QUALITY DATA ANALYSIS EXCERCISE BOOK

Contents

Control charts for IID data

Exercise 1 (max score 14)

A company has recently bought a metal additive manufacturing machine tool and is testing its capability. To this aim, cylindrical specimens have been produced. From previous tests, the machine tool builder say that the diameters of the cylinders should be normally distributed with a mean value of 4 mm and a standard deviation of 0.2 mm.

1) Design a traditional $\bar{X}$ - S control chart in order to have an average number of samples before a false alarm equal to 200 for both the charts with n=5 observation.

2) The following table shows the sample mean and standard deviation values obtained by printing five samples, each of size n=5 (measures are in mm). Is the process in-control?

i	$\bar{X}_i$	S_i	i	$\bar{X}_i$	S_i	i	$\bar{X}_i$	S_i	i	$\bar{X}_i$	S_i	i	$\bar{X}_i$	S_i
1	4,0738	0,1638	*2*	3,9406	0,2148	*3*	4,0430	0,1711	*4*	3,9968	0,1312	*5*	3,8290	0,1555

3) The company thinks that the S charts yields an ARL_0 different from the nominal one. Compute the real value of ARL_0 for this chart and the percentage error between the nominal ARL_0 and the real one.

4) How does the percentage error computed in c) change by using different sample sizes n (show the values for n=2, 5, 10, 20 and 50)?

5) Plot the OC curve of beta against the entity of mean deviation expressed in standard deviation units. Remind that the error beta for a $\bar{X}$ - S control chart is the probability of having no alarm from both the charts under out-of-control conditions. Show the curve (qualitative plot), its formulation and the values for delta = 1, 2 and 3.

Exercise 1 (solution)

1) Design the Xbar-S control chart with known parameters:

$$
\begin{aligned}
&LCS = \mu + k\sigma/\sqrt{n} \qquad && LCS = \mu_S + k\sigma_S = c_4\sigma + k\sqrt{1-c_4^{\,2}}\,\sigma \\
&LC = \mu && LC = \mu_S = c_4\sigma \\
&LCI = \mu - k\sigma/\sqrt{n} && LCI = \mu_S - k\sigma_S = c_4\sigma - k\sqrt{1-c_4^{\,2}}\,\sigma
\end{aligned}
$$

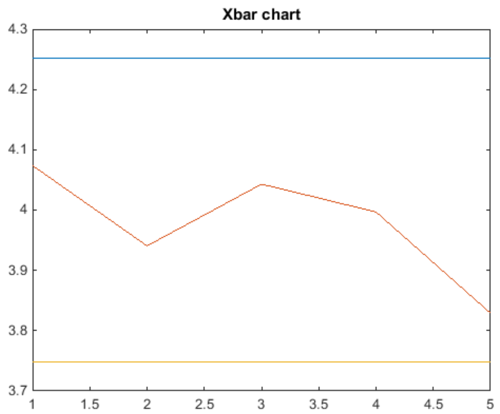
ARL0	200
alpha	0,005
K=z_alpha/2	2,807034
n	5

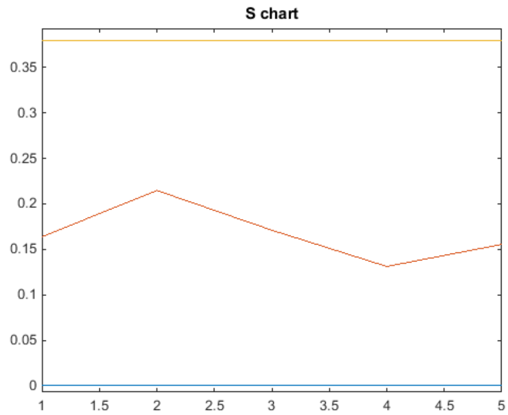
c4(5) 0,94

X-bar chart		
UCL	CL	LCL
4,2511	4	3,7489

S-chart		
UCL	CL	LCL
0,3795	0,188	0

2) Let's check if the new data are IC or not:





The process is in control

3) Being UCL and LCL the limits of the S chart computed at point a). We know that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n\text{-}1}$, thus:

$$1-\alpha = P\left(LCI \le S \le LCS \,|\, \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n\text{-}1} \right) = P\left(\frac{(n-1)}{\sigma^2} LCI^2 \le \frac{(n-1)}{\sigma^2} S^2 \le \frac{(n-1)}{\sigma^2} LCS^2 \right) =$$

$$= P(0 \le \chi^2_{n-1} \le 14{,}402)$$

alpha_LCL	alpha_UCL	alpha	ARL0	error	error %
0	0,0061	0,0061	163,9344	-36,07	-18,03%

The percentage error is about -18%.

4) Let's repeat the estimation by changing the value of the sample size n: this means that we have to include the explicit expression of the control limits in the Type I error formulation:

$$1-\alpha = P\left((n-1)[c_4(n) + k\sqrt{1-c_4(n)^2}\,]^2 \le \chi^2_{n\text{-}1} \le (n-1)[c_4(n) + k\sqrt{1-c_4(n)^2}\,]^2 \right)$$

n	c4	Alpha_LCL	Alpha_UCL	alpha	ARL0	error
2	0,7979	0	0,012776	0,012776	78,27	-121,73
5	0,94	0	0,006109	0,006109	163,70	-36,30
10	0,9727	0,000416	0,004732	0,005148	194,26	-5,74
20	0,9869	0,001016	0,003937	0,004952	201,93	1,93
50	0,994924	0,001612	0,003435	0,005046	198,16	-1,84

The error reduces as the sample size increases.

5) H_1: $\mu_{new} = \mu + \delta\sigma$

Xbar chart:

$$\beta_{\bar{X}} = P(LCL_{Xbar} \leq \bar{X} \leq UCL_{Xbar}|H_1) = P\left(Z \leq \frac{UCL_{Xbar} - \mu_{new}}{\sigma/\sqrt{n}}\right) - P\left(Z \leq \frac{LCL_{Xbar} - \mu_{new}}{\sigma/\sqrt{n}}\right)$$

This is a function of delta.

S chart:

$$\beta_S = P(LCL_S \leq S \leq UCL_S|H_1) = P\left(\frac{(n-1)}{\sigma^2}LCL_S{}^2 \leq \frac{(n-1)}{\sigma^2}S^2 \leq \frac{(n-1)}{\sigma^2}UCL_S{}^2\right)$$

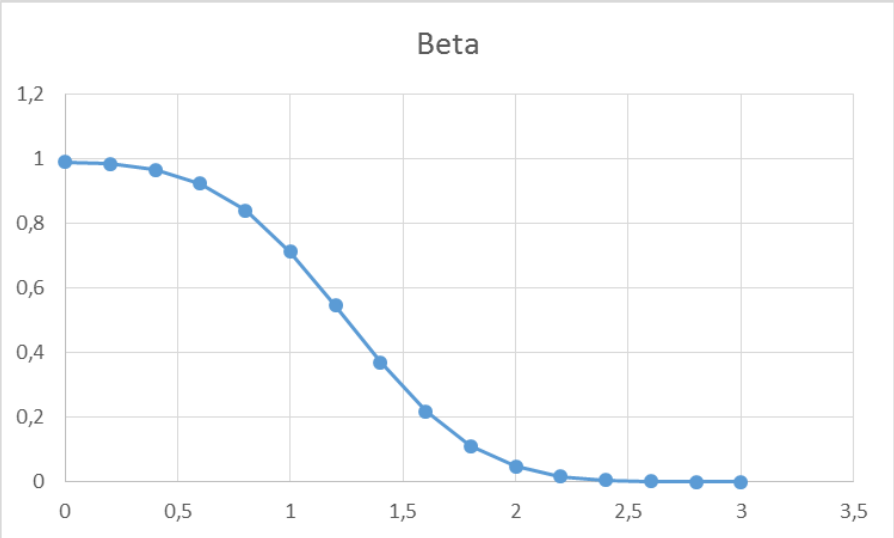
This is constant.

Eventually:

$$\beta = P(no\ alarm|H_1) = P(no\ alarm\ from\ Xbar\ chart|H_1) * P(no\ alarm\ from\ S\ chart|H_1)$$

delta	mu1	Z1	Z2	beta_xbar	CHI1	CHI2	beta_S	beta
0	4	2,806936	-2,807	0,994999	0	14,40468	0,993891	0,98892
0,2	4,04	2,359723	-3,25422	0,990287	0	14,40468	0,993891	0,984237
0,4	4,08	1,912509	-3,70143	0,971987	0	14,40468	0,993891	0,966049
0,6	4,12	1,465295	-4,14864	0,928563	0	14,40468	0,993891	0,92289
0,8	4,16	1,018082	-4,59586	0,845678	0	14,40468	0,993891	0,840512
1	4,2	0,570868	-5,04307	0,715955	0	14,40468	0,993891	0,711581
1,2	4,24	0,123655	-5,49028	0,549206	0	14,40468	0,993891	0,54585
1,4	4,28	-0,32356	-5,9375	0,373136	0	14,40468	0,993891	0,370856
1,6	4,32	-0,77077	-6,38471	0,220421	0	14,40468	0,993891	0,219074
1,8	4,36	-1,21799	-6,83193	0,111615	0	14,40468	0,993891	0,110933
2	4,4	-1,6652	-7,27914	0,047936	0	14,40468	0,993891	0,047644
2,2	4,44	-2,11241	-7,72635	0,017326	0	14,40468	0,993891	0,01722

2,4	4,48	-2,55963	-8,17357	0,005239	0	14,40468	0,993891	0,005207
2,6	4,52	-3,00684	-8,62078	0,00132	0	14,40468	0,993891	0,001312
2,8	4,56	-3,45405	-9,06799	0,000276	0	14,40468	0,993891	0,000274
3	4,6	-3,90127	-9,51521	4,78E-05	0	14,40468	0,993891	4,76E-05



Exercise 2 (max score 3)

Refer to the OC curve computed at point 5) of the previous exercise. The engineers observed that the printed specimens exhibit a proportional relationship between the mean and standard deviation of the diameter, such that $\mu = c\sigma$, and the same factor c applies also in case of large deviations from the nominal size.

How does the OC curve changes if we take into account this information?

Show the curve (qualitative plot), its formulation and the values for delta = 1, 2 and 3.

Exercise 2 (solution)

The proportionaly factor c is 4/0.2 = 20

Thus:

H_1: $\mu_{new} = \mu + \delta\sigma$ and $\sigma_{new} = \mu_{new}/c$

Xbar chart:

$$\beta_{\bar{X}} = P(LCL_{Xbar} \leq \bar{X} \leq UCL_{Xbar}|H_1) = P\left(Z \leq \frac{UCL_{Xbar} - \mu_{new}}{\sigma_{new}/\sqrt{n}}\right) - P\left(Z \leq \frac{LCL_{Xbar} - \mu_{new}}{\sigma_{new}/\sqrt{n}}\right)$$

This is a function of delta.

S chart:

$$\beta_S = P(LCL_S \leq S \leq UCL_S|H_1) = P\left(\frac{(n-1)}{\sigma_{new}{}^2}LCL_S{}^2 \leq \frac{(n-1)}{\sigma_{new}{}^2}S^2 \leq \frac{(n-1)}{\sigma_{new}{}^2}UCL_S{}^2\right)$$

This is a function of delta too.

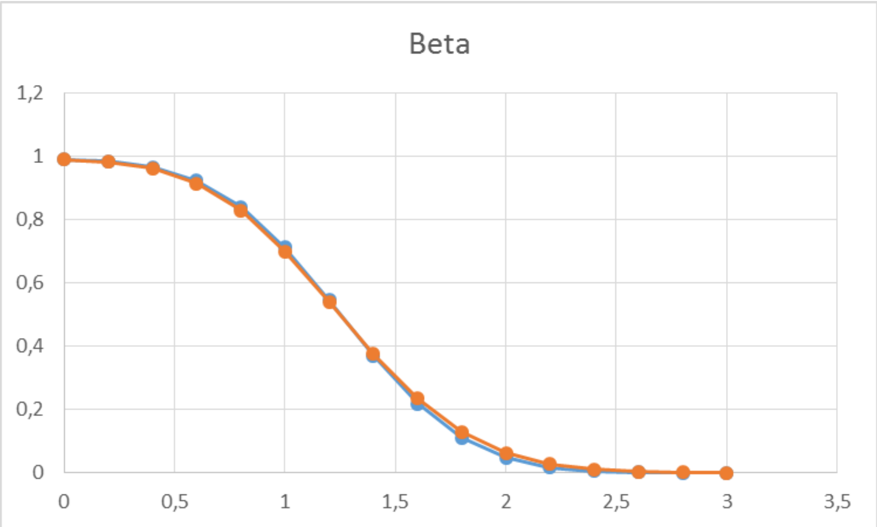
Eventually:

$$\beta = P(no\ alarm|H_1) = P(no\ alarm\ from\ Xbar\ chart|H_1) * P(no\ alarm\ from\ S\ chart|H_1)$$

delta	mu1	Snew	Z1	Z2	beta_xbar	CHI1	CHI2	beta_S	beta
0	4	0,2	2,806936	-2,807	0,994999	0	14,40468	0,993891	0,98892
0,2	4,04	0,202	2,336359	-3,222	0,989627	0	14,12085	0,993081	0,98278
0,4	4,08	0,204	1,875009	-3,62885	0,969462	0	13,84533	0,992195	0,961895
0,6	4,12	0,206	1,422617	-4,02781	0,922548	0	13,57779	0,991228	0,914456
0,8	4,16	0,208	0,978925	-4,41909	0,836186	0	13,31794	0,990178	0,827973

1	4,2	0,21	0,543684	-4,80292	0,70667	0	13,06547	0,98904	0,698924
1,2	4,24	0,212	0,116655	-5,17951	0,546433	0	12,82012	0,987811	0,539773
1,4	4,28	0,214	-0,30239	-5,54906	0,381177	0	12,58161	0,986488	0,376026
1,6	4,32	0,216	-0,71368	-5,91177	0,237713	0	12,34969	0,985068	0,234164
1,8	4,36	0,218	-1,11742	-6,26782	0,131908	0	12,12413	0,983548	0,129738
2	4,4	0,22	-1,51382	-6,6174	0,065036	0	11,9047	0,981926	0,063861
2,2	4,44	0,222	-1,90308	-6,96068	0,028515	0	11,69116	0,980198	0,027951
2,4	4,48	0,224	-2,28538	-7,29783	0,011145	0	11,48332	0,978363	0,010904
2,6	4,52	0,226	-2,66092	-7,62901	0,003896	0	11,28098	0,976419	0,003804
2,8	4,56	0,228	-3,02987	-7,95438	0,001223	0	11,08393	0,974363	0,001192
3	4,6	0,23	-3,39241	-8,27409	0,000346	0	10,89201	0,972195	0,000337

The new curve is shown in orange and superimposed to the previous curve. A very slight different is observed.



Exercise 3 (max score 13)

A company produces plastic tubes. The Quality Assurance procedure consists of picking up a sample of $n = 5$ tubes every hour and recording the mean length of the tubes (in mm).

The table shows the measurements performed in 24 consecutive samplings.

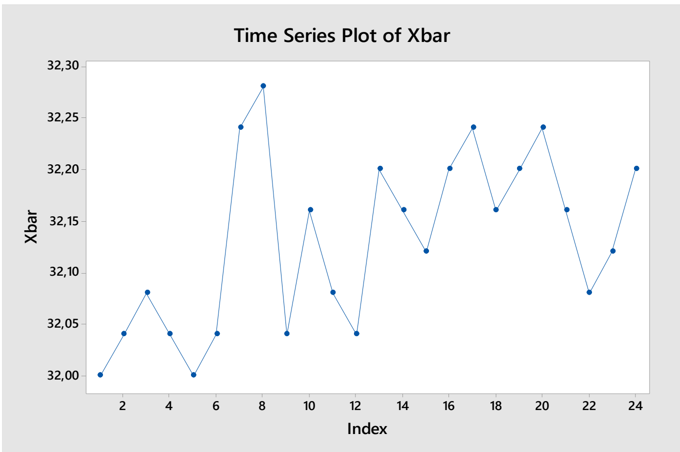
Sample	Xbar	Sample	Xbar	Sample	Xbar	Sample	Xbar
1	32,00	**7**	32,24	**13**	32,20	**19**	32,20
2	32,04	**8**	32,28	**14**	32,16	**20**	32,24
3	32,08	**9**	32,04	**15**	32,12	**21**	32,16
4	32,04	**10**	32,16	**16**	32,20	**22**	32,08
5	32,00	**11**	32,08	**17**	32,24	**23**	32,12
6	32,04	**12**	32,04	**18**	32,16	**24**	32,20

The standard deviation of the mean tube length in samples of size $n = 5$ is assumed to be stable and known:$\sigma_{Xbar} = 0{,}22\ mm$. Assume also that the process target is $32{,}15\ mm$ and the desired mean time between false alarms is equal to 500 hours.

1) Design a traditional control chart for monitoring the process mean.
2) Is the assumed value of σ_{Xbar} appropriate to this process data? Justify with a statistical test, if needed.
3) Design a more appropriate traditional control chart for the process mean based on the conclusions drawn at point 2)

Exercise 3 (solution)

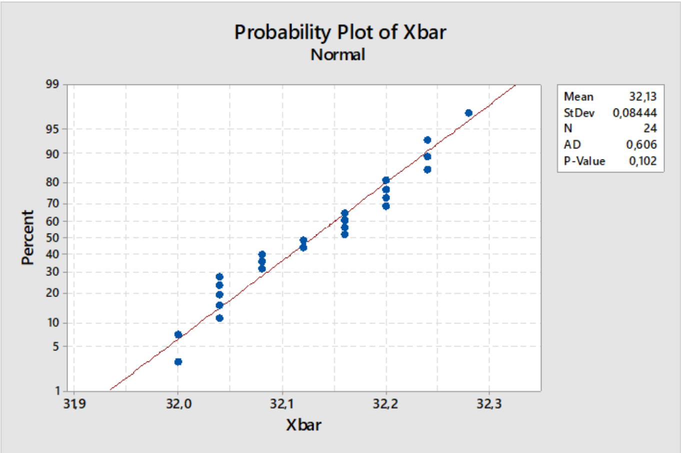
1) Data snooping

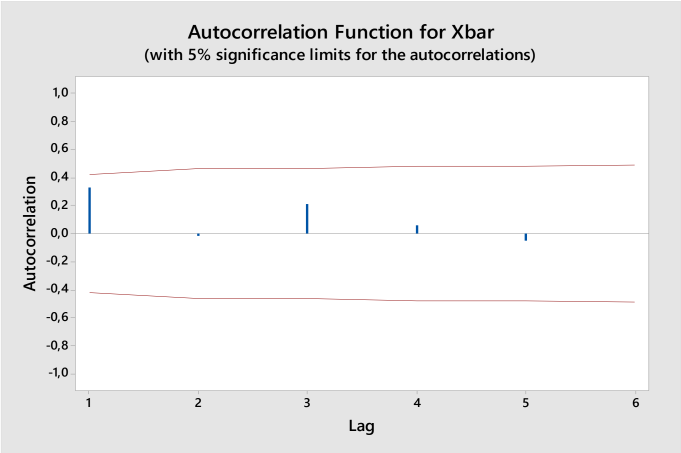


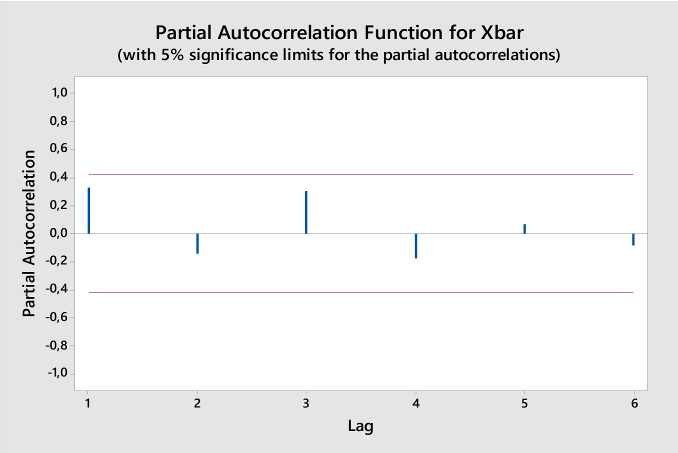
Randomness and normality check:

Runs Test: Xbar

```
Runs test for Xbar

Runs above and below K = 32,13

The observed number of runs = 10
The expected number of runs = 13
12 observations above K; 12 below
P-value = 0,210
```



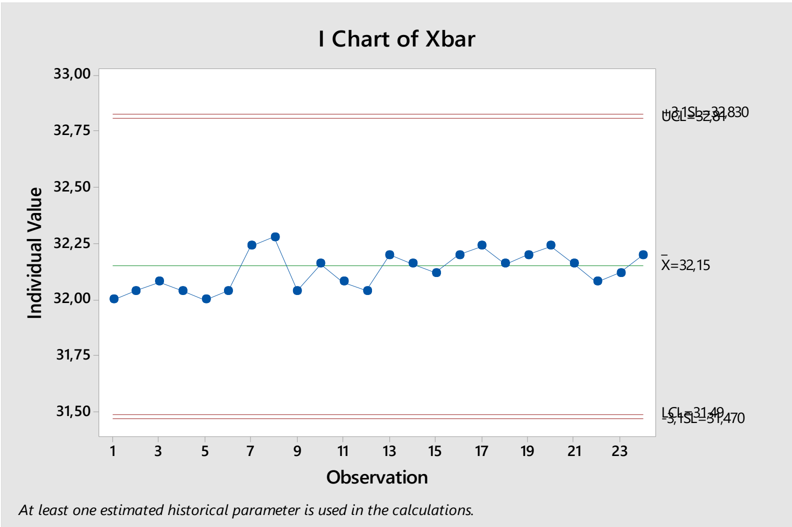




No violation of assumptions. No outlier.

Shewart chart design:

ARL0	500
alpha	0,002
z_alpha/2	3,090232
sigma	0,22
target	32,15



Hugging is present. Probably, the assumed standard deviation is not an appropriate estimate.

2) Hypothesis testing for the variance.

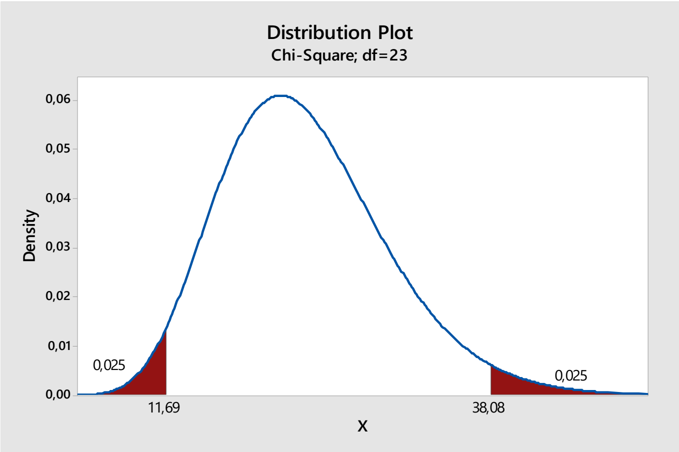
$\sigma^2_{xbar} = 0{,}0484$ (known)

$s^2 = 0{,}0071$ (estimated from data)

The test statistic is:

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_{xbar}^2} = \frac{23 * 0{,}0071}{0{,}0484} = 3{,}373967$$

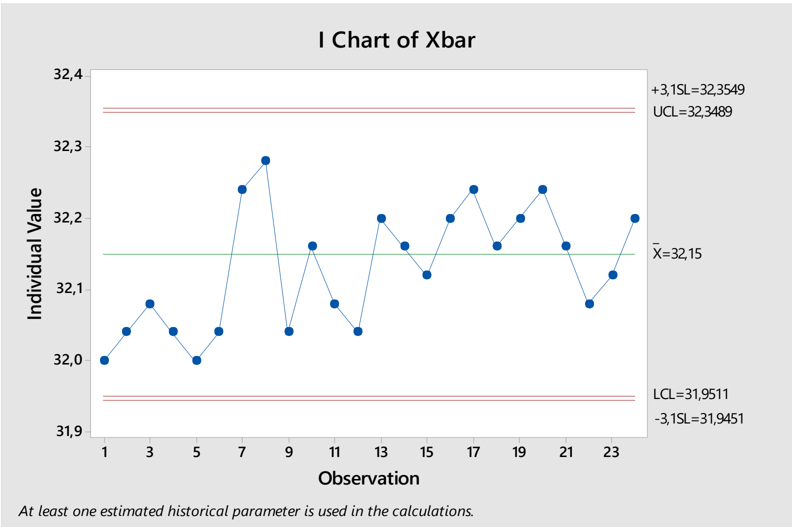
$$\chi^2_{1-\frac{\alpha}{2};n-1} = \chi^2_{0{,}975;23} = 11{,}69$$



$p - value = 0.000$

Thus we can reject the null hypothesis at 5%. A bad estimate of the standard deviation caused the hugging effect observed in the designed chart.

3) A more appropriate control chart is the Shewhart control chart based on the sample standard deviation:



Control charts for non IID Data (with time series models)

Exercise 1 (max score: 15)

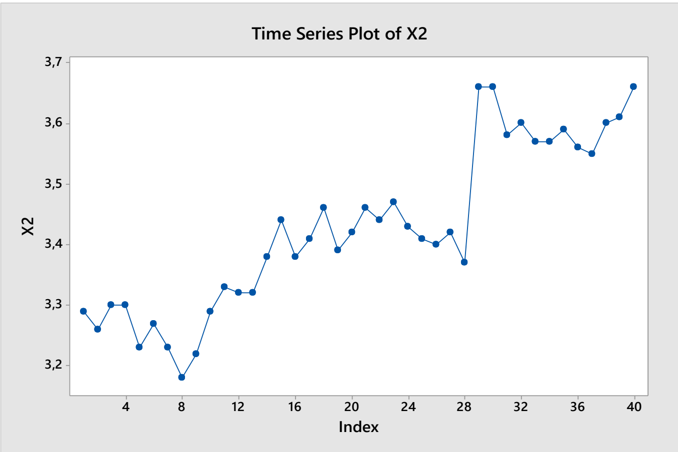
A start-up company based in the Silicon Valley wants to monitor a popularity index related to the number of "likes" on their Facebook page. The index values recorded on a weekly basis for 40 weeks is reported below.

Week	Index	Week	Index	Week	Index	Week	Index
1	3,29	11	3,33	21	3,46	31	3,58
2	3,26	12	3,32	22	3,44	32	3,60
3	3,30	13	3,32	23	3,47	33	3,57
4	3,30	14	3,38	24	3,43	34	3,57
5	3,23	15	3,44	25	3,41	35	3,59
6	3,27	16	3,38	26	3,40	36	3,56
7	3,23	17	3,41	27	3,42	37	3,55
8	3,18	18	3,46	28	3,37	38	3,60
9	3,22	19	3,39	29	3,66	39	3,61
10	3,29	20	3,42	30	3,66	40	3,66

1) Design a suitable control chart to monitor the popularity index. Discuss the results.
2) An additional information is that on the 29^{th} week, the company uploaded a special video on Facebook to celebrate its second anniversary. Was this video upload successful? Only for that day or even in the following ones? How does the control chart design change if this additional information is included? Discuss the results.
3) Using the model estimated at point b), design an interval prediction for the popularity index to be expected next week.

Exercise 1 (solution)

1)

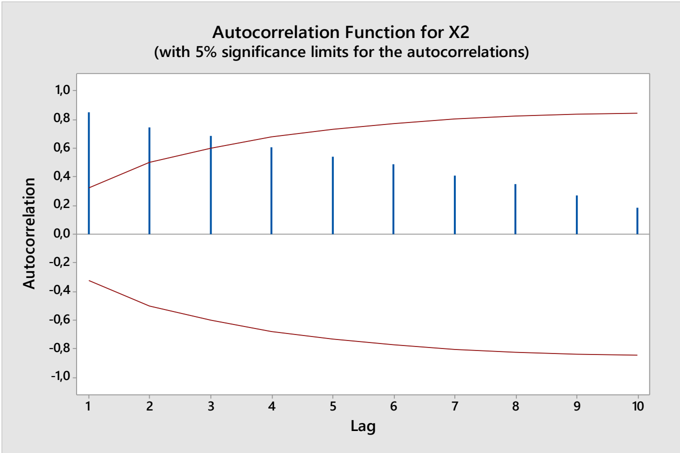


Data seem to be autocorrelated and nonstationary. Runs test confirms the nonrandom pattern observed.

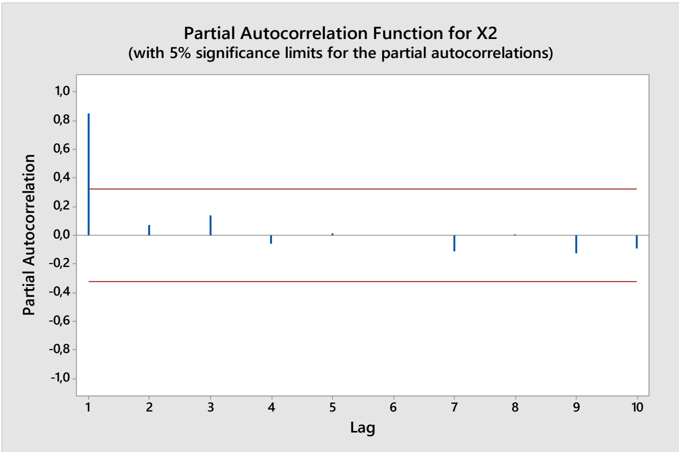
```
Runs test for X2
```

```
Runs above and below K = 3,42575

The observed number of runs = 8
The expected number of runs = 20,8
18 observations above K; 22 below
P-value = 0,000
```

Let’s check for autocorrelation (even if we know that sudden shifts can suggest autocorrelation even if it is not present in the data).



.



It seems that autocorrelation is present. The decrease of the ACF is almost linear. It could even ask for an Integrated component but we know that non-stationarity can be also due to a mean level shift (which seems to characterize data from the 29th observation on). This is why we can use stepwise regression to deepen the analysis and check whether the AR and the week time index are affecting the process.

