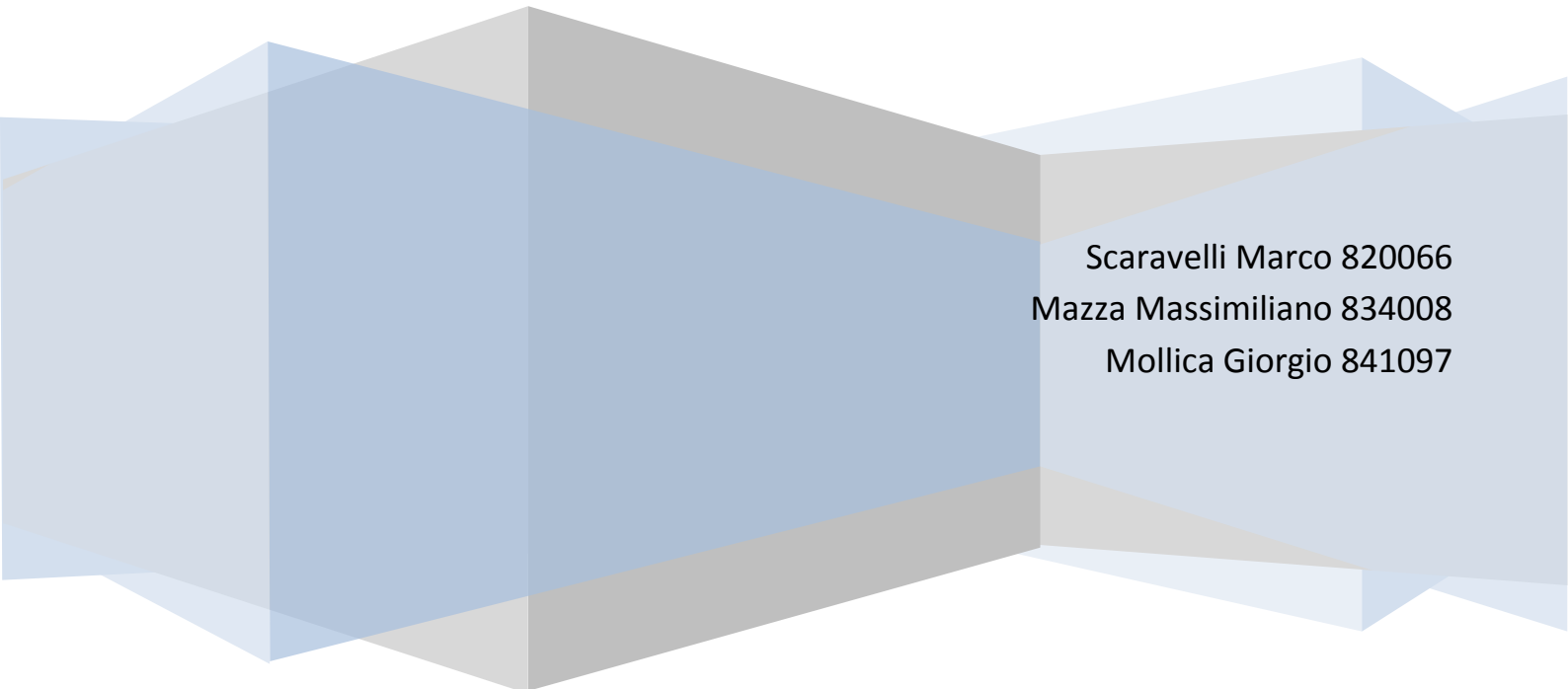


**Group 10**

# **Quality engineering**

## **Project work**



Scaravelli Marco 820066  
Mazza Massimiliano 834008  
Mollica Giorgio 841097

## Sommario

<b>SUMMARY</b>	<b>2</b>
<b>PHASE 1</b>	<b>3</b>
Data Finding	3
Indicator analysis	3
Design	6
<b>PHASE 2</b>	<b>15</b>
<b>CONCLUSION</b>	<b>26</b>
<b>BIBLIOGRAPHY</b>	<b>26</b>

## Summary

*"In God we trust, all others bring data"*

William Edwards Deming

The object of this project was to build a classifier that was able to detect and analyze a sample from a simulated production process. We tried to study several techniques and we compared their features on finding shifts on our process.

Firstly, we have received a sample of 30 images of cylindric section of a metal which is characterized by some pores. Further to this point we had to design a control chart in order to monitor the porosity. Analyzing the images through ImageJ we have created a dataset which contains all the variables that we have considered compelling for construction of our control charts.

Taking one step further in the analysis we have made on the process is divided into two parts:

- *Principal Component Analysis* (PCA) with a multivariate control chart
- *Univariate Control Charts* on three indicators we believe, based on a correlation matrix analysis,

are the most representative for the process

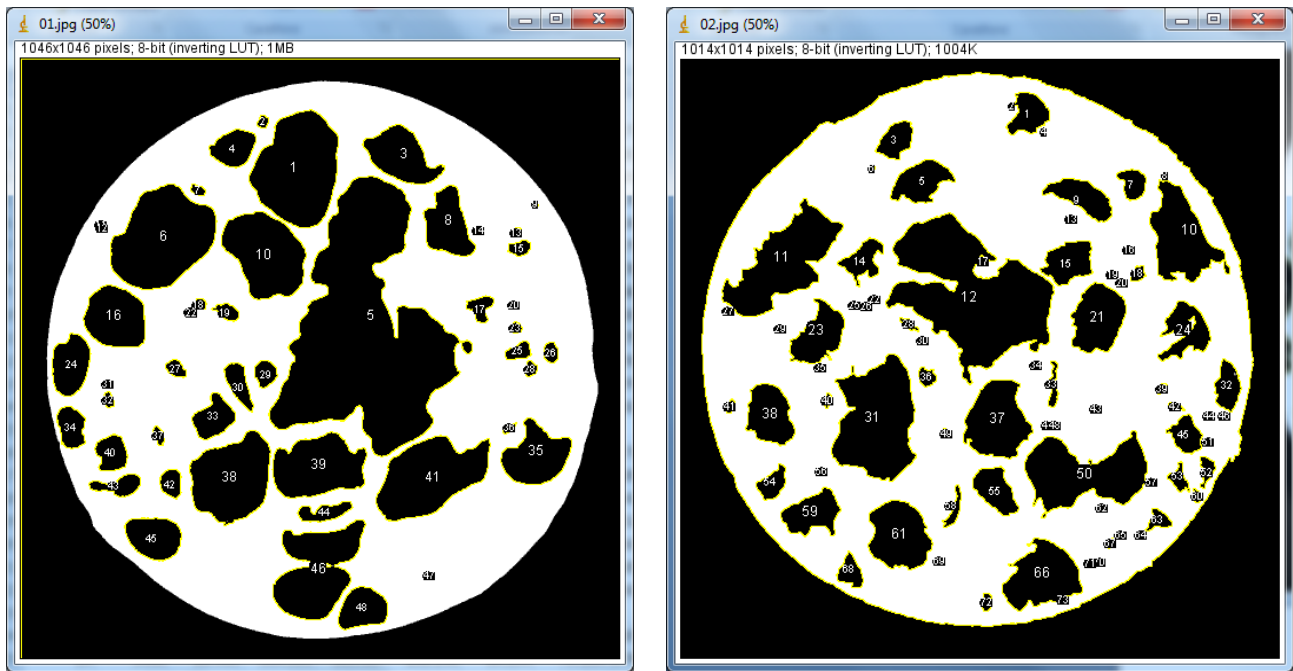
Several corrections and variations of these methods are applied and tested.

Subsequently, we received the other 27 images on which make the control phase using the Control Charts that had been previously designed.

## Phase 1

### Data Finding

We used the software “imageJ” to collect data from the images provided to us. Using the function “Analyze Particles” (after the threshold) on all the 30 sample images we created a dataset for every image with the relative pores (*Dataset\_pores.xls*).



We eliminated pores smaller than  $3 \text{ pixel}^2$  because they could have distorted our next analysis and small pores weren't detected in every image.

We summarized data for each image and then, working on these data, we created a new dataset (*Dataset\_design.xls*), which is the one used to design the Control Charts.

### Indicator analysis

Sample	Count	Pixel^2	Total Area	Average Size	Area	AVGsize/ Pixel^2								
%Area (porosity)	solidity of pores	circ of pores	AR of pores	AVG position of pores	Entropy(E)	Gini	RHT	ln(RHT*1000 )	A/2P	X	Y	X_C	Y_C	

#### Legend:

The yellow indicators are strumental indicator used to construct the others or used in preliminary analysis. The final indicators were chosen on interpretational vision and in order to offer a wide characterization of the images. Moreover indicators were corrected or adjusted in order to pass a normality test.

Sample: image

Count: number of pores

Pixel<sup>2</sup>: area of image in pixel<sup>2</sup> (it changes usually from picture to picture)

Total area: total area of pores

Average size: the average dimension of the pores for each image

Area: is the area of the cylinder section in the image

AVGsize/Pixel<sup>2</sup>: is the average size of the pores in each image divided for the pixels of the same image.

#### **data about pores:**

$$\text{\%Area (porosity): } \frac{\text{Total Area of pores}}{\text{Area of cylinder section}}$$

This indicator represent the definition of porosity (or void fraction) applied in two dimensional space (usually it is used with the volumes).

$$\text{Solidity of pores: } \frac{\sum_{i=1}^N \frac{\text{area of pore}_i}{\text{convex area of pore}_i} \frac{2p_i}{\sum_{i=1}^N 2p_i}}{N}$$

the indicator  $\frac{\text{area of pore}_i}{\text{convex area of pore}_i}$  is directly given from the software ImageJ. In this phase we decided to weight it on the perimeter of the pores ( $\frac{2p_i}{2p_{\text{Total}}}$ ) because, as it is suggested in the guide of software, for small particles this kind of indicator (based on the shape) could be inaccurate and without such expedient the variable would be affected and biased by a large number of small pores. We couldn't use the ratio of the area of the pore over the total because it weighted excessively bigger pores. Thanks to this device, we were able to obtain an higher level from a normality test for this variable and the following ones.

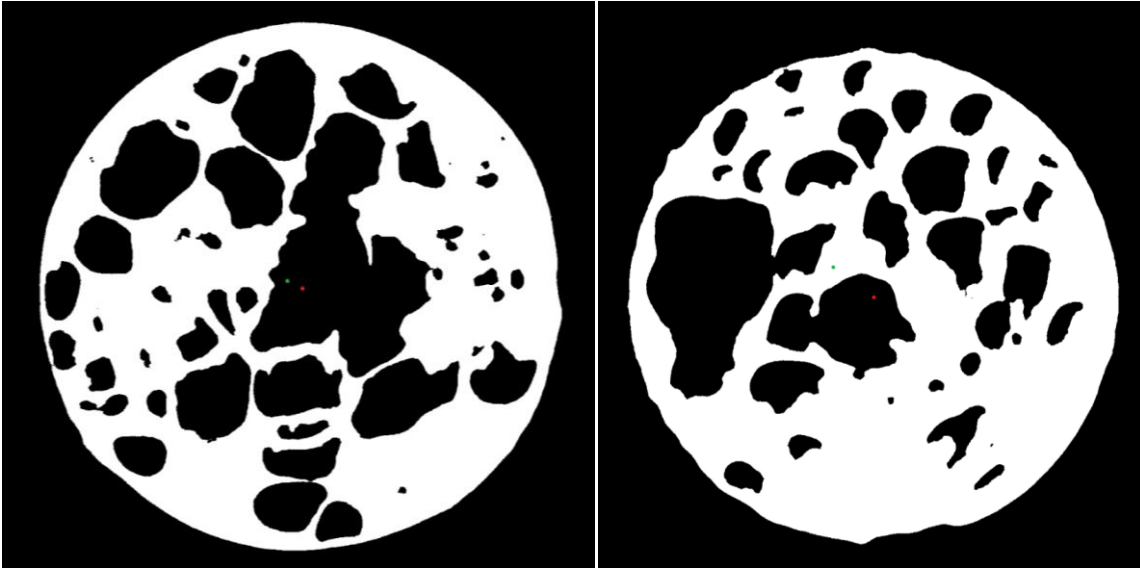
$$\text{Circularity of pores: } 4\pi * \frac{\text{Area of pore}}{\text{Perimeter of pore}^2}$$

with a value of 1,0 indicating a perfect circle. As the value approaches to 0, it indicates an increasingly elongated shape. Values might not be valid for very small particles so we decided to weight the value with the dimension of each pore: as we did with the previous variable, it was fundamental not to use the raw average given by the software but to weight each pore with the ratio of the perimeter pore out of the sum of the perimeters.

