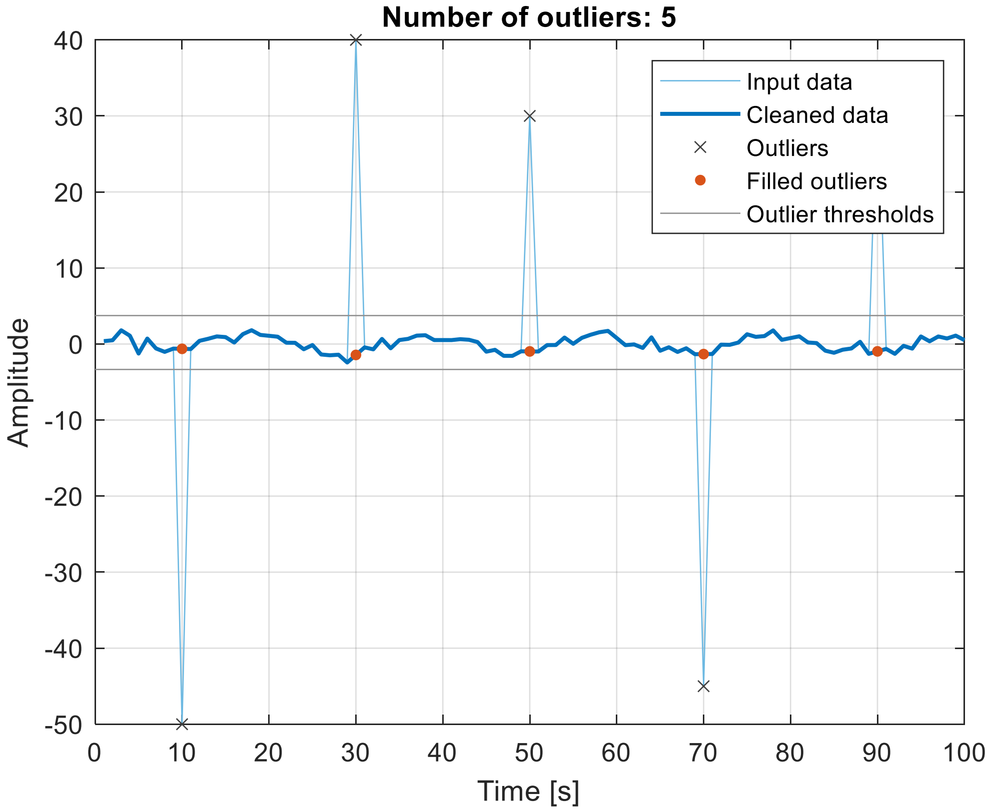
MATHEMATICAL TREATMENTS IN TIME SERIES

OUTLIER

Matlab Ex5Cap2

Fill method for replacing outliers:

- **Clip**: Fills with the lower/upper threshold value for elements smaller/large than the lower/upper threshold determined.
- **Nearest**: Fills with the nearest non-outlier value.
- **Previous**: Fills with the previous non-outlier value.
- **Next**: Fills with the next non-outlier value.
- **Linear**: Fills using linear interpolation of neighboring, non-outlier values.
- **Spline**: Fills using piecewise cubic spline interpolation.
- **Pchip**: Fills using shape-preserving piecewise cubic spline interpolation.
- **Hampel**: Fills using the median of the surrounding values.
- **Makima**: Modified Akima cubic Hermite interpolation.