Regression Analysis: Index versus ar; Week

```
Method

Rows unused  1


Stepwise Selection of Terms
```

α to enter = 0,15; α to remove = 0,15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	0,61018	0,305092	108,01	0,000
ar	1	0,02535	0,025354	8,98	0,005
Week	1	0,03808	0,038085	13,48	0,001
Error	36	0,10169	0,002825		
Total	38	0,71188			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0531487	85,71%	84,92%	81,44%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1,797	0,469	3,83	0,000	
ar	0,439	0,146	3,00	0,005	5,12
Week	0,00628	0,00171	3,67	0,001	5,12

Regression Equation

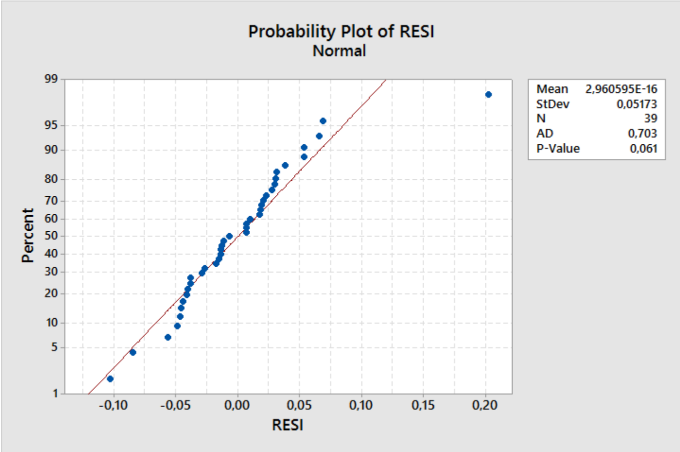
Index = 1,797 +†0,439†ar +†0,00628†Week

Fits and Diagnostics for Unusual Observations

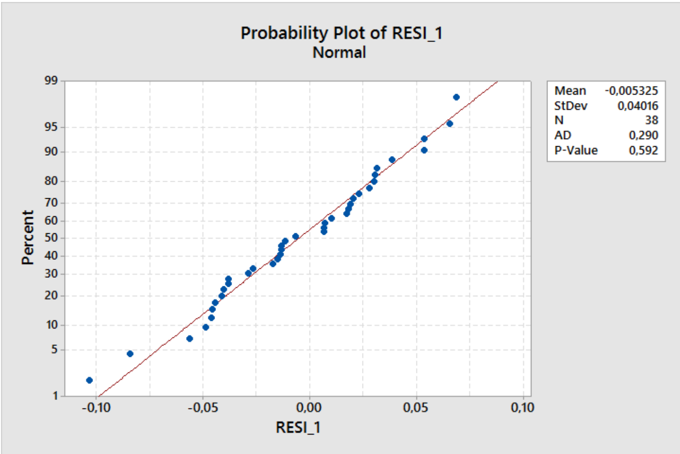
Obs	Index	Fit	Resid	Std Resid	
28	3,3700	3,4733	-0,1033	-2,02	R
29	3,6600	3,4577	0,2023	4,19	R

R Large residual

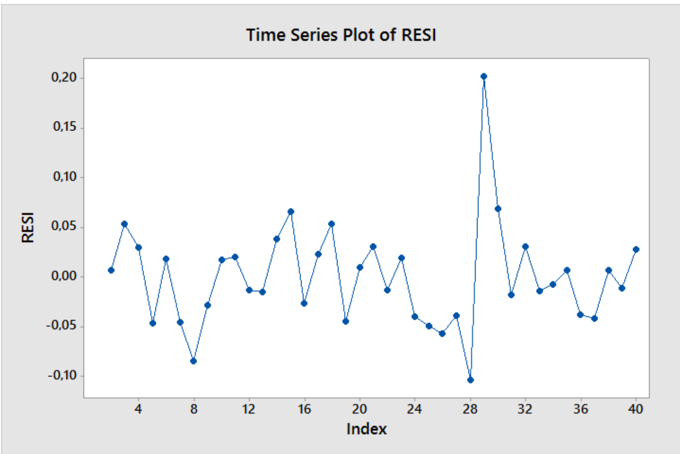
Residuals Check



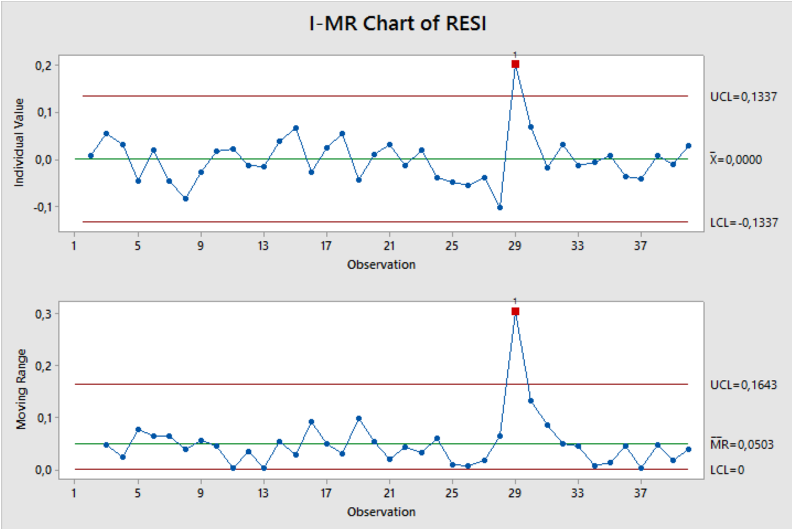
It seems that the weak non-normality is due to one outlying observations. As a matter of fact, by getting rid of the outlying data we have:



The residuals are:



And the control charts result



There is one outlying data at the 29th week.

2) If a new video was posted on FB at week 29th, this information can be used as regressor in a new model. Two different solutions are now compared. In the first, a dummy variable equal to 0 before the 28th week and equal to 1 from week 29 on can be added. This dummy is called "new video_step". The second dummy is equal to 1 only at the 29th week and 0 elsewhere. This dummy is called "new video_impulse". We will now compare the two models.

Let's start with the model where a dummy "new video_step" is possibly included.

Regression Analysis: Index versus Week; ar; new video_step

Method

Rows unused 1

Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0,62715	0,209050	86,36	0,000
Week	1	0,02043	0,020429	8,44	0,006
ar	1	0,01295	0,012951	5,35	0,027
new video_step	1	0,01696	0,016964	7,01	0,012
Error	35	0,08473	0,002421		
Total	38	0,71188			

```
Model Summary

        S    R-sq  R-sq(adj)  R-sq(pred)
0,0492017  88,10%     87,08%      82,41%


Coefficients

Term                 Coef  SE Coef  T-Value  P-Value   VIF
Constant            2,181    0,458     4,77    0,000
Week              0,00486  0,00167     2,90    0,006  5,71
ar                  0,328    0,142     2,31    0,027  5,61
new video_step     0,0788   0,0297     2,65    0,012  3,04


Regression Equation

Index = 2,181 +†0,00486†Week +†0,328†ar +†0,0788†new†video_step


Fits and Diagnostics for Unusual Observations

Obs   Index     Fit    Resid  Std Resid
  8  3,1800  3,2796  -0,0996      -2,13  R
 29  3,6600  3,5063   0,1537       3,77  R  X

R  Large residual
X  Unusual X
```

In this case, considering the family error rate=10%, we should remove tha ar regressor. Let’s fit a second model.

Regression Analysis: Index versus Week; new video_step

```
Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15


Analysis of Variance

Source             DF   Adj SS    Adj MS  F-Value  P-Value
Regression          2  0,63142  0,315709   117,57    0,000
  Week              1  0,10592  0,105918    39,44    0,000
  new video_step    1  0,03324  0,033238    12,38    0,001
Error              37  0,09936  0,002685
Total              39  0,73078
```

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0518208	86,40%	85,67%	84,15%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3,2444	0,0196	165,89	0,000	
Week	0,00733	0,00117	6,28	0,000	2,71
new video_step	0,1035	0,0294	3,52	0,001	2,71

Regression Equation

Index = 3,2444 +†0,00733†Week +†0,1035†new†video_step

Fits and Diagnostics for Unusual Observations

Obs	Index	Fit	Resid	Std Resid	
8	3,1800	3,3031	-0,1231	-2,45	R
29	3,6600	3,5605	0,0995	2,02	R

R Large residual

In the second case, we include the dummy called "new video_impulse" as possible regressor.

Regression Analysis: Index versus Week; ar; new video_impulse

Method

Rows unused 1

Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	0,659751	0,219917	147,66	0,000
Week	1	0,008843	0,008843	5,94	0,020
ar	1	0,053434	0,053434	35,88	0,000
new video_impulse	1	0,049566	0,049566	33,28	0,000
Error	35	0,052126	0,001489		
Total	38	0,711877			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0385918	92,68%	92,05%	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1,004	0,367	2,74	0,010	
Week	0,00328	0,00135	2,44	0,020	6,02
ar	0,687	0,115	5,99	0,000	5,96
new video_impulse	0,2450	0,0425	5,77	0,000	1,18

Regression Equation

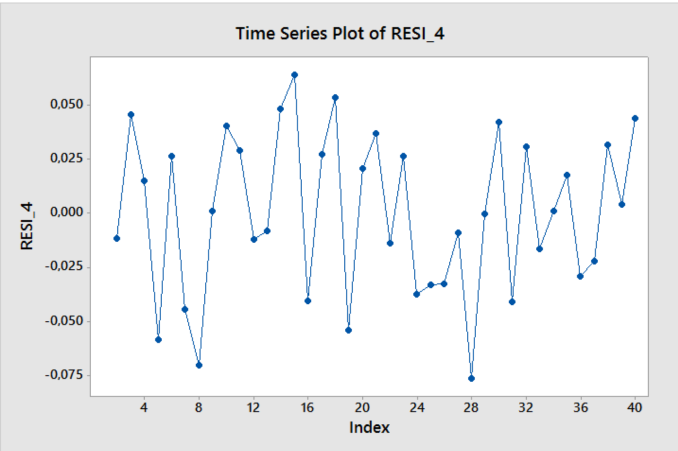
Index = 1,004 +†0,00328†Week +†0,687†ar +†0,2450†new†video_impulse

Fits and Diagnostics for Unusual Observations

Obs	Index	Fit	Resid	Std Resid	
28	3,3700	3,4461	-0,0761	-2,07	R
29	3,6600	3,6600	0,0000	*	X

R Large residual
X Unusual X

It seems that this second model fits better the data. Let's check the residuals.



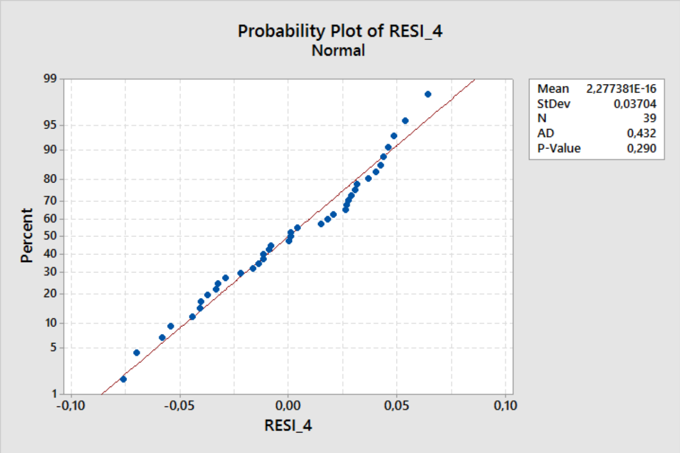
Runs Test: RESI_4

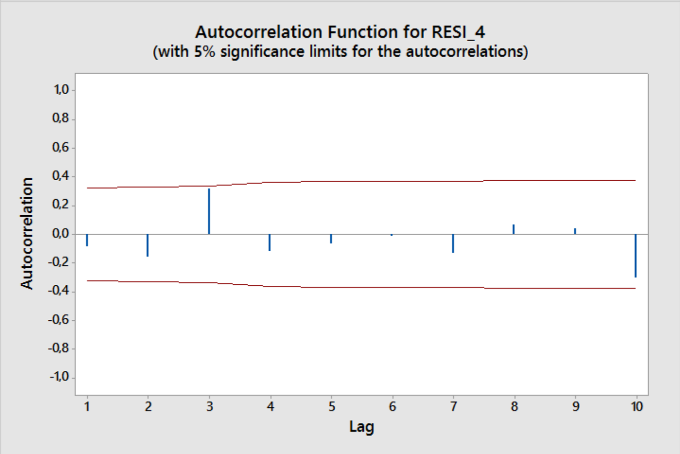
Runs test for RESI_4

Runs above and below K = 2,277381E-16

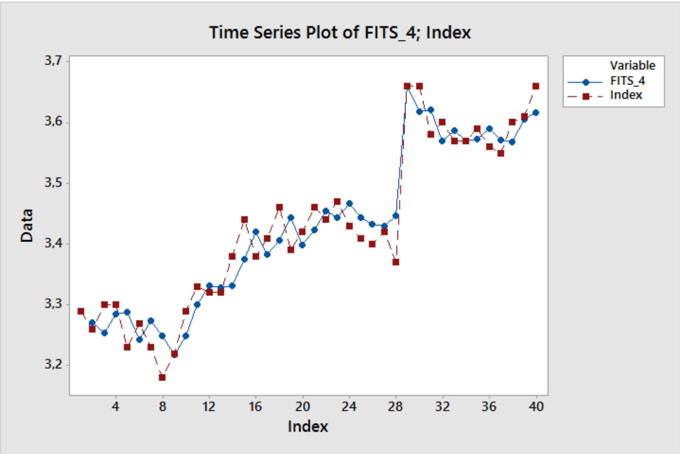
The observed number of runs = 22
The expected number of runs = 20,4872

```
20 observations above K; 19 below
P-value = 0,623
```

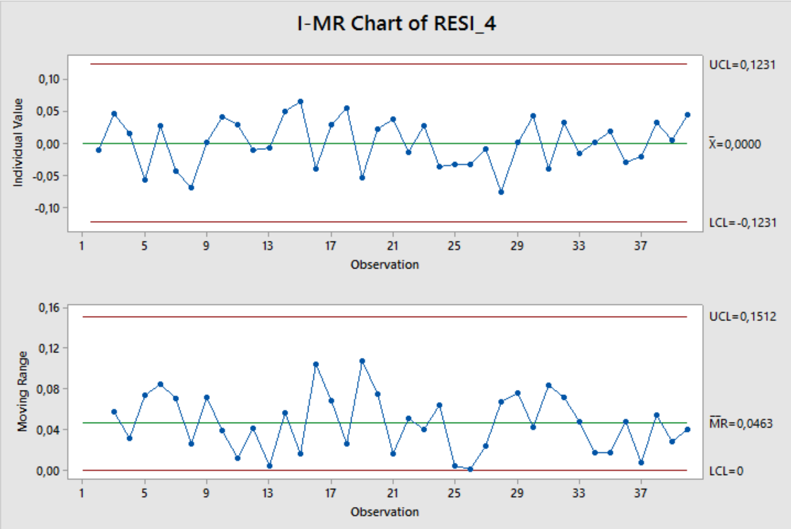




Fitted Value Chart:



Special Cause chart:



3) Interval prediction for the next week (week 41):

Prediction for Index

Regression Equation

Index = 1,004 +†0,00328†Week +†0,687†ar +†0,2450†new video_impulse

Variable	Setting
Week	41
ar	3,61
new video_impulse	0

Fit	SE Fit	95% CI	95% PI
3,61931	0,0131874	(3,59254; 3,64608)	(3,53652; 3,70210)

Exercise 2 (max score 12)

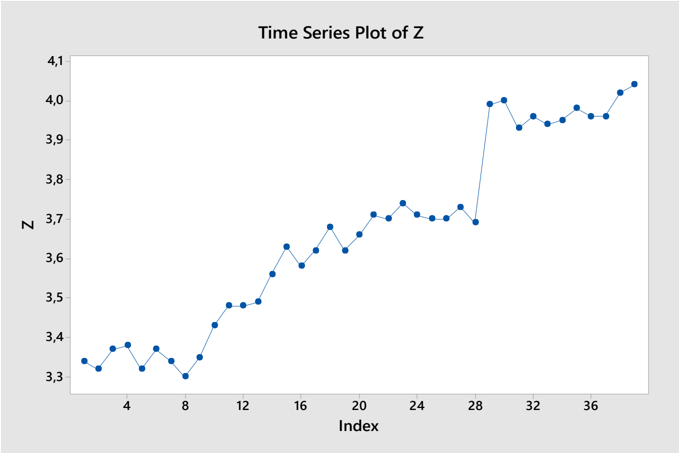
A company has recently started monitoring the percentage of metal powder that is not recovered for recycling and reuse in an Additive Manufacturing process. One value per week has been recorded in the last 39 weeks.

Week	Powder loss (%)	Week	Powder loss (%)
1	3,34	21	3,71
2	3,32	22	3,7
3	3,37	23	3,74
4	3,38	24	3,71
5	3,32	25	3,7
6	3,37	26	3,7
7	3,34	27	3,73
8	3,3	28	3,69
9	3,35	29	3,99
10	3,43	30	4
11	3,48	31	3,93
12	3,48	32	3,96
13	3,49	33	3,94
14	3,56	34	3,95
15	3,63	35	3,98
16	3,58	36	3,96
17	3,62	37	3,96
18	3,68	38	4,02
19	3,62	39	4,04
20	3,66		

1) Design a suitable control chart to monitor the powder loss over time. Discuss the results.
2) Assuming that a known event occurred at week 29, re-design the chart.
3) Estimate an interval prediction for the expected powder loss at week 40.

Exercise 2 (solution)

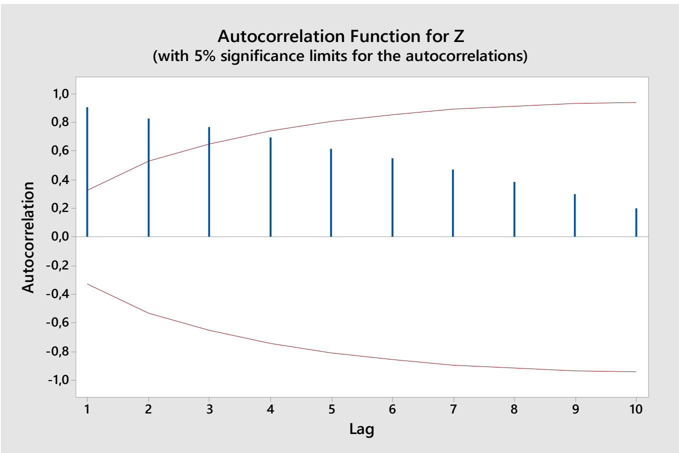
1) Time series plot:

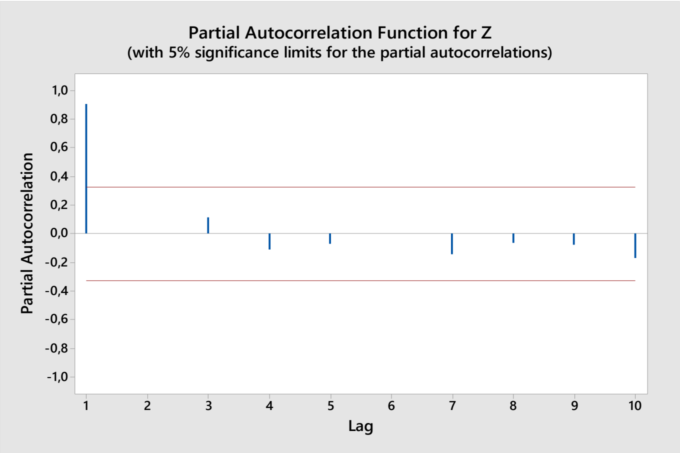


Data seem to be autocorrelated and nonstationary (increasing trend). Runs test confirms the non-random pattern observed.

```
Runs test for Z

Runs above and below K = 3,65974

The observed number of runs = 4
The expected number of runs = 20,3846
21 observations above K; 18 below
P-value = 0,000
```

Let’s check the autocorrelation:





It seems that autocorrelation is present. The decrease of the ACF is almost linear. It may be the consequence of a non-stationary behaviour (trend) but we know that non-stationarity can be also due to a mean level shift (which seems to characterize data from the 29th observation on).

By applying a simple regression model using the week as predictor:

Regression Analysis: Z versus week

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2,0723	2,07235	581,98	0,000
week	1	2,0723	2,07235	581,98	0,000
Error	37	0,1318	0,00356		
Total	38	2,2041			

Model Summary

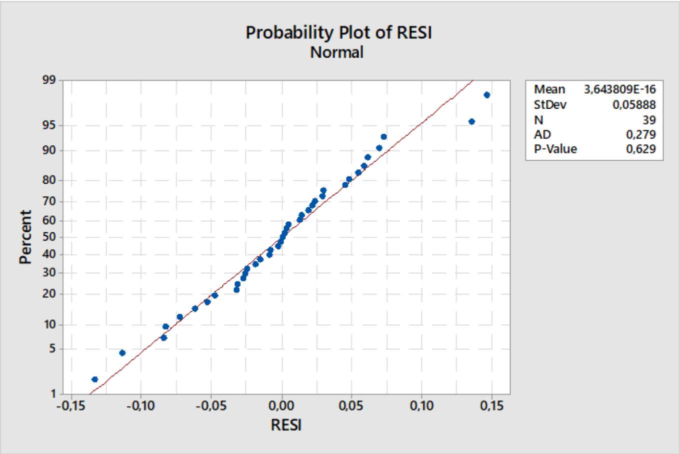
S	R-sq	R-sq(adj)	R-sq(pred)
0,0596727	94,02%	93,86%	93,40%

Coefficients

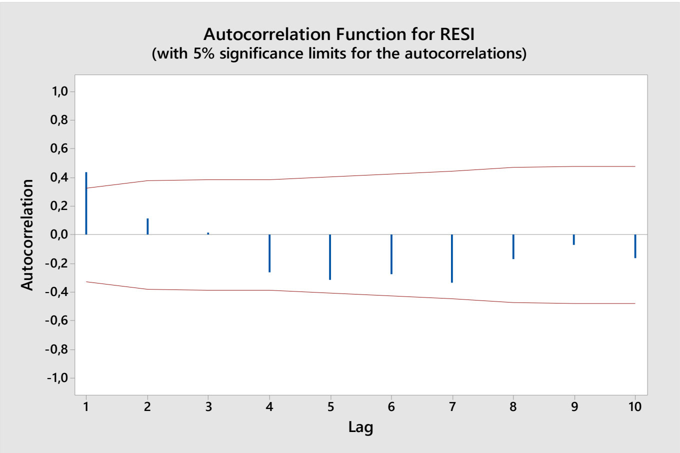
Term Coef SE Coef T-Value P-Value VIF

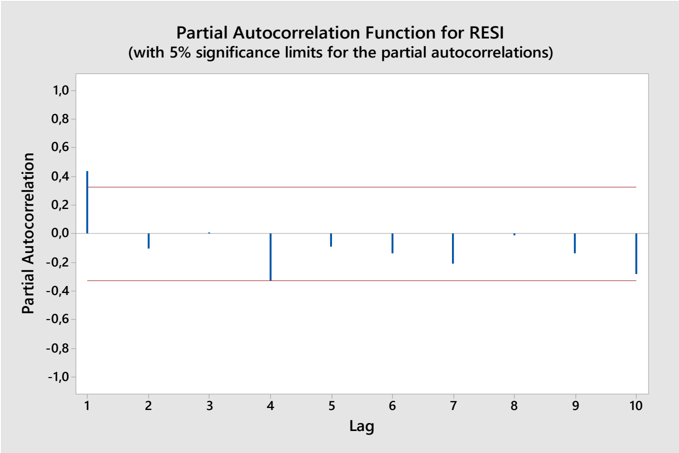
```
Constant     3,2501     0,0195    166,81     0,000
week       0,020482  0,000849     24,12     0,000   1,00


Regression Equation

Z = 3,2501 + 0,020482 week
```

Residuals are normal:



But some autoregressive effect is still present:





By fitting a regression model where both the trend (week) term and an AR(1) term are present we get:

Regression Analysis: Z versus week; AR(1)

Method

Rows unused 1

Analysis of Variance

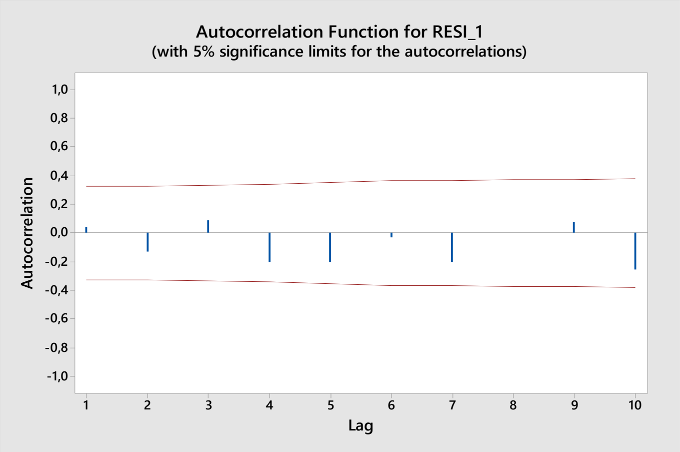
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1,99835	0,999173	346,85	0,000
week	1	0,04029	0,040292	13,99	0,001
AR(1)	1	0,02558	0,025580	8,88	0,005
Error	35	0,10083	0,002881		
Total	37	2,09917			

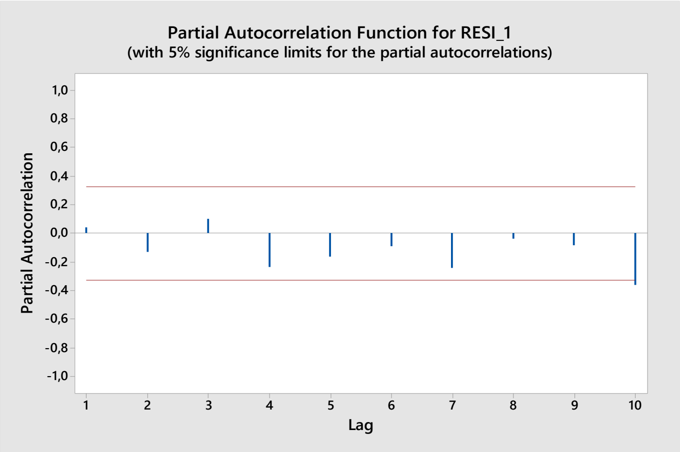
Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0536724	95,20%	94,92%	93,69%

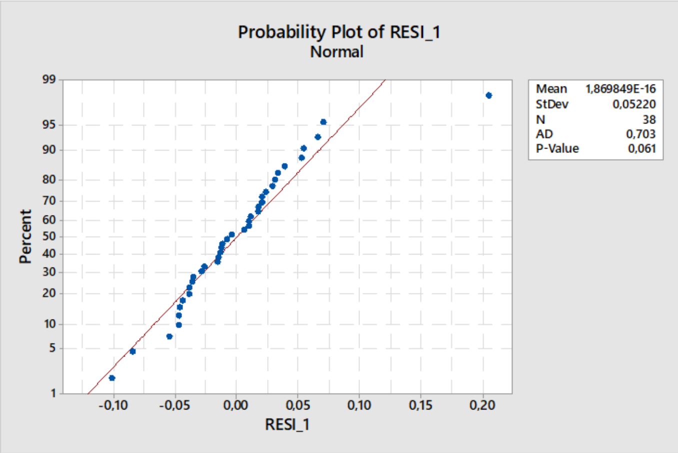
```
Coefficients

Term           Coef  SE Coef  T-Value  P-Value    VIF
Constant      1,819    0,478     3,81    0,001
week        0,01173  0,00314     3,74    0,001  15,61
AR(1)         0,441    0,148     2,98    0,005  15,61

Regression Equation
Z = 1,819 + 0,01173 week + 0,441 AR(1)
```

Now the residuals are not auto-correlated:

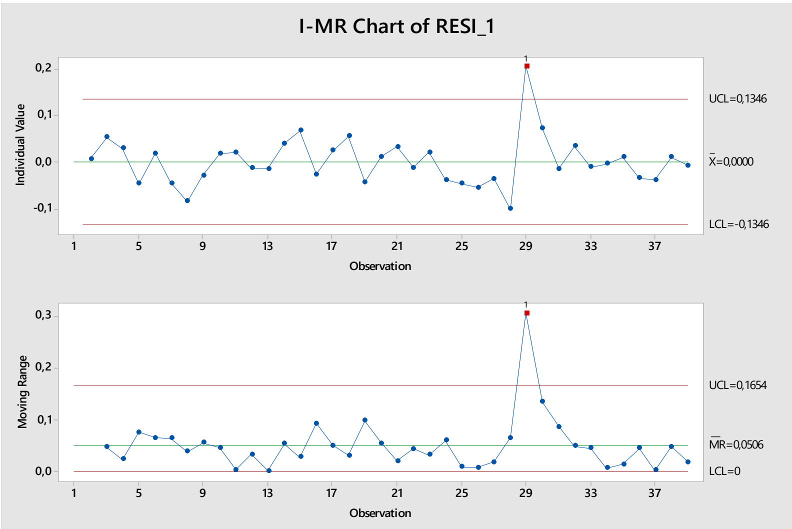




The residuals are barely normal because of the outlying effect of week 29, where the shift affected the original time series. However, the normality assumption can be accepted:



The resulting control chart is:



The control chart signals an alarm at week 29.

2) Considering the information about an assignable cause at week 29, a dummy variable can be included in the model (=0 always apart from week 29, where dummy=1).