AR of pores: is the aspect ratio of the particle's fitted ellipse, i.e. [Major Axis]/[Minor Axis]. This is the average of the aspect ratio of the pores for each image weighted like the previous two indicators.

AVG position of pores: is the distance between the center of cylinder section and the weighted average of the center of pores. It's an indicator of the position of pores: a well distributed image would have the center near the zero. The idea of this variable was taken from the concept of center of mass.

To be more precise in the explanation of this variable, we suggest the following example:



X\_C and Y\_C are the coordinates of center (in red) while X<sub>m</sub> and Y<sub>m</sub> (in green) are estimated using the formula:

X	Y	X_C	Y_C
502,26	515,49	528	529
496,07	509,76	502	493
549,45	555,77	538	544
528,12	507,91	521	538
524,30	504,58	524	549
563,10	497,95	529	517
452,44	550,75	520	518
514,01	549,03	532	520
496,60	518,55	525	542
554,22	529,64	550	543

$$X_m = \sum_{i=1}^N \frac{A_i X_{P_i}}{A}$$

$$Y_m = \sum_{i=1}^N \frac{A_i Y_{P_i}}{A}$$

where A<sub>i</sub> is the area of the pore i, A is the total area of pores, N is the number of pores in the image and X<sub>P<sub>i</sub></sub> and Y<sub>P<sub>i</sub></sub> are the coordinates of the center of the pore i. X<sub>m</sub> and Y<sub>m</sub> can be considered like the coordinates of the symmetric point respect the “center of mass” of the cylinder section.

We used Euclidian distance but, in order to have comparable data, we divided the obtained value for the average radius of each image (because the distance is given in pixels and every image have different dimension in pixels). It represents therefore the percentage difference from the center.

#### CONCENTRATION INDEXES

We derived some variables from some economics courses. They usually represent market share of companies but can also bring value added by explaining some features of the pores. In particular the variables we are showing add some features regarding the numerosity of pores and their relative dimensions. An additional variable we analyzed, Herfindahl-Herschmann Index, was not picked because it was strongly negative correlated (and explained) with Entropy; moreover it suffered of non-normality.

Entropy:  $E = \sum_{i=1}^N \frac{A_i}{A} * \ln \left( \frac{A}{A_i} \right)$

This term is usually used in physics to indicate the degree of disorder of a system, but is used in other situations: this is infact one of the formula used in the economic world to define the concentration of a market. We decided to adapt this indicator to our case: small values of the index indicate high concentration of pores (e.g. if we have only one pore the value is 0).

Gini: 
$$G = \frac{1}{n} * (n + 1 - 2 * \frac{\sum_{i=1}^n (n+1-i) y_i}{\sum_{i=1}^n y_i})$$

It is an index used frequently in economy to evaluate inequality or for compare a market's concentration. Pores are ranked for growing area dimension and then the index is calculated using this formula: "n" is the number of pores, "i" is the rank of pore and  $y_i$  is the ratio  $A_i/A_{tot}$ , between the area of pore and the sum of the areas of pores.

RHT Hall-Tideman and Rosenbluth Index: 
$$RHT = \frac{1}{2 \sum_{i=1}^n i * y_i - 1}$$

This concentration index was developed by Hall and Tideman (1967) and Rosenbluth (Niehans, 1961) . It assumes similar features of Gini, but it gives the opposite rank of the data (bigger pores are weighted differently). Small values of the RHT indicate high concentration. The logarithm of this index was chosen in order to obtain a normal distribution across our sample. Data were multiplied to avoid negative values.

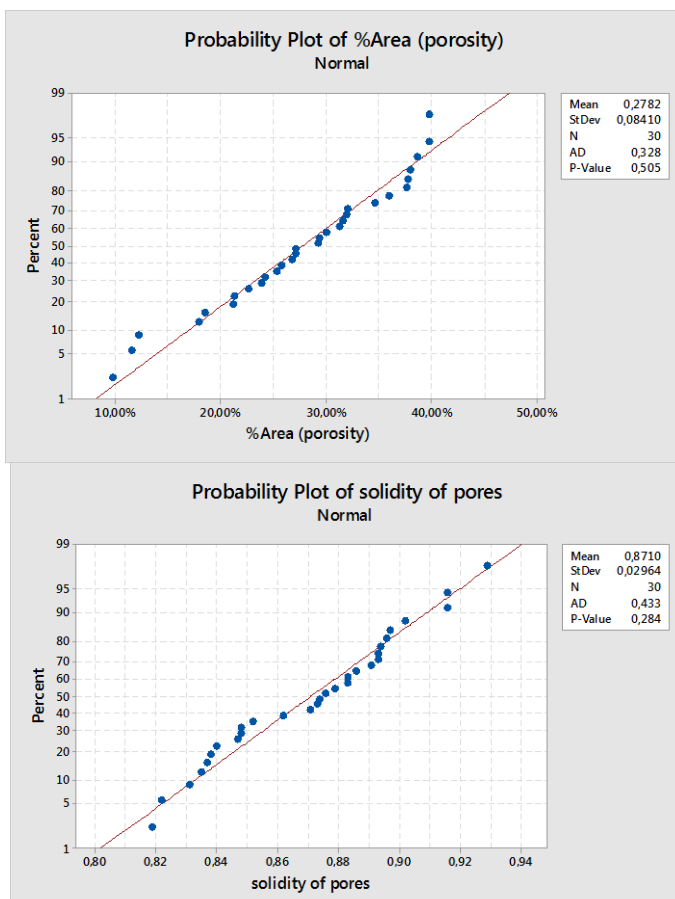
P/A:  $\frac{\sum_{i=1}^N P_i/A_i}{N}$  this index give us the information about the shape of the pore. It is the ratio between perimeter and the area. If the value is high the pore has a concave shape, otherwise the shape is convex.

## Design

### Normality test on data

In order to perform multivariate charts and T2 Hotelling's ones, a strong violation of the assumption of normality could lead to a biased and reliable confidence intervals. We therefore started to perform univariate tests and we for Normality testm, we chose to use Anderson-Darling's one.

The **P-value** is the smallest level of significance that would lead to rejection of the null hypothesis  $H_0$  (in our case the normality). We decide to set a limit equal to 5% of significance, if the Pvalue is higher we can not reject  $H_0$ .



Porosity→accept normality (Pvalue=0,505)

Solidity of pores→accept normality  
(P-value=0,284)

Circularity of pores → accept normality (P-value=0,922)

AR of pores → accept normality  
(Pvalue=0,362)

AVG position of pores → accept normality  
(Pvalue=0,109)

Entropy →accept normality (P-value=0,416)

Gini →accept normality (P-value=0,255)

Ln(RHT) →accept normality (P-value=0,689)

Thanks to the ajustement on some variables (e.g. weighting Circularity, AR and Solidity for the perimeter of each pore, or by taking the

logharitm of RHT), no rejections against normality were possible.

We also saw that data don't appear to be autocorrelated, by checking the Autocorrelation and partial correlation functions. Normality is assured on each variable: however in order to make assumptions and create confidence intervals, it is statistically helpful to test normality on every combination of the variables. We wanted to test if the data could be statistical represented as a multivariate normal.

$$\underline{X} \sim N(\underline{\mu}, \Sigma)$$

```
> mcshapiro.test(Dati)
$wmin
[1] 0.7437174

$pvaleu
[1] 0.3644

$devst
[1] 0.00962523

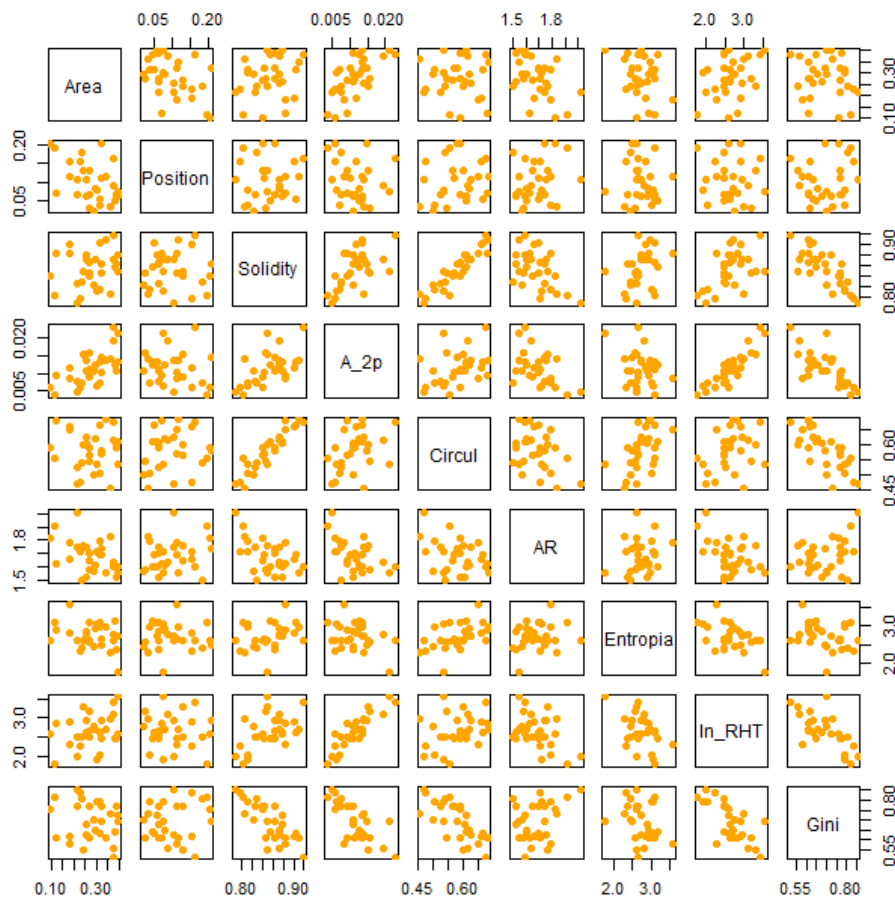
$sim
[1] 2500
```

We used the McShapiro test, in order to test normality of linear combinations of the data and the bivariate distributions.

The p-value is high, and there is no statistical significance to say that my data are not following a multivariate normal distributions. The power of this result enhance the reliability of confidence intervals we are building: not only each variable is normal, but also linear combinations will be.

## Scatterplot

A preliminar analysis of data were performed and Boxplot and Correlation Matrix were analyzed. The scatterplot of data shows the behaviour of each variable and how they covariate compared to the others:



Some patterns and correlated clustered variables can be perceived. However there is always some variation on almost every direction, which doesn't lead use to remove any further variable.

## Variance Inflation Factor

As last, we tested the multicollinearity of our data in order to find whether there were any problems among some variables: VIF is related to the linear regression of a variable from all the others high levels mean that the variable can be already explained by other variables.