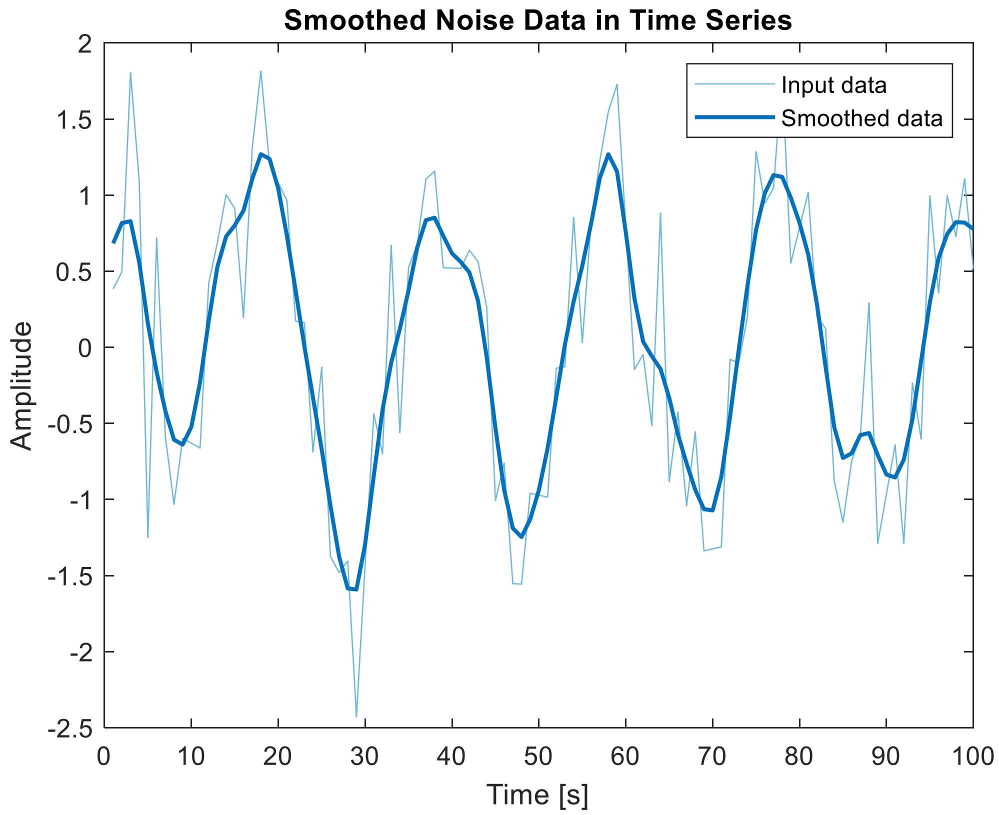


MATHEMATICAL TREATMENTS IN TIME SERIES

NOISE

Matlab Ex6Cap2

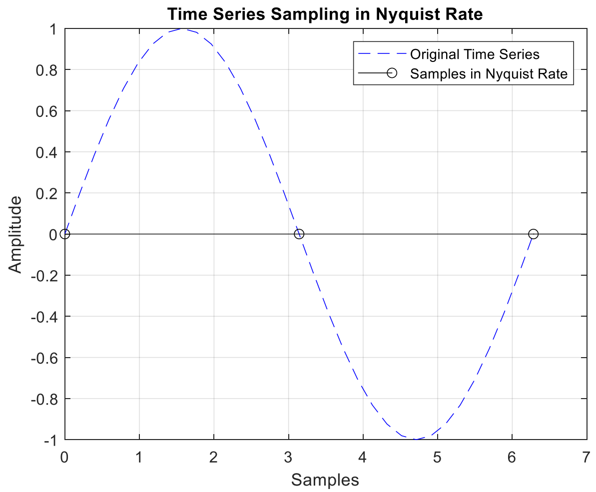
Data smoothing refers to techniques for eliminating unwanted noise or behaviors in data.