The result is:

Regression Analysis: Z versus week; AR(1); Dummy

```
Method


Categorical predictor coding  (1; 0)
Rows unused                   1
```

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	2,04920	0,683068	464,79	0,000
week	1	0,00909	0,009093	6,19	0,018
AR(1)	1	0,05442	0,054419	37,03	0,000
Dummy	1	0,05086	0,050858	34,61	0,000
Error	34	0,04997	0,001470		
Total	37	2,09917			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0,0383357	97,62%	97,41%	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1,003	0,369	2,72	0,010	
week	0,00607	0,00244	2,49	0,018	18,50
AR(1)	0,694	0,114	6,09	0,000	18,21
Dummy					
1	0,2489	0,0423	5,88	0,000	1,19

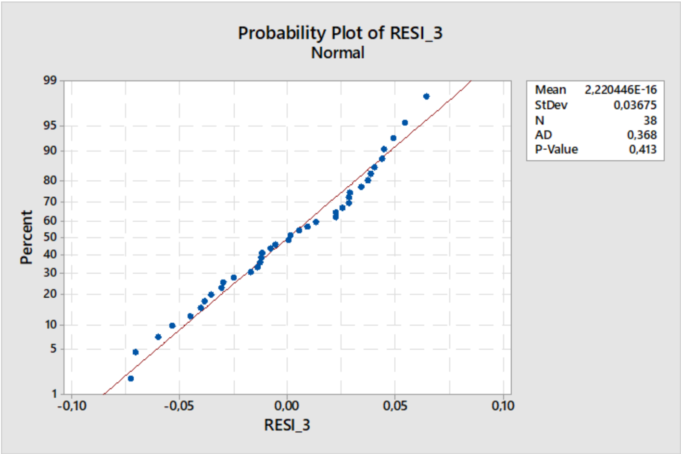
Regression Equation

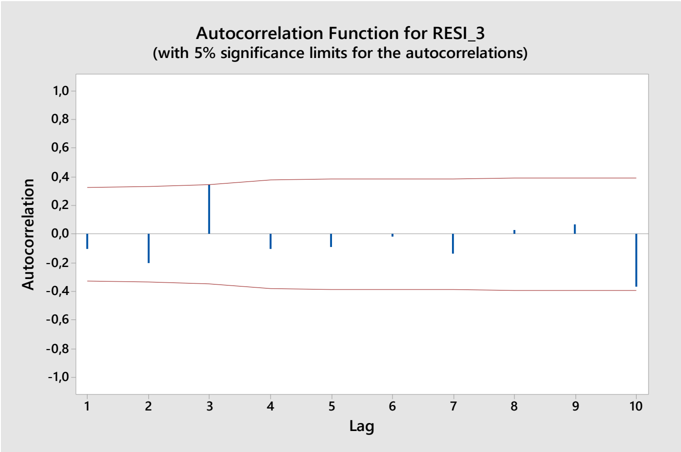
Dummy

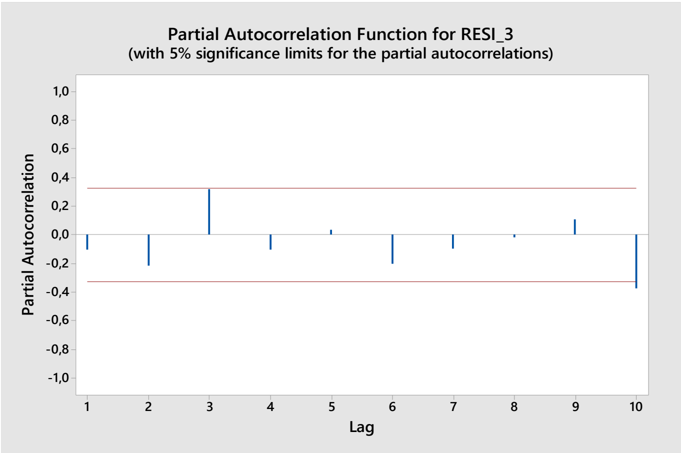
0 Z = 1,003 + 0,00607 week + 0,694 AR(1)

1 Z = 1,252 + 0,00607 week + 0,694 AR(1)

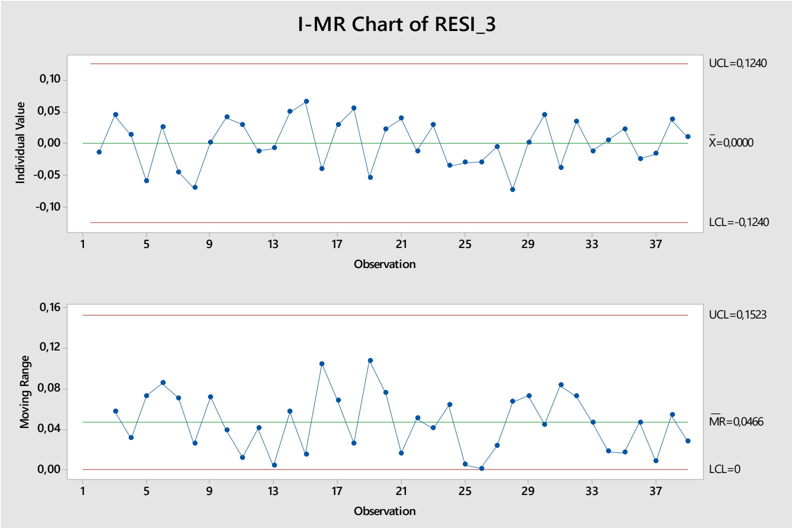
Residuals are normal and independent:







The resulting control chart is:



No alarm.

3)

Prediction for Z

```
Regression Equation


Z = 1,003 + 0,00607 week + 0,694 AR(1) + 0,000000 Dummy_0
+ 0,2489 Dummy_1



Variable  Setting
week           40
AR(1)        4,04
Dummy           0



    Fit     SE Fit        95% CI                 95% PI
4,05090  0,0130959  (4,02428; 4,07751)  (3,96857; 4,13323)
```

Exercise 3 (max score 13)

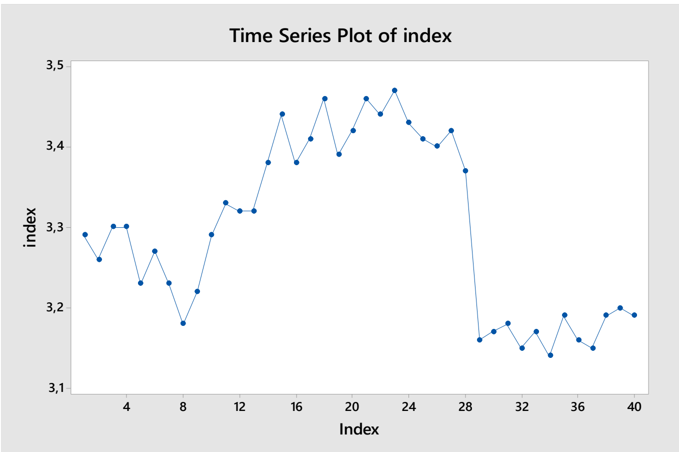
A company that develops high-precision machine tools is able to keep under control the health condition of their systems by monitoring a synthetic index based on sensor signals acquired during repeated operation cycles. The table reports the values of the synthetic index measured in 40 consecutive replicates of the same cycle performed by a single machine tool.

Cycle	Index	Cycle	Index	Cycle	Index	Cycle	Index
1	3,29	11	3,33	21	3,46	31	3,18
2	3,26	12	3,32	22	3,44	32	3,15
3	3,30	13	3,32	23	3,47	33	3,17
4	3,30	14	3,38	24	3,43	34	3,14
5	3,23	15	3,44	25	3,41	35	3,19
6	3,27	16	3,38	26	3,40	36	3,16
7	3,23	17	3,41	27	3,42	37	3,15
8	3,18	18	3,46	28	3,37	38	3,19
9	3,22	19	3,39	29	3,16	39	3,20
10	3,29	20	3,42	30	3,17	40	3,19

1) Design a suitable control chart to monitor the health condition of the machine tool. Discuss the results.
2) A maintenance intervention was performed after the 28^{th} cycle. How does the control chart design change if this additional information is included? Discuss the results.
3) Did the maintenance intervention have a significant effect on the synthetic index? Use a statistical test, if needed.
4) Using the model estimated at point b), design a prediction interval for the synthetic index to be expected in the next cycle.

Exercise 3 (solution)

1)

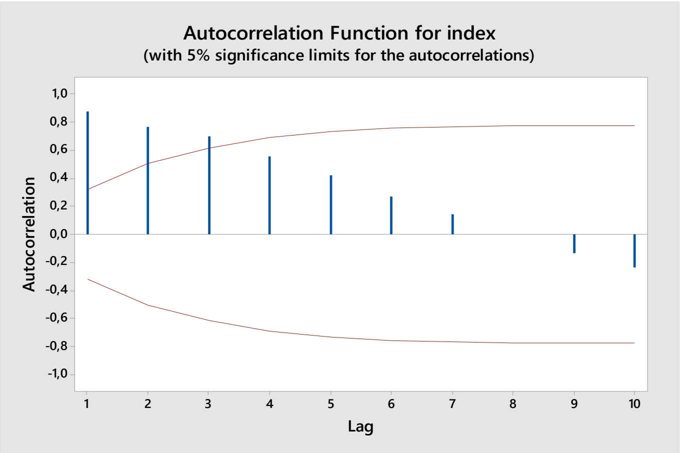


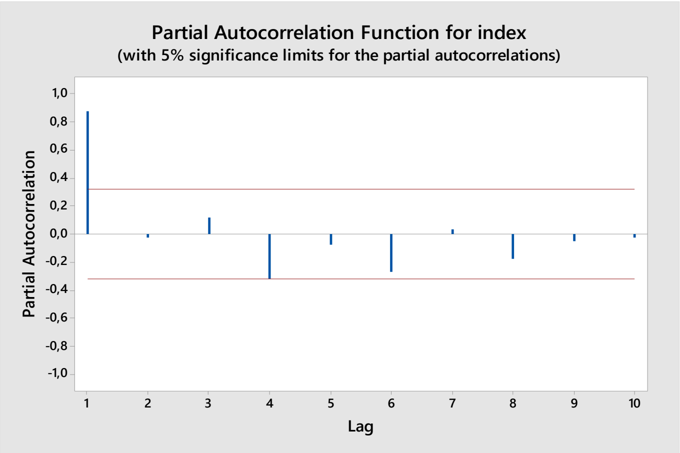
Data seem to be autocorrelated and nonstationary. Runs test confirms the non-random pattern observed.

Runs Test: index

```
Runs test for index


Runs above and below K = 3,29675


The observed number of runs = 5
The expected number of runs = 21
20 observations above K; 20 below
P-value = 0,000
```

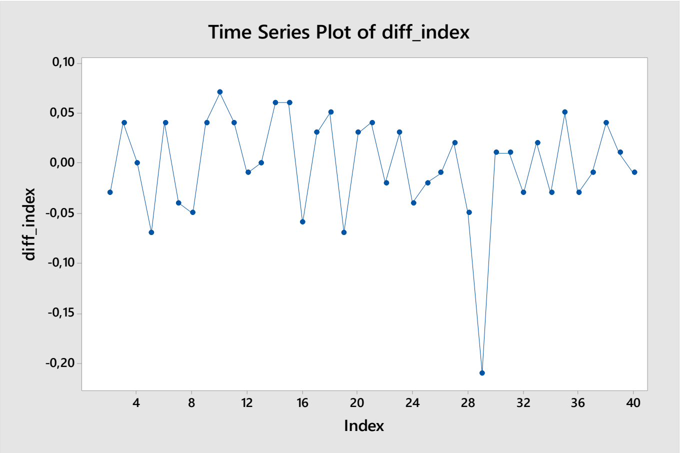
Let's check for autocorrelation (even if we know that sudden shifts can suggest autocorrelation even when it is not present in the data).

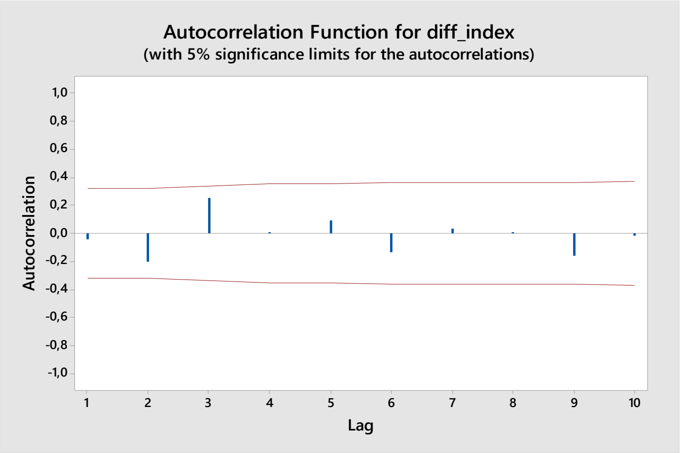


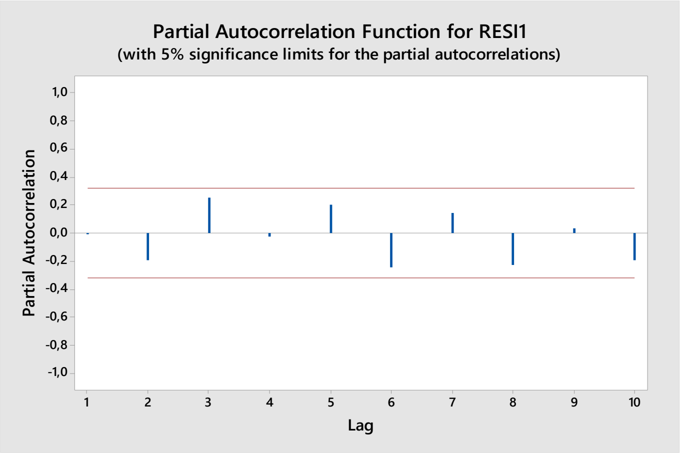


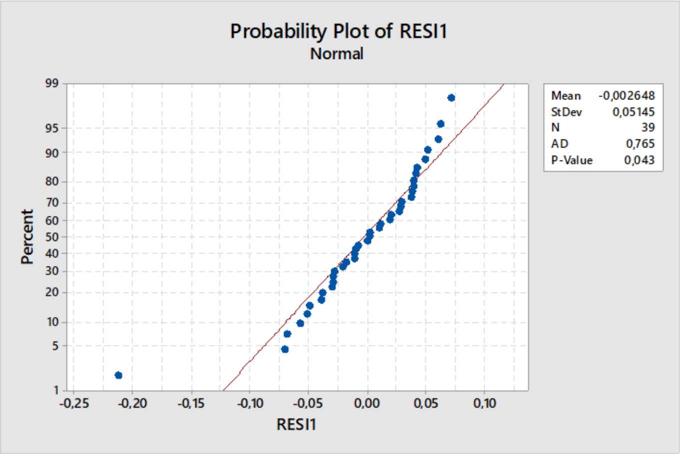
The decrease of the ACF is almost linear. It could even ask for an Integrated component (we also know that nonstationarity can be due to a mean level shift, which seems to characterize data from the 29th observation on).

By applying the differencing operator we get a process that is not auto-correlated and barely normal. Thus, the best model for this time series is a random walk (the same result could be achieved by fitting an AR(1) model to the original data).

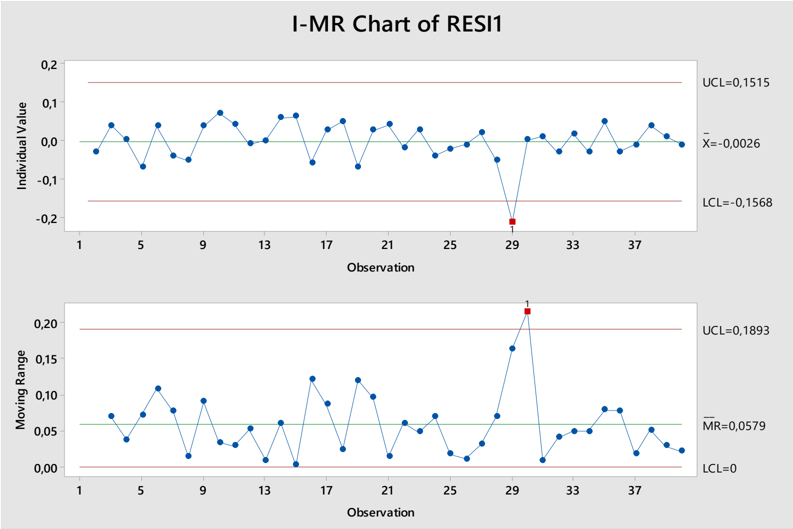








The resulting control chart is:



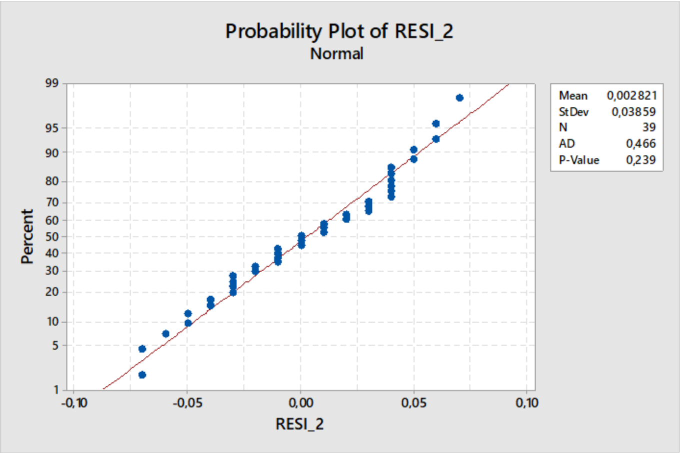
There is one outlying data at the 29th cycle. In the absence of information about assignable causes, the design step is over.

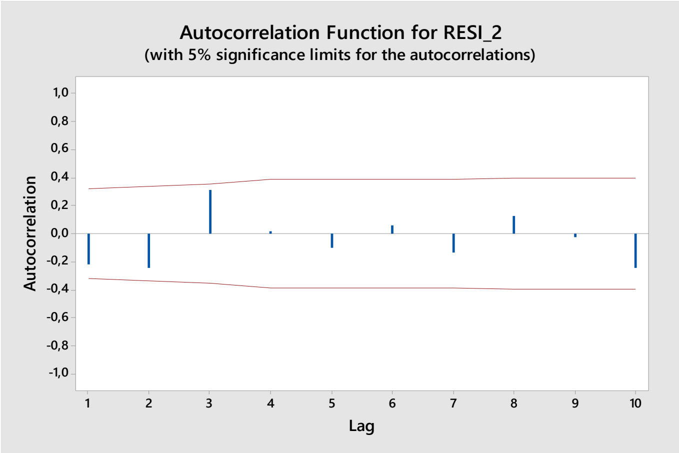
2) The knowledge about the maintenance intervention after cycle 28 represents an assignable cause, which allows introducing a dummy variable into the random walk model:

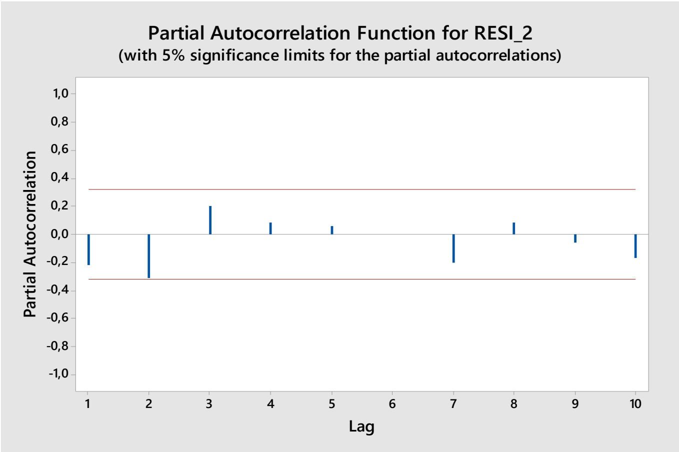
```
Analysis of Variance


Source          DF     Adj SS     Adj MS  F-Value  P-Value
Regression       1   0,044100   0,044100    29,45    0,000
  dummy          1   0,044100   0,044100    29,45    0,000
Error           38   0,056900   0,001497
  Lack-of-Fit    1   0,000318   0,000318     0,21    0,651
  Pure Error    37   0,056582   0,001529
Total           39   0,101000
```

```
Model Summary

        S     R-sq  R-sq(adj)  R-sq(pred)
0,0386958  43,66%     42,18%           *


Coefficients

Term       Coef  SE Coef  T-Value  P-Value   VIF
dummy
  1     -0,2100   0,0387    -5,43    0,000  1,00


Regression Equation

diff_index = 0,0 dummy_0 - 0,2100 dummy_1
```

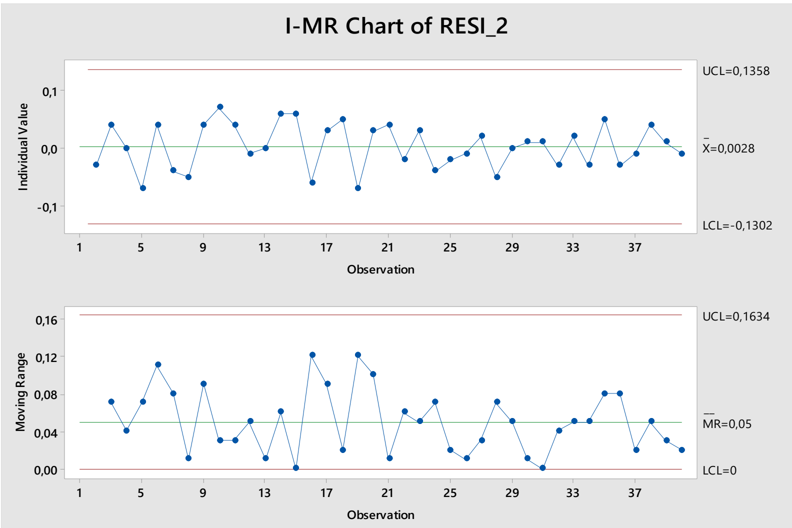
Residuals checking:







No violation is present. The resulting control chart is:



3) Statistical test for the mean change after cycle 28:

```
Two-sample T for index
```

dummy1	N	Mean	StDev	SE Mean
0	28	3,3507	0,0834	0,016
1	12	3,1708	0,0193	0,0056

```
Difference = μ (0) - μ (1)
Estimate for difference:  0,1799
95% lower bound for difference:  0,1516
T-Test of difference = 0 (vs >): T-Value = 10,76  P-Value = 0,000
DF = 32
```

The new mean of the synthetic index after the maintenance intervention is significantly lower than the mean before the intervention.

4) Interval prediction for the next cycle (cycle 41):

Prediction for Index

```
95% PI
(-0,0783356; 0,0783356)
```

Exercise 4 (max score 13)

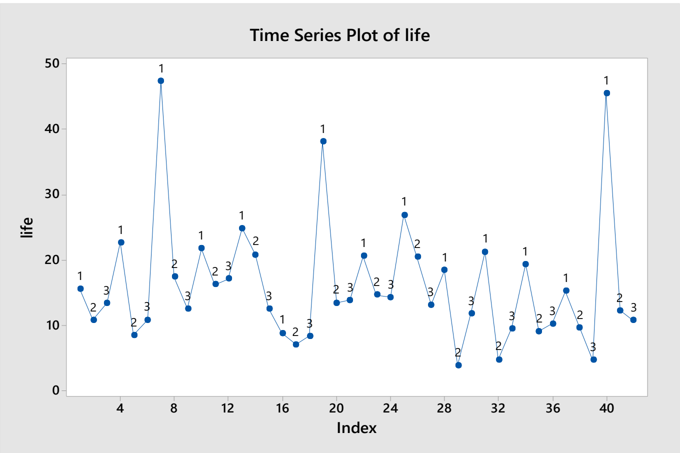
A company produces a critical component of the landing gear of the Airbus A320 (one part per week). The milling process has a very long duration and three copies of the same tool are sequentially used during the roughing phase. Whenever a tool copy reaches the end of life (based on a tool wear criterion), the actual tool life is recorded (in hours) as it can be used as a proxy of the stability of the process. The tool life values are reported in the table below for each tool copy used over a time period of 14 weeks.

life	tool_copy	week	life	tool_copy	week
15,6135	1	1	20,625	1	8
10,893	2	1	14,7045	2	8
13,446	3	1	14,331	3	8
22,7025	1	2	26,9425	1	9
8,571	2	2	20,511	2	9
10,9035	3	2	13,1865	3	9
47,3775	1	3	18,45	1	10
17,457	2	3	3,9105	2	10
12,6105	3	3	11,8515	3	10
21,7695	1	4	21,3015	1	11
16,2825	2	4	4,836	2	11
17,127	3	4	9,5775	3	11
24,8805	1	5	19,3695	1	12
20,871	2	5	9,1395	2	12
12,5835	3	5	10,317	3	12
8,8785	1	6	15,3045	1	13
7,143	2	6	9,6975	2	13
8,433	3	6	4,7145	3	13
38,139	1	7	45,54	1	14
13,506	2	7	12,294	2	14
13,8855	3	7	10,8225	3	14

Design a suitable control chart to determine if the process was in-control during the monitored period.

Exercise 4 (solution)

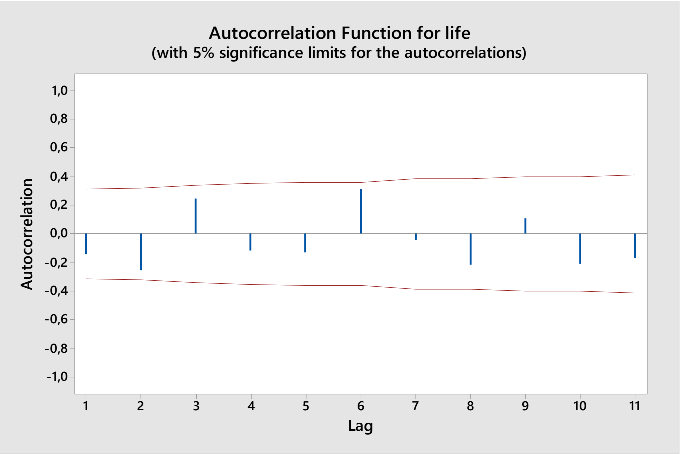
Time series plot:

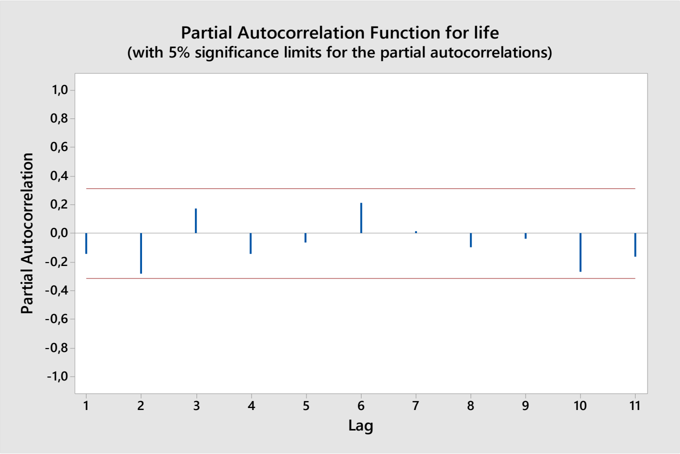


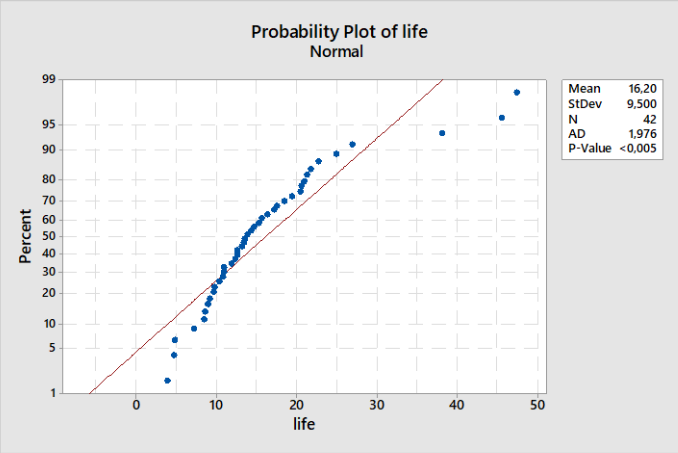
No evident trend is present, but there seems to be an effect of the tool copy (especially copy 1). Check of randomness:

Runs Test: life

```
Runs test for life


Runs above and below K = 16,2024


The observed number of runs = 21
The expected number of runs = 20,8095
16 observations above K; 26 below
P-value = 0,950
```

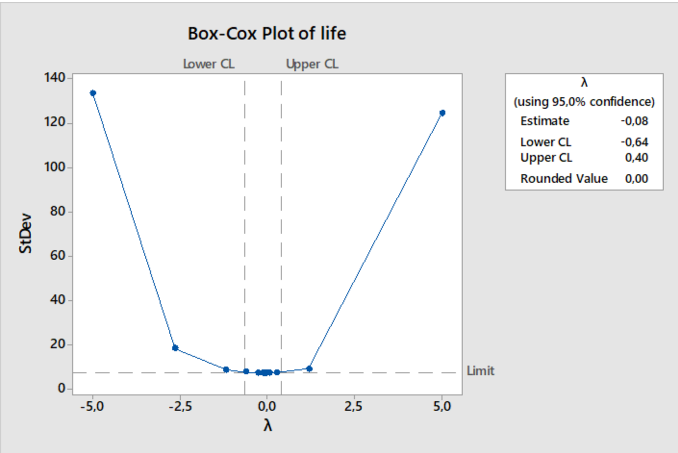


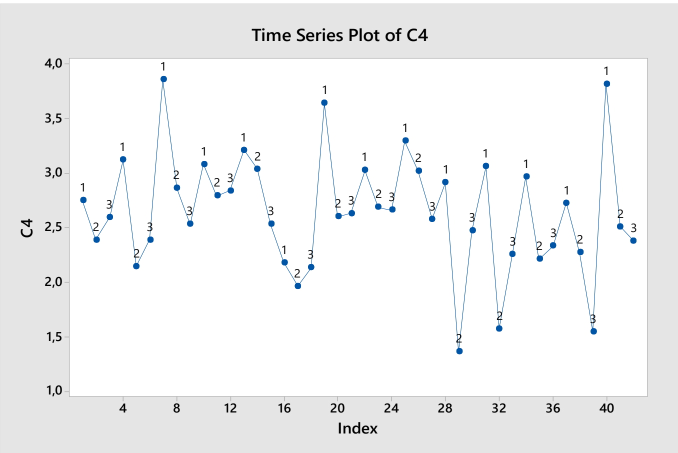


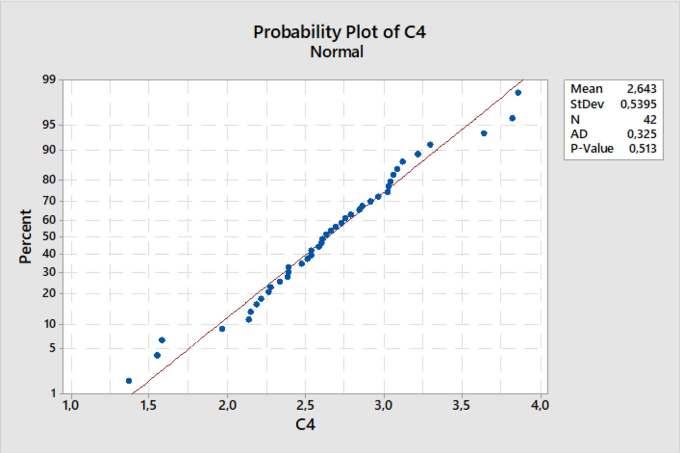


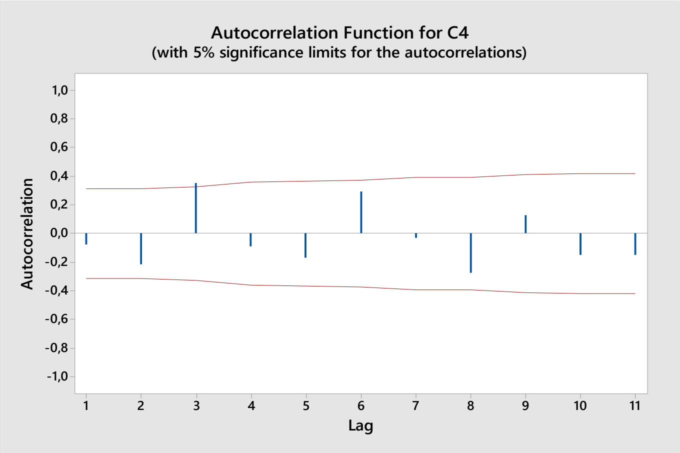
No violation of the randomness assumption is signaled, but the normality is violated. The cause is the systematic effect of the tool copy.

