QUALITY DATA ANALYSIS

30/08/2022

General recommendations:

- For exams in presence: to access the software on the provided laptops, go on browser → Favourites → Managed favourites → Virtual Desktop and enter your Polimi credentials.
- write the solutions in CLEAR and READABLE way on paper and show (qualitatively) all the relevant plots;
- avoid (if not required) theoretical introductions or explanations covered during the course;
- always state the assumptions and report all relevant steps/discussion/formulas/expression to present and motivate your solution;
- when using hypothesis tests provide the numerical value of the test statistic and the test conclusion in terms of p-value.
- Exam duration: 2h 10min
- **Multichance students should skip: point b) in Exercise 1, point a) in Exercise 2**

Exercise 1 (15 points)

The concentration of a contaminant (measured in ppm) in the production of synthetic rubber is monitored over time. Table 1 shows the measurements collected in 50 consecutive samples.

Table 1

Sample	Concentration	Sample	Concentration
1	8,36	26	8,83
2	12,72	27	7,54
3	8,60	28	12,35
4	7,72	29	6,27
5	5,97	30	8,68
6	5,43	31	11,27
7	4,32	32	10,37
8	5,58	33	12,3
9	4,59	34	10,62
10	6,94	35	13,89
11	3,56	36	15,08
12	7,71	37	13,3
13	1,57	38	19,47
14	7,32	39	18,25
15	3,95	40	17,26
16	6,31	41	19,67
17	-0,23	42	17,15
18	-0,35	43	18,17
19	3,23	44	19,59
20	7,38	45	16,13
21	4,73	46	17,08
22	7,96	47	13,49
23	11,58	48	13,55
24	14,99	49	16,34
25	8,21	50	10,95

a) Being known that a negative value is the result of a temporary miscalibration of the measuring device, fit a suitable model to these data;
b) Based on the result of point a), estimate the 95% prediction interval for the contaminant concentration in the next sample.
c) Based on the result of point a), design an appropriate control chart for these data with $ARL_0 = 250$.
d) From historical data, it is known that the most appropriate model for this process yielded a standard deviation of residuals equal to $\sigma_\varepsilon = 2.5$. Determine, with a statistical test, if the model fitted at point a) is such that the standard deviation of residuals is greater than this value (report also the p-value of the test). Discuss the result.

Exercise 2 (15 points)

A company produces aluminum laminates. The quality control department has recently introduced a statistical monitoring tool to keep under control the planarity of the laminates. It consists of an $\bar{X}$ control chart designed such that the number of samples before a false alarm is equal to 250.

a) Estimate and draw the curves of ARL_1 as a function of the mean shift δ expressed in standard deviation units with a sample size $n = 4$ and $n = 8$, respectively (show the two curves for $\delta \in [0\ 2]$ and report the ARL_1 values for $\delta = 1$ and $\delta = 2$).
b) Estimate and draw the curves of ARL_1 as a function of the sample size n for two values of the shift, $\delta = 1$ and $\delta = 2$, where δ is expressed in standard deviation units (show the two curves for $n \in [2\ 20]$ and report the ARL_1 values for $n = 3$ and $n = 6$).
c) The head of the quality control department is interested in selecting an optimal sample size n to minimize the lack of quality costs in the presence of a mean shift equal to $\delta = 2$ standard deviation units. Knowing that samples are gathered every 4 hours, the cost of planarity measurements for each laminate is $C_1 = 2$ € and an extra cost equal to $C_2 = 15$ € is due for each hour spent in the out-of-control state, determine the optimal sample size that minimizes the overall expected costs (assume the cost of the process in its in-control state as a reference baseline). Discuss the results.

Exercise 3 (3 points)

A company that produces thermal cameras is interested in monitoring the calibration curves of their devices. The calibration curve can be modelled by a linear model $y = \beta_0 + \beta_1 x + \varepsilon_t$ where the regressor x is the infrared counts measured by the sensor, whereas y is the temperature shown as output by the camera. All calibration curves are generated by using the same infrared counts levels for the regressor; moreover, the intercept $\hat{\beta}_0 = b_0$ and the slope $\hat{\beta}_1 = b_1$ are estimated using ordinary least squares.

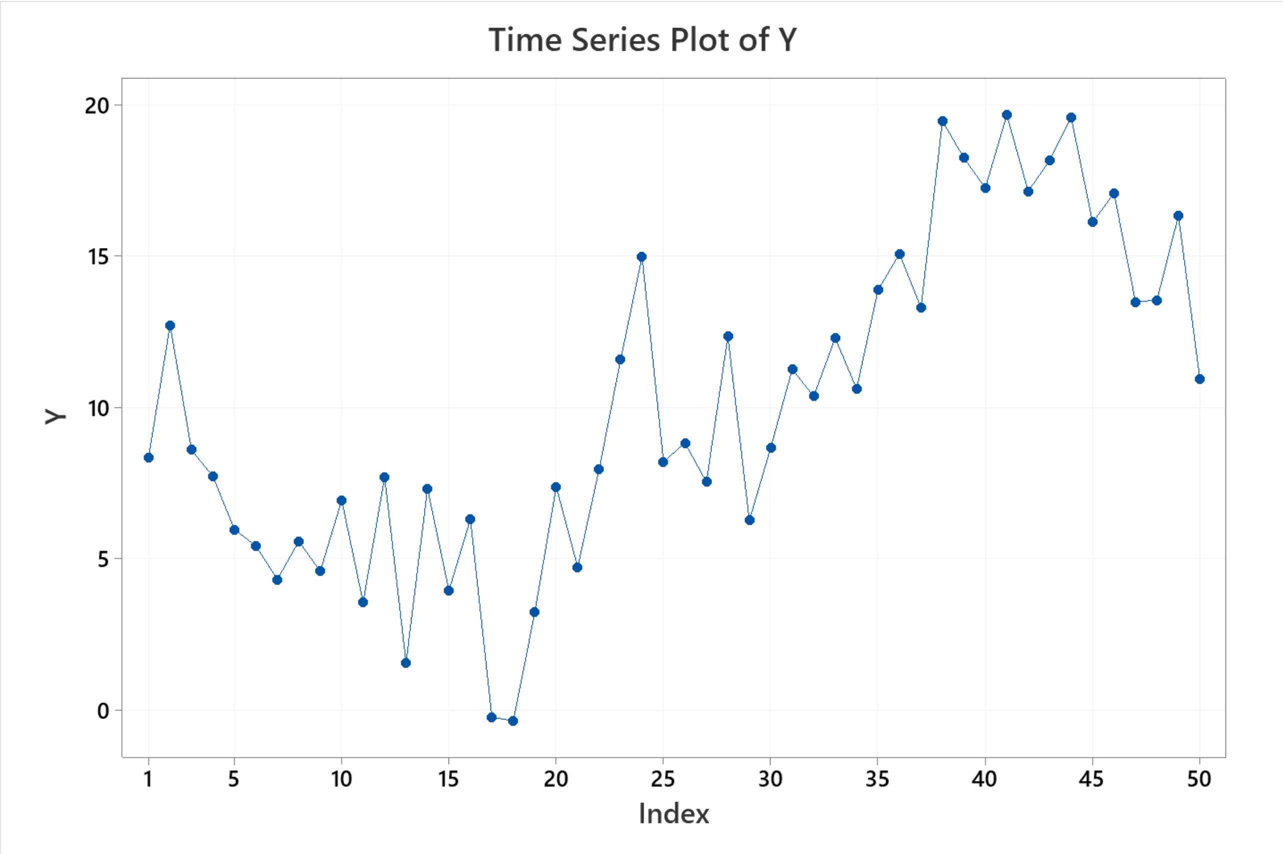
Assuming $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, and assuming that β_0 and β_1 and σ_ε^2 are known, write down the expression of the control limits of a control chart for monitoring the slope of calibration curves.