Values > 10.0 may indicate a collinearity problem

Area	8.970
Position	2.060
Solidity	18.747
A_2p	28.903
Circul	19.117
AR	2.372
Entropia	23.016
ln_RHT	45.937
Gini	33.170

$VIF(j) = 1/(1 - R(j)^2)$ , where  $R(j)$  is the multiple correlation coefficient between variable  $j$  and the other independent variables

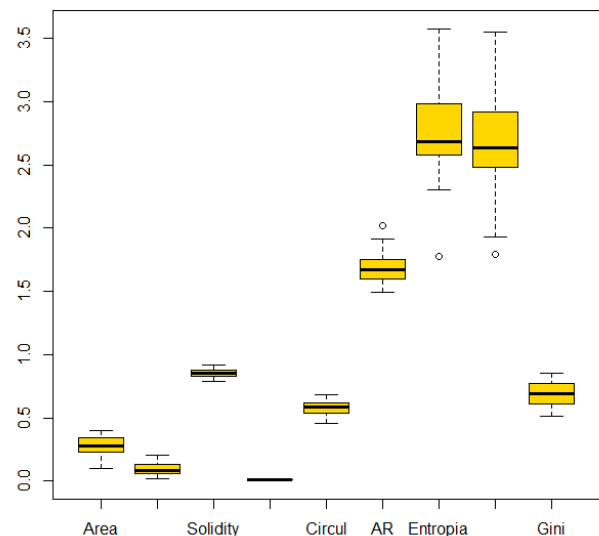
At this point several methods could be used to solve this problem: a solution could have been of eliminating some variables through a backward selection. However, we didn't want to lose any precious information available to characterize pores and we decided to transform data through the method of Principal Components. We will then select the directions that explain mostly the variability.

## Principal Components Analysis

We performed the Principal Components Analysis by using the software R.

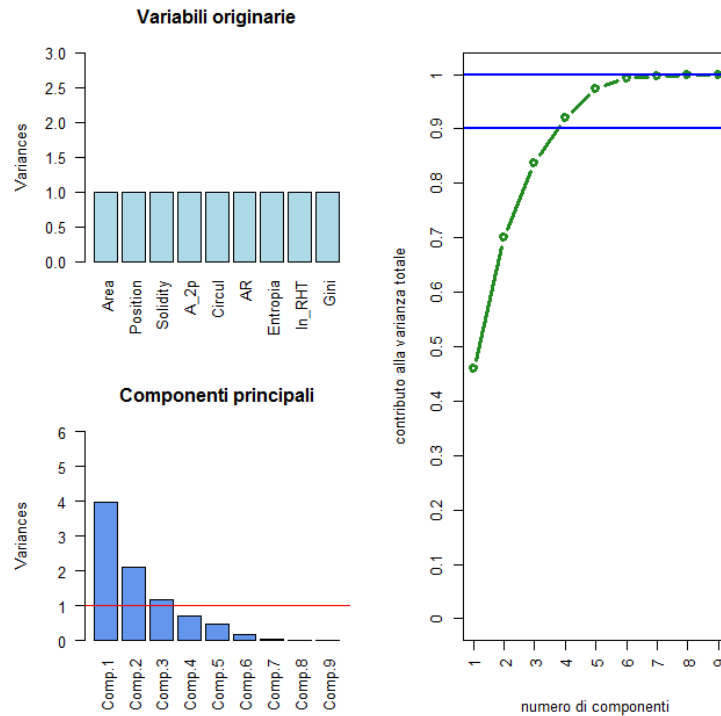
A boxplot of the 9 variables shows how their variability differs deeply and why they cannot be compared immediately.

We had to standardize the variables directly in the code in order to make comparable variation respect to the mean; in fact if some variables have a large variance and other small, PCA (maximizing variance) will prefer the large variances' variables. [For example if you change one variable from km to cm (increasing its variance), it may go from having little impact to dominating the first principle component]. If you want your PCA to be independent of such rescaling, standardizing the variables will do that. In this way we obtained loadings and scores (result\_R.xls file).

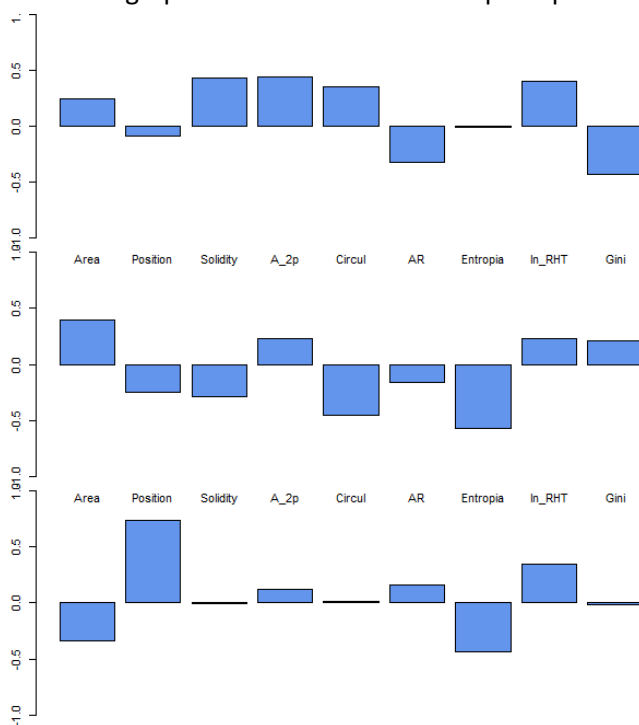


Our objective with PCA, was to work on a lower dimension space and at the same time to do variable selection and eliminating multicollinearity: we focused on the first three principal components because they catch more than 84% of the total variability of our data. In this way, it will be possible to work with lower dimensions data from the starting 9 variables while keeping variability.





In the next graph we show the first three principal components loadings that we have chosen to work on:



*I principal component:* it detects physical aspects of pores, such as the shape or how much they are concentrated.

*II principal component:* it gives focus on the portion of Area and the Entropia which affects the figure.

*III principal component:* it mainly characterizes the position (average distance from the center) of pores in the figure

By the time we had obtained these results, we decided to focus our analysis on three PC which maximize the percentage of variance explained (higher than 84% from graph). At this point the analysis was carried out on MINITAB and R through a Hotelling  $T^2$  Control Chart on the first three scores. (file T\_2-CC).

Setting  $ARL_0=100$  we have an  $\alpha = 1/100$  ( $p=3$ ,  $n=30$ ).

(We put in `t_squared options` → `limits` → `display additional components limit` → `percentile=0.99`)

Due to the low number of observations, a Chi Squared approximation could have led to significant errors.

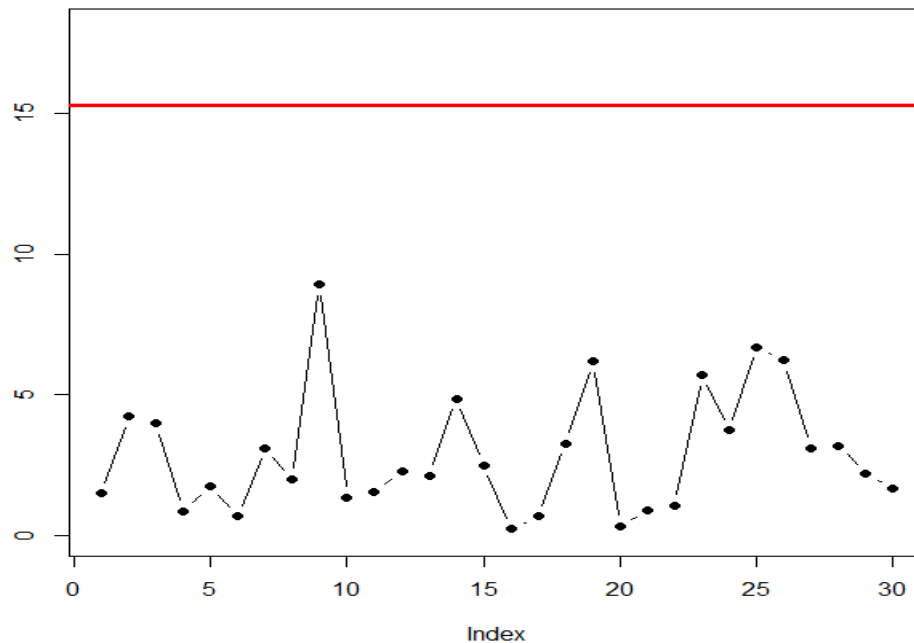
We therefore have used the Exact Fisher value for calculating the Upper Control Limit:

$$UCL = \frac{p(n-1)(n+1)}{n^2 - np} F_{p, n-p} = 15,32$$

with:

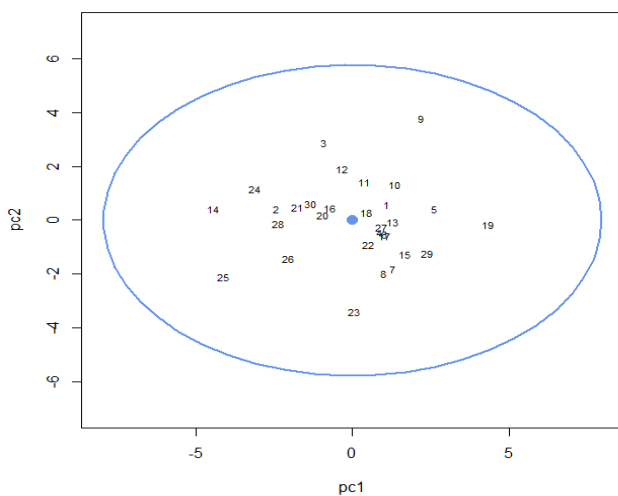
$$T_i^2 = \frac{\hat{y}_1^2}{\hat{\lambda}_1} + \frac{\hat{y}_2^2}{\hat{\lambda}_2} + \frac{\hat{y}_3^2}{\hat{\lambda}_3}$$

where  $y_i$  are the principal components and  $\lambda$  the eigenvalues.

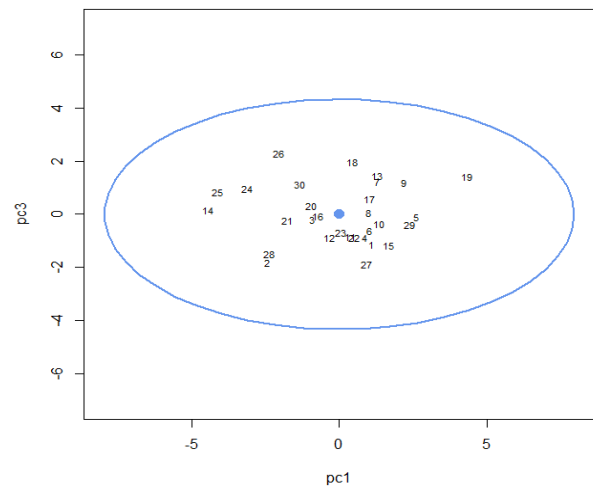


Data seem in control and the chart seem acceptable for future approssimation and evaluating future values. What we obtained is an ellipsoid prevision region and the T2 chart represents the results in a one dimension space. However, it is interesting to show the features of our Hotelling Chart in a 2-dimension combination of the principal components: once we obtain results, we will be able to plot them on our two dimensional graphs and analyzing on which component an eventual outlier will be out of boundaries.