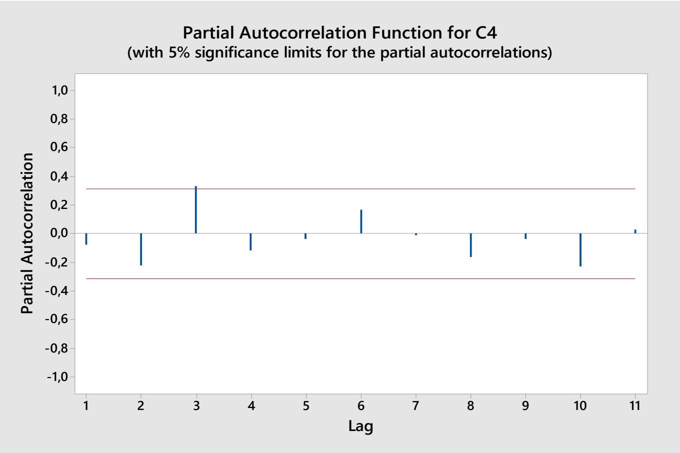
If we transform the data via Box-Cox and check the randomness of the transformed data, a weak effect of the tool copy arises in the ACF and PCAF functions (lag 3):











Let's define a dummy variable = 1 for tool copy 1 and = 0 for other tool copies and fit a regression model.

Regression Analysis: life_BC versus dummy

```
Method

Categorical predictor coding  (1; 0)

Analysis of Variance
```

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4,748	4,7482	26,42	0,000
dummy	1	4,748	4,7482	26,42	0,000
Error	40	7,187	0,1797		
Total	41	11,936			

Model Summary

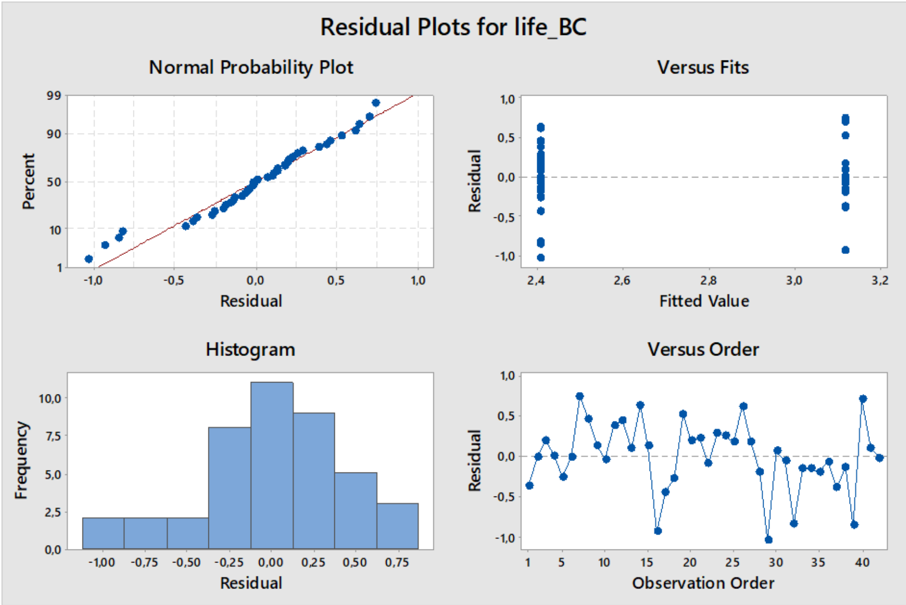
S	R-sq	R-sq(adj)	R-sq(pred)
0,423894	39,78%	38,28%	33,41%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2,4048	0,0801	30,02	0,000	
dummy					
1	0,713	0,139	5,14	0,000	1,00

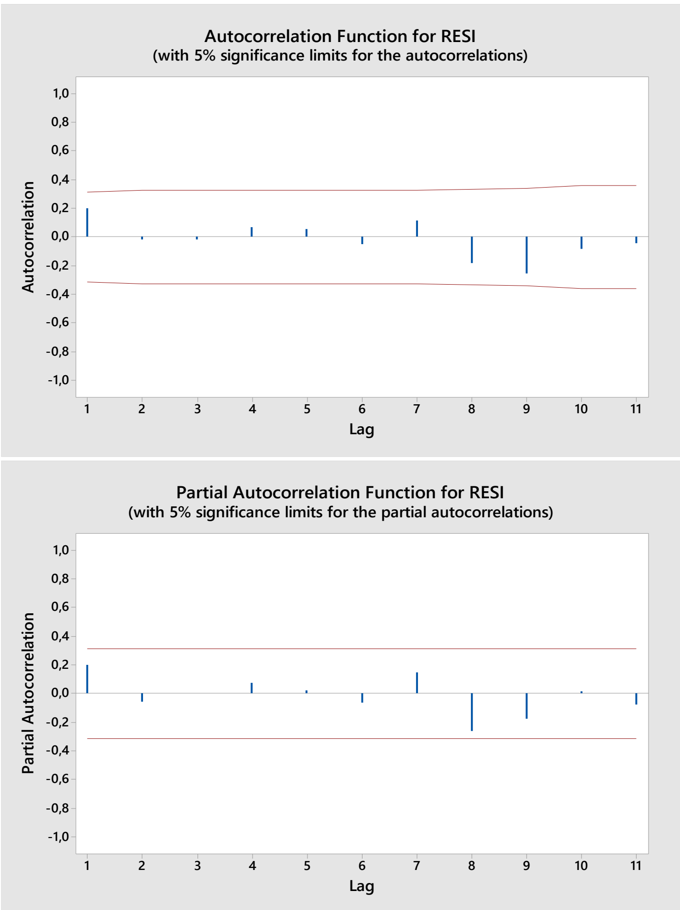
Regression Equation

life_BC = 2,4048 + 0,0 dummy_0 + 0,713 dummy_1



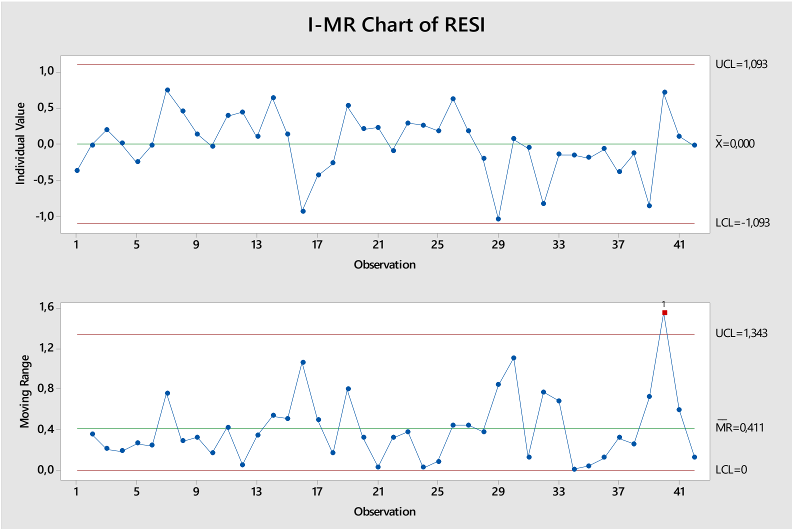
Normality test on residuals: p-value = 0.129

ACF and PCAF of residuals:

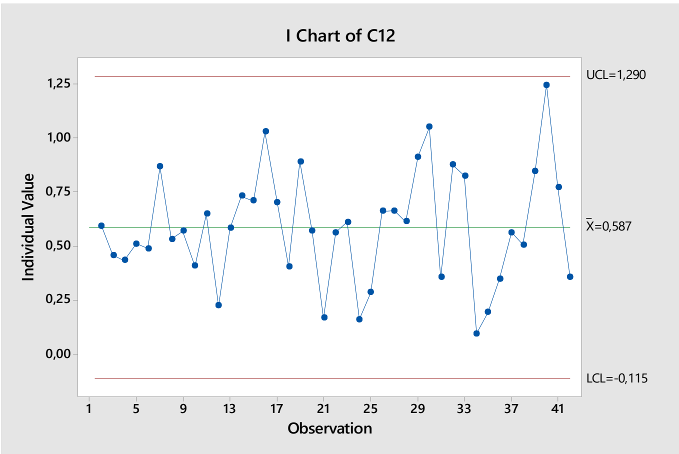


The residuals are ok. We can proceed with the control chart design.

Note: we can also check if the week (time) is significant. The result confirms that only the dummy variable is significant.



There is an out-of-control in the MR chart. To be sure that it is not caused by the violation of distributional assumptions for the MR statistic we can apply a control chart on the transformed MR via Box-Cox.



Now no out-of-control is signaled. However, there seems to be a funnel effect in the MR time series. Attention should be paid to this pattern.

Exercise 5 (max score 13)

A company wants to use SPC to monitor its bottle filling process. They took a random sample of one bottle from the production line and carefully measured the amount of liquid in the bottle. The results of 30 samplings are reported in the table below:

t	data	t	data	t	data
1	500.12	11	498.49	21	499.69
2	500.90	12	498.39	22	499.24
3	499.85	13	498.56	23	499.49
4	500.44	14	498.85	24	500.90
5	501.60	15	498.87	25	501.90
6	501.78	16	500.57	26	501.04
7	499.75	17	501.24	27	499.88
8	499.44	18	500.91	28	501.65
9	499.47	19	498.96	29	502.46
10	500.56	20	499.37	30	501.09

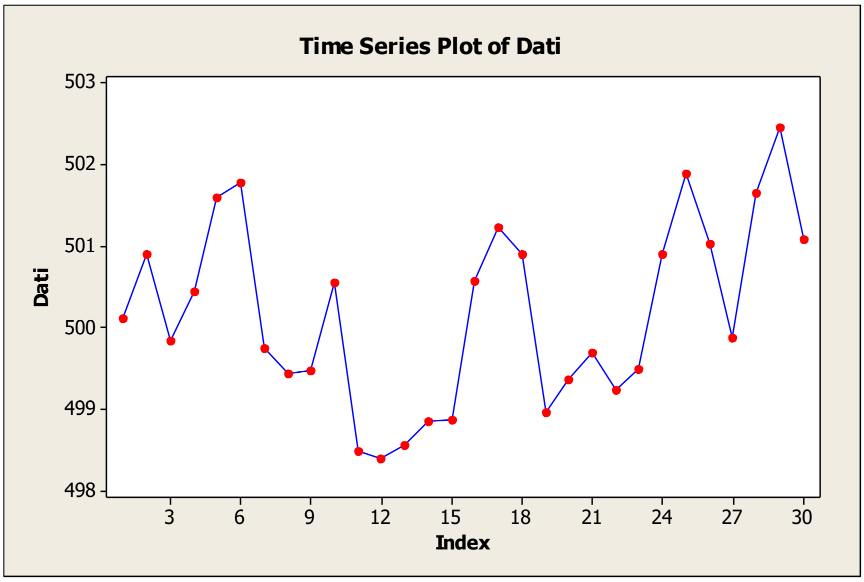
Assuming ARL_0 set to 200:

1) design an appropriate control chart to monitor the data.

Hint: among the possible models, choose the simplest one.

Exercise 5 (solution)

1)



From the time series plot there is no evident pattern (apart from some possible meandering)

Let’s make a runs test:

Runs Test: Dati

Runs test for Dati

Runs above and below K = 500,182

The observed number of runs = 12

The expected number of runs = 15,9333

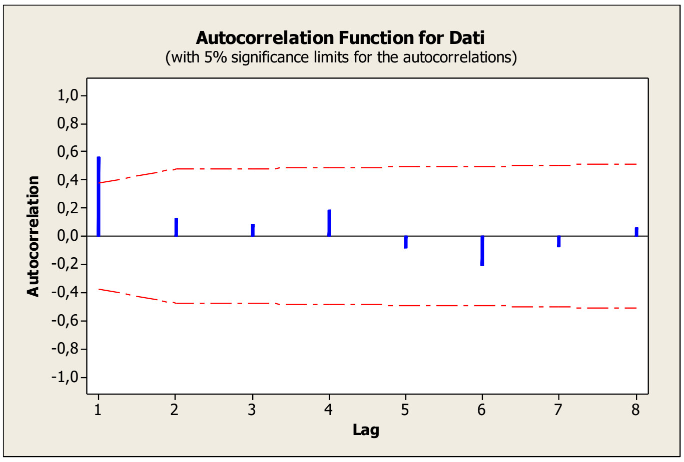
14 observations above K; 16 below

P-value = 0,142

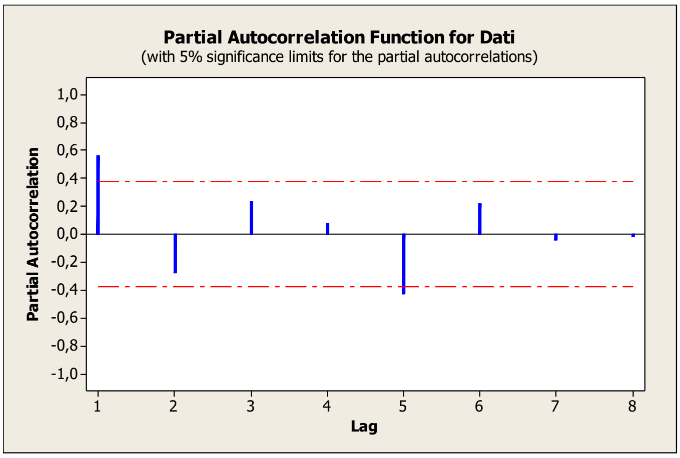
Given the p-value, we can not reject the null hypothesis (process randmoness).

By the way, let’s check the ACF and PACF too.

ACF



PACF



The ACF suggests MA(1) to be a possible model, whereas the PACF highlights a correlation at lags 1 and 5.

Let's fit the model with lowest complexity: MA(1).

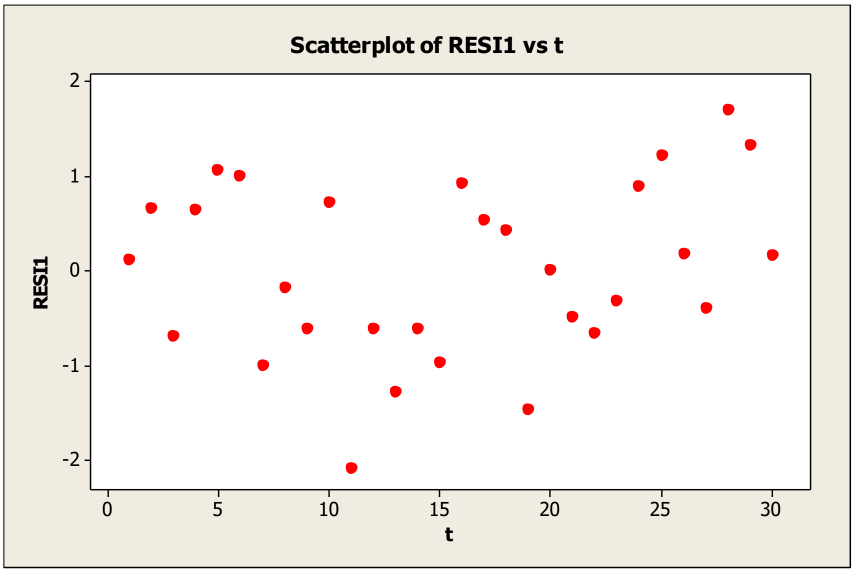
Type	Coef	SE Coef	T	P
MA 1	-0,5601	0,1559	-3,59	0,001
Constant	500,174	0,265	1885,97	0,000
Mean	500,174	0,265		

Number of observations: 30

Residuals: SS = 24,4933 (backforecasts excluded)

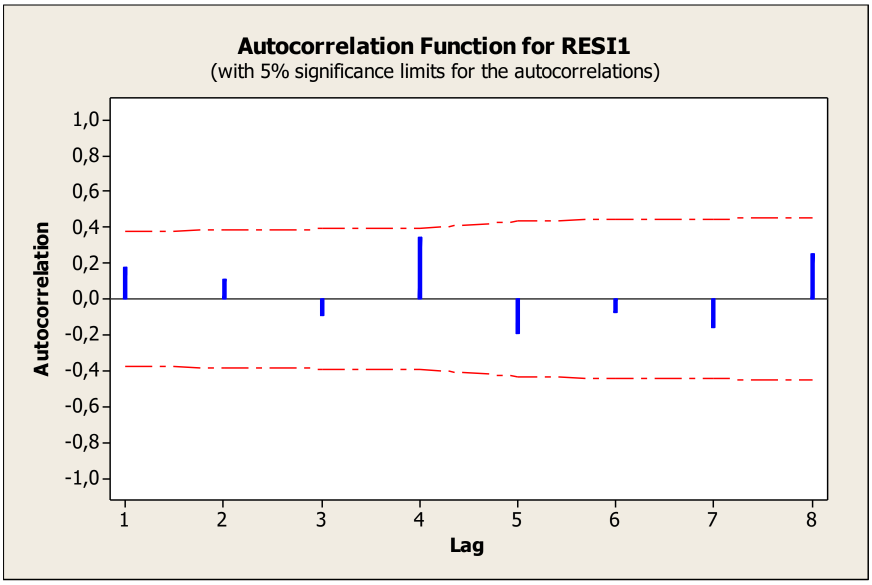
MS = 0,8748 DF = 28

Residuals versus order.

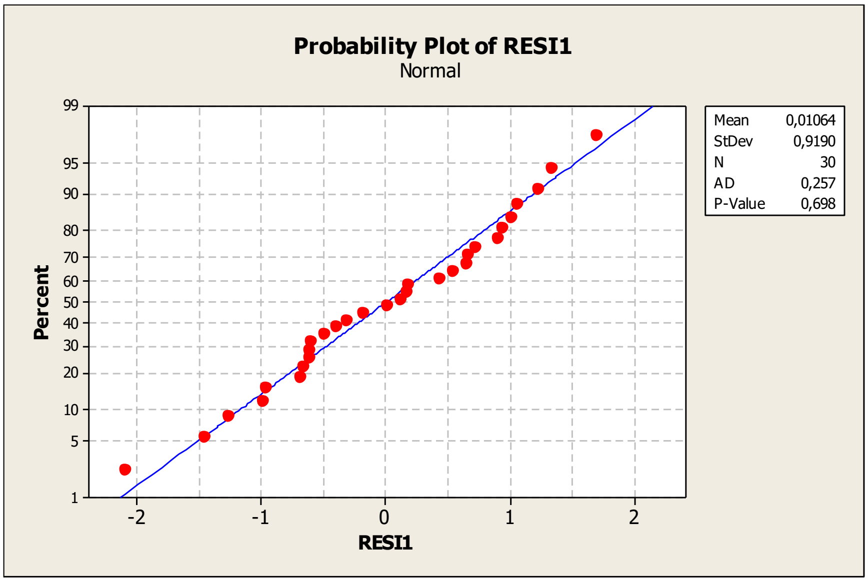


Residuals seem to be randomly distributed.

Let's plot the ACF of model residuals:



By running the normality test on the residuals:

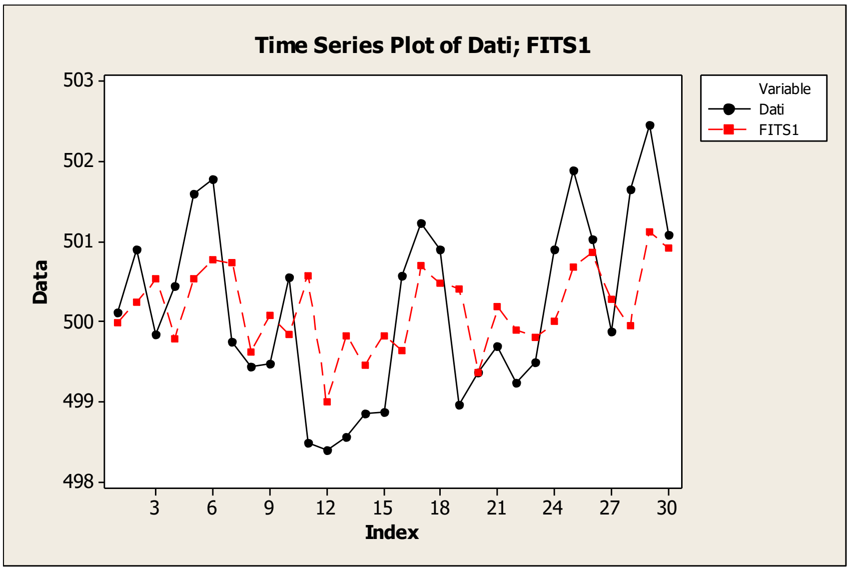


At α=0,05 we can not reject the null hypothesis about data normality.

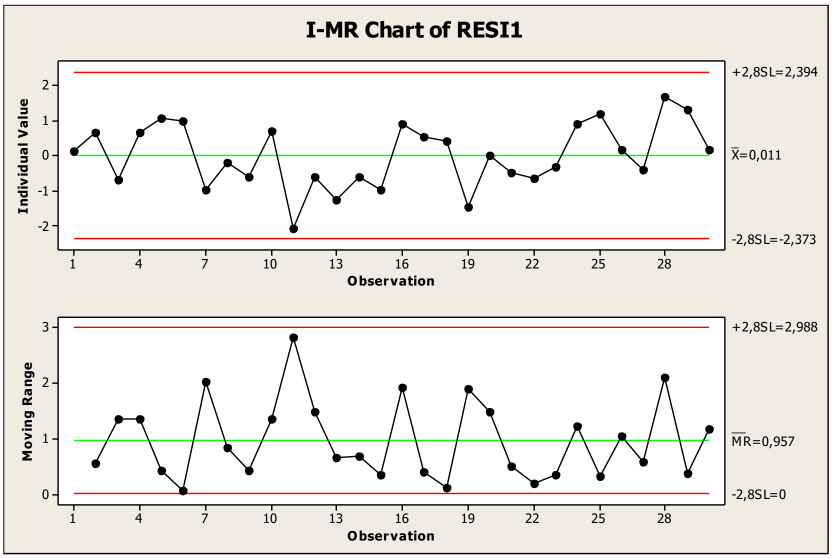
After residual checking we can design the FVC and the SCC.

The Type I error is α: $\alpha = \frac{1}{ARL_0} = \frac{1}{200} = 0{,}005$, and hence: $z\alpha/_2 = 2{,}81$.

FVC



SCC



The charts signal no OOCs and no strange pattern.

Exercise 6 (score 15)

The following data refer to the thickness of a sheet produced (shown from the left to the right and then from the top to the bottom):

100	109	103	109	96	112	103	101	101	100	96	93	111	109
107	104	113	95	101	105	92	100	77	74	69	58	72	63
69	73	54	51	56	49	49	49	74	56	80	57		

1) Design the appropriate monitoring system. Assume ARL_0=100 for monitoring both the process level and variability.

2) As the following new data are collected,

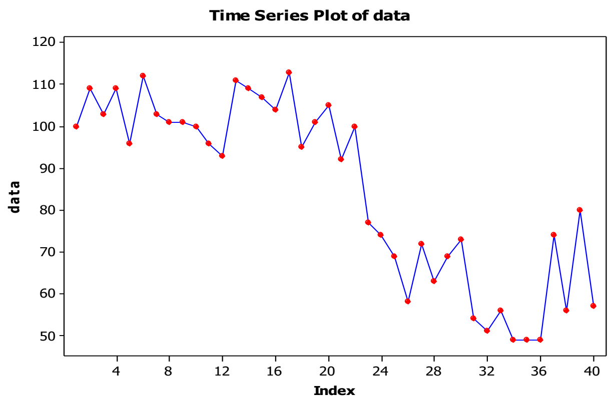
49	69	78	68	70	77	77	85	73	86

is the process in control?

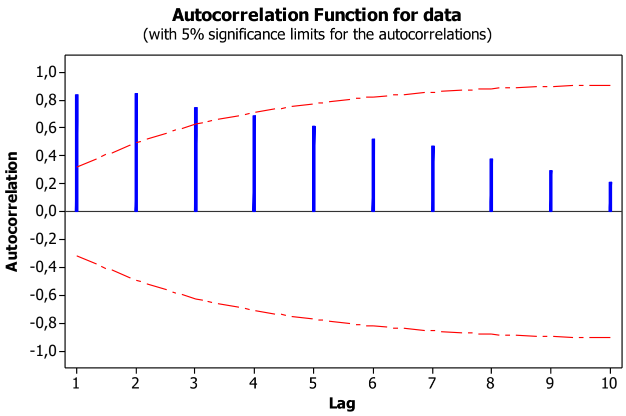
3) How does the design step (carried out in a) change if one assumes that an assignable cause is available for the first iteration of the control charts design?

Exercise 6 (solution)

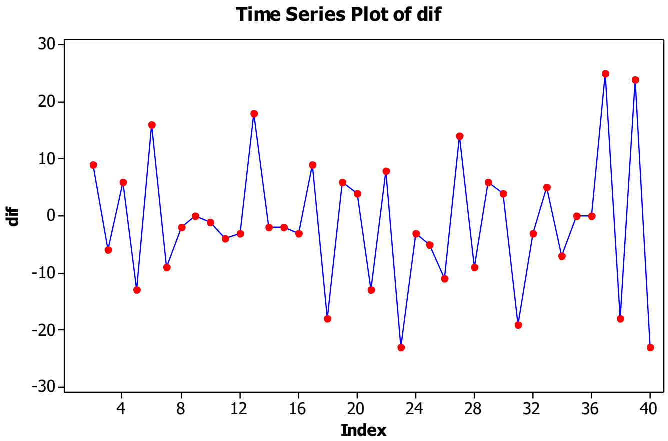
1) Data "snooping":



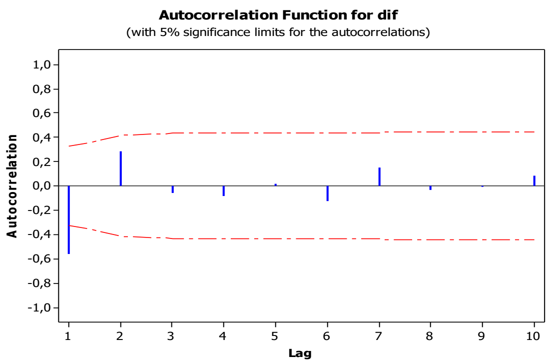
The process seems to be not stationary. This is confirmed by the ACF:

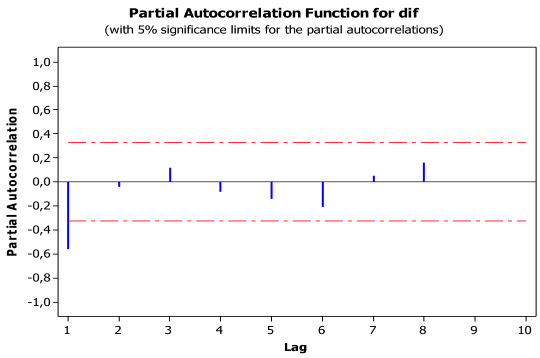


Let’s apply the differencing operator:



The differenced time series looks stationary now. Let’s try to identify an ARMA model for this time series:





It is not clear if the model is ARIMA(1,1,0) or ARIMA(0,1,1). Let’s try to fit both the models and choose the one with minimum variance of residuals (in both cases the constant term is not significant):

ARIMA(1,1,0)

```
Final Estimates of Parameters
```

Type		Coef	SE Coef	T	P
AR	1	-0,6174	0,1385	-4,46	0,000

Differencing: 1 regular difference

Number of observations: Original series 40, after differencing 39

Residuals: SS = 3450,45 (backforecasts excluded)

MS = 90,80 DF = 38

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	6,3	15,3	22,4	*
DF	11	23	35	*
P-Value	0,855	0,885	0,951	*

ARIMA(0,1,1)

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
MA	1	0,4550	0,1444	3,15	0,003

Differencing: 1 regular difference

Number of observations: Original series 40, after differencing 39

Residuals: SS = 3942,80 (backforecasts excluded)

MS = 103,76 DF = 38

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	7,9	14,6	36,3	*
DF	11	23	35	*
P-Value	0,723	0,907	0,410	*

The Ljung-Box test (lag12) confirms that the residuals of both models are not autocorrelated; the residual variance estimate leads us to prefer the first model.