Solutions

Exercise 1

a)

Time series plot of the temperature series:



It is present a meandering pattern. Negative values were observed in sample 17 and 18.

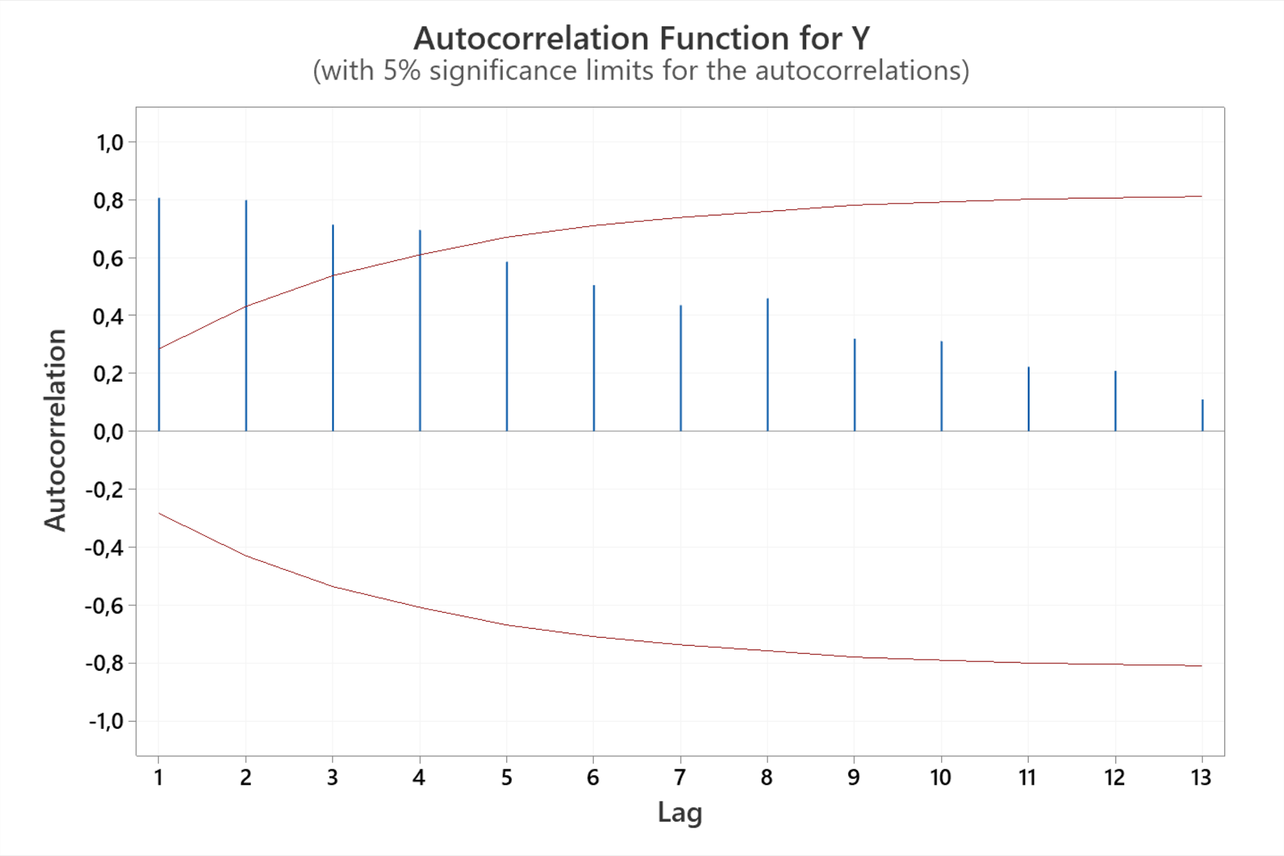
Runs test: null hypothesis is not accepted:

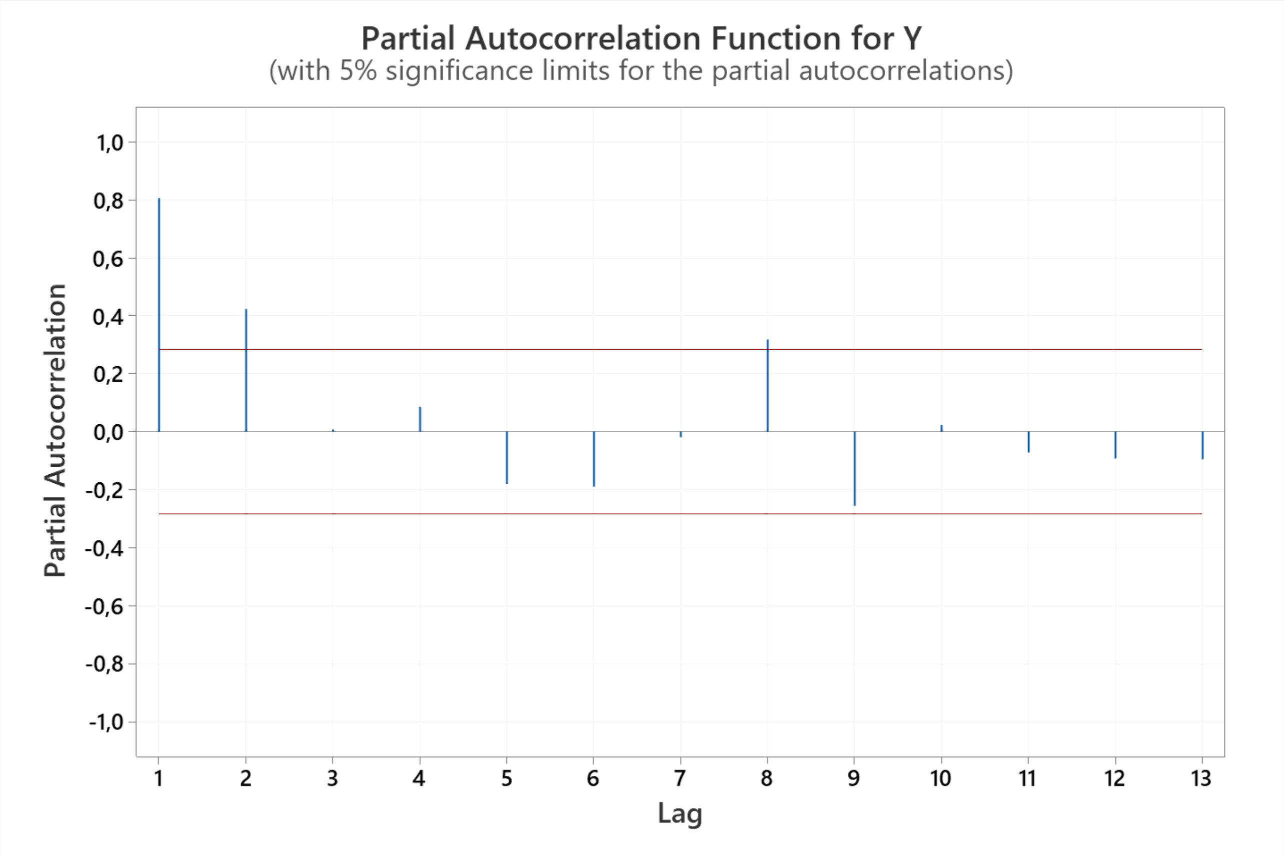
Test

Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

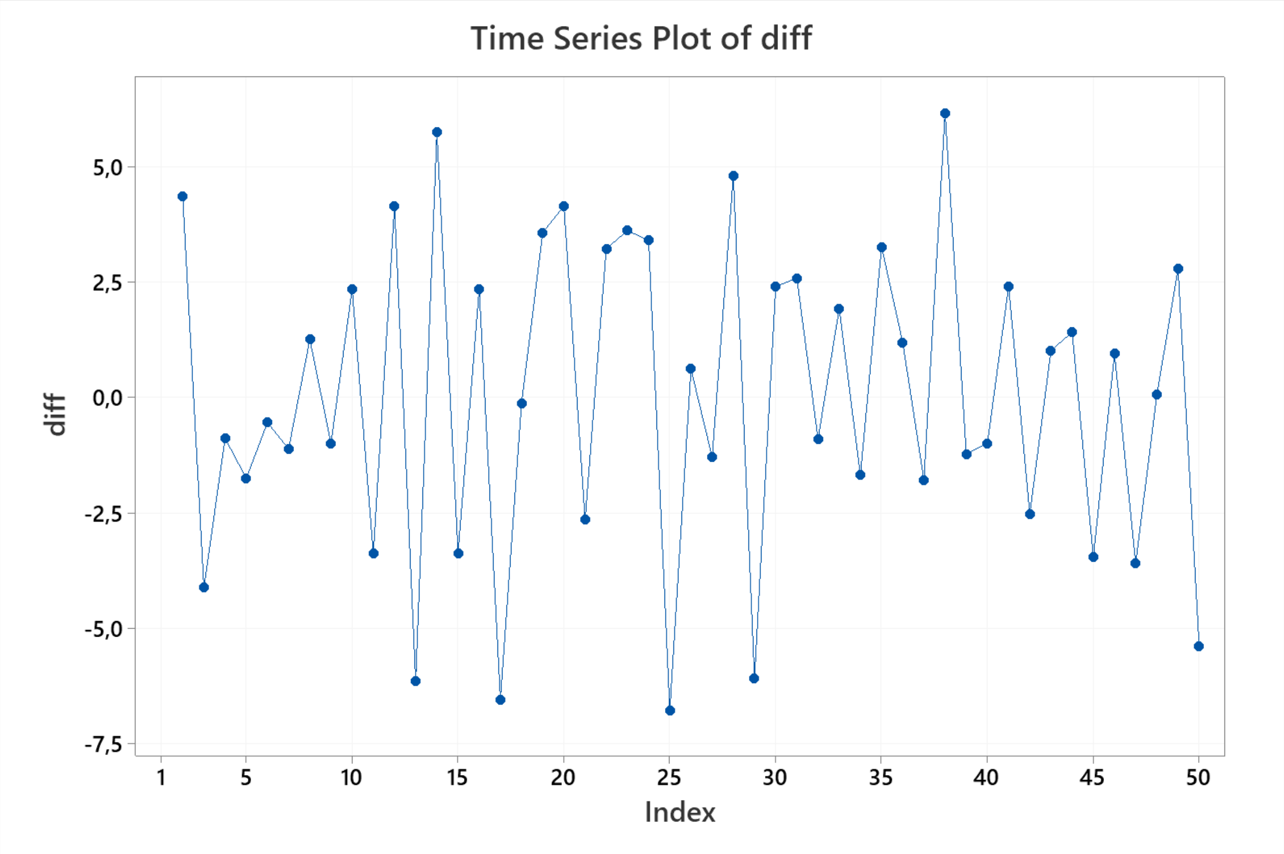
Number of Runs		
Observed	Expected	P-Value
8	25,96	0,000

Sample autocorrelation and partial autocorrelation functions:

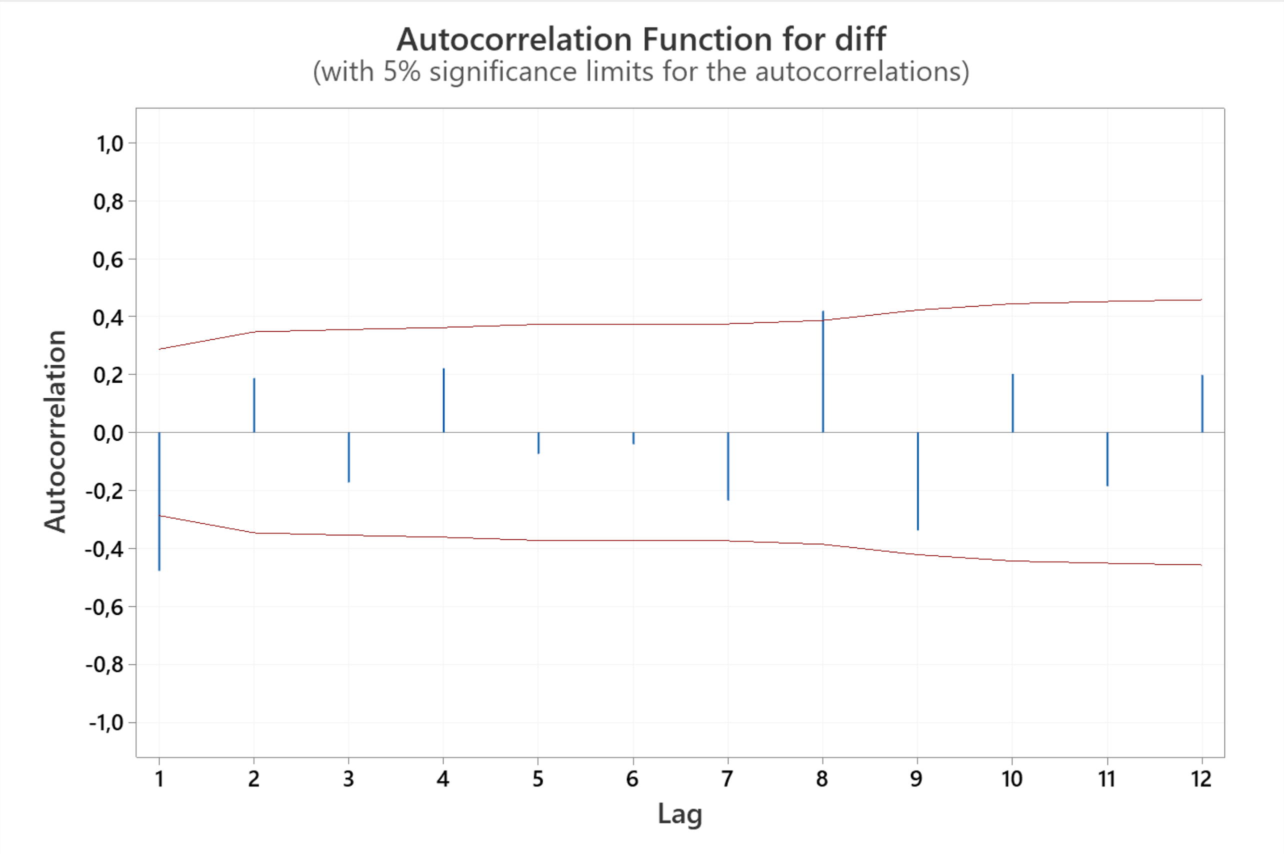


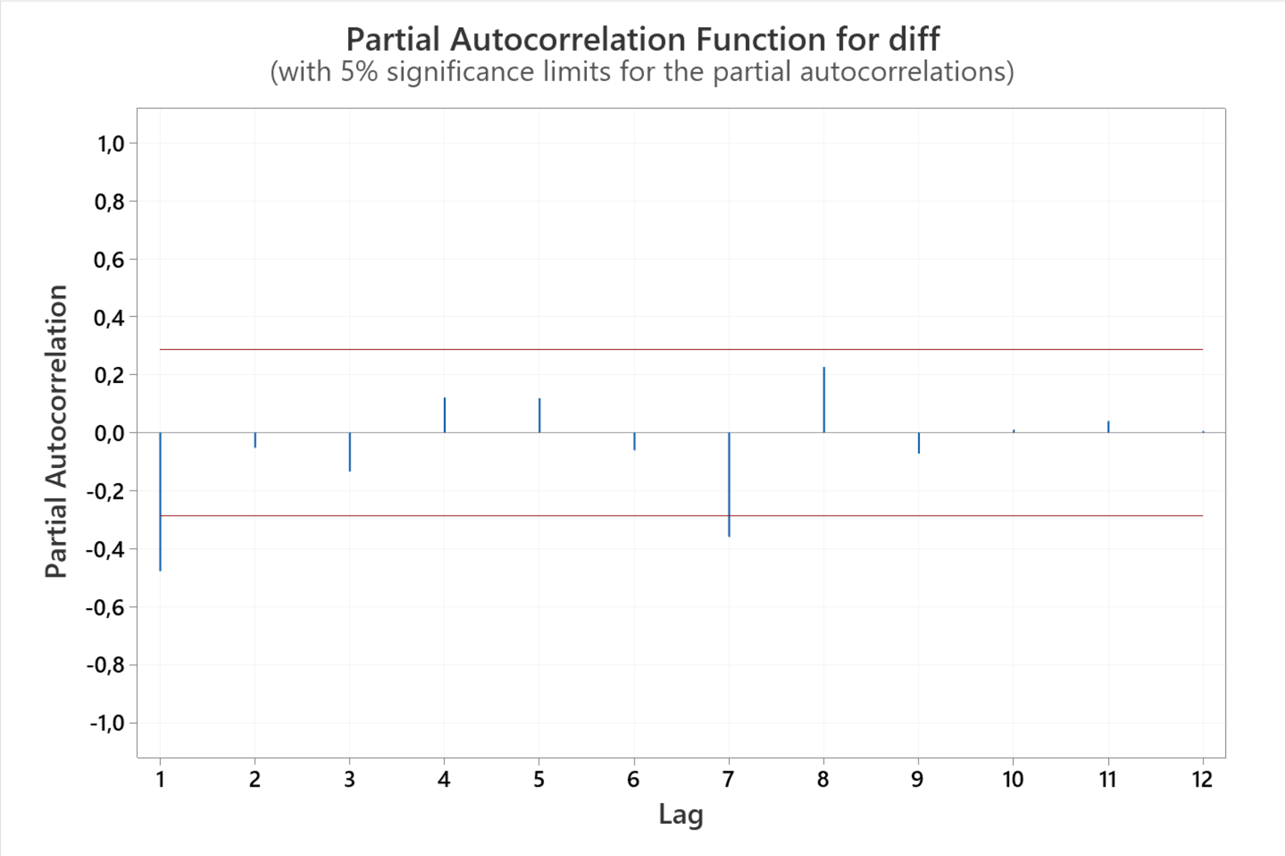


A slow decay of the SACF is present, which suggests a non-stationarity of the process. By differencing the timeseries we get:



The SACF and SPACF of the data after the differencing operation are the following:





A suitable model for the temperature time series is therefore an ARIMA(1,1,0). However, we should keep in mind that two negative values are present, caused by a temporary miscalibration of the sensor. Thus, a dummy variable that is equal to 1 for these two samples and 0 for all other samples can be included in the model.

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,251 - 0,546 AR1
1	diff = -4,47 - 0,546 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0,251	0,413	(-0,581; 1,083)	0,61	0,547	
AR1	-0,546	0,125	(-0,797; -0,295)	-4,38	0,000	1,02
dummy						
1	-4,72	2,04	(-8,83; -0,62)	-2,32	0,025	1,02

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,79333	32,91%	29,92%	387,706	25,92%	240,66	247,22

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	172,21	32,91%	172,21	86,106	11,04	0,000
AR1	1	130,36	24,91%	149,58	149,580	19,17	0,000
dummy	1	41,85	8,00%	41,85	41,854	5,36	0,025
Error	45	351,12	67,09%	351,12	7,803		
Total	47	523,33	100,00%				

The constant term is not significant, thus we may remove it:

WORKSHEET 1

Regression Analysis: diff versus AR1; dummy

Method

Categorical predictor coding	(1; 0)
Rows unused	2

Regression Equation

dummy	
0	diff = 0,0 - 0,540 AR1
1	diff = -4,46 - 0,540 AR1

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
AR1	-0,540	0,123	(-0,789; -0,292)	-4,37	0,000	1,02
dummy						
1	-4,46	1,98	(-8,44; -0,48)	-2,25	0,029	1,02