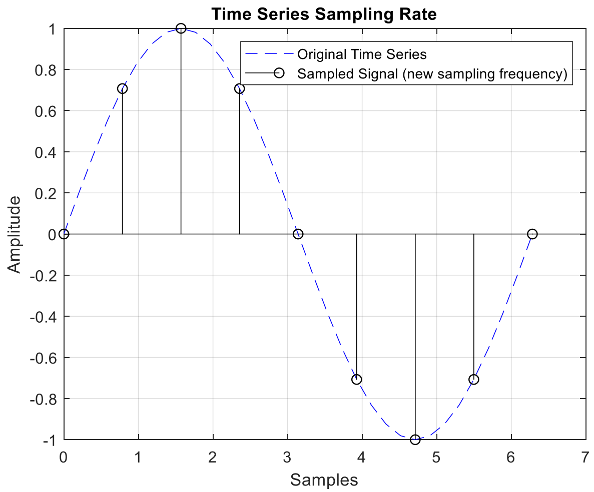
MATHEMATICAL TREATMENTS IN TIME SERIES

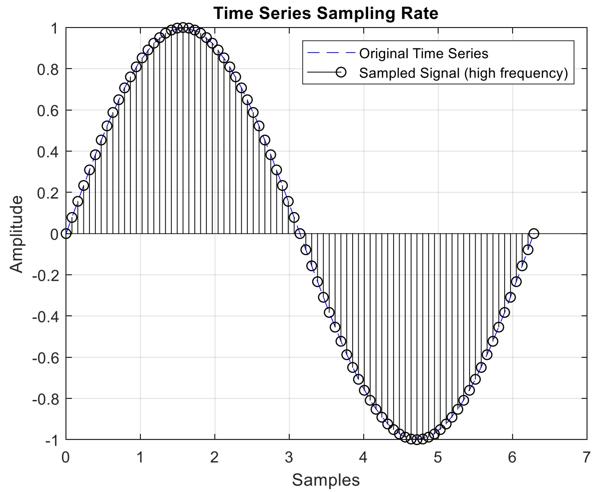
INADEQUATE SAMPLING RATE

Solutions to the sampling rate problem are resampling or re-acquiring data in the process with a more suitable sampling rate.

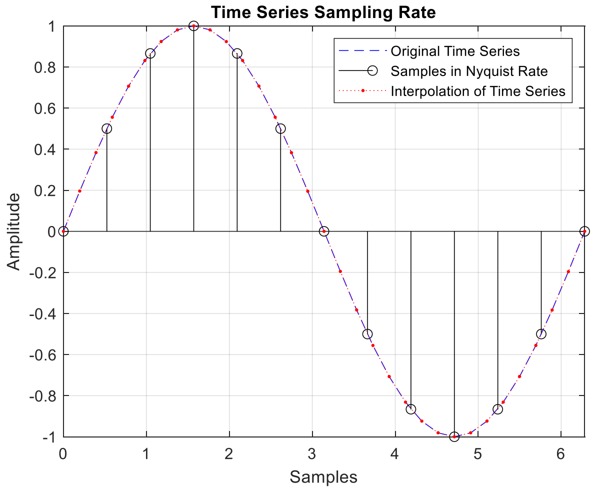


Re-acquiring data





Downsampling or interpolation



MATHEMATICAL TREATMENTS IN TIME SERIES

SENSOR ERROR

Some sensor errors can be handled:

- Noise: use of filters or other smoothing technique;
- Measures outside the scope of the variable: cut in the range of the scope;
- Null data: treat as missing data or outliers;
- Saturation: cut in scope range or outliers;
- Calibration: use of correction factor or re-acquiring data.

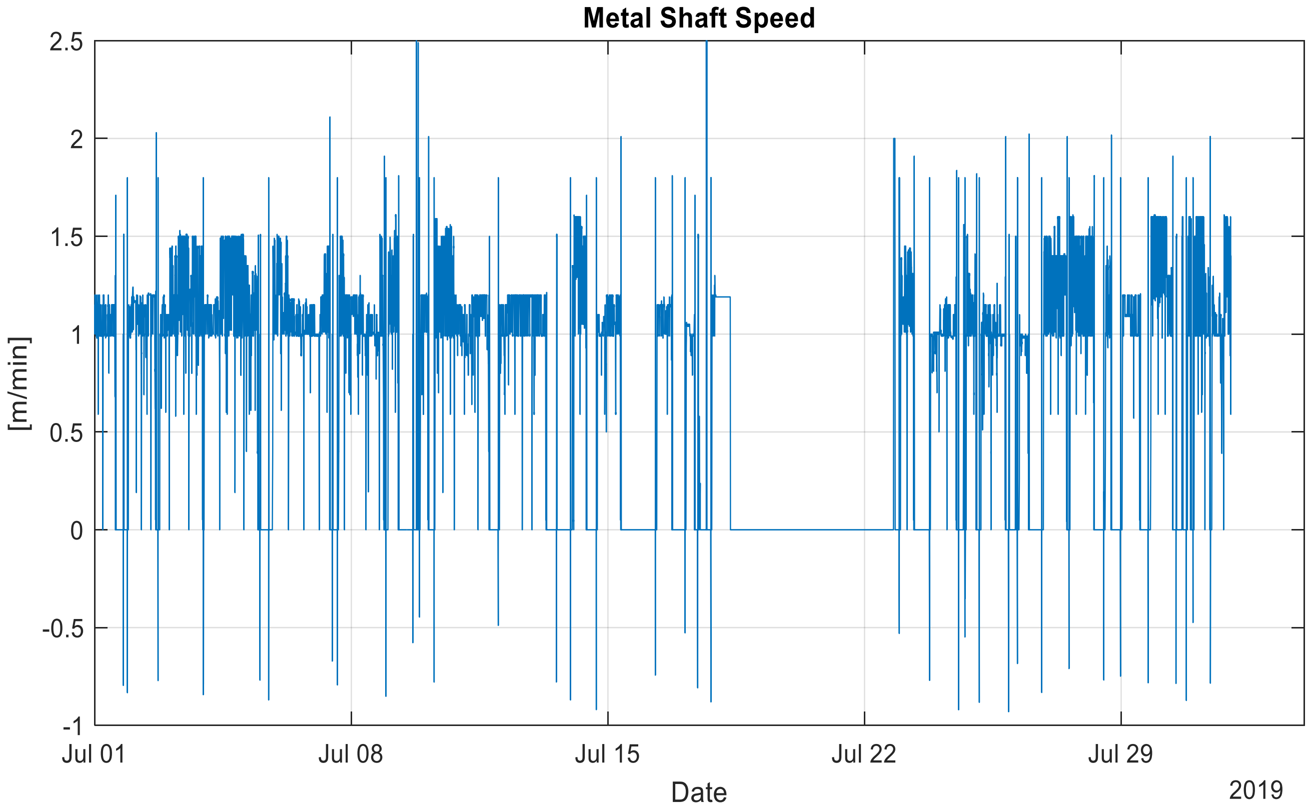
Incorrect values within scope is difficult to detect and treat.