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
RESI2	39	1	-1,51	1,51	9,41	-18,06	-7,56	-1,23	5,57	25,00
RESI3	39	1	-1,80	1,60	10,02	-21,70	-10,51	-2,54	4,84	24,35

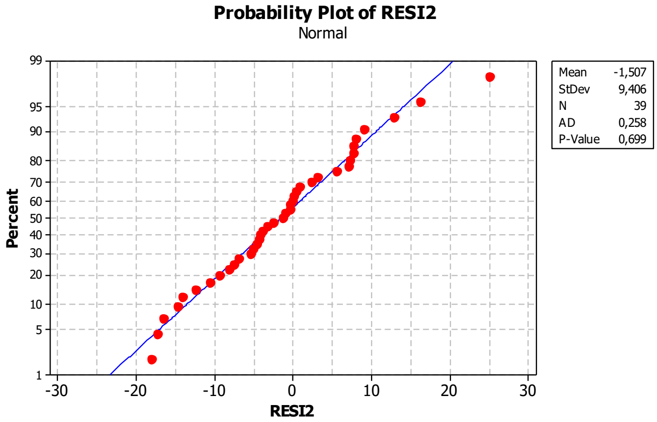
Notice: we are not stating that there is statistical evidence of a significant difference of residual variances (to do that we should perform an hypothesis test).

We choose the ARIMA (1,1,0) model.

The model is:

$$\nabla X_t = -0,6174 \nabla X_{t-1} + \varepsilon_t$$
$$(X_t - X_{t-1}) = -0,6174 (X_{t-1} - X_{t-2}) + \varepsilon_t$$
$$X_t = (1 - 0,6174) X_{t-1} - 0,6174 X_{t-2} + \varepsilon_t \qquad (*)$$

Assumption checking:



Runs test for RESI2

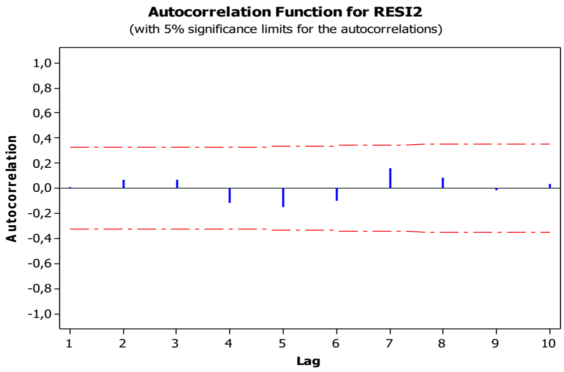
Runs above and below K = -1,50714

The observed number of runs = 22

The expected number of runs = 20,4872

20 observations above K; 19 below

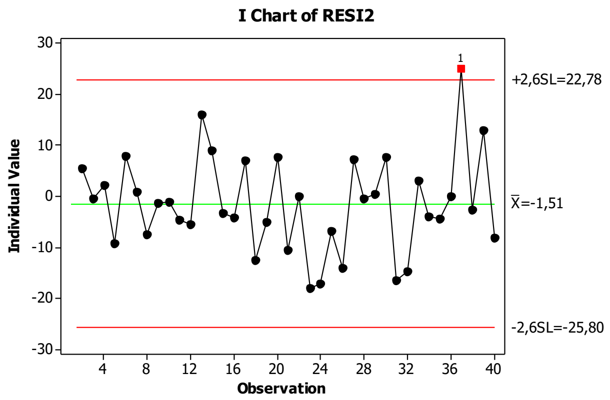
P-value = 0,623

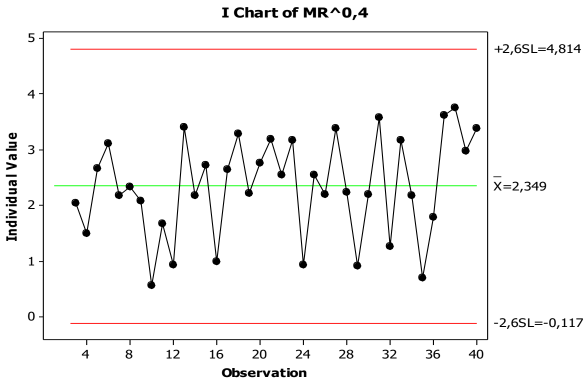


Control chart on residuals:

k= 2,576=$z_{a/2}$ where a=1/ARL_0=1/100

for the MR chart we can use a transformation with l=0.4



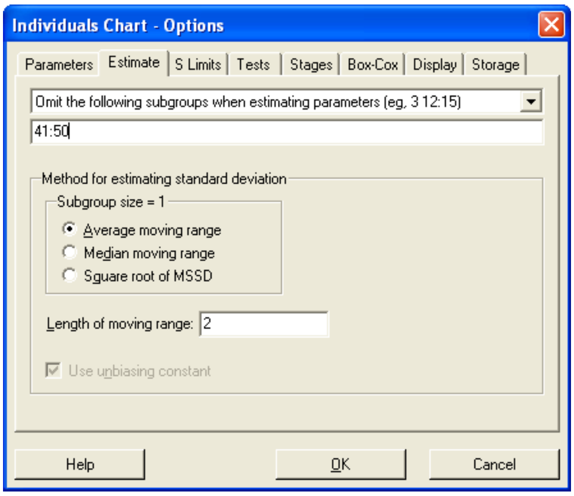


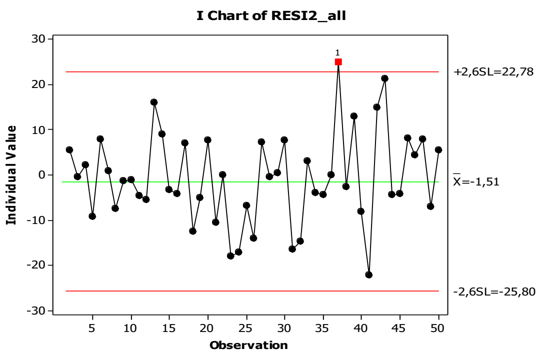
One single OOC observation; no assignable cause is assumed.

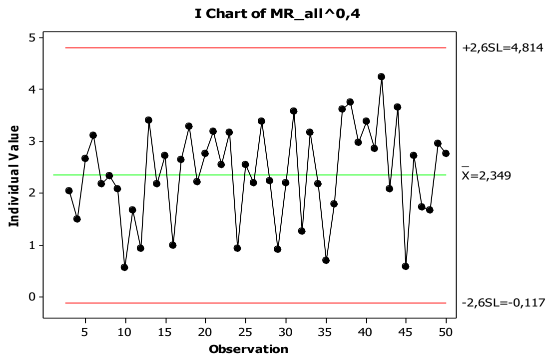
2) Let's use the previous equation (*) to compute the residuals e_t, being known the model in Phase I; we get:

t	*data*	$FIT_t = (1-0,6174)x_{t-1} - 0,6174x_{t-2}$	$e_t = x_t - FIT_t$
41	49	71,2002	-22,2002
42	69	53,9392	15,0608
43	78	56,652	21,348
44	68	72,4434	-4,4434
45	70	74,174	-4,174
46	77	68,7652	8,2348
47	77	72,6782	4,3218
48	85	77	8
49	73	80,0608	-7,0608
50	86	80,4088	5,5912

The control limits must not change:







New observations are IC.

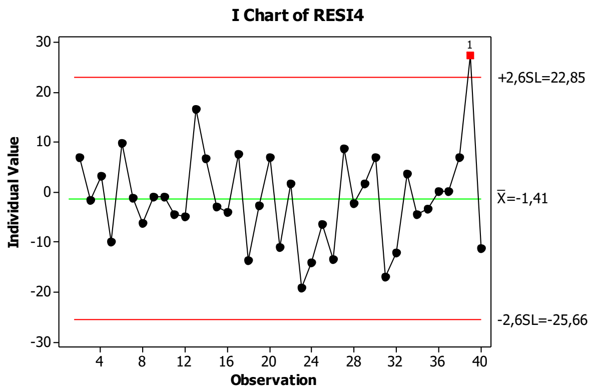
3) Analogously to the use of dummy variables for special cause charts with non random patterns modelled via regression, the value of the OOC observation can be substituted by the corresponding fit value; thus, it is possible to re-estimate the ARIMA(1,1,0) coefficients. Then, the control chart can be re-designed.

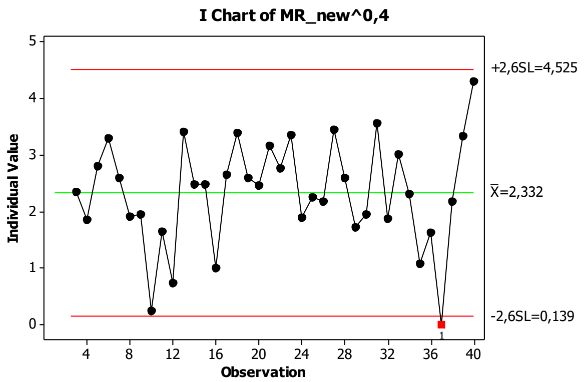
ARIMA Model: data (37a oss= fit37)

```
Relative change in each estimate less than 0,0010

Final Estimates of Parameters

Type         Coef   SE Coef       T       P
AR    1   -0,4845    0,1540   -3,15   0,003
```

```
Differencing: 1 regular difference
Number of observations:  Original series 40, after differencing 39
Residuals:    SS =  3442,81 (backforecasts excluded)
              MS =  90,60  DF = 38


Modified Box-Pierce (Ljung-Box) Chi-Square statistic


Lag             12     24     36   48
Chi-Square     2,7    7,9   17,9    *
DF              11     23     35    *
P-Value      0,994  0,998  0,993    *
```

The coefficient of ARIMA(1,1,0) model seems to be not robust to the presence of a single strange data. This may be due to the reduced number of data in Phase I. Assumptions are verified. The resulting charts are:





Two new OOCs.

Exercise 7 (max score 15)

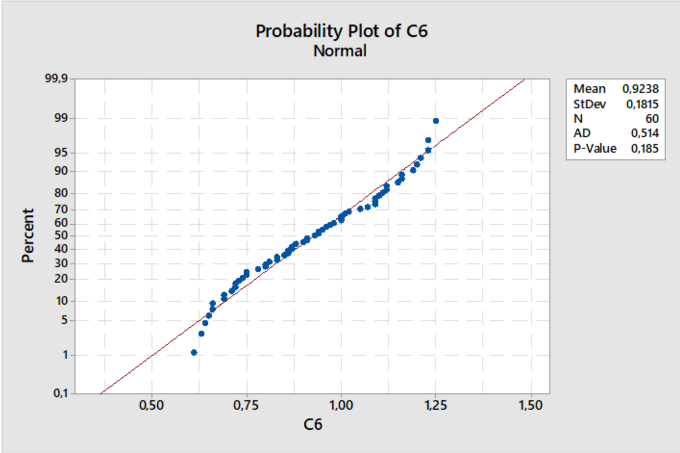
X-ray Tomography machines consist of two fundamental components: an X-ray source and a series of detectors. The latter includes many cells that collect the in-coming X-ray radiation that passed through the scanned object. In order to check the quality of the detector, the machine builder applies the following approach. First, the detector is divided into four distinct zones (1: upper left, 2: upper right, 3: bottom-left, 4: bottom-right): for each zone, a randomly chosen cell is hit by a known amount of X-ray radiation. The ratio between emitted and detected rays, called R, is recorded. The following table shows the R values collected from 15 detectors (inspected detectors are in sequential order).

Detector	**Zone 1**	**Zone 2**	**Zone 3**	**Zone 4**
1	1.19	0.96	0.81	0.64
2	1.16	1.00	0.91	0.65
3	0.91	1.11	0.87	0.88
4	1.05	1.15	0.75	0.61
5	1.00	0.97	0.83	0.75
6	0.94	0.93	0.86	0.86
7	1.09	0.85	1.02	0.71
8	1.01	0.98	0.69	0.72
9	1.00	1.16	0.73	0.72
10	1.07	0.94	0.95	0.87
11	1.12	1.12	0.66	0.66
12	1.09	1.20	0.63	0.78
13	1.21	1.25	0.90	0.80
14	1.23	1.10	0.69	0.74
15	1.09	1.23	0.80	0.83

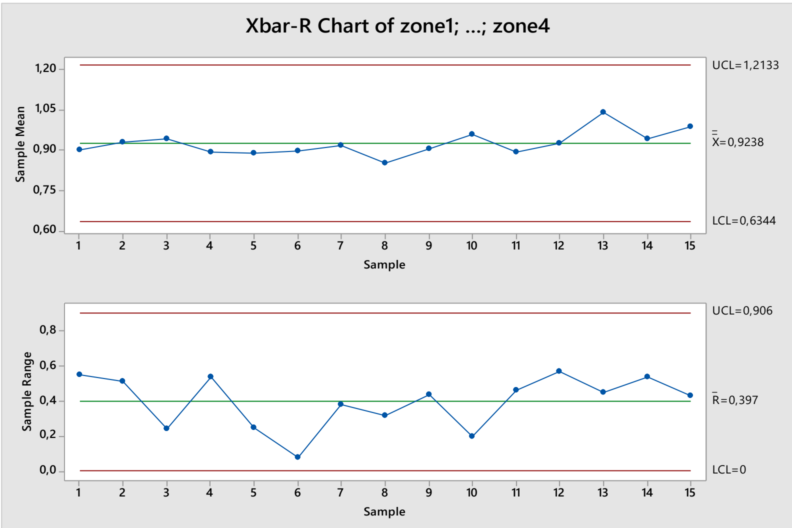
1) Design a traditional control chart to monitor the R descriptor in the n=4 considered zones. Which problems arise by using this approach? Without using the information about the cell position within the detector, which approach do you suggest to use in order to avoid the problems observed by applying the traditional chart?
2) Design a control chart that includes, if necessary, the information about the cell position within the detector. How do the results change with respect to point a)?

Exercise 7 (solution)

1) Normality test:

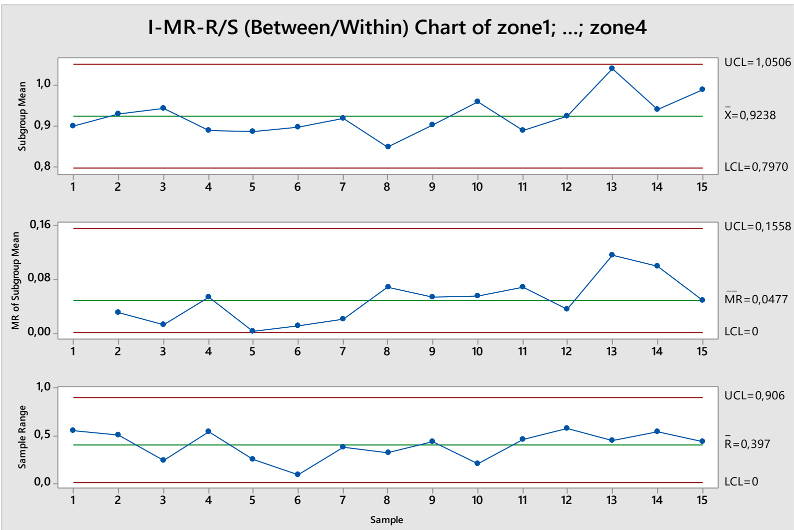


Traditional control chart:



Hagging is present.

An I-MR-R control chart is a more suitable choice:



2) By considering the cell location as a dummy variable, a stepwise regression can be performed:

Regression Analysis: R versus Z1; Z2; Z3; Z4

Method

Categorical predictor coding (1; 0)

Stepwise Selection of Terms

α to enter = 0,15; α to remove = 0,15

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	2	1,31355	67,57%	1,31355	0,65677	59,38	0,000
Z3	1	0,27456	14,12%	0,69520	0,69520	62,85	0,000
Z4	1	1,03899	53,45%	1,03899	1,03899	93,93	0,000
Error	57	0,63047	32,43%	0,63047	0,01106		
Lack-of-Fit	1	0,00147	0,08%	0,00147	0,00147	0,13	0,719
Pure Error	56	0,62900	32,36%	0,62900	0,01123		
Total	59	1,94402	100,00%				

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)

```
0,105171  67,57%     66,43%  0,697137      64,14%
```

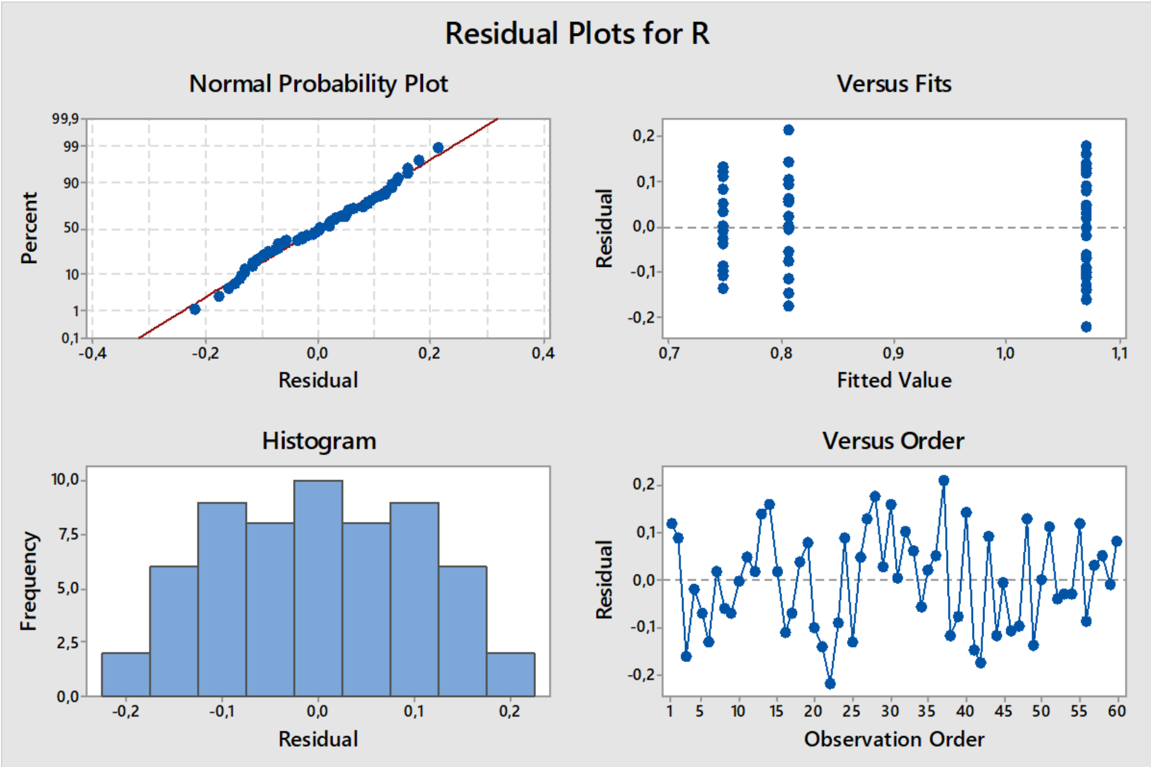
Coefficients

```
Term           Coef  SE Coef        95% CI          T-Value  P-Value
VIF
Constant   1,0703   0,0192  ( 1,0319;  1,1088)    55,74    0,000
Z3
  1       -0,2637   0,0333  (-0,3303; -0,1971)    -7,93    0,000
1,13
Z4
  1       -0,3223   0,0333  (-0,3889; -0,2557)    -9,69    0,000
1,13
```

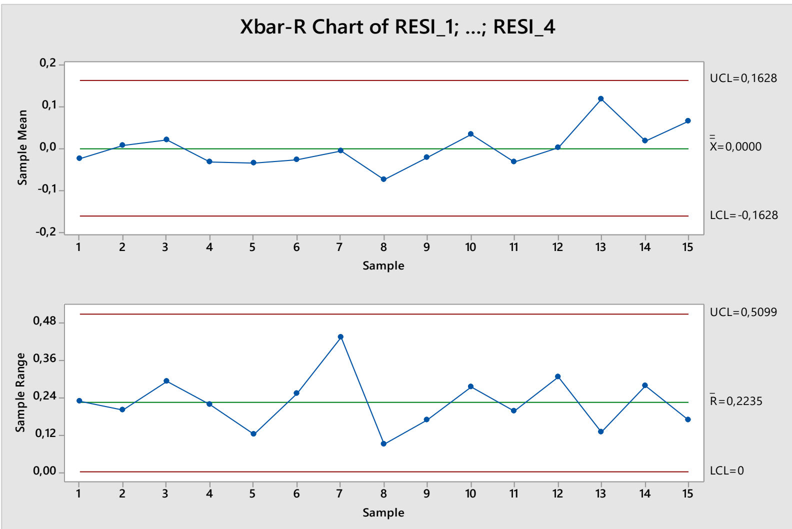
Regression Equation

```
R = 1,0703 + 0,0 Z3_0 - 0,2637 Z3_1 + 0,0 Z4_0 - 0,3223 Z4_1
```

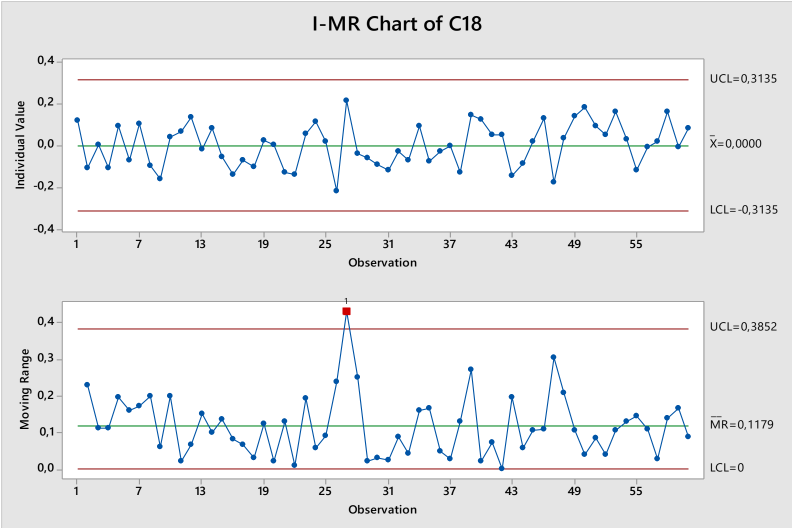
Locations 3 and 4 are significant. Residual check highlights no violation.



An Xbar-R chart can be designed to monitor the four residuals (one residual for each location):



Otherwise, an acceptable (but not fully appropriate approach) consists of applying an I-MR chart on the residuals treated as individual observations:



This approach yields an alarm (obs 27).

Note: using multiple charts, one for each position is not the most correct approach to take into account the effect of the location factor, but it was accepted.

Exercise 8 (max score 12)

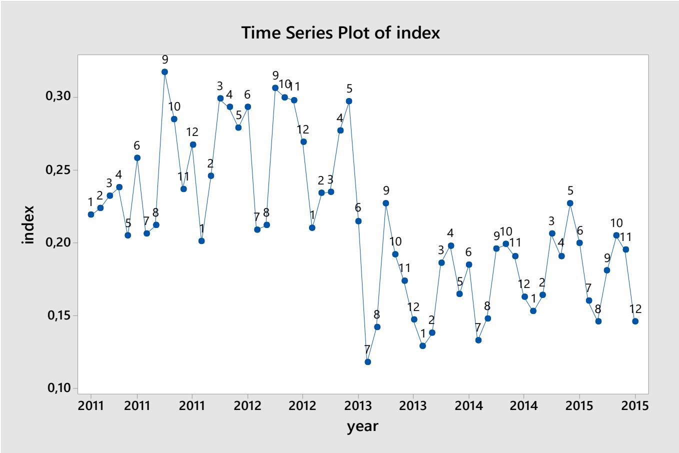
A manufacturing company in China started monitoring the soil pollution in the area surrounding one of its major plants in 2011. A normalized pollution index that ranges between 0 (no pollution) and 1 (alert level pollution) was recorded on a monthly basis and the values are shown in the table below.

Year	Month	Index	Year	Month	Index
2011	1	0,219	2013	7	0,118
2011	2	0,224	2013	8	0,142
2011	3	0,232	2013	9	0,227
2011	4	0,238	2013	10	0,192
2011	5	0,205	2013	11	0,174
2011	6	0,258	2013	12	0,147
2011	7	0,206	2013	1	0,129
2011	8	0,212	2014	2	0,138
2011	9	0,317	2014	3	0,186
2011	10	0,285	2014	4	0,198
2011	11	0,237	2014	5	0,165
2011	12	0,267	2014	6	0,185
2012	1	0,201	2014	7	0,133
2012	2	0,246	2014	8	0,148
2012	3	0,299	2014	9	0,196
2012	4	0,293	2014	10	0,199
2012	5	0,279	2014	11	0,191
2012	6	0,293	2014	12	0,163
2012	7	0,209	2015	1	0,153
2012	8	0,212	2015	2	0,164
2012	9	0,306	2015	3	0,206
2012	10	0,300	2015	4	0,191
2012	11	0,298	2015	5	0,227
2012	12	0,269	2015	6	0,200
2013	1	0,210	2015	7	0,160
2013	2	0,234	2015	8	0,146
2013	3	0,235	2015	9	0,181
2013	4	0,277	2015	10	0,205
2013	5	0,297	2015	11	0,195
2013	6	0,215	2015	12	0,146

1. Identify and fit a suitable model for pollution index. If necessary, exploit the following information: in June 2013 an extraordinary flood occurred in that Chinese province.
2. Design a suitable control chart. Comment the results.

Exercise 8 (solution)

1) Time series plot:



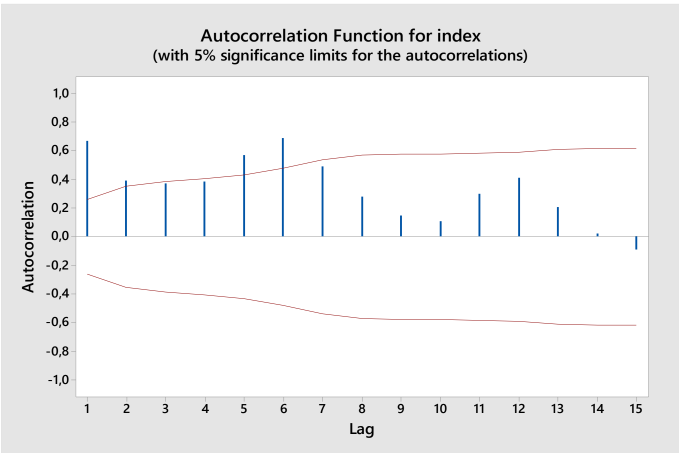
The time series looks not random; there is a seasonality of the index, with a possible jump of the mean. Meandering may also be present.

Runs test:

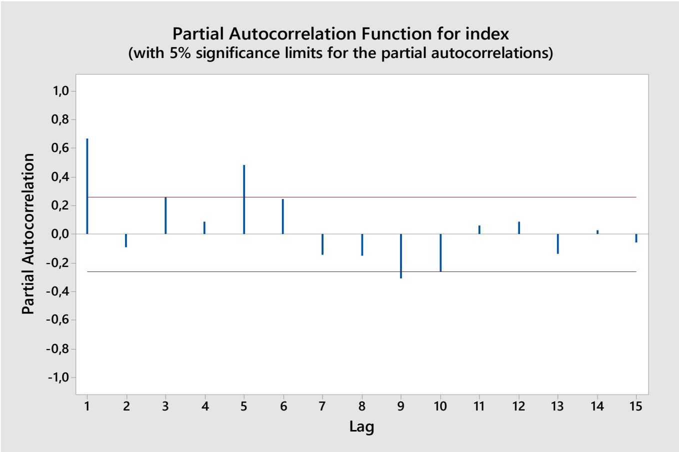
```
Runs test for index


Runs above and below K = 0,212967


The observed number of runs = 16
The expected number of runs = 30,1667
25 observations above K; 35 below
P-value = 0,000
```

There is a strong statistical evidence to reject the null hypothesis of randomness.