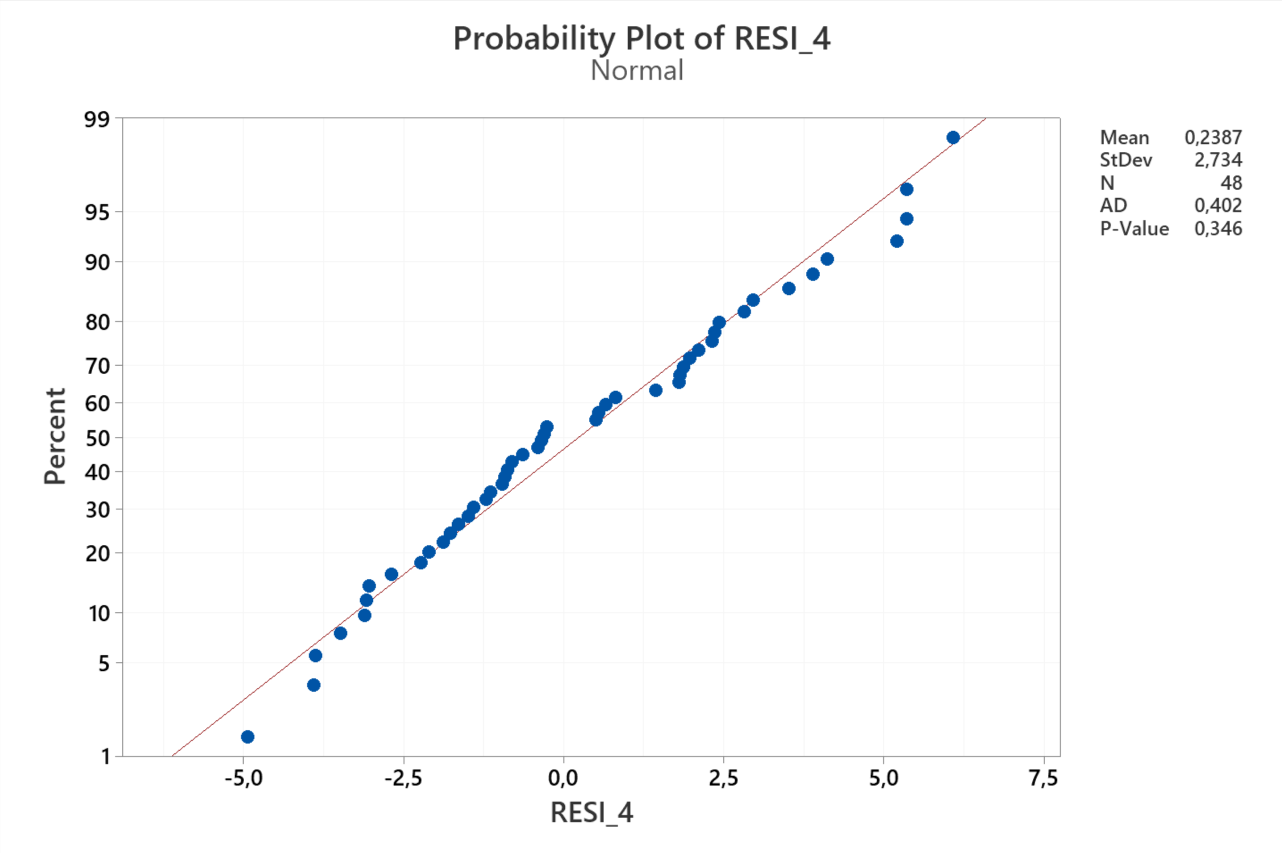
Model Summary

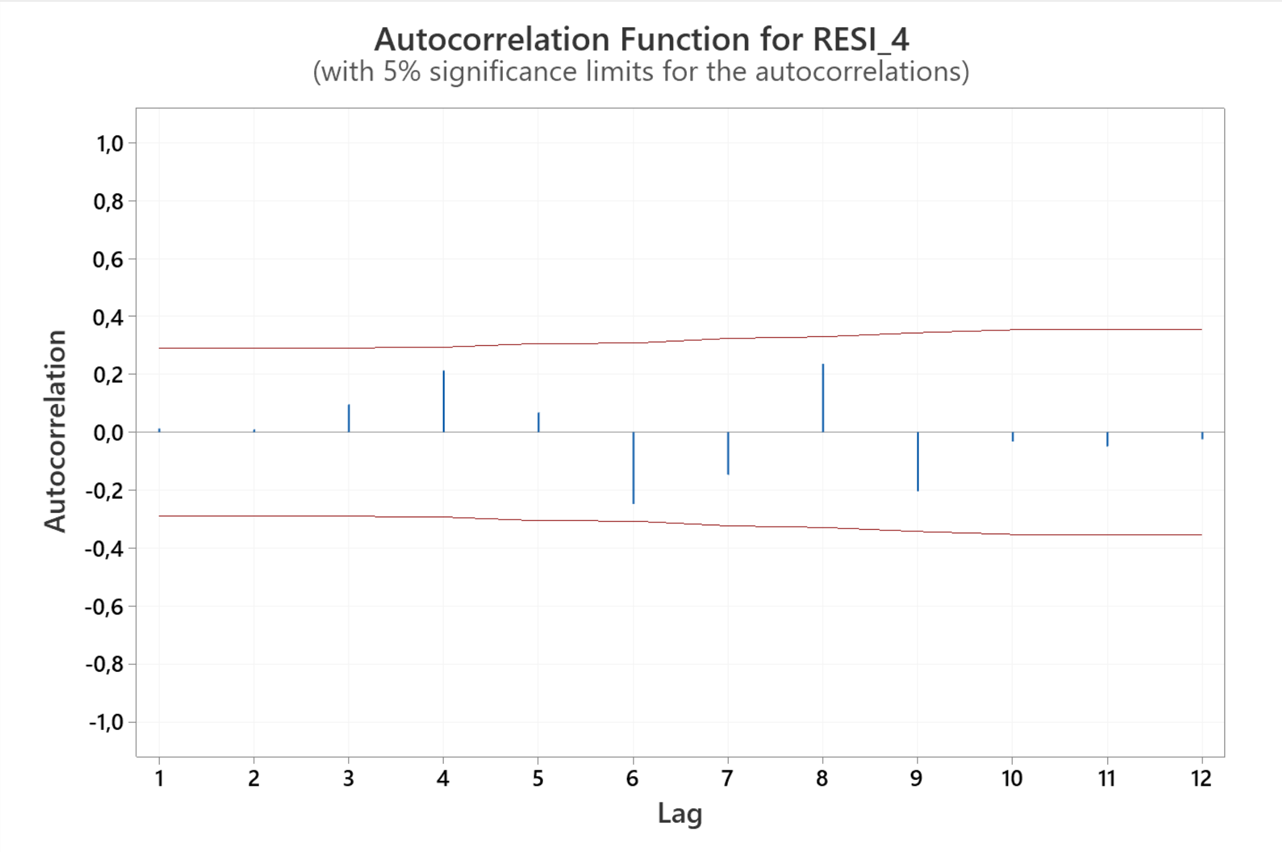
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
2,77408	32,37%	29,43%	374,471	28,45%	238,67	243,74

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	169,41	32,37%	169,41	84,703	11,01	0,000
AR1	1	130,32	24,90%	147,23	147,227	19,13	0,000
dummy	1	39,09	7,47%	39,09	39,087	5,08	0,029
Error	46	353,99	67,63%	353,99	7,696		
Total	48	523,40	100,00%				

Check of residuals:


Probability Plot of RESI_4
Normal
Percent
RESI_4
Mean 0,2387
StDev 2,734
N 48
AD 0,402
P-Value 0,346


Autocorrelation Function for RESI_4
(with 5% significance limits for the autocorrelations)
Autocorrelation
Lag

Test

Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs		
Observed	Expected	P-Value
29	24,83	0,221

The residuals are normal and independent. The model is adequate.

b)

The 95% prediction interval for the differenced time series for observation 51 is the following:

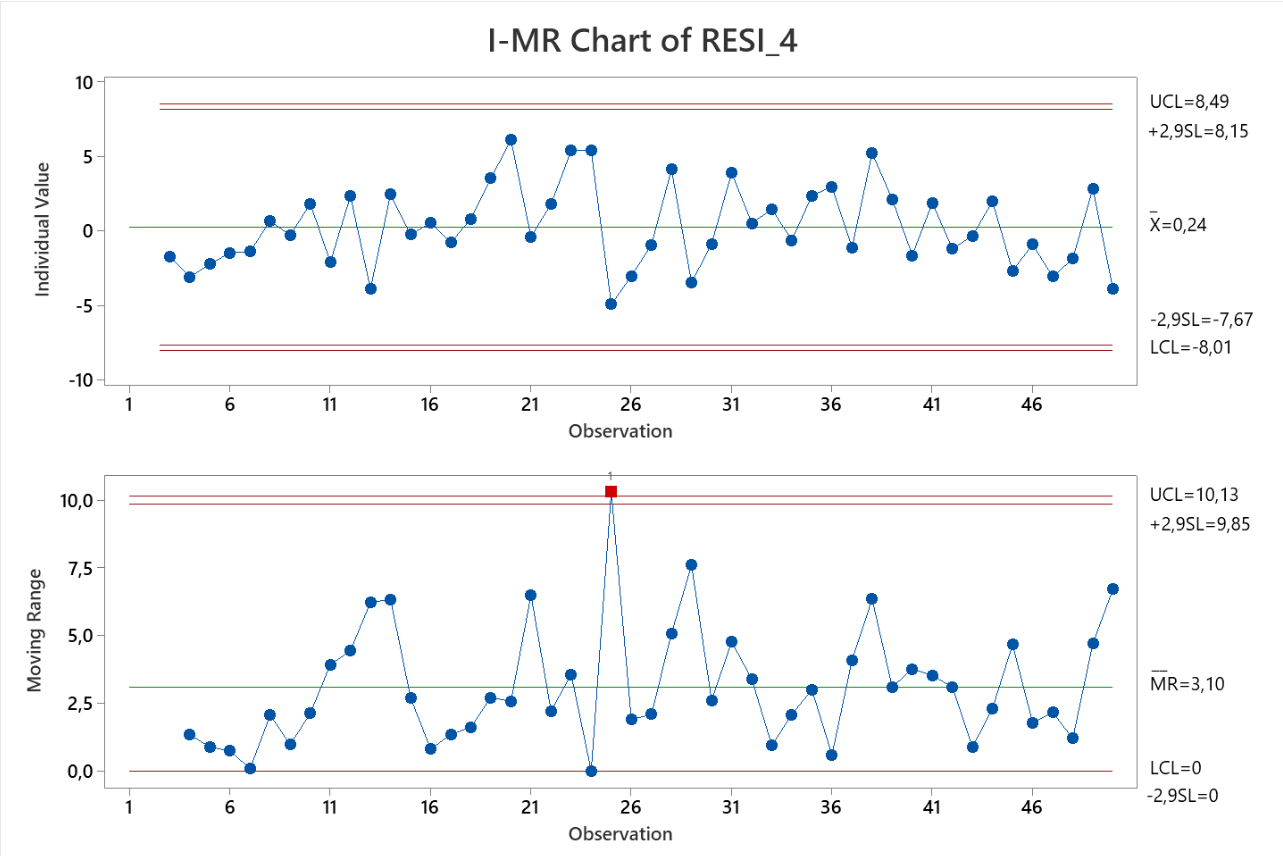
95% PI
(-2,83103; 8,65381)

This is a prediction interval on the differenced data. To obtain the prediction interval on the original data (contaminant concentration in ppm) we must sum the value of the variable at the 50th sample, i.e., $Y = 10{,}95$, thus:

$$8.119\ ppm\ \leq Y \leq 19.604\ ppm$$

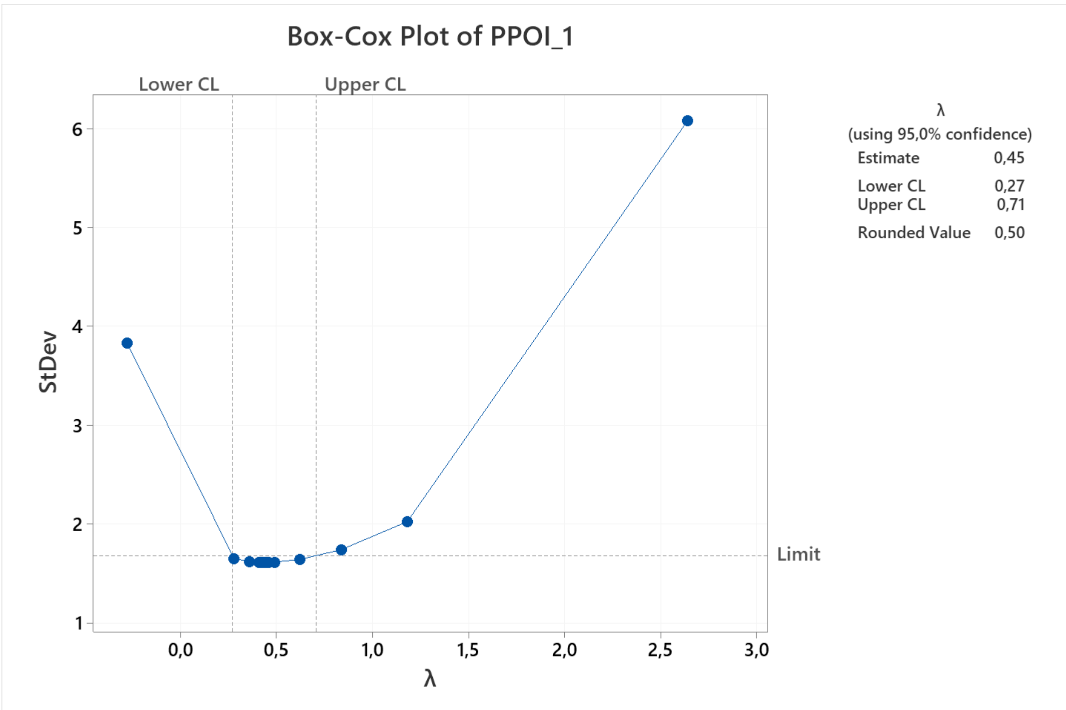
c)

The Type I error corresponding to $ARL_0 = 250$ is $\alpha = 0{,}004$, which corresponds to $k = z_{\alpha/2} = 2{,}878$. The resulting I-MR control chart for the model residuals is the following:

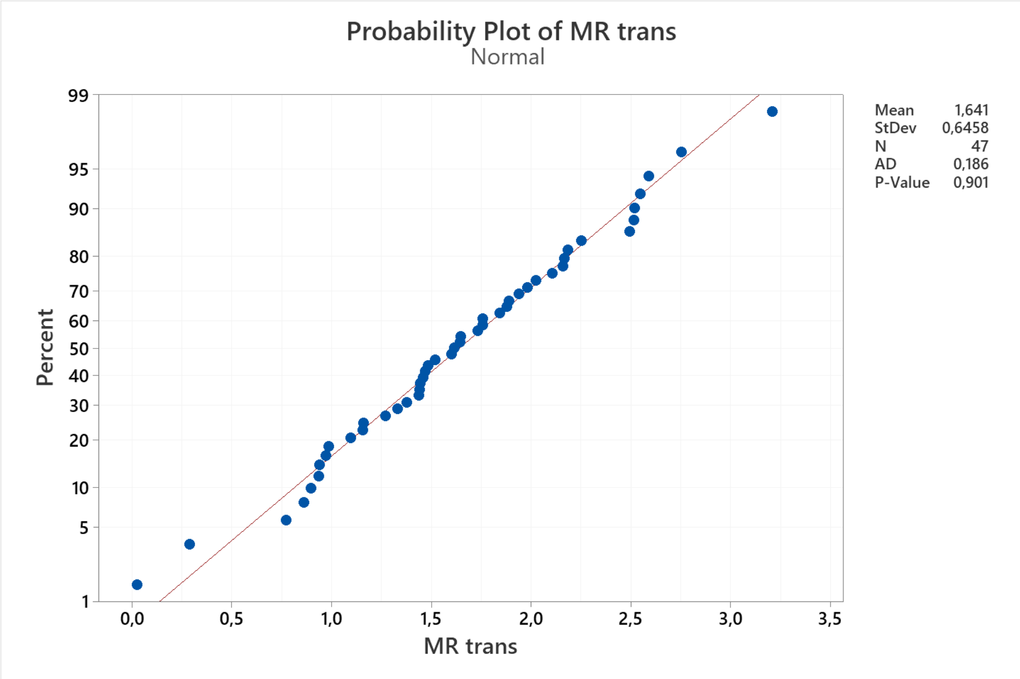


Sample 25 yields an OOC in the MR control chart. It is possible to verify if this OOC is the consequence of a violation of assumptions in the MR chart. One possible way is to transform MR data to normality and redesign the chart as follows:

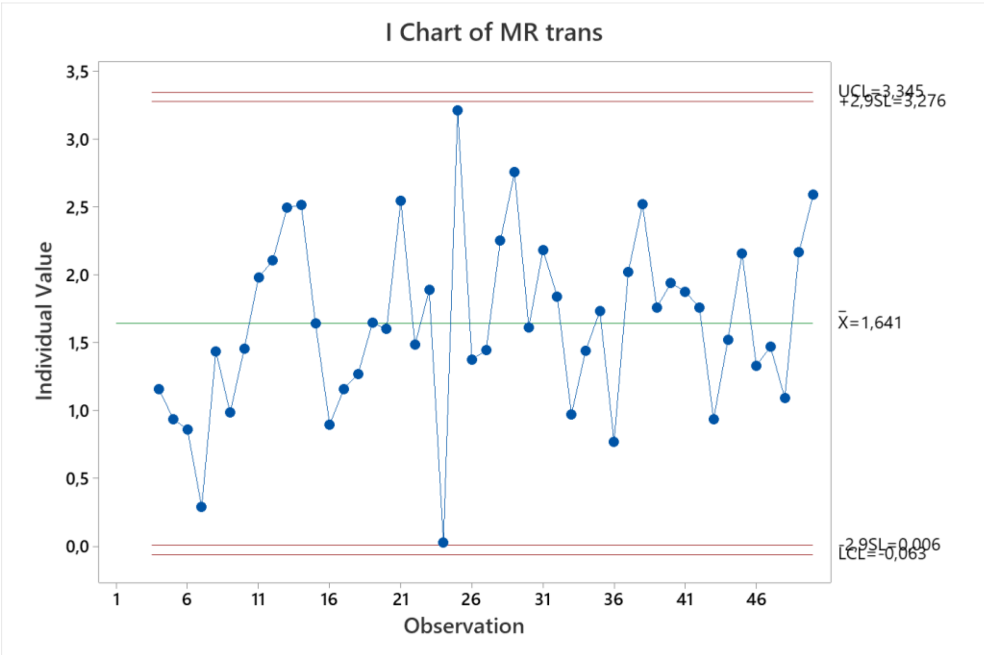
Box-Cox transformation:



Normality of MR statistic after transformation:



New MR control chart:



The OOC in the MR control chart was caused by a violation of assumptions of the chart itself.

The process is in-control.

d)

Since model residuals are normal and independent, it is possible to perform a one sample chi-squared test as follows.

By estimating the standard deviation of the model residuals as $\hat{\sigma}_\varepsilon = \sqrt{MSE} = 2.774$.

The test is such that:

$H_0: \sigma_\varepsilon = 2.5$

$H_1: \sigma_\varepsilon > 2.5$

The test statistic is $X^2 = \frac{(n-p)\hat{\sigma}_\varepsilon{}^2}{\sigma_\varepsilon{}^2} \sim X^2_{n-p}$, where $p = 2$ is the number of model terms, and $n - p = 46$.

Under H_0 we get $X^2 = 56.636$. The corresponding p-value is 0.135.

At 95% confidence, the standard deviation of residuals of the model fitted in point a) is not statistically larger than the one observed on historical data.

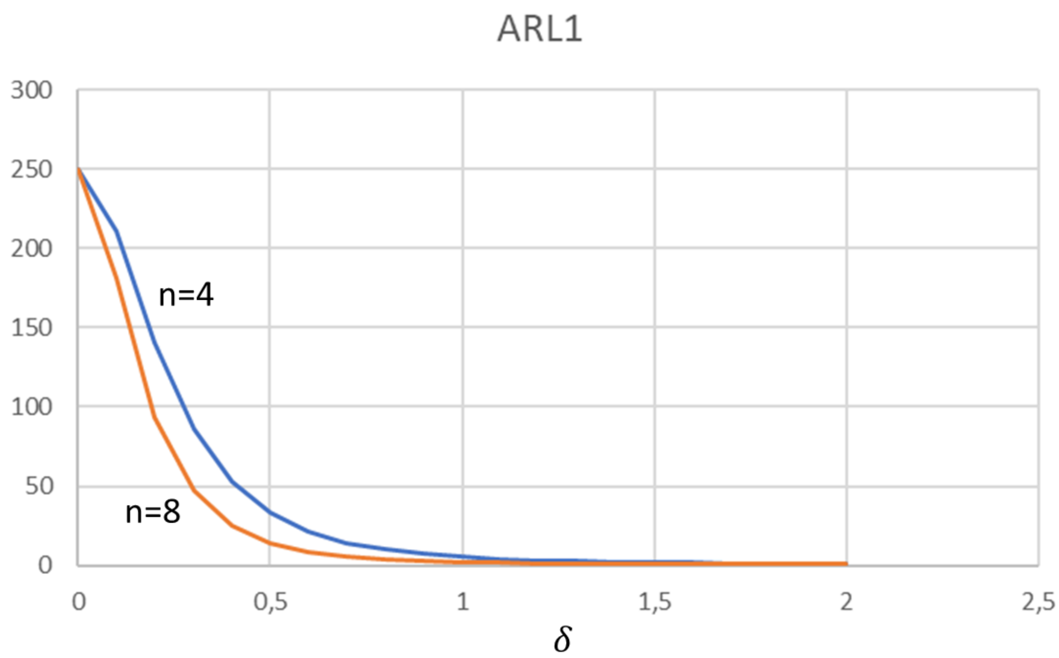
Exercise 2

The value of $K = z_{\alpha/2}$ with $\alpha = \frac{1}{250} = 0.004$ is: $K = 2.878$.