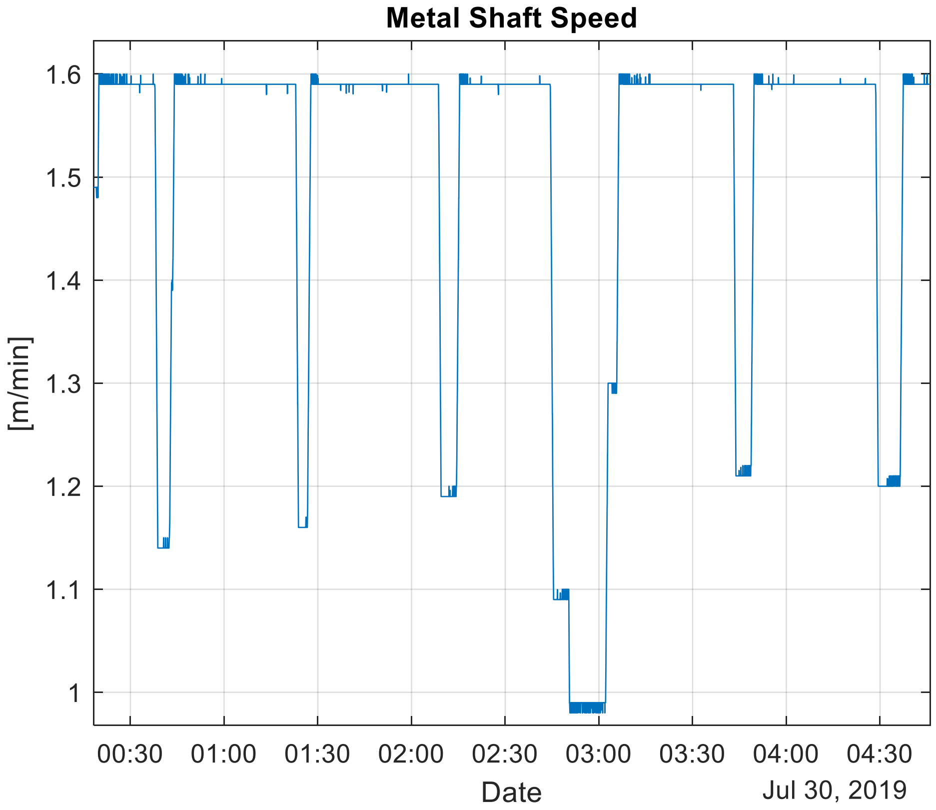
EXAMPLES

EXAMPLES

ANOMALIES

1) The graphs below shows the measurement by the speed sensor of an industrial process. The scope values for the variable are 0 [m/min] and 2.5 [m/min]. Identify issues related to acquired data.





EXAMPLES

ANOMALIES

2) The graphs below shows the measurement by the carrier car weight sensor of an industrial process. The operational range of values for the variable is 30 [ton] to 65 [ton]. Identify issues related to acquired data.