ACF:



PACF:



There is no evident pattern to suggest the choice of an ARIMA model. One possible model includes a jump by using a dummy variable such that:

Dummy = 0 (before the flood)

Dummy = 1 (after the flood)

In order to cope with the seasonality of the model and the meandering, the model should also include an autoregressive term (e.g., AR(1)) and each month as regressor (a dummy corresponding to each month can be used, corresponding to use the "month" as categorical regressors). The resulting model is:

```
Method


Categorical predictor coding  (1; 0)

Rows unused                    1



Analysis of Variance


Source          DF    Seq SS  Contribution    Adj SS    Adj MS  F-
Value  P-Value

Regression      13  0,138891        87,52%  0,138891  0,010684
24,27    0,000

  dummy          1  0,094452        59,51%  0,014025  0,014025
31,86    0,000

  AR1            1  0,005333         3,36%  0,002425  0,002425
5,51    0,023
```

month	11	0,039107	24,64%	0,039107	0,003555	8,08	0,000
Error	45	0,019812	12,48%	0,019812	0,000440		
Lack-of-Fit	44	0,019751	12,45%	0,019751	0,000449	7,42	0,285
Pure Error	1	0,000061	0,04%	0,000061	0,000061		
Total	58	0,158703	100,00%				

Model Summary

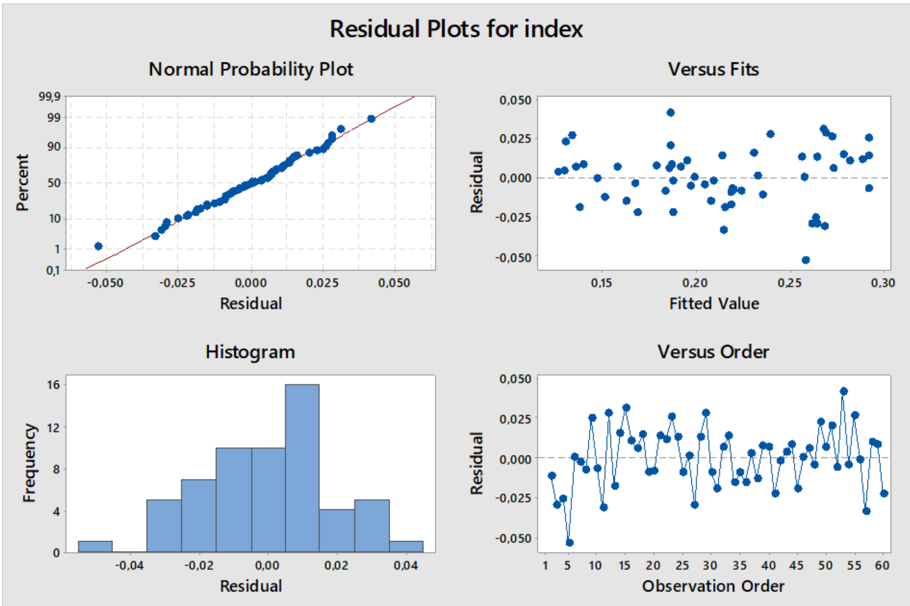
S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
0,0209824	87,52%	83,91%	0,0334708	78,91%

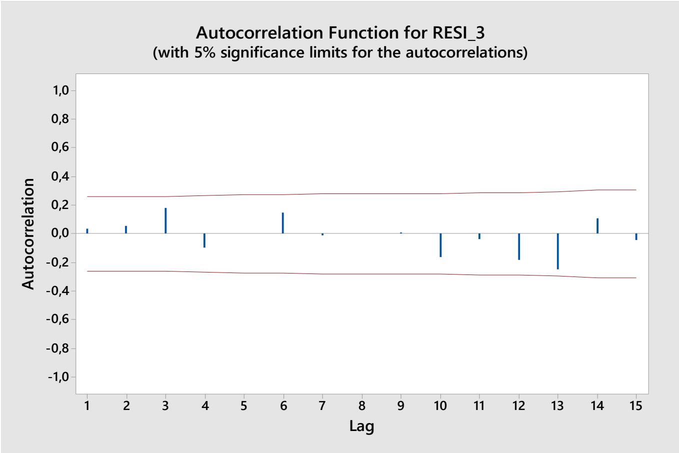
Coefficients

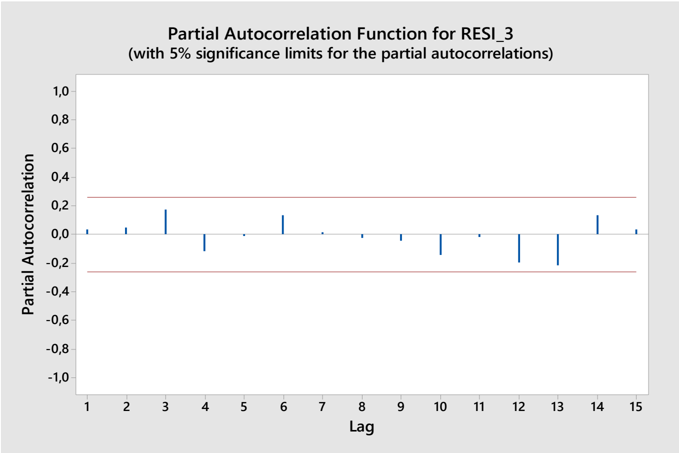
Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	0,1441	0,0314	(0,0808; 0,2074)	4,59	0,000	
dummy	-0,0593	0,0105	(-0,0805; -0,0381)	-5,64	0,000	3,69
AR1	0,278	0,118	(0,039; 0,516)	2,35	0,023	4,91
month						
2	0,0301	0,0147	(0,0004; 0,0598)	2,04	0,047	2,26
3	0,0553	0,0142	(0,0266; 0,0840)	3,88	0,000	2,11
4	0,0546	0,0142	(0,0261; 0,0832)	3,86	0,000	2,09
5	0,0477	0,0143	(0,0189; 0,0765)	3,34	0,002	2,12
6	0,0565	0,0145	(0,0272; 0,0858)	3,88	0,000	2,20

7	-0,0073	0,0144	(-0,0364;	0,0217)	-0,51	0,615
2,16						
8	0,0175	0,0148	(-0,0123;	0,0474)	1,18	0,242
2,28						
9	0,0891	0,0146	(0,0597;	0,1184)	6,11	0,000
2,21						
10	0,0595	0,0149	(0,0294;	0,0895)	3,99	0,000
2,31						
11	0,0448	0,0146	(0,0154;	0,0742)	3,07	0,004
2,21						
12	0,0290	0,0142	(0,0004;	0,0576)	2,04	0,047
2,10						

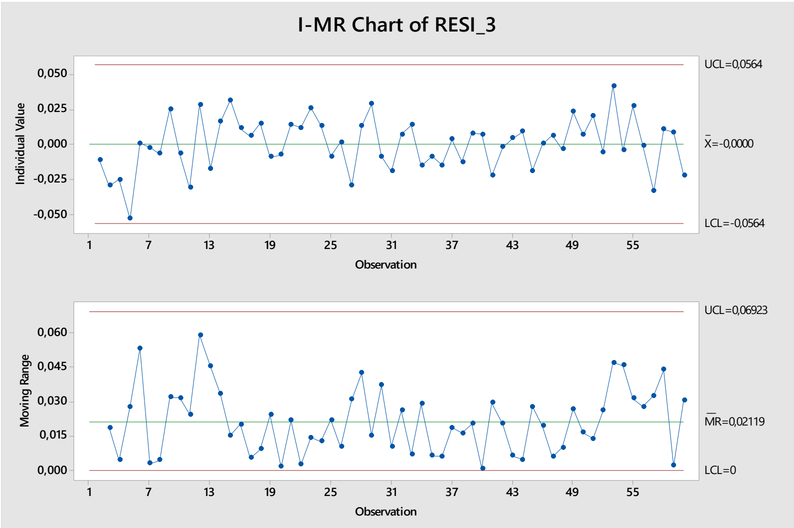
The model is significant (also the month term is significant), there is no lack of fit, and the R2 adjusted is higher than the previous model. The residuals are now normal (p-value=0,93) and random (runs test p-value=0,497). Non significant months could also be removed, leading to a reduced model.







2) I-MR chart on the model residuals:



The process is in-control. The flood produced a shift of the mean but not a modification of the variability of the index.

Exercise 9 (max score 16)

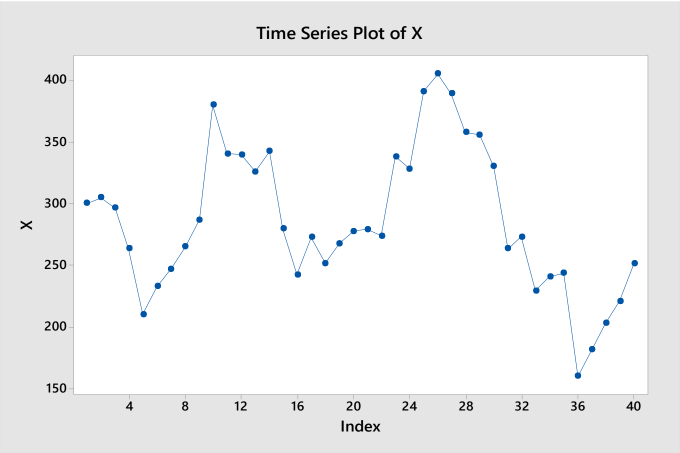
A waterjet cutting system is used to cut titanium laminates for the aerospace industry. In order to monitor the stability of the process, the average water pressure (in MPa) was measured in each pumping cycle by using a pressure transducer. The measured values in 40 consecutive cycles are reported below (read from left to right and from the top to the bottom).

300.7	305.0	296.7	263.7	210.3	233.0	247.3	265.3	287.0	380.3	341.0	340.0
326.3	342.7	280.0	242.7	273.0	252.0	268.0	278.0	279.7	274.0	338.7	328.7
391.3	405.7	389.7	358.0	356.3	330.3	263.7	273.3	229.7	241.0	244.0	160.3
182.0	203.3	221.3	251.7								

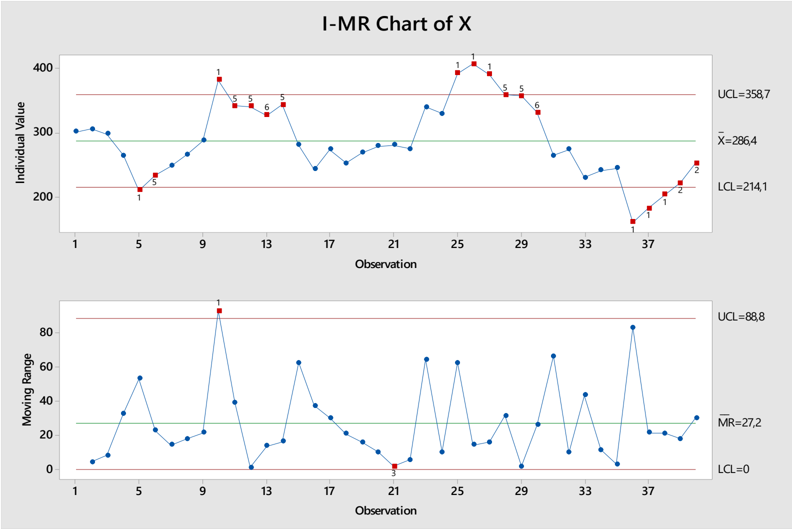
1) Design a traditional control chart (assuming a NID behaviour) with run-rules and comment the result.
2) Identify and fit a suitable model
3) Design a control chart based on the model fitted at point b)

Exercise 9 (solution)

1) Time-series plot:



The process seems not NID, but let's design the traditional chart as requested in point 1):



Run rules:

Test Results for I Chart of X

TEST 1. One point more than 3,00 standard deviations from center line.

Test Failed at points: 5; 10; 25; 26; 27; 36; 37; 38

TEST 2. 9 points in a row on same side of center line.

Test Failed at points: 39; 40

TEST 5. 2 out of 3 points more than 2 standard deviations from center line (on one side of

CL).

Test Failed at points: 6; 11; 12; 14; 25; 26; 27; 28; 29; 37; 38; 39

TEST 6. 4 out of 5 points more than 1 standard deviation from center line (on one side of

CL).

Test Failed at points: 13; 14; 26; 27; 28; 29; 30; 36; 37; 38; 39; 40

TEST 8. 8 points in a row more than 1 standard deviation from center line (above and below

CL).

Test Failed at points: 30; 40

Test Results for MR Chart of X

TEST 1. One point more than 3,00 standard deviations from center line.

Test Failed at points: 10

TEST 3. 6 points in a row all increasing or all decreasing.

Test Failed at points: 21

2)

Runs-test:

Runs test for X

Runs above and below K = 286,392

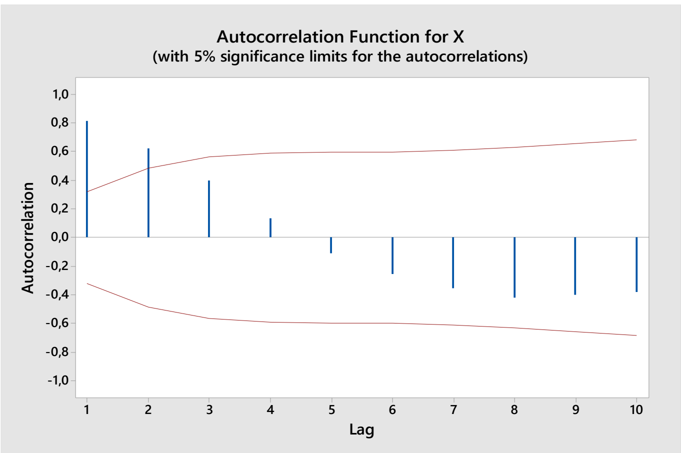
The observed number of runs = 6

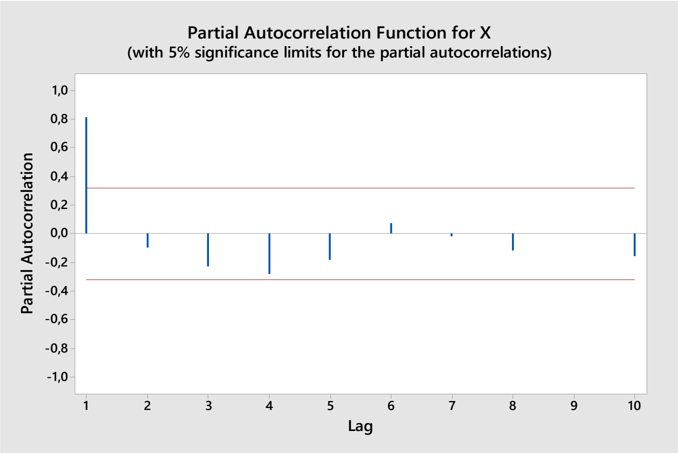
The expected number of runs = 20,55

17 observations above K; 23 below

P-value = 0,000

ACF and PACF:





The process is not random, and a suitable model may be AR(1).

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	89221	89221	73,40	0,000
AR1	1	89221	89221	73,40	0,000
Error	37	44976	1216		
Lack-of-Fit	36	42991	1194	0,60	0,794
Pure Error	1	1984	1984		
Total	38	134196			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
34,8648	66,49%	65,58%	63,84%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	50,9	28,0	1,82	0,077	

AR1	0,8185	0,0955	8,57	0,000	1,00

Regression Equation

X = 50,9 + 0,8185 AR1

The constant term is not significant. Let's remove it.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3275826	3275826	2541,14	0,000
AR1	1	3275826	3275826	2541,14	0,000
Error	38	48986	1289		
Lack-of-Fit	37	47002	1270	0,64	0,781
Pure Error	1	1984	1984		
Total	39	3324812			

Model Summary

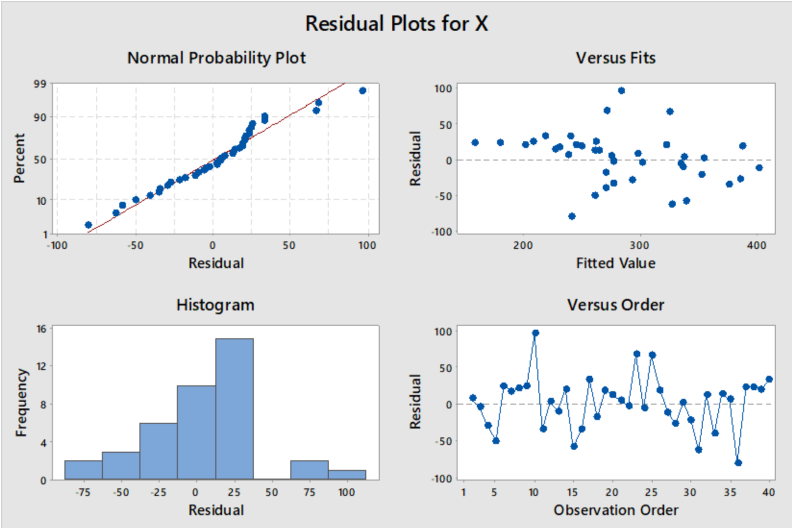
S	R-sq	R-sq(adj)	R-sq(pred)
35,9043	98,53%	98,49%	98,45%

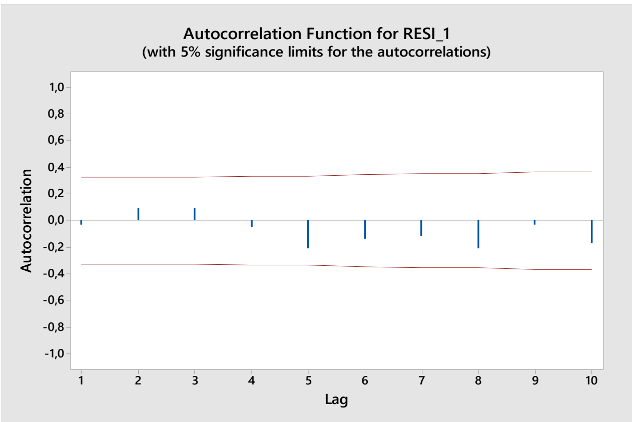
Coefficients

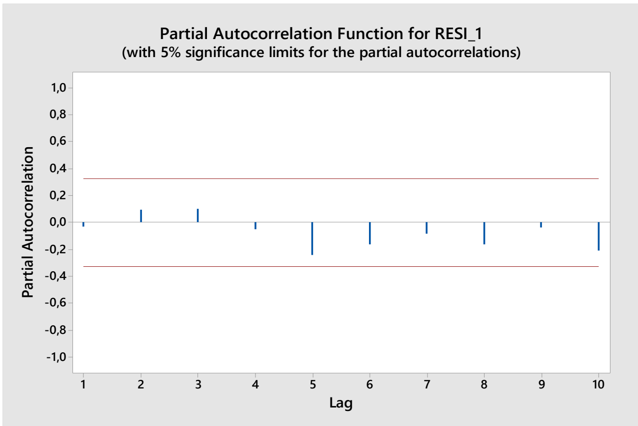
Term	Coef	SE Coef	T-Value	P-Value	VIF
AR1	0,9886	0,0196	50,41	0,000	1,00

```
Regression Equation

X = 0,9886 AR1
```

The model without constant meet the assumptions (normality of residuals: p-value=0.118, runs-test: 0.294, ACF & PCAF ok, no strange pattern, lack of fit ok).

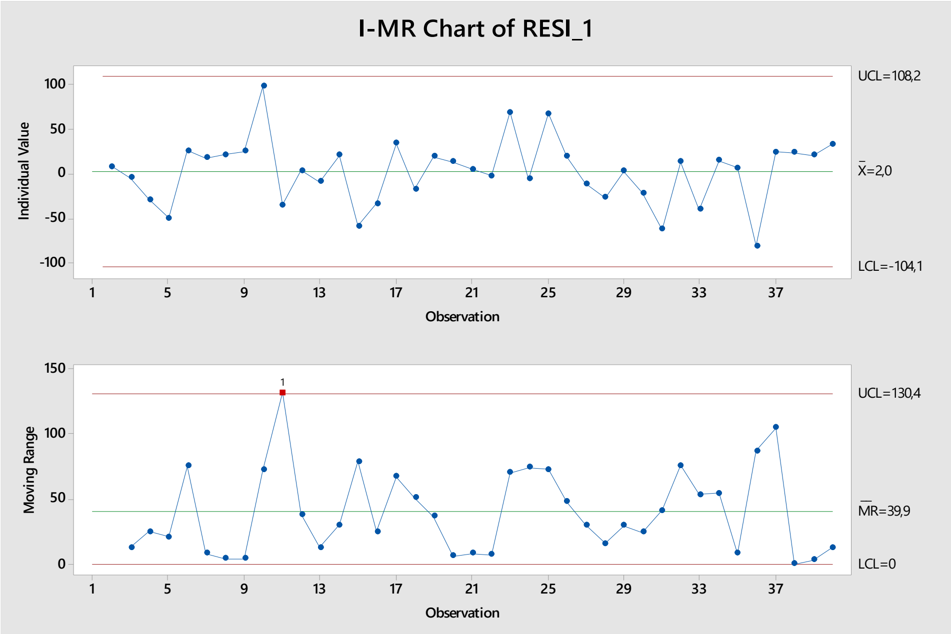




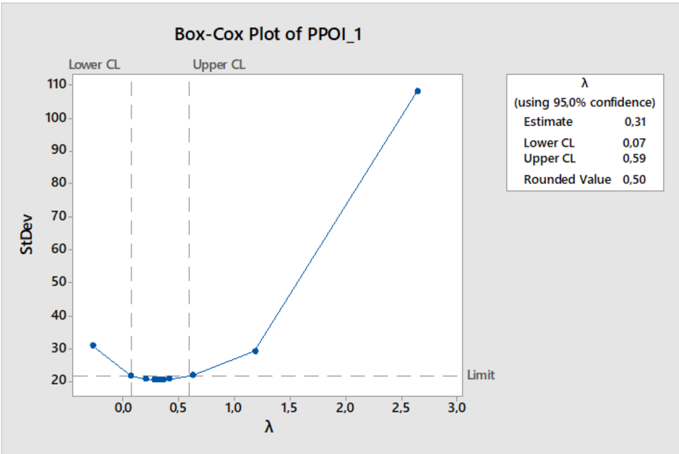


The model is approximately a random walk.

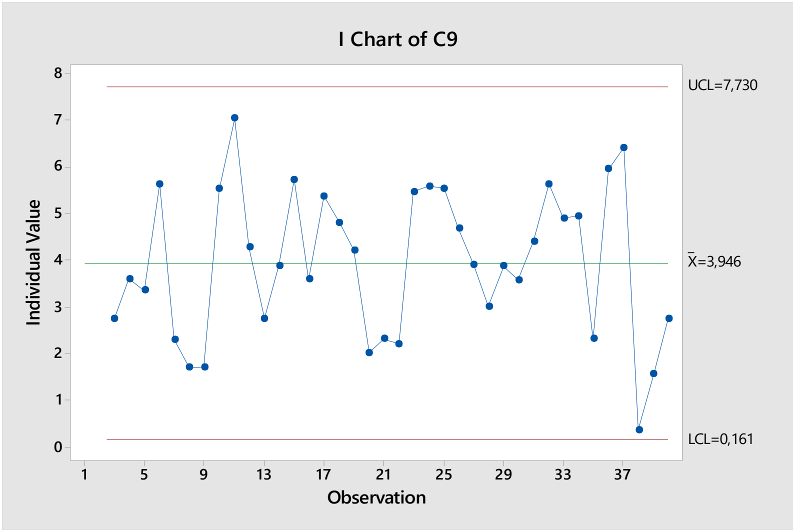
3) Special cause control chart:



One out-of-control point is signalled by the MR chart only. It can be caused by the fact that the MR statistic follows an half-normal distribution. Let transform it to normality with the Box-Cox transformation:



The result yields a value close to the known transformation (λ=0.4). By using λ=0.4, the new MR chart is:



The process is in control.

Exercise 10 (max score 13)

In a shop floor, one assembly cycle involves five sequential stations where different operators perform different operations. The duration (in minutes) of each assembly step was monitored during six consecutive cycles. The maximum duration allowed by the company for one single assembly step is 90 minutes.

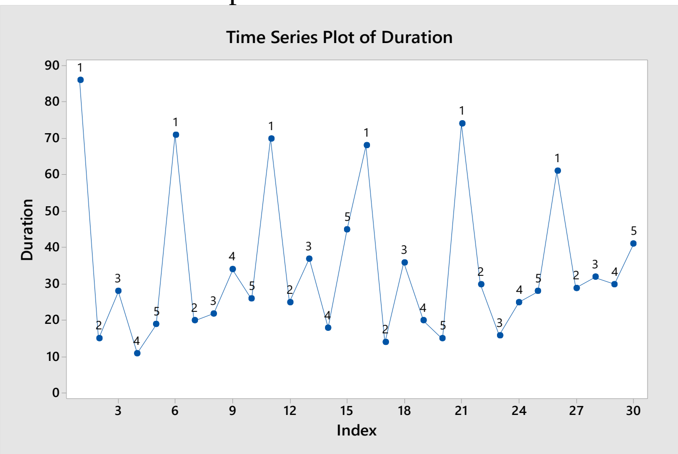
The data are reported in the table:

duration	operator	duration	operator	duration	operator	duration	operator	duration	operator	duration	operator
86	1	71	1	70	1	68	1	74	1	61	1
15	2	20	2	25	2	14	2	30	2	29	2
28	3	22	3	37	3	36	3	16	3	32	3
11	4	34	4	18	4	20	4	25	4	30	4
19	5	26	5	45	5	15	5	28	5	41	5

1) Design a suitable statistical control system in order to guarantee ARL_0=100
2) Different conclusions can be drawn if the real distribution of the MR statistic is used?
3) Which is the expected number of assembly steps whose duration exceed the allowed one in each cycle?

Exercise 10 (solution)

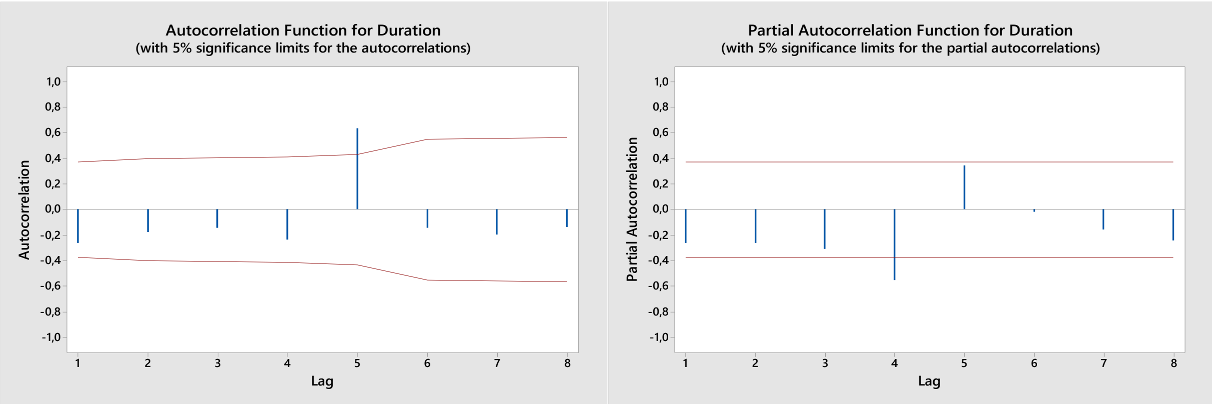
1) Graphical analysis and check of assumptions:



There is a systematic effect of operator 1.

```
Runs test for Duration


Runs above and below K = 34,8667
```

```
The observed number of runs = 17
The expected number of runs = 14,3333
10 observations above K; 20 below
* N is small, so the following approximation may be invalid.
P-value = 0,263
```



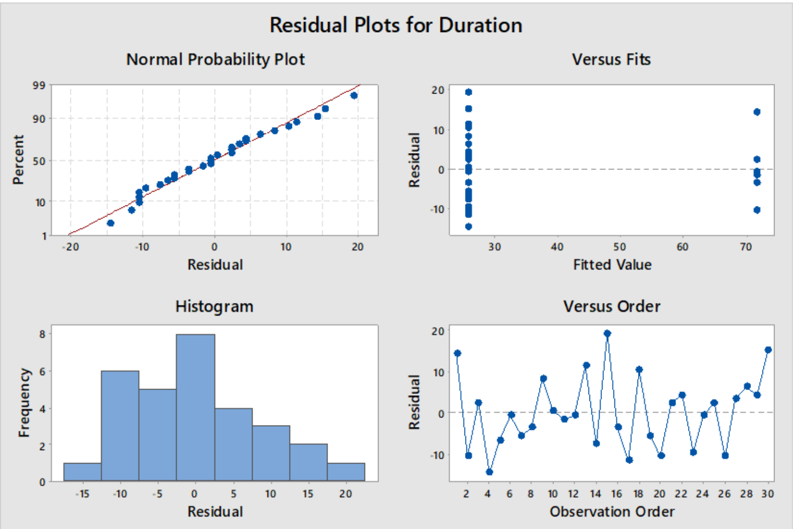
The auto-correlation functions confirms that there is a periodic effect of lag 5. Let's apply a regression model with a dummy variable X=1 when operator = 1 and X=0 otherwise.