*Plot of PC1 and PC2*

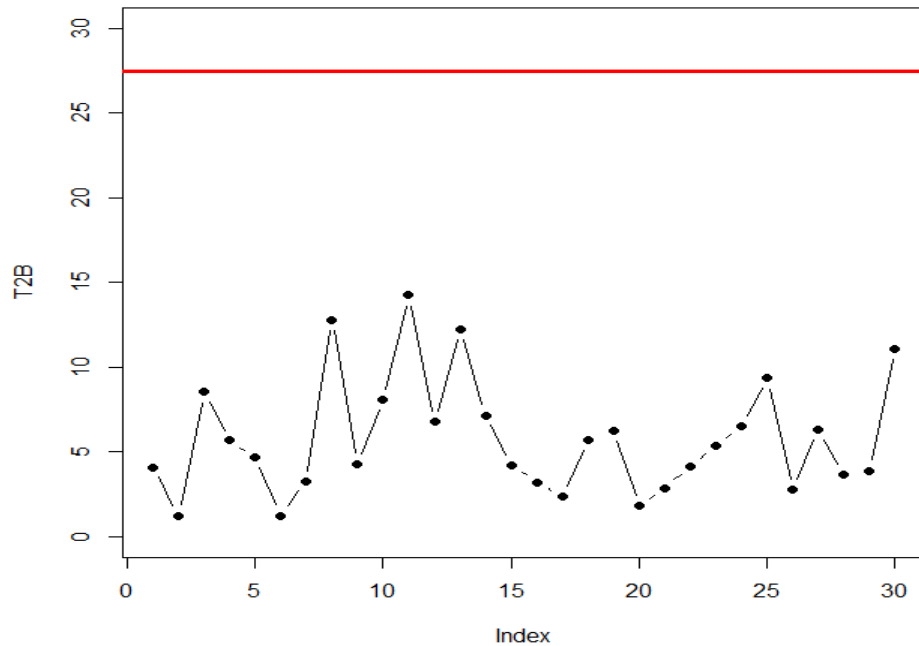


*Plot of PC1 and PC3*



We built an ulterior chart, as suggested in Montgomery and other text books, on the remaining six principal components, due to the fact that they offer an helpful representation of the residuals and they should therefore carefully be monitored. ( $p=6, n=30$ )

$$UCL = \frac{p(n-1)(n+1)}{n^2 - np} F_{p, n-p} = 27,47$$



This chart is important in the preliminary phase in order to detect any outlier that is disturbing the creation of the limits. No particular value was detected and we could proceed therefore further.

### *Subgrouped T2 Control Chart*

Alt (1985) proposed a grouping methodology which helps to catch significant changes in the process by grouping variables.

We will use this method in parallel with the previous one and will compare the solutions. This method gives more relevance to several out of control units.

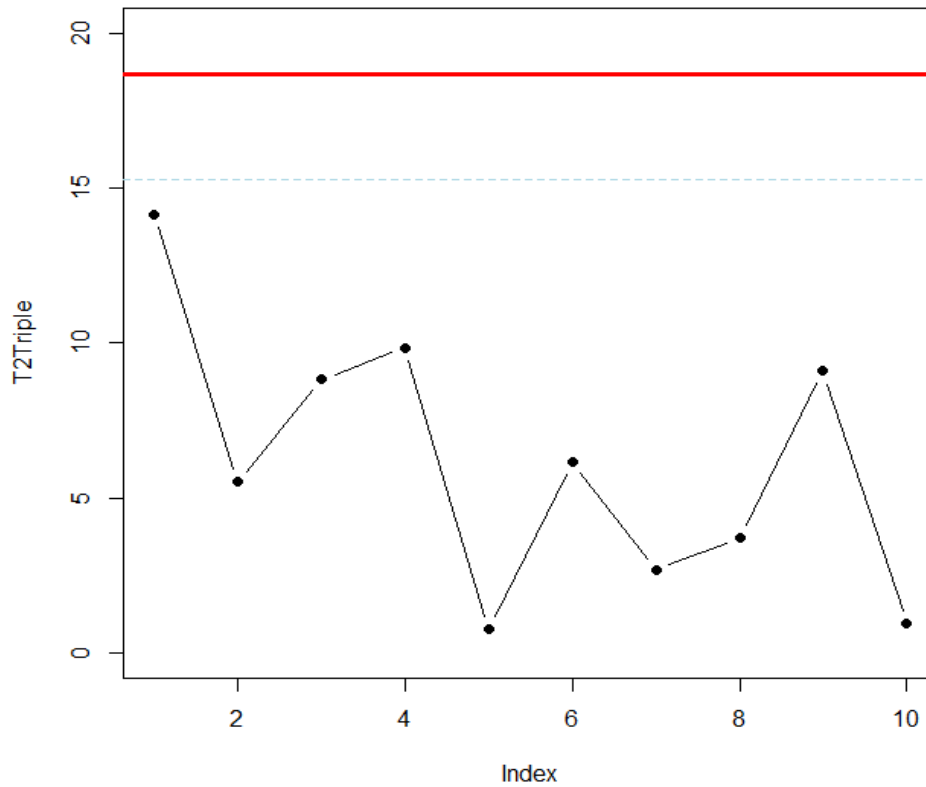
We constructed the chart firstly transforming the data: we took data in groups of three (in sequence because we didn't have any correlation and data were already randomized). We took the average of each 3 data for each of the three first principal components. Then, we built the prevision confidence intervals in a T2 control chart with the Alt formula. In this situation we had:  $p=3$ ,  $m=10$ ,  $n=3$  and  $\alpha$  is the usual 1%.

The final control limit of Phase II (used for monitoring future production) is the following:

An additional temporary limit to detect was tested (the so-called phase 1) but no out of bounds value were detected as it can be seen in the light blue line in the next chart.

$$UCL = \frac{p(m+1)(n-1)}{mn - m - p - 1} F_{p, n-p} = 18,67$$

T2 chart for the grouped means values.



This approach is taken in consideration parallel to the first one explained before (not simultaneously).

### Univariate Control Charts

In order to achieve completeness, we wanted to analyze the process from a last different point of view choosing to adopt a univariate approach on three variables which represent some main characteristics of the process. Variable selected are *Porosity*, *Avg position of pores* and *Gini's coefficient of concentration*.

This approach, despite its lack of completeness, balances with its easiness in understanding on which characteristic our data is differing compared to the training set.

We performed the analysis on MINITAB in the file Univariate\_CC.

Setting  $ARL_0=100$  we have an  $\alpha_{Overall} = 1/100$

From the Bonferroni's inequality (we are considering three features simultaneously):  $\alpha_{Overall} \leq n * \alpha$

$$\alpha = \frac{1}{300} = 0.0033$$

We used the Z approximation, due to the fact that normality is high for each variable and the numerosity of the observations is sufficient to approximate the univariate t-distribution to a normal.

$$Z_{\alpha/2} = 2.71338$$

$$\left\{ \begin{array}{l} UCL = \bar{x} + Z_{\alpha/2} \frac{\overline{MR}}{d_2} \\ CL = \mu_i \\ LCL = \bar{x} - Z_{\alpha/2} \frac{\overline{MR}}{d_2} \end{array} \right.$$

Where  $d_2$  is equal to 1.128

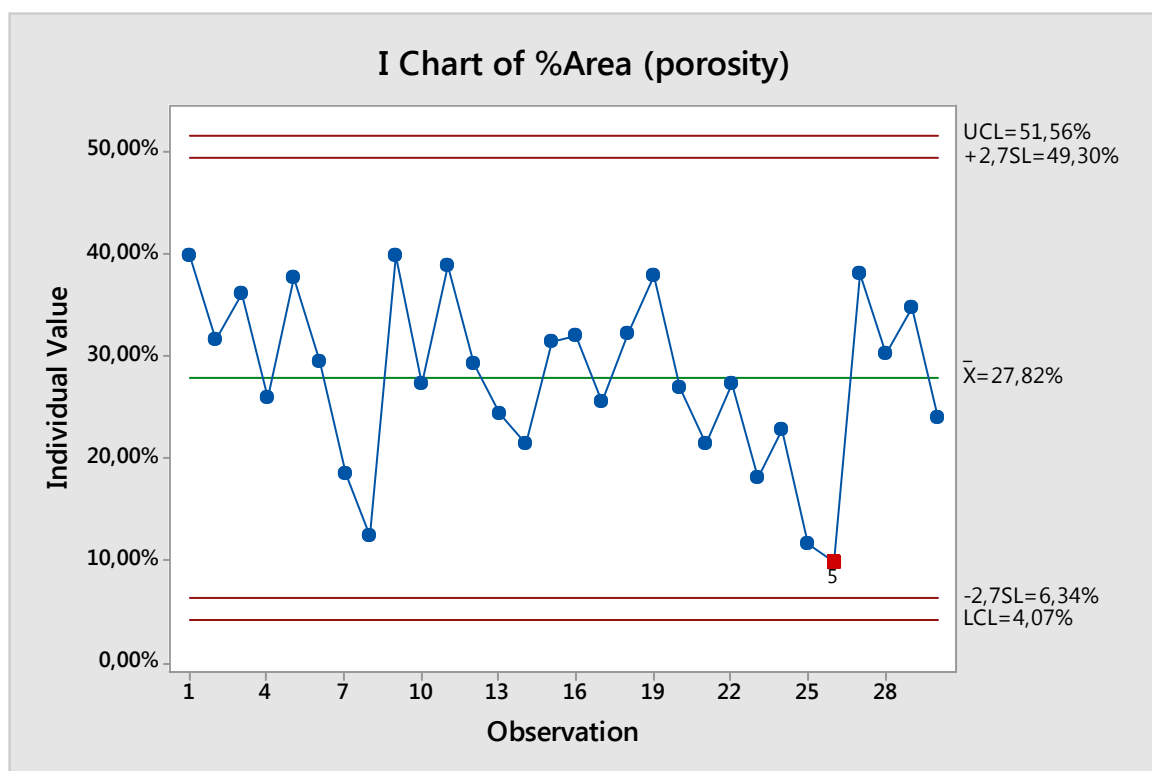
On the variable Position, we decided to use a univariate control limit, because we believe that is more consistent with the process: infact the value 0 would be indeed optimal for our data (perfectly centered and well balanced). We built a One way control chart, with only the right tail taken in consideration:

The UCL for this variable is bigger than the two sided version:  $Z_{\alpha} = 2.9352$

The upper control limit for the distance, is therefore:

$$UCL = \bar{x} + Z_{\alpha} \frac{\overline{MR}}{d_2} = 0.2181$$

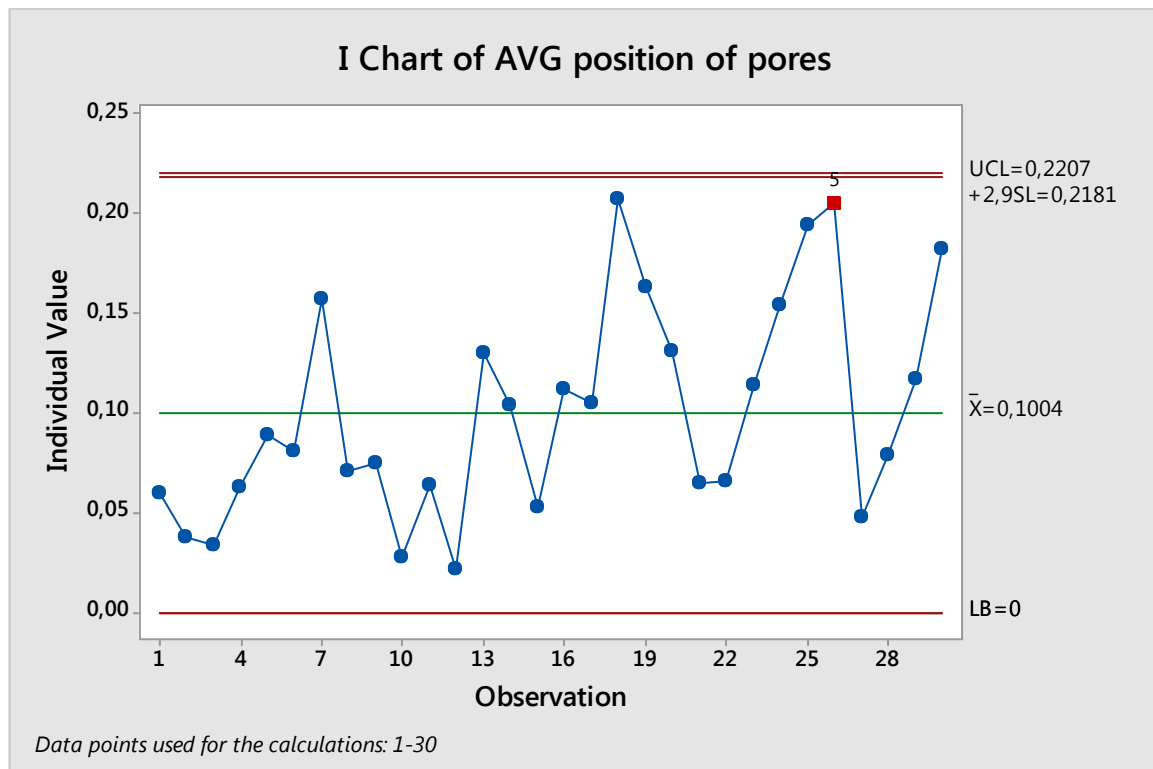
We decided to perform all the tests that MINITAB provide for this type of charts, in order to understand irregular behavior of the data, which control limits can't detect (like small shift or run rules). Minitab offers some rules on how to identify an out of specification process.