The Type II error as a function of the mean shift in standard deviation units is given by:

$\beta = \Pr\left(Z \le K - \delta\sqrt{n}\right) - \Pr\left(Z \le -K - \delta\sqrt{n}\right)$, where $\delta = \dfrac{\mu_1 - \mu_0}{\sigma}$

Being, $ARL_1(\delta) = \frac{1}{1-\beta}$. The $ARL_1(\delta)$ curves for $n = 4$ and $n = 8$ are the following:



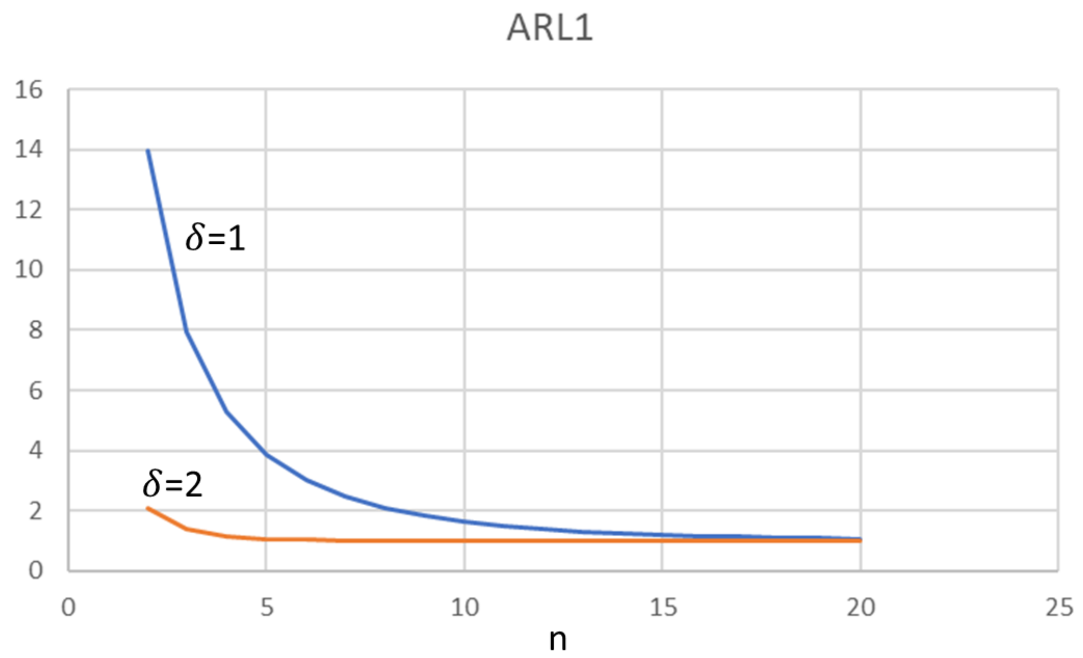
	$\delta = 1$	$\delta = 2$
ARL_1 with n=4	5.26	1.15
ARL_1 with n=8	2.08	1.00

b)

Being fixed δ, the type II error can be estimated as a function of n with the same expression used in the previous case:

$\beta = \Pr(Z \leq K - \delta\sqrt{n}) - \Pr(Z \leq -K - \delta\sqrt{n})$

The resulting $ARL_1(n)$ curves for the two given mean shifts are the following:



	$n = 3$	$n = 6$
ARL_1 with $\delta = 1$	7.94	2.99
ARL_1 with $\delta = 2$	1.39	1.02

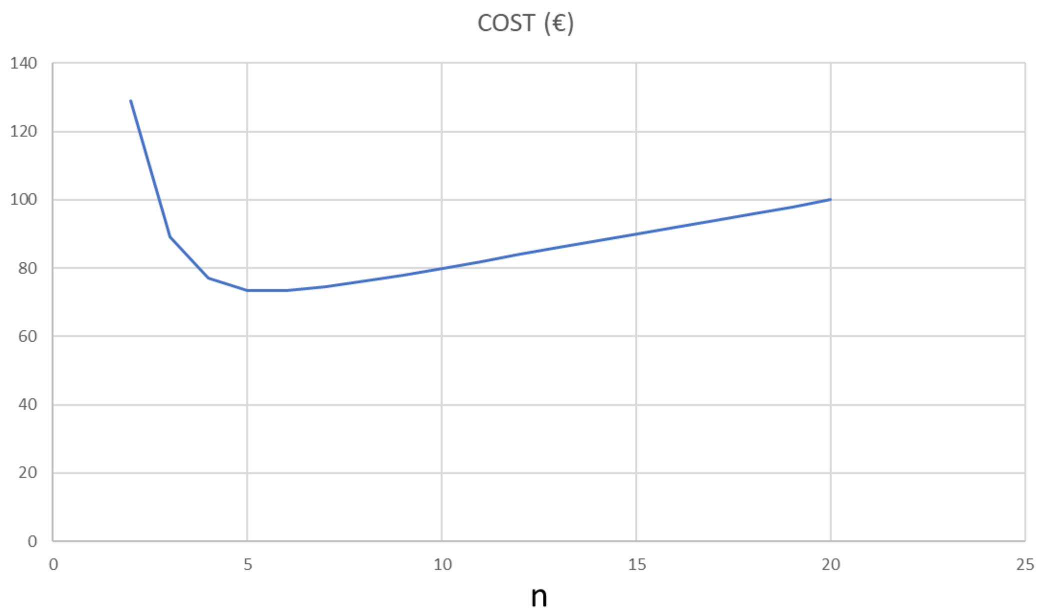
c)

The function to be minimized is the following:

$$C(n) = C1 * n + C2 * ATS(n) = 2 * n + 15 * ATS(n)$$

Where ATS = $h \cdot ARL_1$, where h is the time between the collection of two consecutive samples, i.e., $h = 4\ h$.

The cost function for $\delta = 2$ is shown below:



The late detection cost predominates at smaller values of n, whereas the inspection cost predominates at larger values of n. The optimal values of the sample size is n=6.

Exercise 3

The estimated slope b_1 is a random variable such that:

$E(b_1) = \beta_1, V(b_1) = \frac{\sigma_\varepsilon^2}{S_{xx}}$ where:

- σ_ε^2 is the variance of the normal error term
- $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$

By using the Shewhart's scheme and assuming known parameters, the control chart for b_1 can be designed as follows:

$$UCL = \beta_1 + z_{\alpha/2}\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}$$

$$CL = \beta_1$$

$$LCL = \beta_1 - z_{\alpha/2}\sqrt{\frac{\sigma_\varepsilon^2}{S_{xx}}}$$

Where α is the Type I error.

The control charts can be used to monitor the stability over time of the calibration curves' slope for different sensors. It can be possibly combined with a control chart on $\hat{\sigma}_\varepsilon^2$, to monitor the model residuals as well.