Regression Analysis: Duration versus Dummy

```
Method

Categorical predictor coding  (1; 0)


Analysis of Variance

Source       DF  Adj SS   Adj MS  F-Value  P-Value
Regression    1   10157  10156,8   128,76    0,000
  Dummy       1   10157  10156,8   128,76    0,000
Error        28    2209     78,9
Total        29   12365
```

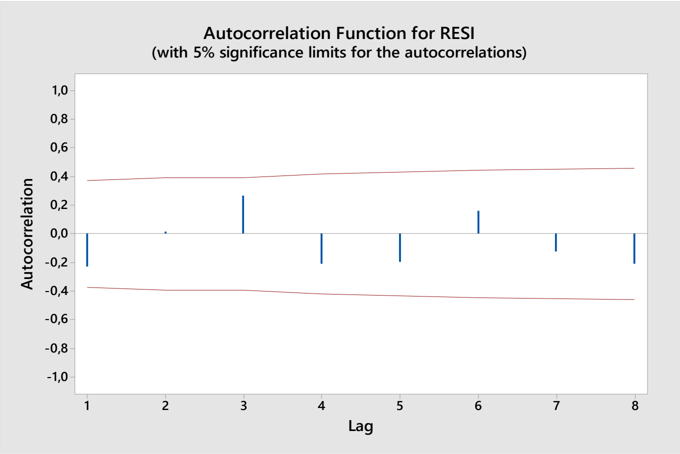
```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
8,88149  82,14%     81,50%      79,58%

Coefficients
Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  25,67     1,81    14,16    0,000
Dummy
  1        46,00     4,05    11,35    0,000  1,00


Regression Equation

Duration = 25,67 + 0,0 Dummy_0 + 46,00 Dummy_1
```

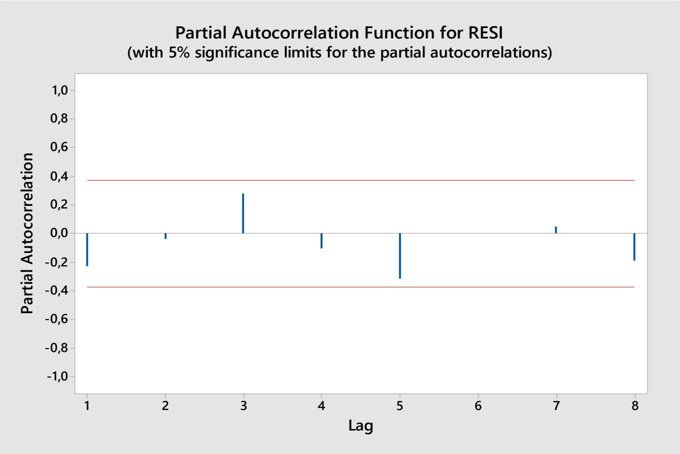
The dummy variable is significant. The check of residuals is ok.




```
Runs test for RESI

Runs above and below K = -1,89478E-15
```

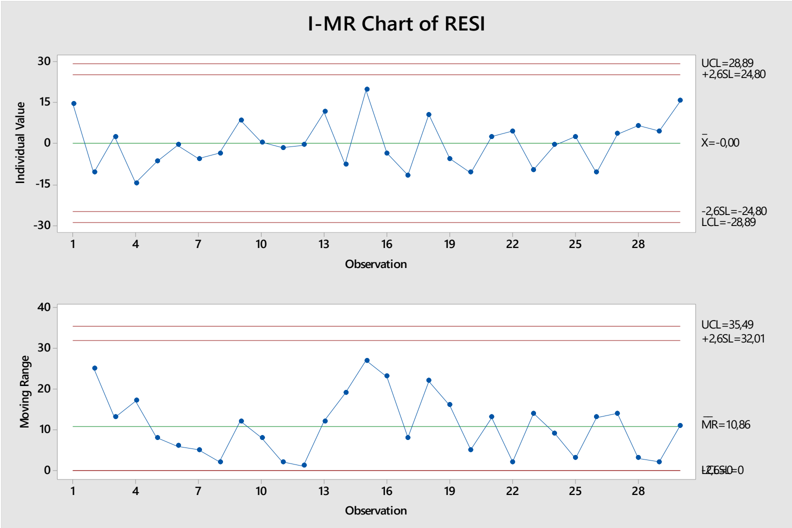
```
The observed number of runs = 17
The expected number of runs = 15,9333
14 observations above K; 16 below
P-value = 0,690
```



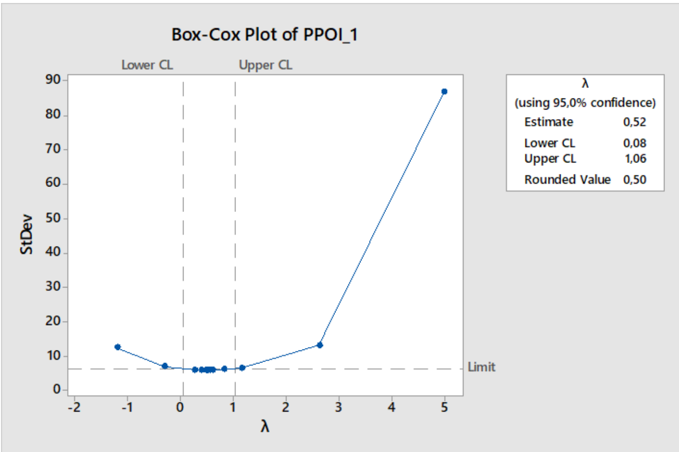


ARL0=100 implies: z_alpha/2 = 2.5758

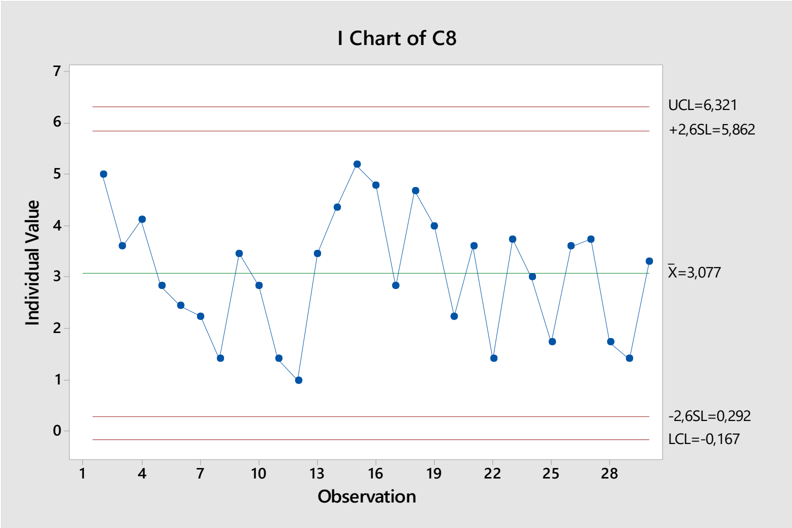
The resulting control chart on the residual is the following. The process is in-control.



2) Regarding the MR chart, three options are available: (1) re-design the chart by using the half-normal distribution, (2) re-design the chart by using the known transformation with lambda=0.4, (3) re-design the chart by using the Box-Cox transformation. Here we apply the method (3), but other methods are equivalent.



The modified MR chart is the following:



No change in the conclusions about the in-control state of the process.

3) In order to compute the probability of too long assembly operations we need to use the regression equation:

```
Duration = 25,67 + 0,0 Dummy_0 + 46,00 Dummy_1
```

With normal residuals having zero mean and standard deviation given by:

sigma_res = mean(MR_res)/d2(2) = 10.86/1.128= 9.63

The distribution of the cycle operation duration is:

duration~N(25.67,9.63^2) if the operator is 2, 3, 4 or 5

duration~N(25.67+46,9.63^2) if the operator is 1

Thus, the probability that the duration of one assembly operation exceeds the maximum allowed duration (90 minutes) depends on the operator:

mean	sigma	P(duration>90)	
25.67	9.63	1.19E-011	For operators 2, 3, 4, 5
71.7	9.63	0.0287	For operator 1

The expected number of tool long assembly operations in each cycle is:

4/5*P(duration>90|operator >1)+1/5* P(duration>90|operator = 1)= 1/5* 0.0287=**0.00574**.

Multivariate Control Charts and PCA

Exercise 1 (max score: 11)
Two synthetic indexes extracted by processing an acoustic emission signal acquired during a micro-milling process were used to monitor the stability of the process. Data are reported in the table below.

Observation	X1	X2	Observation	X1	X2	Observation	X1	X2
1	0,03	10,04	11	0,37	9,43	21	0,08	10,91
2	0,02	10,90	12	0,31	10,60	22	0,30	12,73
3	0,19	8,87	13	0,00	9,35	23	0,60	10,39
4	0,00	8,86	14	0,59	11,48	24	0,00	10,20
5	0,01	8,76	15	0,15	10,21	25	0,56	9,04
6	0,10	10,77	16	0,03	10,03	26	0,14	8,01
7	0,30	10,51	17	0,01	10,04	27	0,28	9,01
8	0,31	10,51	18	0,31	9,74	28	1,38	9,97
9	0,19	10,62	19	0,30	8,43	29	0,09	10,30
10	0,00	10,28	20	0,00	10,10	30	0,14	9,74

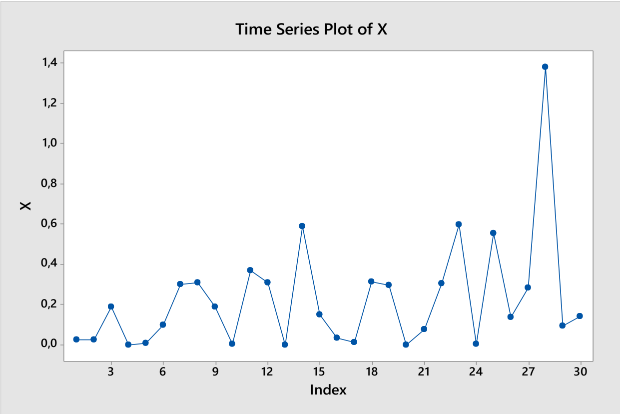
1) Design a multivariate control chart based on the long-term estimate on the variance-covariance matrix (alpha=0.0027). Discuss the result.
2) Re-design the chart by using the short-term estimate and alpha=0.02. Discuss the result assuming the existence of assignable causes for possible out-of-control observations.
3) A new signal acquisition procedure was applied, by moving the sensor location. The new data collected are reported below:

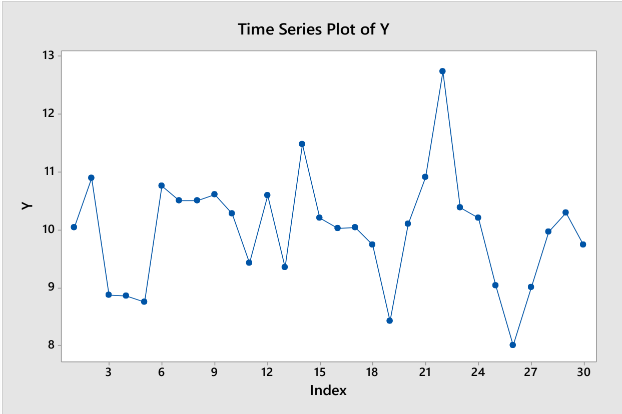
X1	X2
0,18	12,37
0,03	12,30
0,08	12,01
0,27	11,87
0,31	12,50
0,40	11,72
0,11	12,63
0,00	12,21
0,01	11,33
0,01	12,79

Is the process still in control after the sensor location change when control chart designed at point a) is assumed?

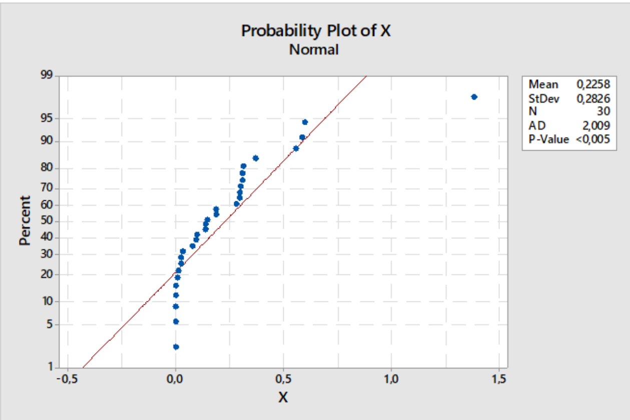
Exercise 1 (solution)

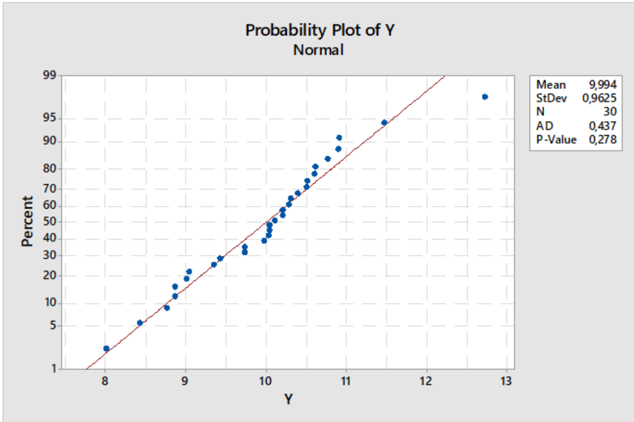
1) Time series X:



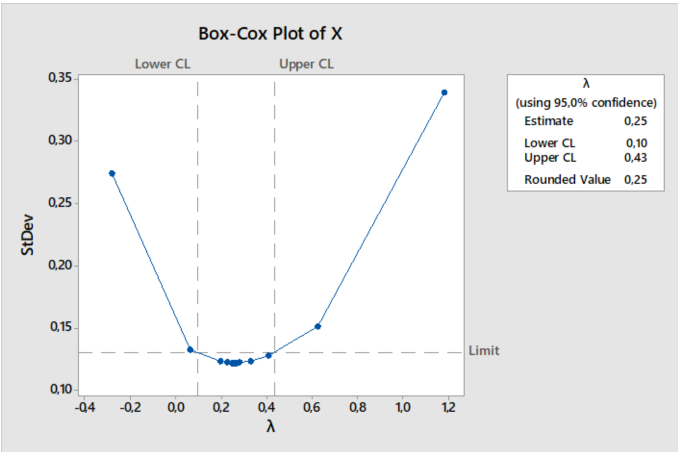


Normality test X e Y:

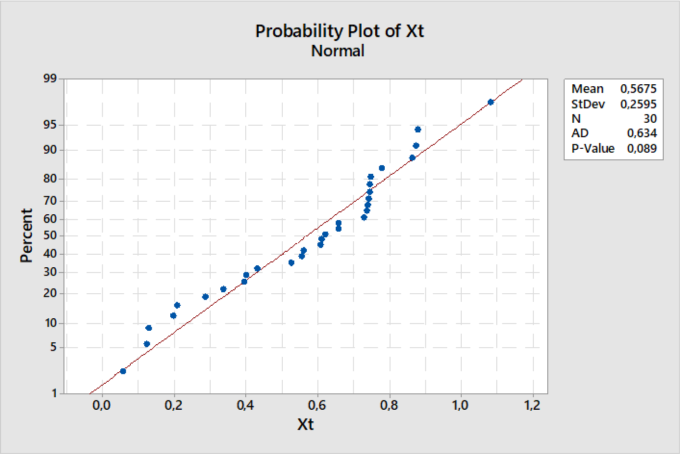




Box-Cox su X:



Normality X trasformata:



Randomness:

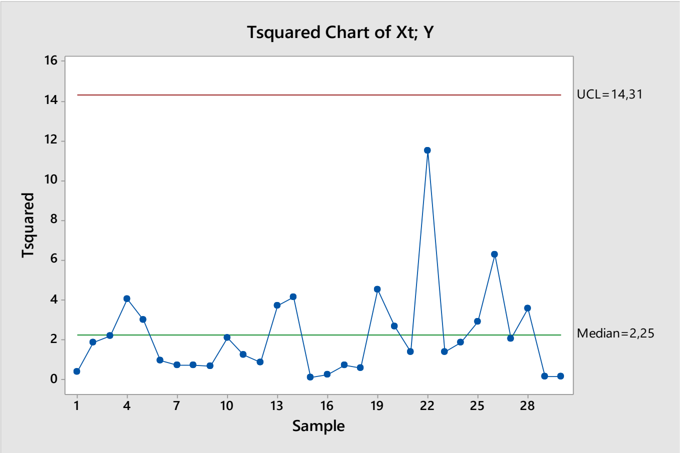
Runs Test: Xt

```
Runs test for Xt

Runs above and below K = 0,567499

The observed number of runs = 16
The expected number of runs = 15,7333
17 observations above K; 13 below
P-value = 0,920
```

Runs Test: Y

```
Runs test for Y

Runs above and below K = 9,99416

The observed number of runs = 12
The expected number of runs = 15,4
18 observations above K; 12 below
P-value = 0,187
```

T2 (long term) – alfa=0.0027:



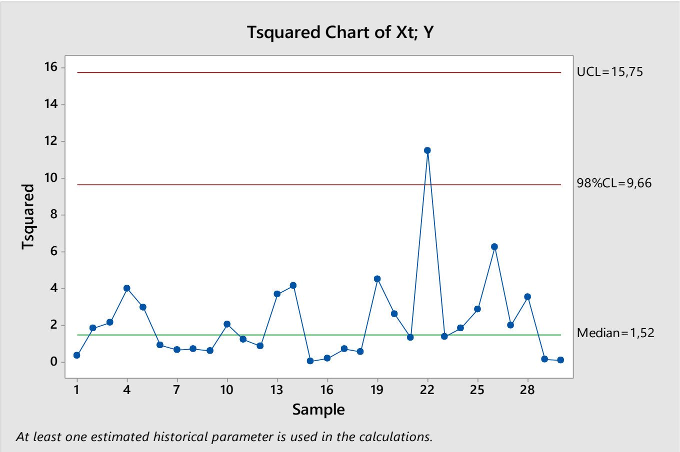
2) Variance – covariance matrix short-term:

S=0.5*(v'*v)/29

S =

0.0765	0.0289
0.0289	0.6512

T2 (short term) – alfa=0.02:

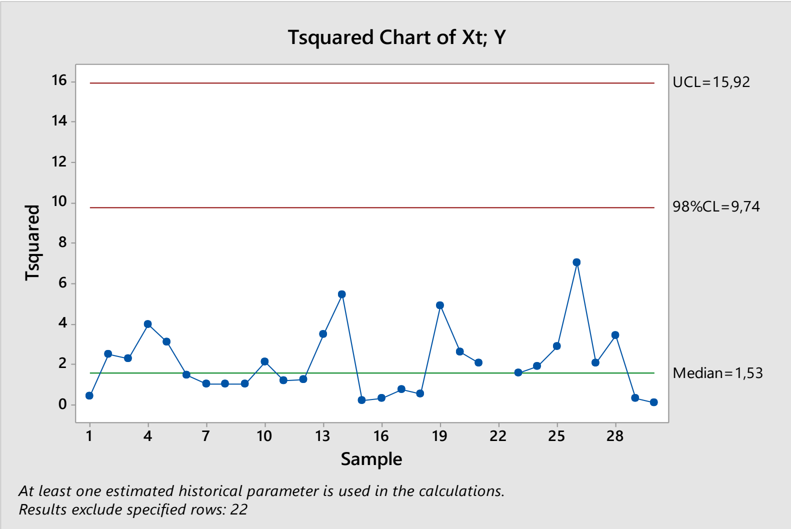


I assume assignable cause, remove the data and restate the variance-covariance matrix:
Short-term variance-covariance matrix without outlier:
S=0.5*(v'*v)/28

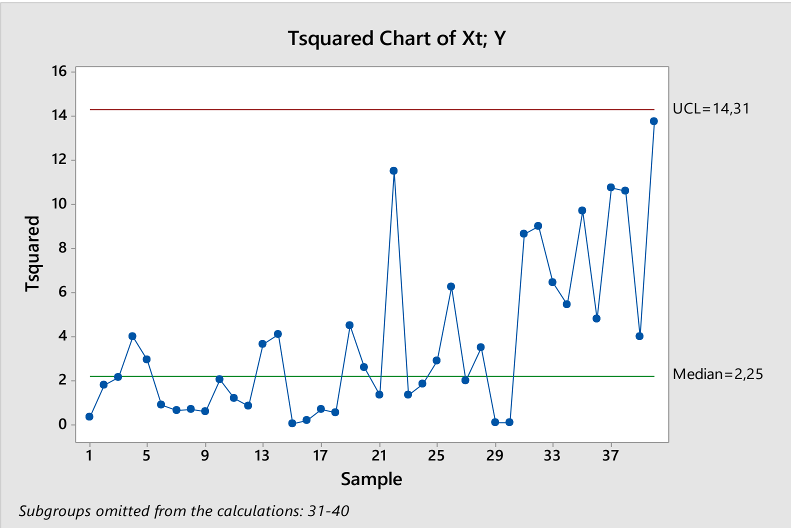
S =

0.0803	0.0247
0.0247	0.5279

T2 short term without outlier:



3) T2 with new data



Exercise 3 (max score 13)

In order to determine the health state of a machine tool, a check-up analysis is repeated 30 times (once per week). During this analysis, a repeatable batch of operations is performed and three sensor signals are acquired, and the mean values of those signals is stored. The signals are the following:

X: vibration rms [m^2/s],

Y: spindle torque [Nm], and

Z: spindle temperature [°C]).

The end-user wants to know if the health conditions of the machine were stable during the entire monitoring period.

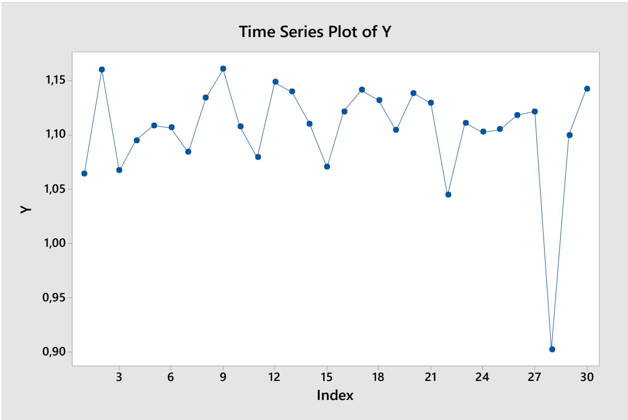
	X	Y	Z
1	0,94	1,06	22,47
2	1,31	1,16	28,33
3	1,06	1,07	25,78
4	1,16	1,09	25,18
5	1,32	1,11	25,85
6	1,40	1,11	21,81
7	1,00	1,08	24,92
8	1,44	1,13	29,47
9	1,34	1,16	26,42
10	1,21	1,11	25,49
11	1,19	1,08	23,40
12	1,40	1,15	25,89
13	1,17	1,14	26,35
14	1,23	1,11	26,41
15	1,23	1,07	24,17
16	1,37	1,12	25,05
17	1,48	1,14	30,56
18	1,31	1,13	20,52
19	1,19	1,10	30,49
20	1,33	1,14	26,60
21	1,26	1,13	27,99

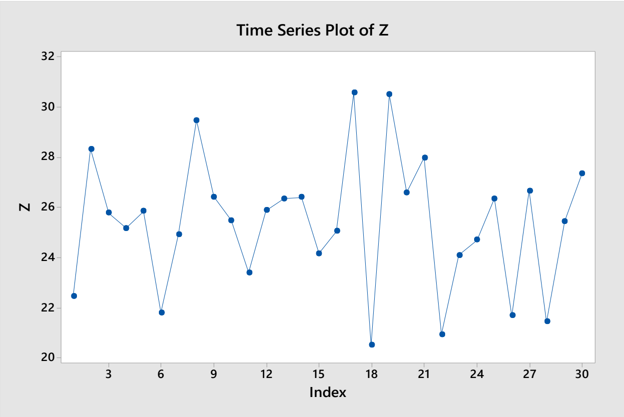
22	1,22	1,04	20,96
23	1,39	1,11	24,11
24	1,28	1,10	24,71
25	1,25	1,11	26,35
26	1,32	1,12	21,70
27	1,27	1,12	26,67
28	1,04	0,90	21,44
29	1,19	1,10	25,46
30	1,20	1,14	27,37

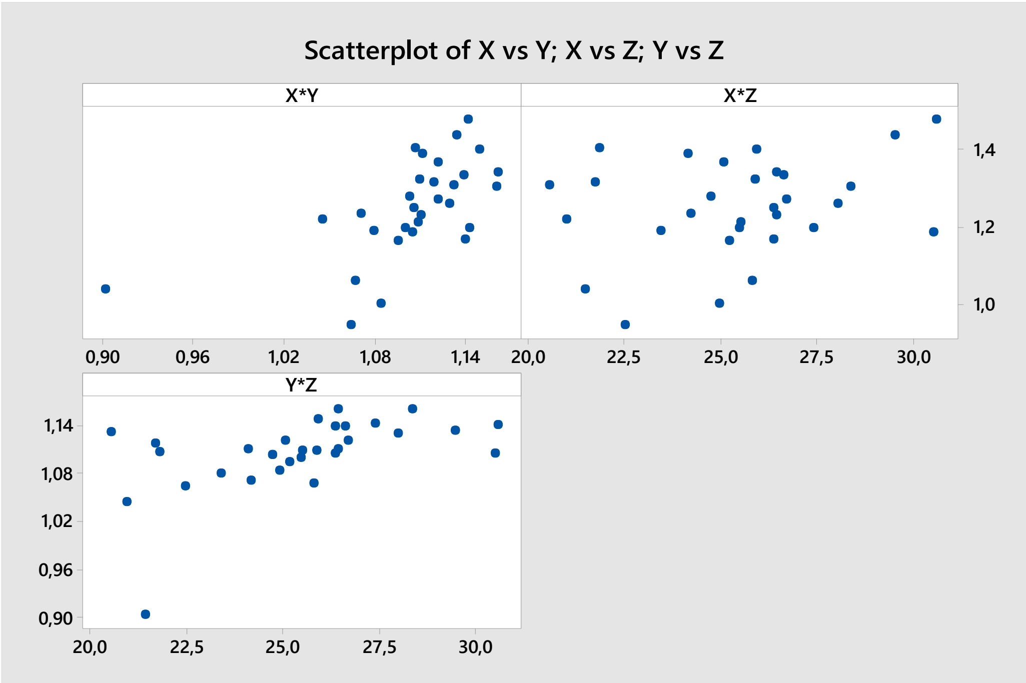
1) Assuming that no additional information was collected together with the signal data, propose a method to reduce the dimensionality of the problem in such a way to capture at least 85% of the overall variability. How many principal components (PCs) are needed? Discuss the results (include the plots of the loadings and, if possible, their interpretation).
2) Design a T2 chart based on long-term variance-covariance estimator and discuss the result (show the chart - qualitatively - and report the value of the control limit).
3) Design a T2 chart based on short-term variance-covariance estimator and discuss the result (show the estimated var-covar matrix, a qualitative plot of the chart, and report the value of the control limit).

Exercise 3 (solution)

1) Time-series plots and scatter plots:





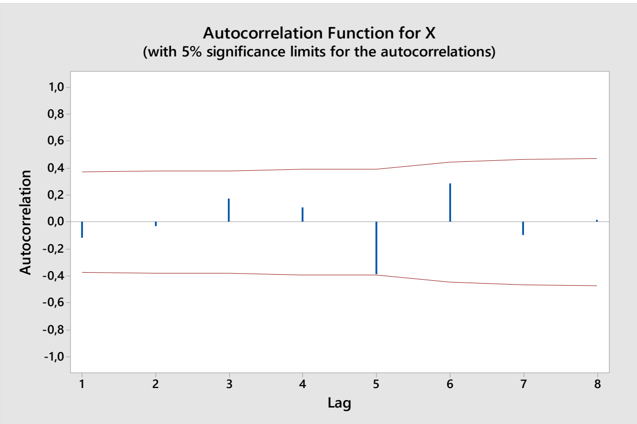


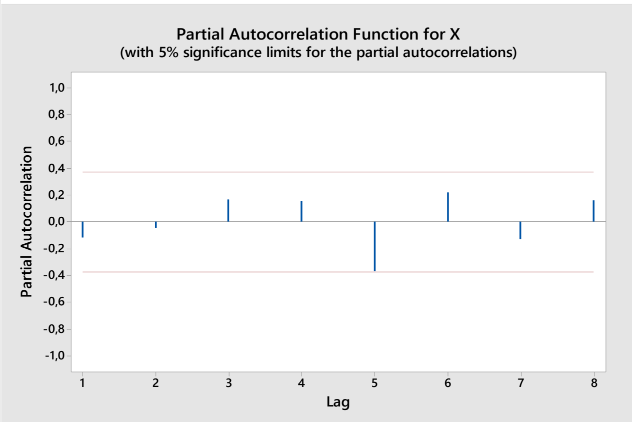
There is a suspect on an outlier in the signal Y. No information is available about possible causes, thus, let's first check the assumptions.

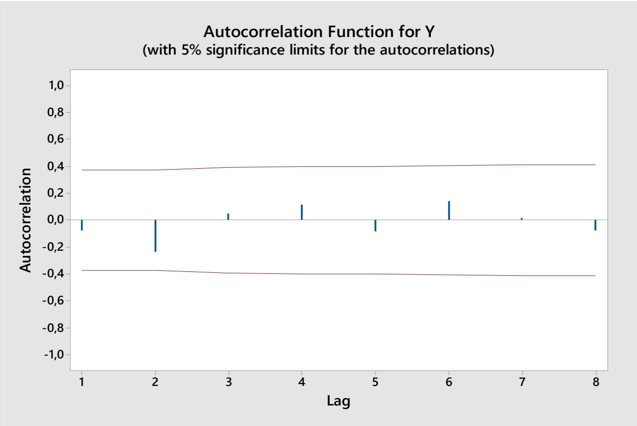
Randomness:

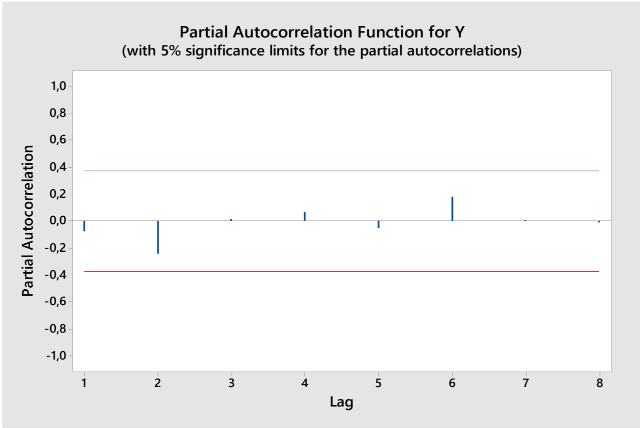
```
Runs test for X


Runs above and below K = 1,24952


The observed number of runs = 17
The expected number of runs = 16
15 observations above K; 15 below
P-value = 0,710
```

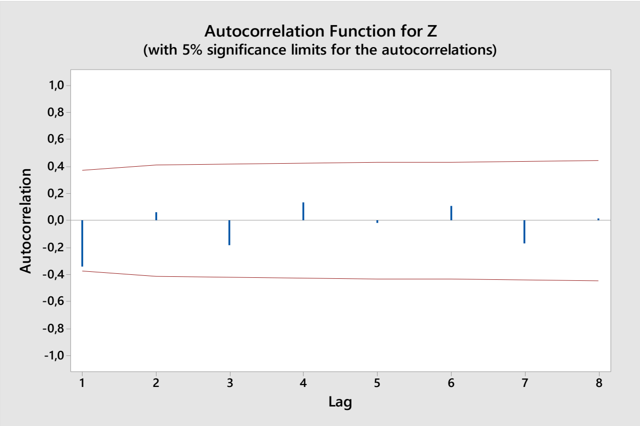
```
Runs test for Y

Runs above and below K = 1,10513

The observed number of runs = 18
The expected number of runs = 14,9333
19 observations above K; 11 below
P-value = 0,219

Runs test for Z

Runs above and below K = 25,3968

The observed number of runs = 18
The expected number of runs = 15,7333
17 observations above K; 13 below
P-value = 0,391
```

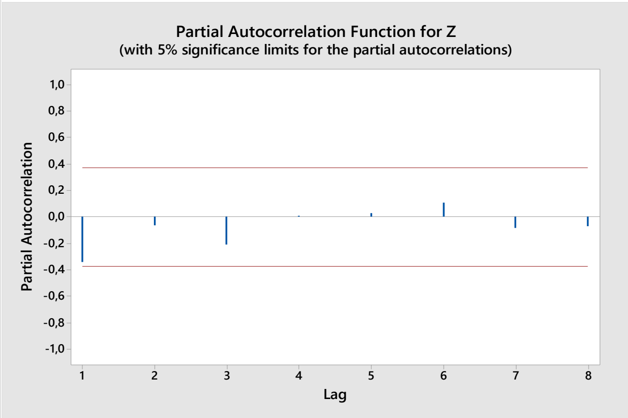






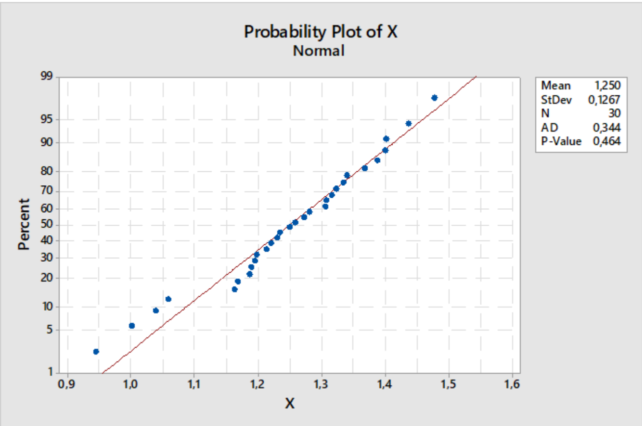


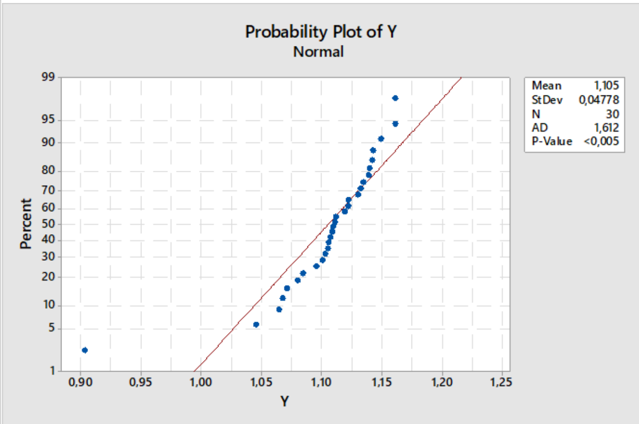


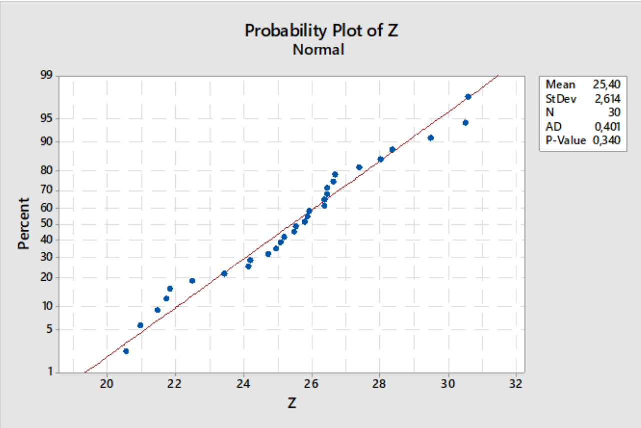


There is no statistical evidence of violations of the randomness assumption.