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

CL) .

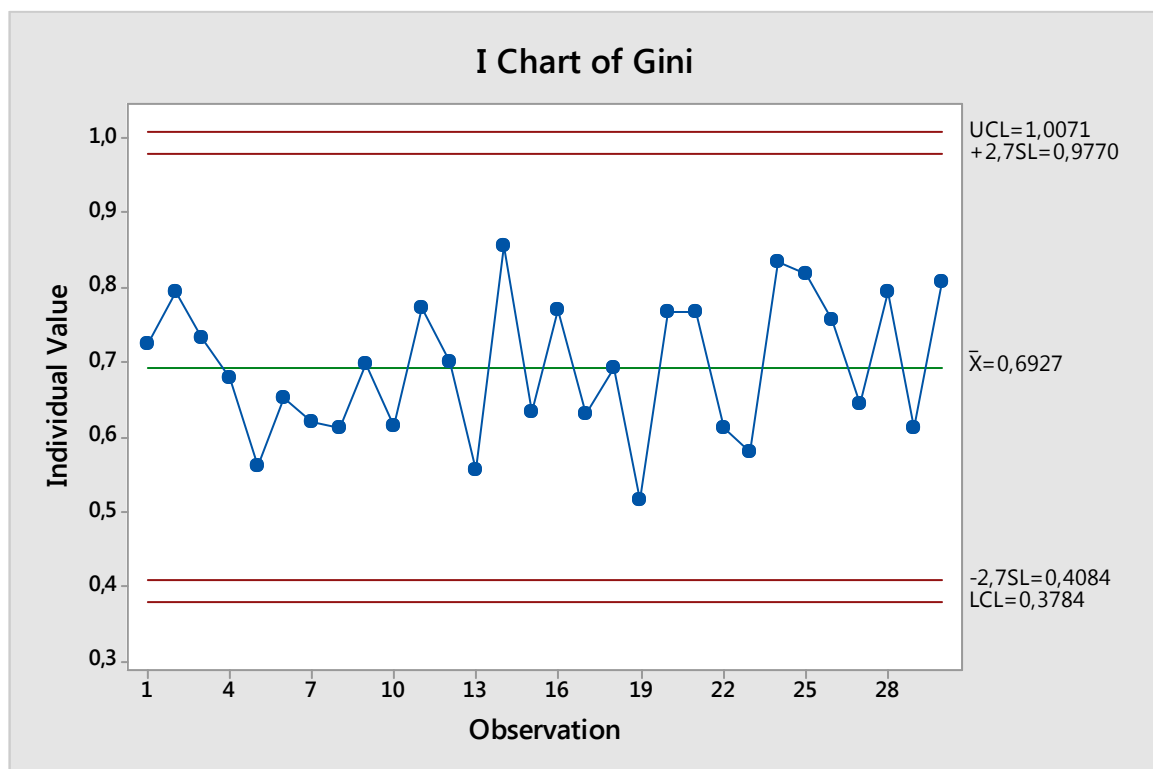
Test Failed at points: 26



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

CL) .

Test Failed at points: 26



Among the two control charts, the data 26 is the second within the dataset for which we obtained a deviation from the center line of more than two standard deviation. Likewise, since we do not have any information that would allow us to consider this observation as an out of control, eventually we have

decided not to modify the design of the charts and proceed with the phase two, where we tested the previously built charts.

## Phase 2

By the time we had obtained the new dataset and worked on the image to obtain comparable data, we firstly performed an analysis for a better understanding of what should be expected from our control charts and which process characteristic should be detected to be out of control.

We will commence the analysis of the new data through an analysis which compares Mean Vectors from two populations. Firstly we defined a mean vector for both the datasets, the training set and the new dataset. Each mean is reported with the standard deviation on that data.

Training Set			Dataset2		
	Mean	Std. Dev.		Mean	Std. Dev.
Area	0.2781	0.0841	Area	0.1883	0.0466
Position	0.1004	0.0540	Position	0.1641	0.0894
Solidity	0.8533	0.0346	Solidity	0.8774	0.0260
A_2p	0.01146	0.00458	A_2p	0.004426	0.000978
Circul	0.5798	0.0637	Circul	0.7120	0.0488
AR	1.679	0.121	AR	1.670	0.220
Entropia	2.733	0.340	Entropia	2.171	0.449
ln_RHT	2.653	0.421	ln_RHT	3.203	0.332
Gini	0.6927	0.0927	Gini	0.6972	0.0753

We obtained the variance-covariance matrix from data and we estimated the pooled covariance as:

$$S_{pooled} = \frac{n_1 - 1}{n_1 + n_2 - 2} S_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S_2$$

where  $n_1=30$   $n_2=27$  and  $S_1$  and  $S_2$  are obtained by the data of the two datasets.

In order to perform this analysis, we should have the multivariate normality: using as in precedence the mcshapiro multivariate test, we obtain a p-value of only 4% but it is still sufficiently high to proceed. Moreover the two variances matrixes occur to have omogenous structures.

We constructed Bonferroni univariate intervals for the mean difference vector:

$$M_{diff} = M_1 - M_2$$

where  $M_1$  and  $M_2$  are the mean vectors for the two groups. Then, by using t-di student intervals with Bonferroni corrections (with  $k=9$ ), we were able to construct confidence intervals for the mean differences.

$$Lim. Inf = \bar{M}_j - t_{n.cor.} \left( \frac{\alpha}{2k} \right) \frac{S_{pool,j,j}}{\sqrt{n.cor.}} ; \quad Lim. Sup = \bar{M}_j + t_{n.cor.} \left( \frac{\alpha}{2k} \right) \frac{S_{pool,j,j}}{\sqrt{n.cor.}}$$

In the next table we can see for each variable the mean of the difference, its upper and its lower bound with a 99% confidence. Whenever all three coefficients appear to have the same sign, a shift in the mean of the variable has statistically occurred: in case of no shift the value 0 should be inside the confidence interval. Area, A/2p and Entropia's means result as being statistically higher in the first dataset compared to the second. Circularity and log(R-H-T) 's means result as being statistically lower in the first dataset. Regarding the other variables nothing particular can be declared. Many variables seem therefore to have deeply changed.

	inf	mean	sup
Area	0.0133	0.0898	0.1663
Position	-0.1446	-0.0637	0.0172
Solidity	-0.0583	-0.0241	0.0101
A_2p	0.0032	0.0070	0.0108
Circul	-0.1956	-0.1321	-0.0687
AR	-0.1853	0.0092	0.2037
Entropia	0.1233	0.5621	1.0009
ln_RHT	-0.9740	-0.5506	-0.1271
Gini	-0.0988	-0.0045	0.0897

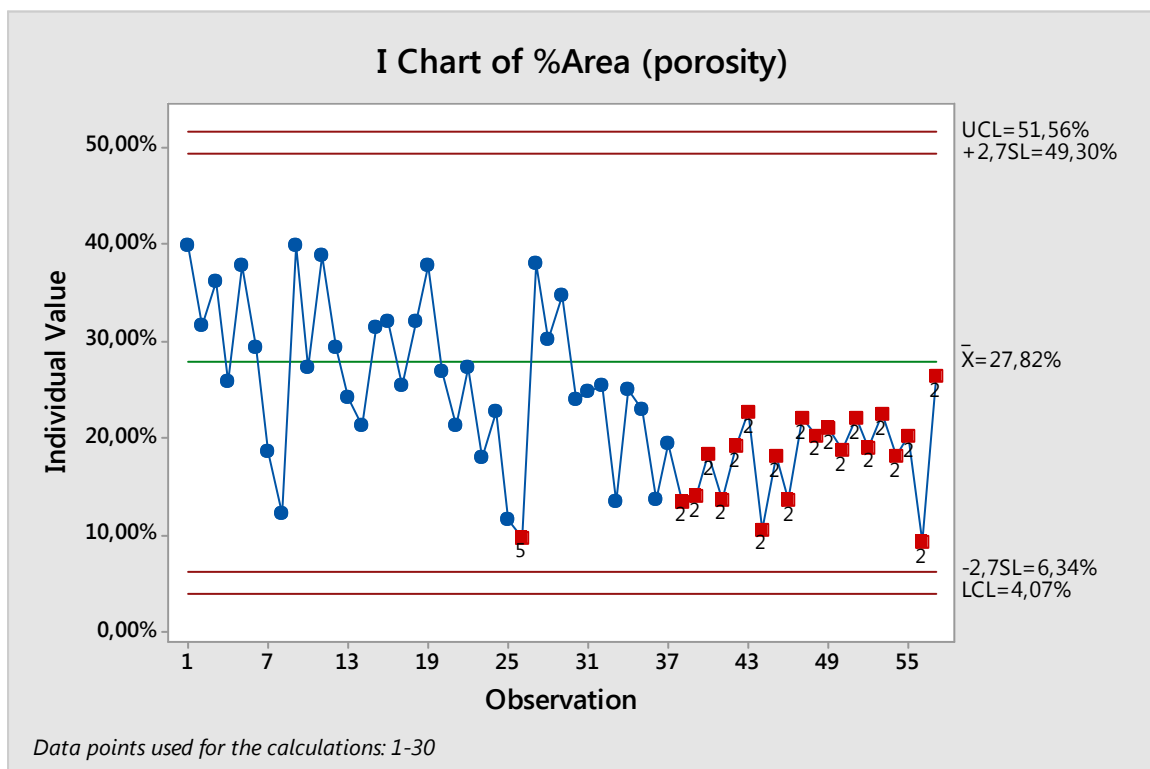
One last observations can be made regarding Circularity and Gini: those 2 variables are centered and almost no difference seems to appear, given our datasets (in fact, lower and upper limit are almost symmetric and the mean difference value is around 0). In general, the two datasets appear to differ significantly regarding many variables and features, especially given the high confidence level established.

The two dataset come from different populations: we are now going to test whether our classifiers were able to detect any structural change in the data.

As already mentioned, we have designed three univariate CC, in which we have used all the tests that MINITAB makes available in order to analyze the process: the results for the 27 data of the control phase are shown below.

### Univariate control charts

#### CONTROL CHART FOR POROSITY (%Area)



#### Test Results for I Chart of %Area (porosity)

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

Test Failed at points: 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 53; 54;

55; 56; 57

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

Test Failed at points: 39; 40; 41; 42; 44; 45; 46

These results signal a downward shift for the distribution mean.



Despite no value is below the control limits, data are often under the previous calculated average.

TEST 4. 14 points in a row alternating up and down.

Test Failed at points: 56; 57

Instead here we highlighted a possible correlation between the data which seem oscillating.

Since it is pretty clear that there was a shift in the process, and the Individual CC is not the best mean to analyze small shifts, we could move to a more appropriate control chart as offered by the EWMA Control Chart. We have chosen EWMA CC because, if well implemented, it is faster in detecting small shift reporting an out of control and it is also more robust than a CUSUM CC against departures from normality.

EWMA: POROSITY

Control charts for small shifts: the EWMA control chart  
**Exponentially Weighted Moving Average (EWMA)**

- For samples of size  $n \geq 1$ , the EWMA for process mean is defined as:

$$z_i = \lambda \bar{x}_i + (1 - \lambda)z_{i-1}$$

where  $0 < \lambda \leq 1$  is a constant and  $z_0 = \mu_0 = \text{target}$  (or  $z_0 = \hat{\mu}_0 = \bar{\bar{x}}$ )

- The control limits are computed as follows:

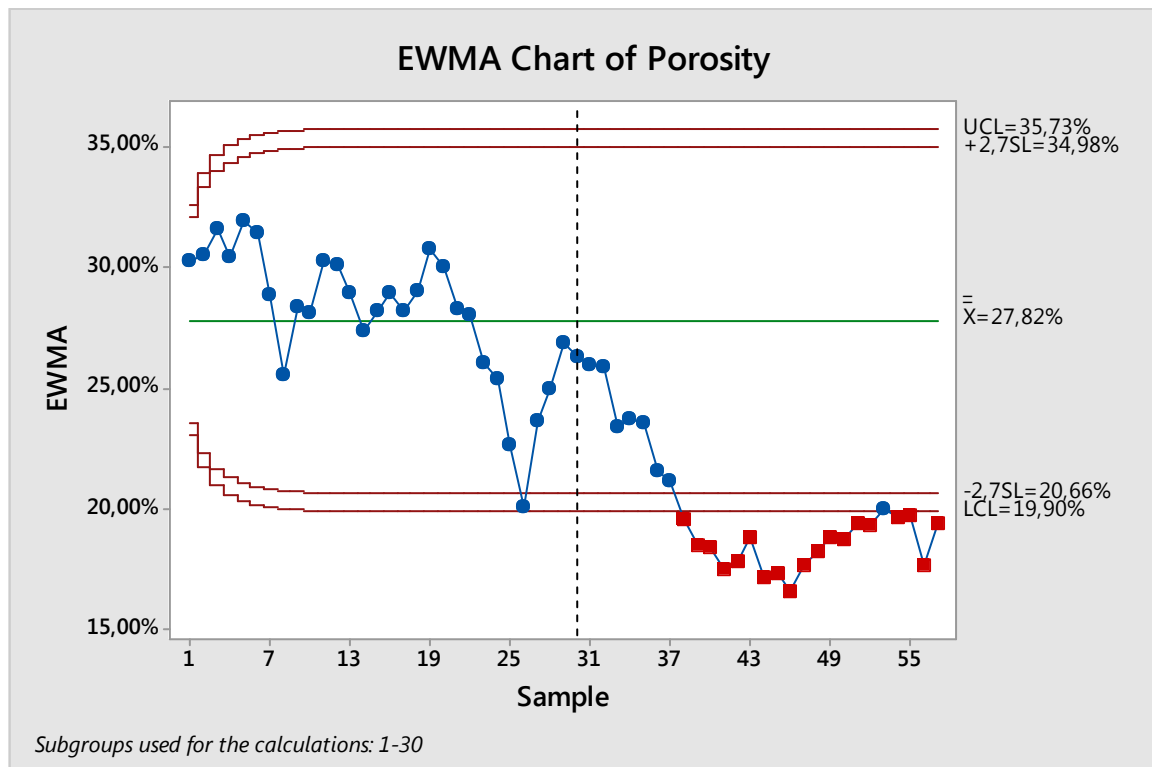
$$V_i = z_i \quad \begin{cases} LCL = \mu_{V_i} - L\sigma_{V_i} \\ CL = \mu_{V_i} \\ UCL = \mu_{V_i} + L\sigma_{V_i} \end{cases}$$

Where  $L$  is a constant (typically  $L=3$ )

For single measurements ( $n=1$ ):  $z_i = \lambda x_i + (1 - \lambda)z_{i-1}$

$$\begin{cases} LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{2-\lambda}} \\ CL = \mu_0 \\ UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{2-\lambda}} \end{cases}$$

We used  $\lambda=0,2$  and  $L=3$ .



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

Test Failed at points: 38; 39; 40; 41; 42; 43; 44; 45; 46; 47; 48; 49; 50; 51; 52; 54; 55;

56; 57

As it can be seen in the 38 observation, the control chart immediately signals an Out-Of-Control, thus it gives an alarm. The process has gone below its level and the pores occupy less area then before.

In order to verify if the process became autocorrelated, we have performed this analysis which includes the Runs Test and the ACF. Of course this analysis concerns only the second sample of data.

## Runs Test: Porosity\_2

Runs test for Porosity\_2

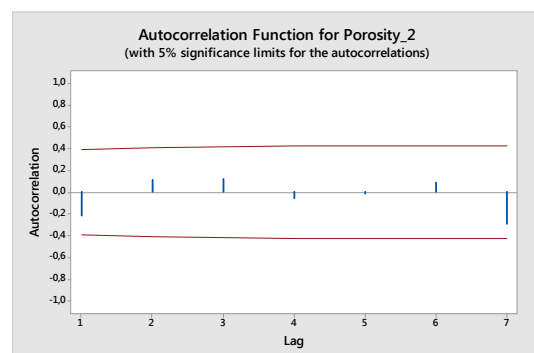
Runs above and below  $K = 0,188324$

The observed number of runs = 15

The expected number of runs = 14,3333

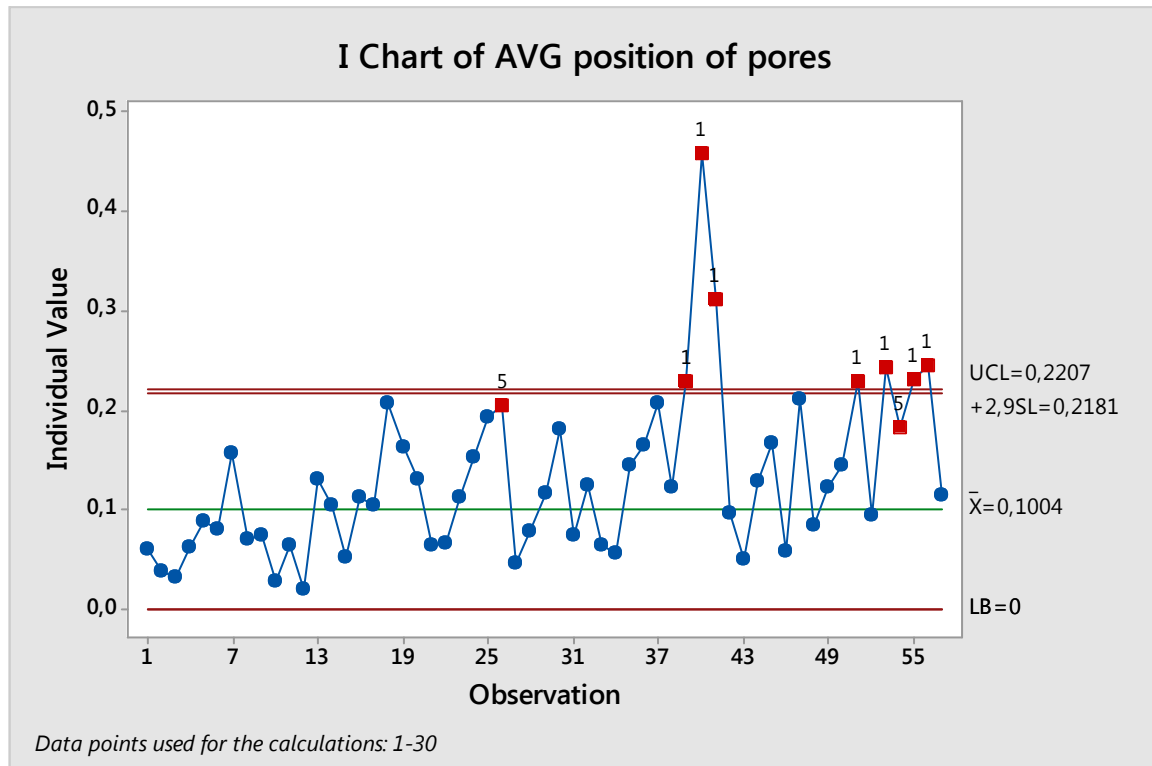
15 observations above  $K$ ; 12 below

P-value = 0,791



Both tests don't give statistical evidence to assert that the process is autocorrelated.

## CONTROL CHART ON POSITION OF PORES



Once again we were used all the additional tests provided by MINITAB, with the following results:

### Test Results for I Chart of AVG position of pores

TEST 1. One point more than 3,00 standard deviations from center line.  
Test Failed at points: 39; 40; 41; 51; 53; 55; 56

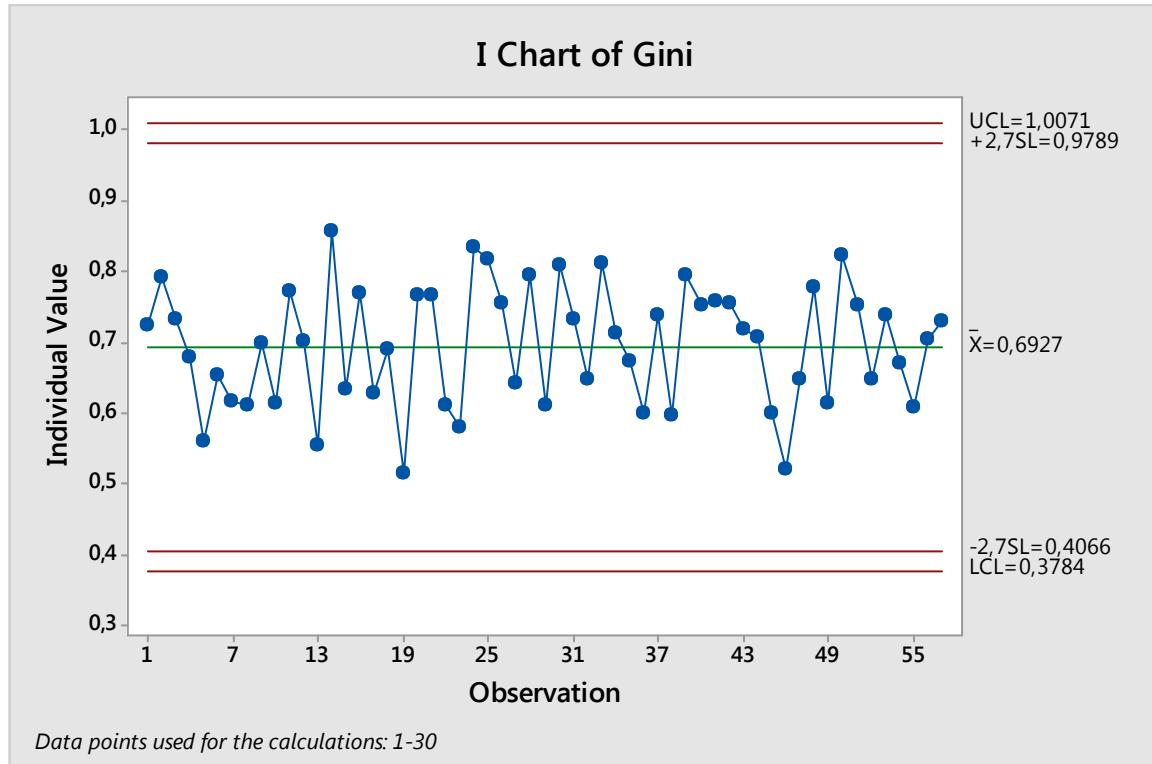
### Several points exceed the upper control limit

TEST 5. 2 out of 3 points more than 2 standard deviations from center line (on one side of CL).  
Test Failed at points: 26; 39; 40; 41; 53; 54; 55; 56

TEST 6. 4 out of 5 points more than 1 standard deviation from center line (on one side of CL).  
Test Failed at points: 39; 40; 41; 54; 55; 56

The process is evidently out of control, at least for the 40th data. Some images have now pores which are not anymore as centered and balanced as before.

## CONTROL CHART ON GINI'S INDEX OF CONCENTRATION



In this case the chart does not show any change in the process: it doesn't highlight out of control. This feature could have indeed been predicted by the analysis we made at the beginning regarding the group mean analysis. If we had decided to pick another variable which could represent the concentration of the bubble (for example  $\ln(R-H-T)$ ), a different output would have appeared and a shift would have been detected. However, since we decided to use this variable in the first assignment, we chose to stick to it and we analyze and now we are going to present and compare the other methods we have used.

### Principal Components Analysis

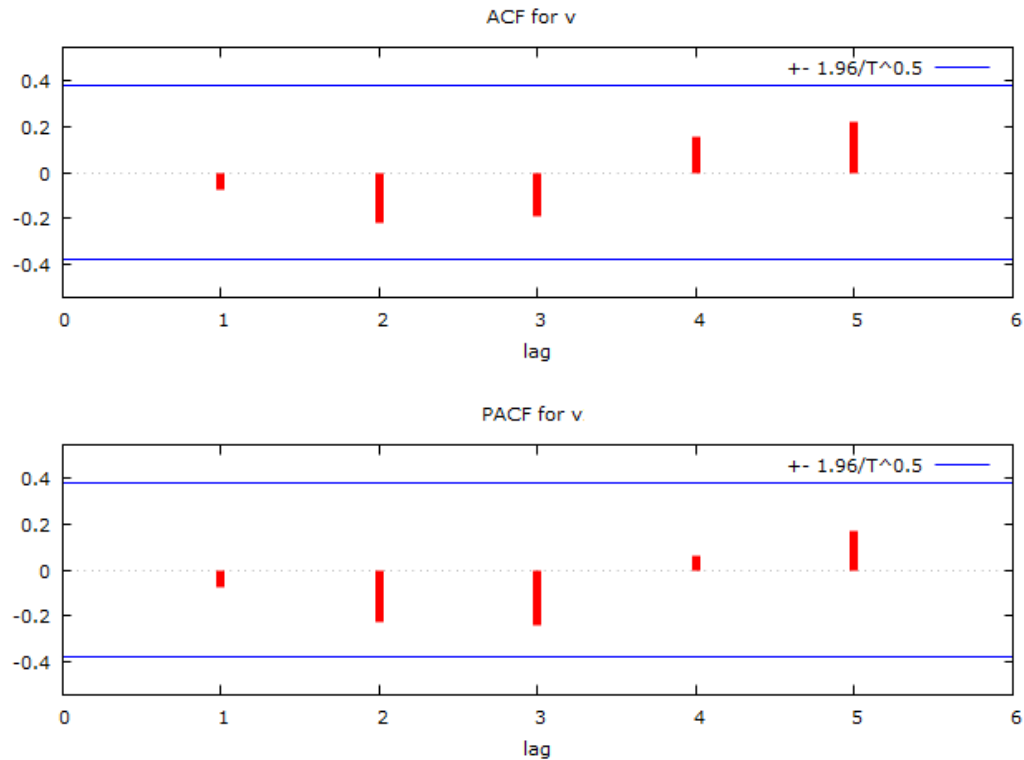
For the PCA analysis we proceeded in this way: we firstly standardized the data with the coefficients calculated before.

$$z_{ij} = \frac{(x_{ij} - M_j)}{sd_j}$$

Where  $M_j$  and  $sd_j$  are, respectively, the mean and the standard deviation for each variable calculated through the training set during phase 1.

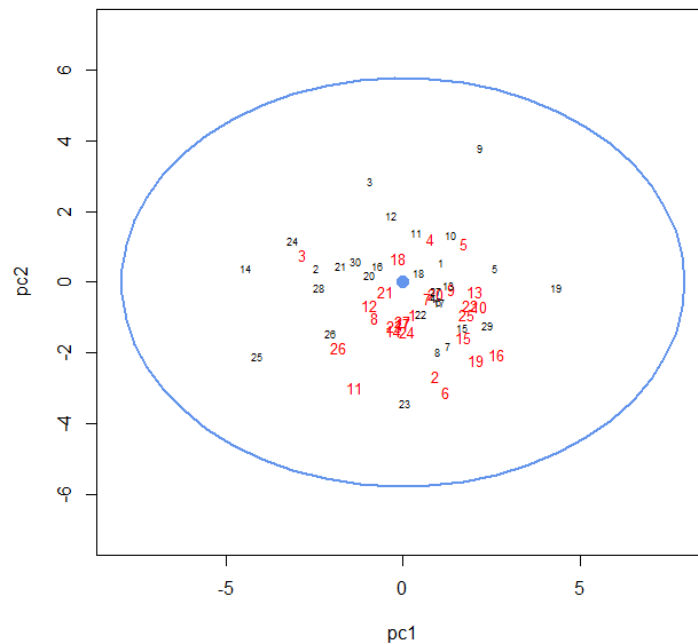
At this point, we used the Loadings Vectors of the three principal components ( $L1, L2, L3$ ), in order to transform the new standardized data to be comparable to the principal components calculated previously in the training set.

The obtained matrix, composed by three columns and 27 observations, was tested in the autocorrelation of residuals for each variables for different lags: we wanted to know if the process was still changing or whether it was stable. The multivariate normality test didn't show any proofs against normality (p-value 20%). To test the eteroschedasticity we performed ACF and PACF for each variable. Results were positive, and the p-value (calculated until the fifth lag) never goes under 40% (Ljung-Box Q Test).

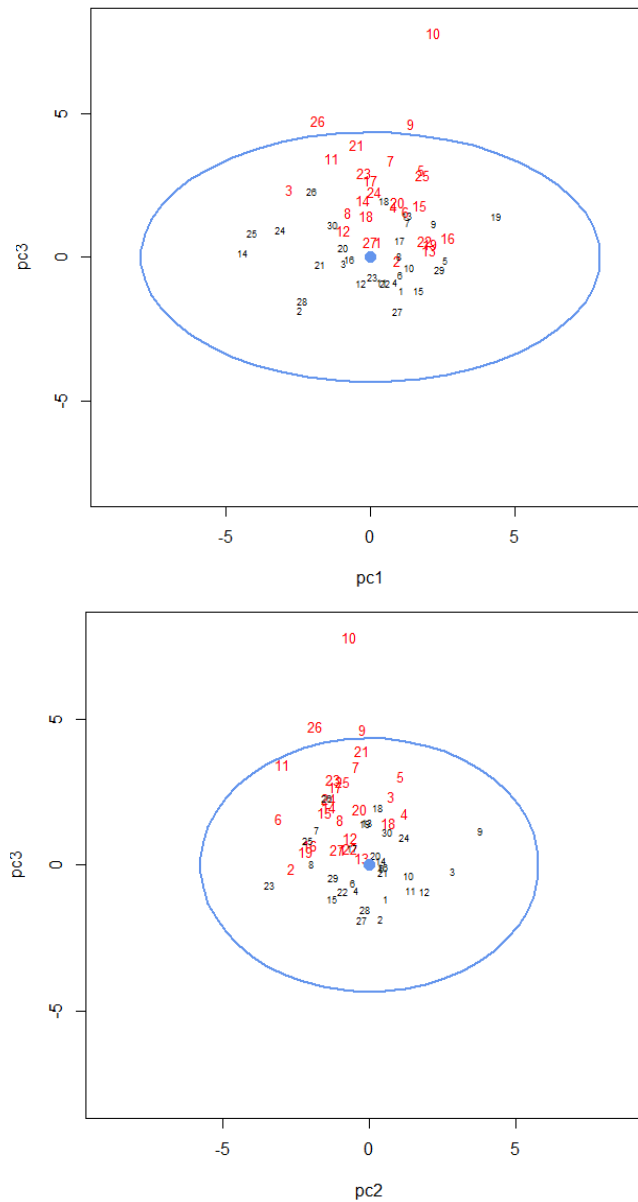


Those results made us positively continue our analysis and we wanted to test our data in the control charts previously constructed.

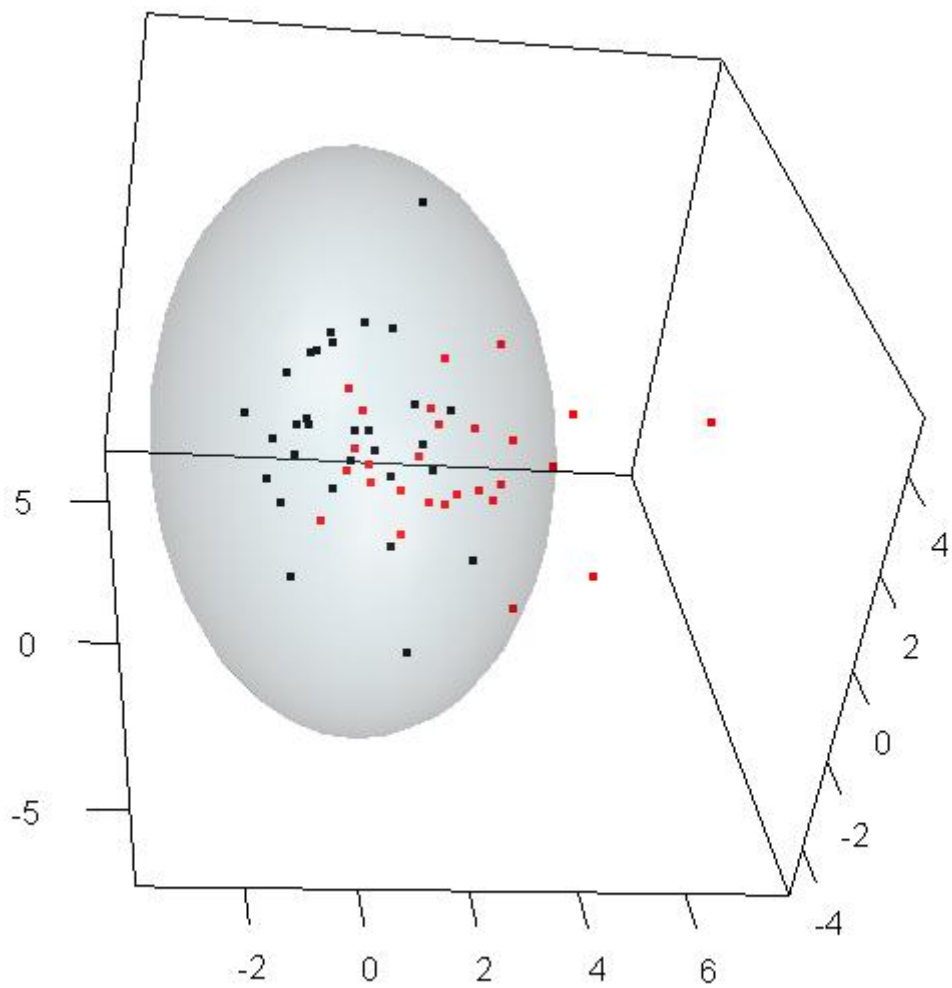
In the ellipsoid representation charts we can see that for the two first principal components data were inside the prevision ellipse: moreover it seems that the variability was indeed decreased across the process. In red one can see the new data and in black we can still see (in smaller characters) the training set numbered images we used to build the control chart.



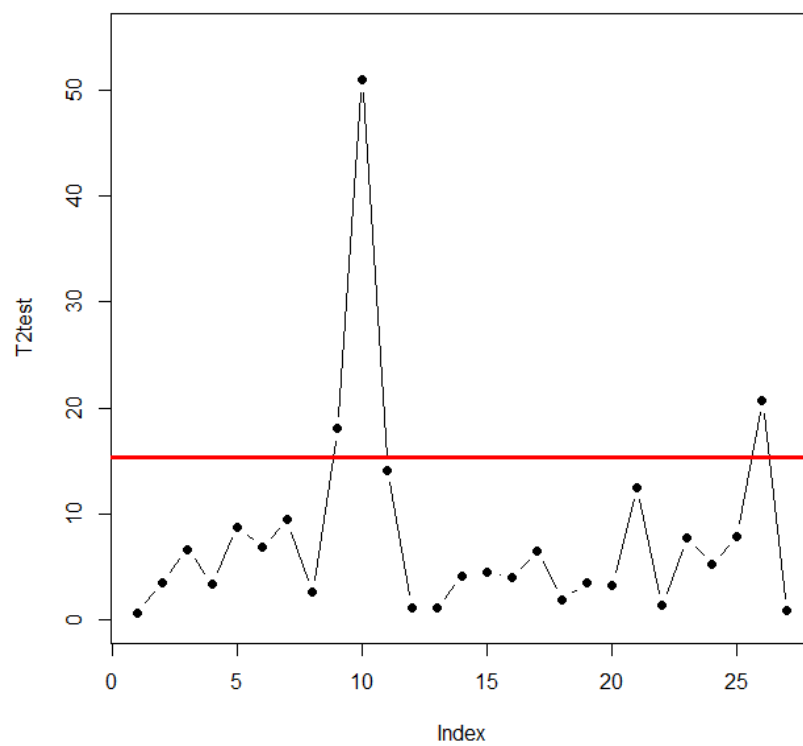
The data in the second principal component seem to be consistently below the average (meaning, for example, that the area is on average smaller in the second dataset); however the control chart constructed in this way doesn't detect structural changes. On the direction of the third principal component (related to the distance from the center), the control chart seems to detect some outliers (in particular the image number 9, 10 and 26).



Data can also be represented on a three dimensional graph, through a package built in the software R. It is helpful in order to detect the values for each observations on the three principal components. This 3D representation of the ellipsoid containing the dataset of the three principal components shows a significant shifts of the data compared to the previous situation (data in red are the new observations).



We finally plot the T2 chart for the new observations with the limits previously calculated.



This T2 chart shows that the observations 9,10 and 26 went out of the bounds, showing a drift in the process. However many information is lost because there is no memory of the previous value: the reasoning as regards the values of the second and third principal components systematically over or under the average is lost in the chart. In order to detect such shift (small but constant) an EMWA chart or a CUSUM could have had detected more powerfully the shift in means.

At this moment, despite we didn't take into account when we had to design the control charts in the first phase, we wanted to monitor the variability of the process: as noted before it seems that in the first principal component there is no shift in the average but a lower variability seems to have occurred. Our object was to construct an EWMS (Exponentially Weighted Mean Square Error) control chart that could help us detect this trend.

We used the approach proposed by MacGregor and Harris (1993), with the statistics  $S_i^2$ :

$$S_i^2 = \lambda * (x_i - \mu)^2 + (1 - \lambda)S_{i-1}^2$$

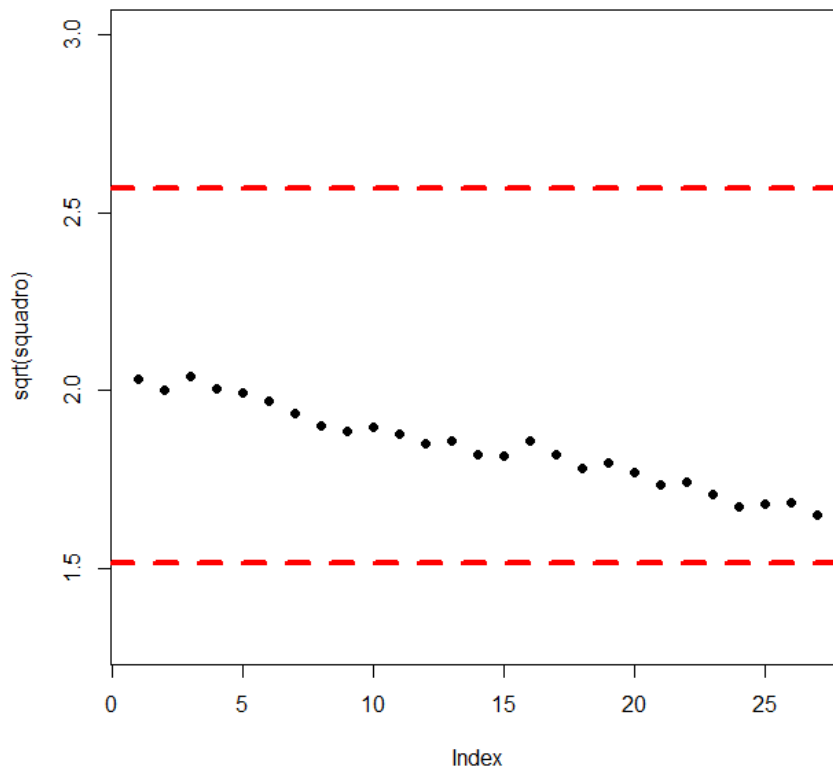
$\sigma_0$  is the standard deviation obtained from the training set on the first principal component.

With a selected  $\lambda$  of 0.05 (and usual alpha), the control limits for large numbers of observations are:

$$UCL = \sigma_0 * \sqrt{\frac{\chi_{v,\alpha/2}^2}{v}} = 2.567$$

$$LCL = \sigma_0 * \sqrt{\frac{\chi_{v,1-\alpha/2}^2}{v}} = 1.515$$

With  $v=(2-\lambda)/\lambda$ . The limits obtained are for the  $\sqrt{S_i^2}$ , which are represented in the next graph.



The process for the EMWS seems to be still in control from the chart, despite it can be seen the decreasing trend of the data. With a bigger testing sample, it would probably have been detected a decrease in the standard deviation of the process in the first principal component (even if there were no big changes in the mean in this variable). This aspect could be actually positive for the process of the new dataset because it means that the change in the process was able to reduce the overall variation.



## Subgrouping Method

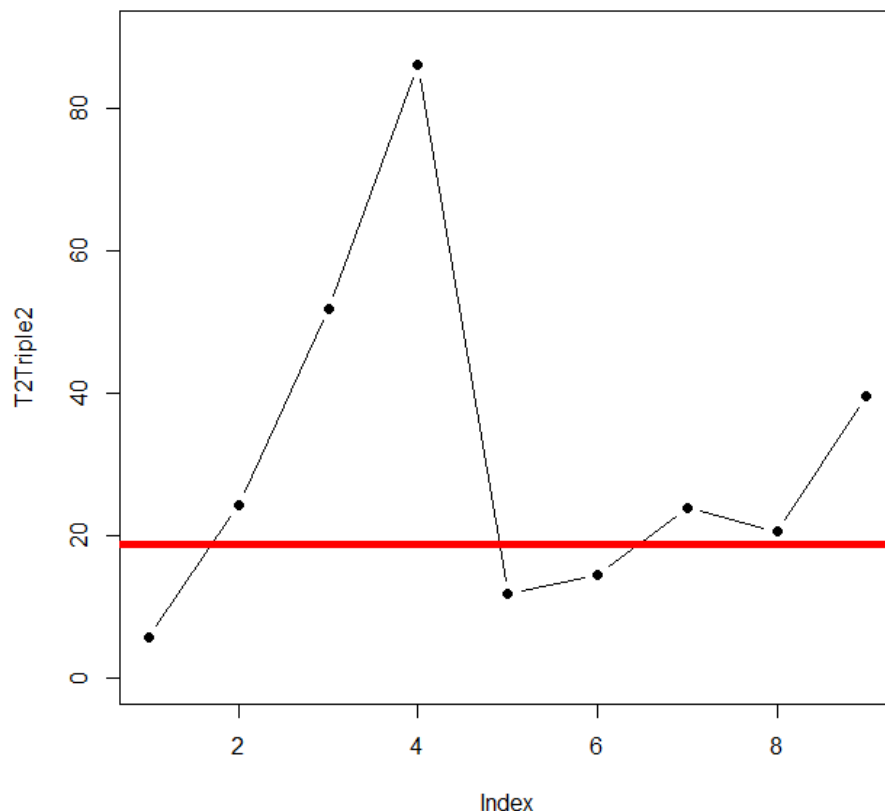
We finally grouped data, by using the method proposed by Alt (1985): as in the training set, we grouped each 3 sequential observations and we calculated for each subgroup the mean vector. Data obtained were:

	Comp. 1	Comp. 2	Comp. 3
[1,]	-0.524628011	-0.9353913	0.9250481
[2,]	1.267597786	-0.2650972	2.1300074
[3,]	0.440496384	-0.5557948	3.1833930
[4,]	-0.038499219	-1.4474531	4.0565755
[5,]	1.158013778	-1.0604656	1.3541231
[6,]	0.841291801	-0.8520748	1.6072412
[7,]	0.845708153	-0.9242453	2.0948368
[8,]	0.597206668	-1.0949983	1.9262638
[9,]	-0.006845233	-1.2955280	2.7205737

Here it is clear how a shift in the process occurred: this data should have been originated from a standardized averaged process, therefore data should be centered in zero with a certain degree of variability. However, data in Comp.2 are all under the zero level and data in Comp.3 are all beyond zero showing a particular behavior and cumulated trend. This behavior shows a clear shift in comparison to the previous mean that however (due to the low difference in respect of the previous standard deviation and the decreased variability) weren't much detected in the first multivariate control chart we have built.

An EMWA or a CUSUM in the second and third principal component would have surely detect this drift.

However also the approach chosen, inspired by Alt (1985) strongly detect the variations in the process: 6 values (out of 9) are indeed out of control and the process shift can be easily perceived. In this graph, we see the T2 values with the limits we previously calculated during the first phase.



## Conclusion

Overall, different methods have been valued and compared (all with a 99% confidence interval). Each of them detect different features and have pros and cons. Small shifts like in our sample are always easily detected by EMWA or CUSUM charts: therefore using just a simple normal approach could have not been helpful in detecting new changes.

Overall we have found that in the first principal component not significant changes were detected, and actually it seems that the variability was reduced, which is, in general, positive for any industrial process.

The process seems to have changed on several features: the area of the pores occupy now a smaller portion of the overall figure: moreover the concentration indexes show that now pores's concentration is higher and few big pores occur over many smaller pores. The average distance of the pores seems to be slightly more distant, on average, from the center. Finally new pores have a more round shape and the circularity of the pores seem to have increased.

Overall we don't know whether the change in the process was positive or negative for the production: our classifiers detected a shift on the overall process, but it is always up to the statistical engineer to understand and explain what are the effects and the implications of these changes.

For example, since the variability has decreased in this second sample, it is possible that a manufacturer could prefer the parts to be produced with the new method (in a six-sigma optic of reducing variation). In general from this project it was clear how different strategies should be applied in order to detect different features of the process. For example, by comparing the PCA with the Univariate Control charts you have a trade off among easy to use and immediate explanation with loss of variation and information. Likewise, we always suggest monitoring simple variables and pre-analyzing data in order to achieve a better understanding of data and not fall in a black-box system that you cannot master completely.

## Bibliography

Statistical Quality Control, Montgomery

Applied multivariate statistical analysis, Jonhson and Wickern

Multivariate quality control, Alt Frank and Kamlesh Jain ,

Encyclopedia of Operations Research and Management Science, Springer US 2001 544-550

Slides Teaching class, Quality engineering 2015