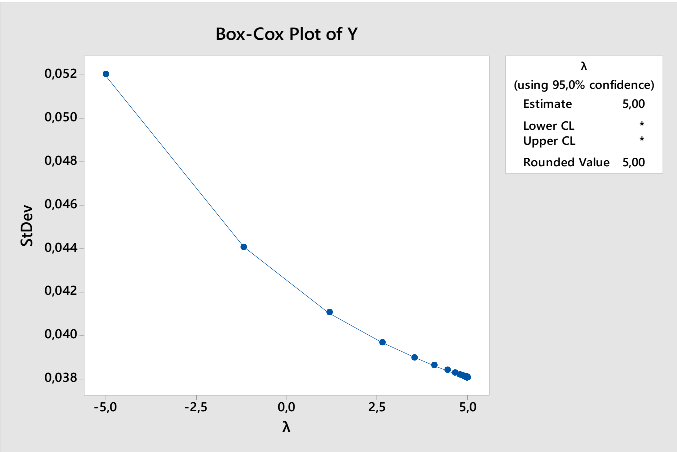
Normality (marginal):





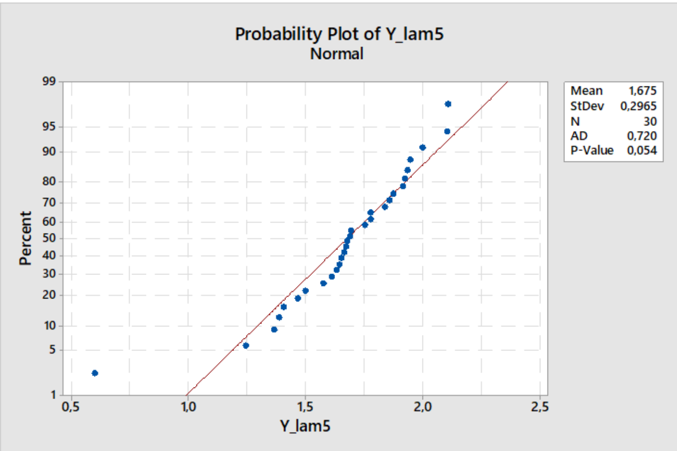


The signal Y violates the marginal normality assumption. Let’s try to apply Box-Cox:



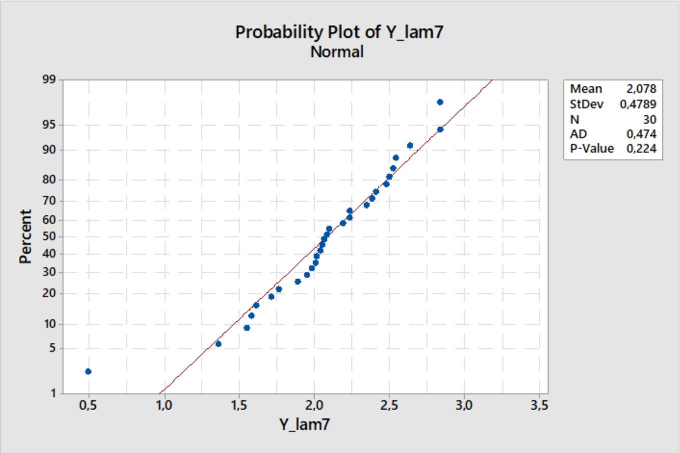
The Box-Cox method does not converge. This means that: (i) a power transform is not suitable to transform the data, (ii) the suitable power is larger than $\lambda = 5$.

If we set $\lambda = 5$, normality is barely acceptable at 5%:



If we set $\lambda > 5$, e.g.: $\lambda = 7$, the closeness to normality increases.

In both cases there is no need to remove the outlier.



In the following, we will consider the transformation $\lambda = 5$.

Apply the PCA based on Correlation Matrix, as the three signals are defined on different scales.

Principal Component Analysis: X; Y_lam5; Z

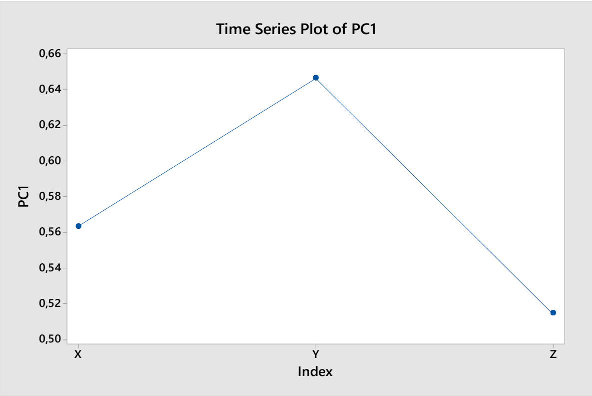
Eigenanalysis of the Correlation Matrix

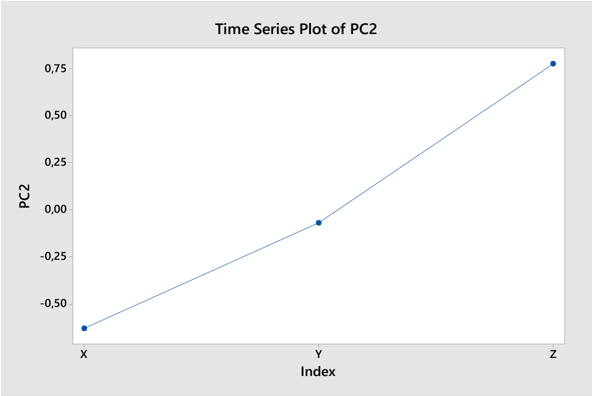
Eigenvalue	1,9807	0,7261	0,2932
Proportion	0,660	0,242	0,098
Cumulative	0,660	0,902	1,000

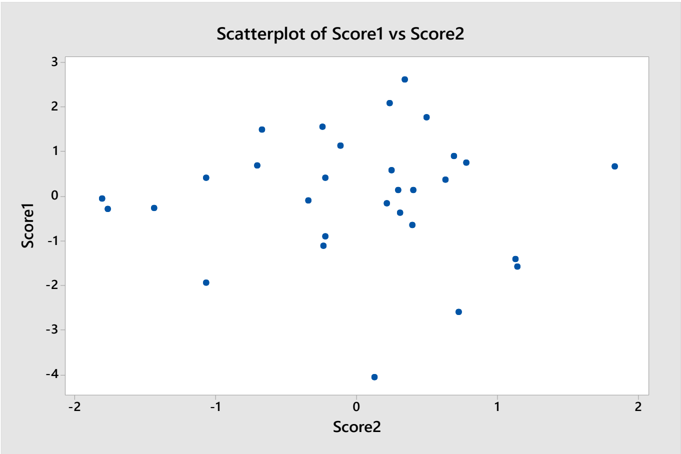
Variable	PC1	PC2	PC3
X	0,563	-0,629	0,536
Y_lam5	0,646	-0,069	-0,760
Z	0,515	0,775	0,368

In order to retain at least 85% of the overall variability, the first two PCs are retained.

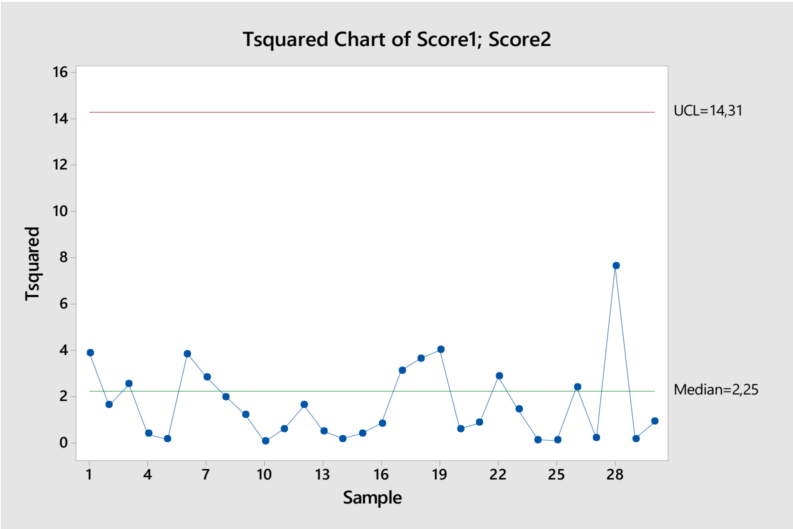
Loadings:





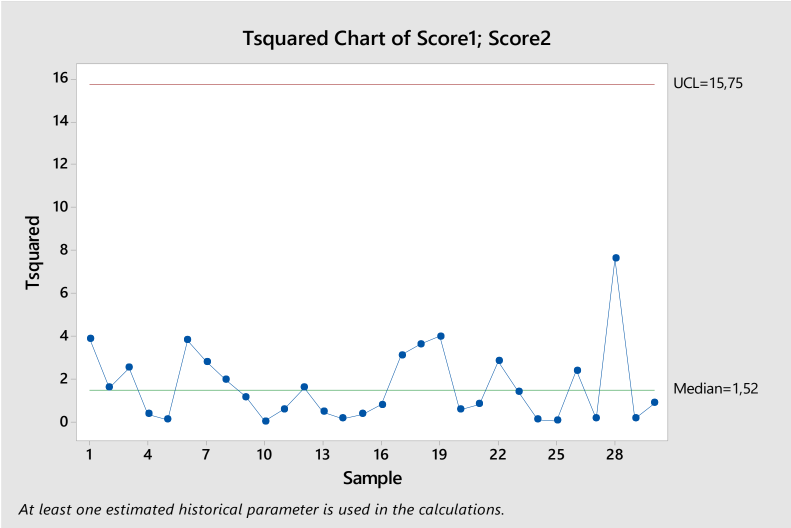


2) T2 chart with long term variance-covariance estimator:



3) T2 chart with short term variance-covariance estimator:

S=[2.1869 0.0870; 0.0870 0.8582].



The process is in-control (stable health conditions of the machine), but cycle 28 deserves some attention.

Exercise 4 (max score 14)

In order to monitor the stability of an additive manufacturing process, two variables, X and Y are measured via in-situ sensors. They represent, respectively, the diameter of the melt pool and the height of the plasma vapor generated by the process. Both the two measurements are performed in three different locations. During a small portion of the process, the following measurements were collected:

	X			Y		
Layer 1	15,64	15,33	17,72	22,73	22,03	22,70
Layer 2	14,29	16,71	17,65	21,00	21,18	23,09
Layer 3	15,39	16,26	17,57	22,42	19,14	22,77
Layer 4	16,49	16,38	18,65	25,33	24,67	23,78
Layer 5	15,03	14,69	17,95	21,79	24,54	23,12
Layer 6	16,08	15,73	17,89	26,72	24,05	22,27
Layer 7	14,78	16,06	17,29	21,09	25,85	20,41
Layer 8	15,48	14,90	17,61	21,11	22,78	22,62
Layer 9	14,00	16,91	18,46	17,96	25,78	24,83
Layer 10	14,67	16,98	18,30	20,43	26,29	23,24

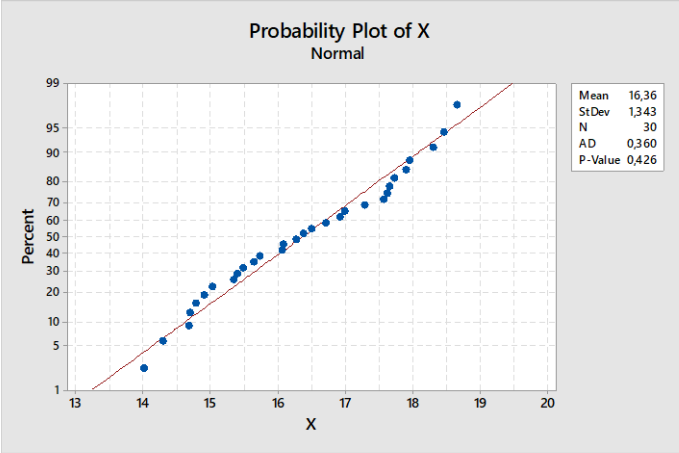
1. Design a control chart for the mean, assuming that the variance-covariance matrix is known (ARL_0=100):

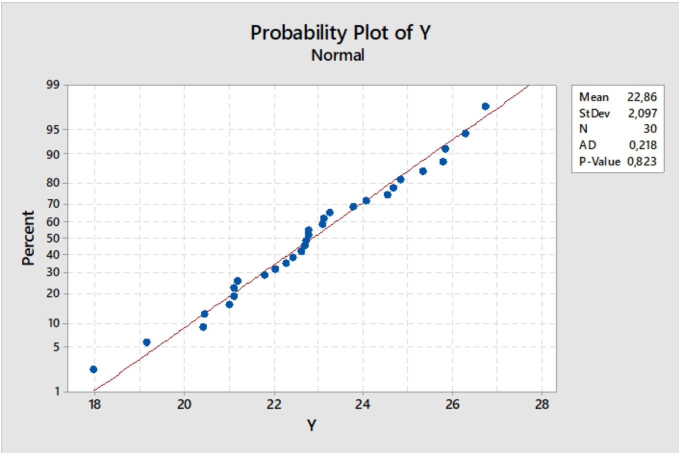
   SIGMA = [2,45 1,12
   1,12 4,92]

2. The head of the quality assurance department thinks that the variability of the 9th sample is out-of-control. Verify whether he is right or not (consider the same ARL_0 of the previous point)
3. Design a control chart for the linear combination of the sample means of the two variables that maximises the amount of explained variability, with ARL_0=100. Specify the weights of the linear combination. *Hint: use, for your analysis, the correlation matrix of the sample means.*
4. What is the average number of samples one has to wait before the chart designed at point 3 signals a shift of the mean equal to 1.5 units of standard deviation?

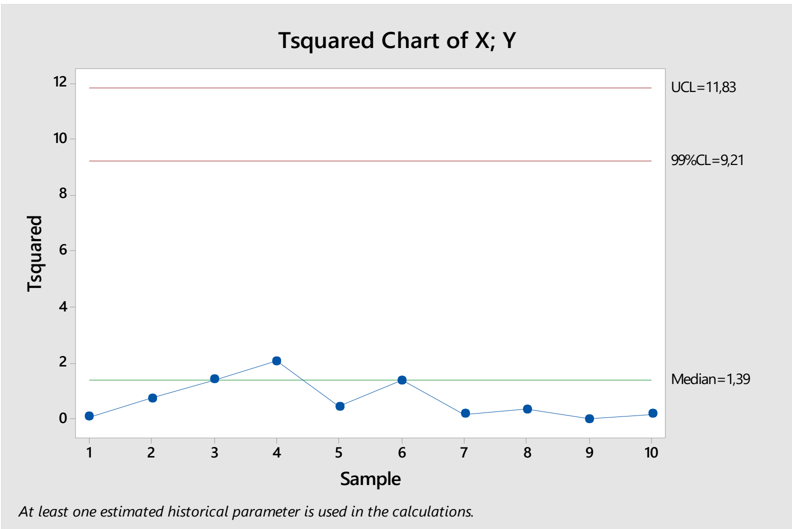
Exercise 4 (solution)

1) Let’s check about data normality (marginal normality is OK, we assume that joint normality is OK as well).





The T2 control chart with known SIGMA and ARL_0=100 is the following:



2) In order to determine if the variability of the 9^{th} sample is in-control or not, it is possible to use a Whishart control chart. The variance-covariance matrix is known. Then, we can estimate the sample variance-covariance matrix for the 9^{th} sample and the Whishart statistic as follows:

Control statistic (k-th sample):

$$W_K = pn(\ln(n)-1) - n \cdot \ln\left(\frac{|\underline{A}_K|}{|\underline{\Sigma}|}\right) + tr\left[\underline{\Sigma}^{-1}\underline{A}_K\right] \qquad \underline{A}_K = (n-1)\underline{S}_K$$

We also know that:

$$W_K \sim \chi^2\left(\frac{p(p+1)}{2}\right)$$

In this case we have:

$$A_9 = \begin{bmatrix} 10,25 & 17,31 \\ 17,31 & 36,43 \end{bmatrix}$$

$|A_9| = 73,77$

$W_9 = 4,17$

The upper control limit is the alfa-percentile of the chi-square distribution, i.e., UCL = 11,34

Thus, the variability of the 9^{th} sample is in-control.

3)The linear combination that maximises the explained variability is the first Principal Component. In this case we have two different quantities, thus we may apply the PCA to the correlation matrix of the sample means of the variables.

Principal Component Analysis: Xmean; Ymean

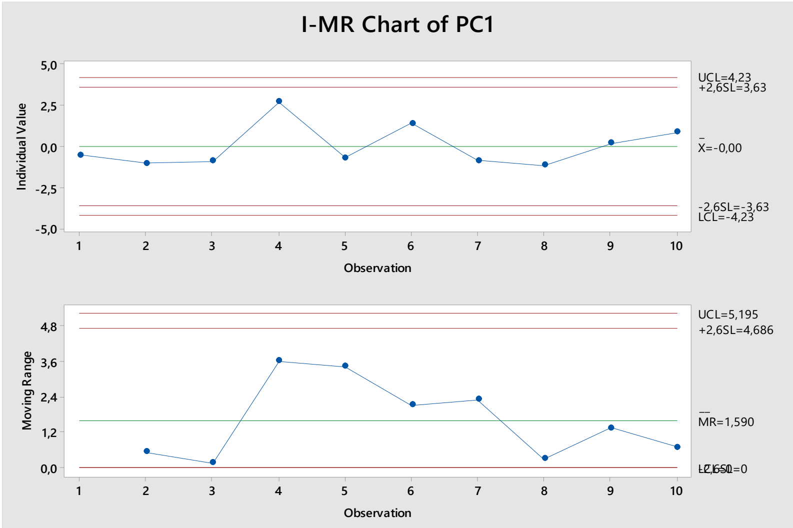
Eigenanalysis of the Correlation Matrix

Eigenvalue	1,6382	0,3618
Proportion	0,819	0,181
Cumulative	0,819	1,000

Variable	PC1	PC2
Xmean	0,707	0,707
Ymean	0,707	-0,707

The first PCs has equal weights, w1 = 0,707 and w2 = 0,707.

Since we have now a univariate individual variable, we may design an I-MR chart (with K = $z_{alfa/2}$ = 2,576).



4) The standard deviation is $\overline{MR}/d_2$.
We can compute the Type II error as $\beta = \Pr\,(LCL \leq I \leq UCL|H_1)$, for $\delta = \frac{\mu_1 - \mu_0}{\sigma} = 1.5$

The result is $\beta = 0{,}859$

The corresponding ARL_1 is ARL_1=7,09

Exercise 5 (max score 11)

A medical prosthesis is produced via metal additive manufacturing. The quality of the process is monitored by measuring four quality characteristics, namely X1, X2, X3 and X4. In each process run, one prosthesis is randomly picked up and the four quality characteristics of interest are measured. The following table shows the measurements acquired on 15 consecutive process runs.

X1	X2	X3	X4
11,9	9,6	8,1	6,4
11,6	10	9,1	6,5
9,1	11,1	8,7	8,8
10,5	11,5	7,5	6,1
10	9,7	8,3	7,5
9,4	9,3	8,6	8,6
10,9	8,5	10,2	7,1
10,1	9,8	6,9	7,2
10	11,6	7,3	7,2
10,7	9,4	9,5	8,7
11,2	11,2	6,6	6,6
10,9	12	6,3	7,8
12,1	12,5	9	8
12,3	11	6,9	7,4
10,9	12,3	8	8,3

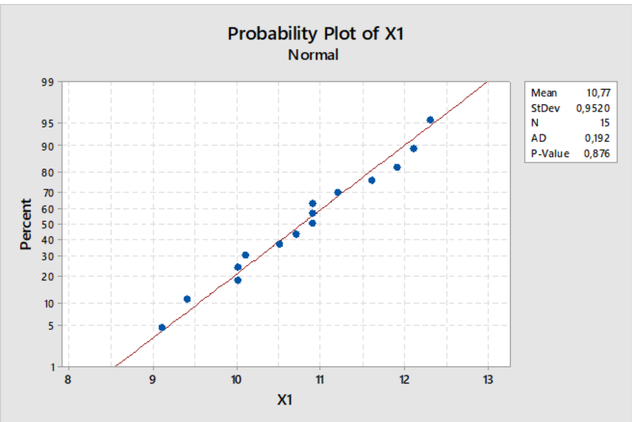
Assume the mean vector to be known and equal to $\mu = [10{,}0\ 10{,}0\ 8{,}0\ 8{,}0]'$.

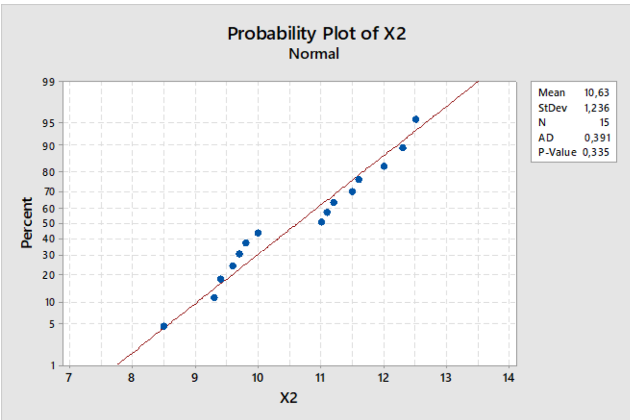
1) Design an Hotelling's T^2 chart where the variance-covariance matrix is estimated by using the short-term estimator and ARL_0=500.
   (note: Show the variance-covariance matrix)

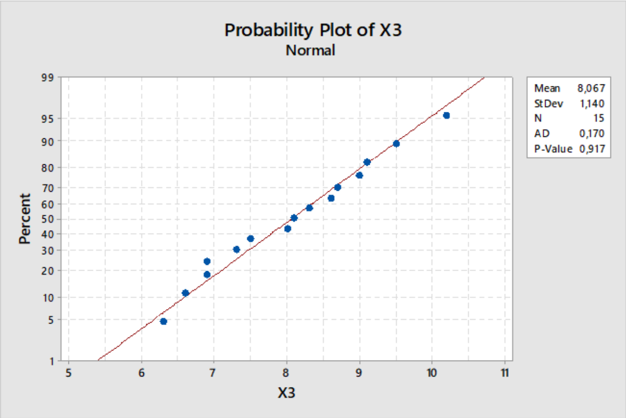
Exercise 5 (solution)

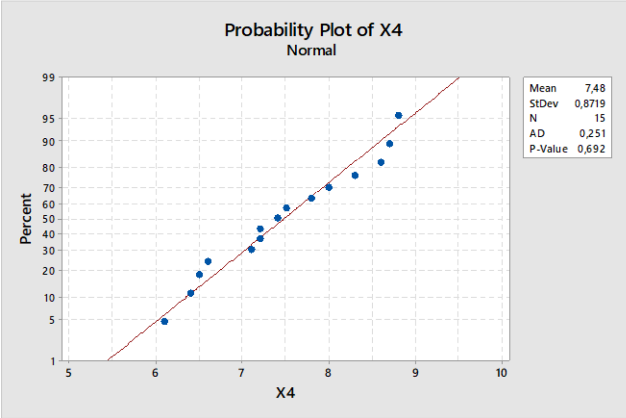
1) Check of assumptions.

Data are marginally normal (we assume multivariate normality).









There is evidence of lack of auto-correlation for each variance, and hence we can assume independence.

```
Runs test for X1

Runs above and below K = 10,7733

The observed number of runs = 5
The expected number of runs = 8,46667
8 observations above K; 7 below
* N is small, so the following approximation may be invalid.
P-value = 0,062

Runs test for X2

Runs above and below K = 10,6333

The observed number of runs = 6
The expected number of runs = 8,46667
8 observations above K; 7 below
* N is small, so the following approximation may be invalid.
P-value = 0,184
```

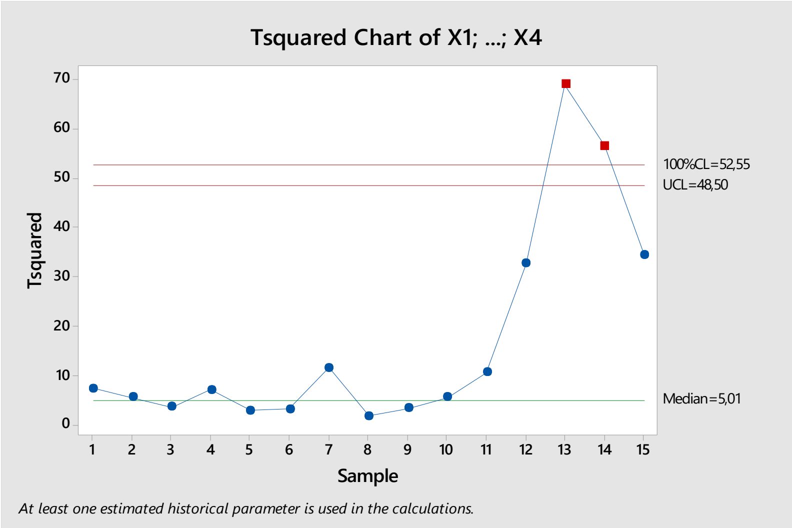
```
Runs test for X3

Runs above and below K = 8,06667

The observed number of runs = 8
The expected number of runs = 8,46667
8 observations above K; 7 below
* N is small, so the following approximation may be invalid.
P-value = 0,802

Runs test for X4

Runs above and below K = 7,48

The observed number of runs = 10
The expected number of runs = 8,46667
7 observations above K; 8 below
* N is small, so the following approximation may be invalid.
P-value = 0,409
```

The short-term variance-covariance matrix is:

S =

0.5743	-0.2139	0.1750	-0.5268
-0.2139	0.8361	-0.4036	-0.1461
0.1750	-0.4036	1.5425	0.4629
-0.5268	-0.1461	0.4629	0.9761

The resulting T2 control chart is the following:



There are two points out-of-control. No assignable cause is assumed to be present. The control chart design is over.

Exercise 6 (max score 4)

Consider the dataset of Exercise 5 and assume that the variance-covariance matrix is known and equal to:

SIGMA =

0.5900	-0.2139	0.1750	-0.5268
-0.2139	0.8400	-0.4036	-0.1461
0.1750	-0.4036	1.5600	0.4629
-0.5268	-0.1461	0.4629	0.9800

Design a control chart for the linear combinations of the four variables that explain at least 80% of the data variability, with ARL_0=500. Specify the weights of the linear combination.

Exercise 6 (solution)

By applying the eigen-decomposition to the known SIGMA matrix we have:

Eigenvalues: λ_1 = 1,9751, λ_2 = 1,2782, λ_2 = 0,632, λ_2 = 0,0748

The first two principal components explain about 82.15% of the variability.

The weights of the first two principal components are:

```
U1:
-0.0093
-0.3490
0.8269
0.4408

U2:
0.6534
-0.3087
0.2253
-0.6534
```

In order to compute the scores, the original data must be projected onto the direction spanned by the first principal component.

The result is:

Z1 =	Z2 =
6.0587	2.4554
6.7929	2.2959
7.1152	-1.2703
4.7802	1.0149
6.6916	0.5093
7.5697	-0.4104
8.4970	2.1573
5.3659	0.4244
5.0694	-0.1065
8.3111	0.5456
4.3545	1.0354
4.3590	-0.2593
6.4943	0.8481
5.0149	1.3608
5.8805	-0.2956

The Hotelling's T2 control chart on the first 2 principal components is: