Motto: *Is it normal to have normal data ?*

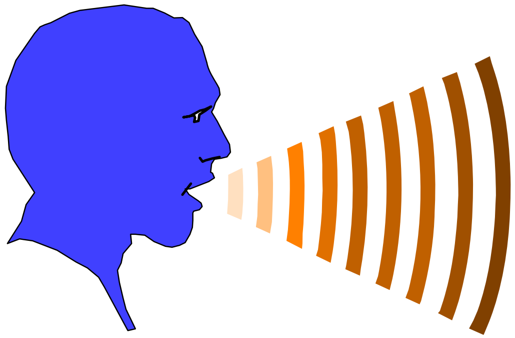
Data transformation in multivariate quality control

George E. P. Box

J. Militký and M. Meloun

The Technical University of Liberec
Liberec, Czech Republic

University of Pardubice
Czech Republic



Quality & Productivity Research Conference
IBM T. J. Watson Research Center, Yorktown Heights, NY June 3-5, 2009

OUTLINE

- Multivariate quality control
- Box Cox transform
- Cotton fibers basic properties
- Simple measures of cotton quality
- Application of utility value concept for creation of complex quality criterion („quality index“).
- Dependencies between cotton fiber properties and strength of yarns.

QC and data

The products quality is characterized by several properties expressing the ability of a product to fulfill functions it was designed for.

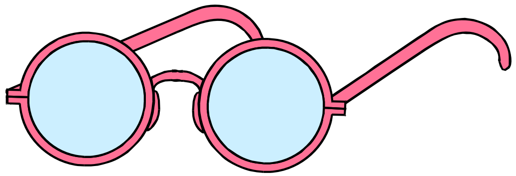
The main problem with utilization of these properties for quality characterization is multivariate character of information, various units and lack of proper aggregation.

It is well known that a majority of methods used in quality control are based on the normality assumption.

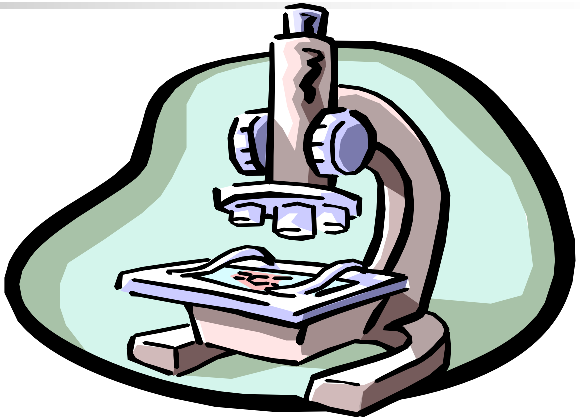
Main procedures that are dependent on this assumption include:

- The computation of quality control charts
- Determination of the capability of processes
- Prediction of product performance

Real Data



(No -) normal distribution,
(No -) homogeneous –
outliers,
(No -) constant variance,
(No -) additive effects,
(No -) autocorrelation,
(No -) sufficient sample size
(No -) certainty of models,
(No -) artificial parameters
Physical sense

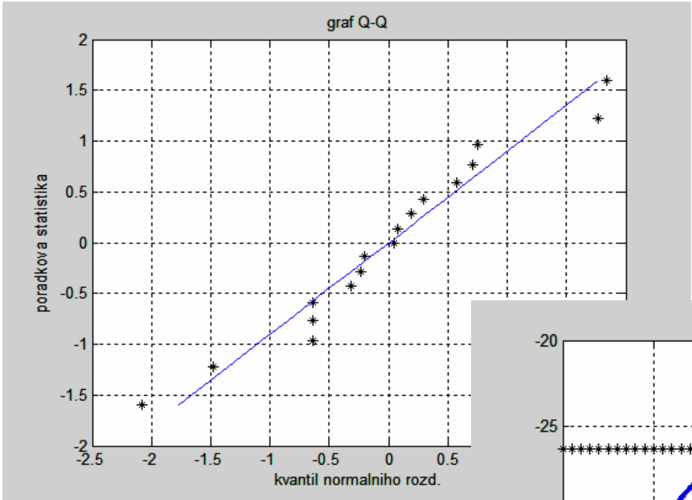


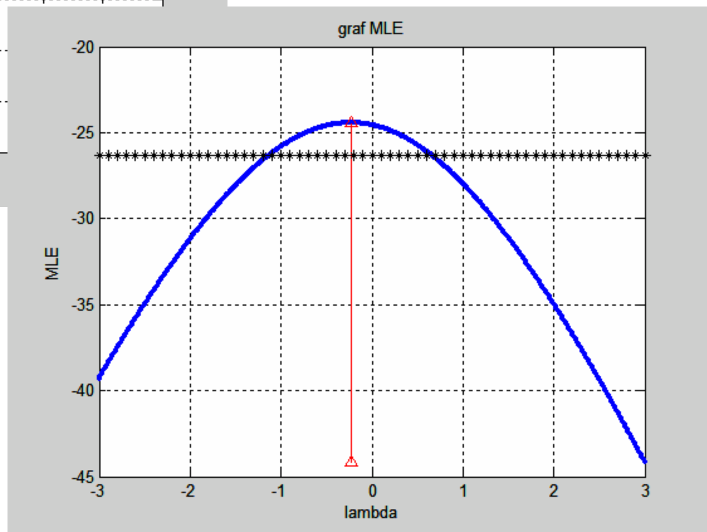
Classical statistical methods

Based on the idealized assumptions about data distribution and their statistical behavior.

Basic assumptions:

- independence,
- normality,
- constant variance,
- no outliers,
- errors additivity,
- sufficient sample size





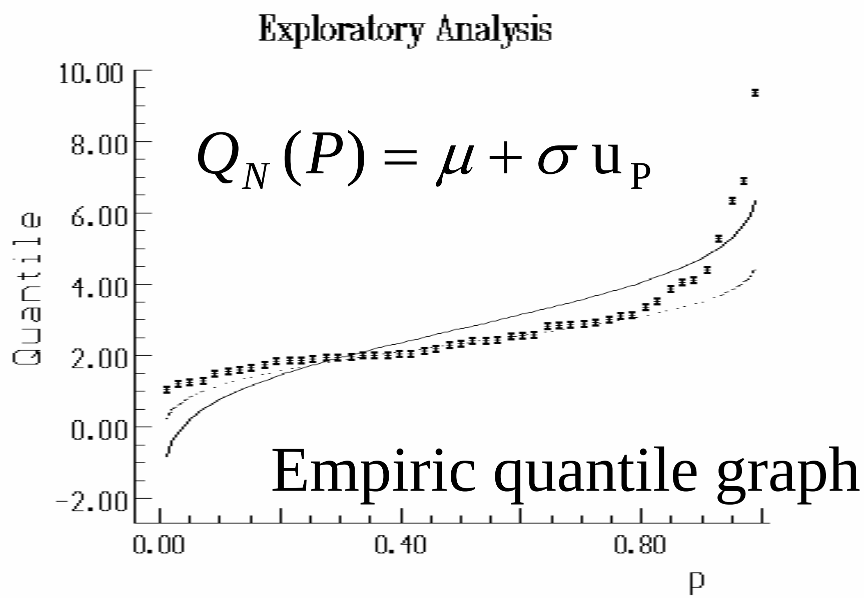
Computer assisted analysis

- Exploration of data peculiarities
- Analysis of assumptions about triplet: data, model, method
- robust methods
- adaptive methods
- **data transformation**
- Simulation based techniques

"data - model - statistical method"

Order statistics $x_{(i)}$ vs. P_i

$$P_i = \frac{i}{N+1}$$



Empiric quantile graph

Problems

$$D(h(x)) = \left[\frac{dh(x)}{dx}\right]^2 \sigma^2$$

- Measurements on the various materials (variation due material is much higher in comparison with measurement precision.
- Measurement under conditions of constant relative error (CV). Variance is increasing with increase of mean value.
- Data are skew to the right (materials with low quality parameters are removed during processing).

For removal of data asymmetry the proper transformation *h(x)* is often selected. Constant CV case: ***h(x)=ln(x)***

→ $$D(h(x)) = \left(\frac{\sigma}{x}\right)^2 = \delta^2$$

For additive model of measurement model with constant variance this transform leads to the non constant variance

Possible solutions

- Selection of suitable data distribution and use general methods (as maximum likelihood) for parameter estimation and testing.
- Improving of data distribution and using standard methods valid for normal data.
- Using of computer assisted (distribution free) methods as Bootstrap (in fact we have some assumption about sample distribution – multinomiality)

Modification of data distribution via proper transform is useful for situation when data are combined into complex criterion of quality.

Box Cox transform

$$\sigma_c^2 = \frac{1}{N}\sum\left[h(x_i) - h(\mu)\right]^2$$

Box Cox power transform (positive data)

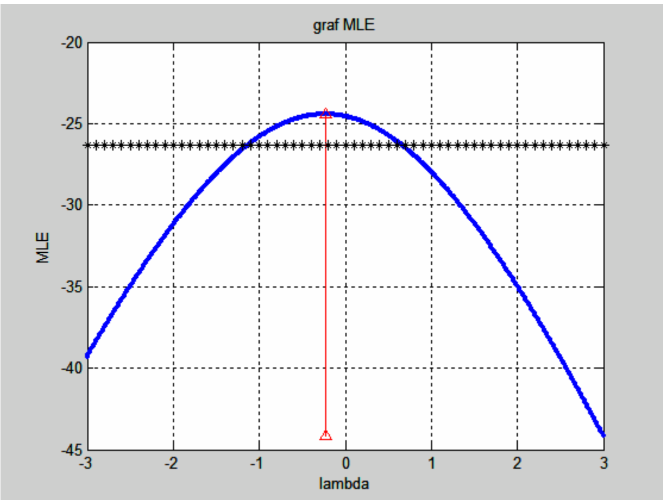
$$h(y) = \frac{y^{\lambda} - 1}{\lambda} \qquad \frac{d\,h(y)}{d\,y} = y^{\lambda - 1}$$

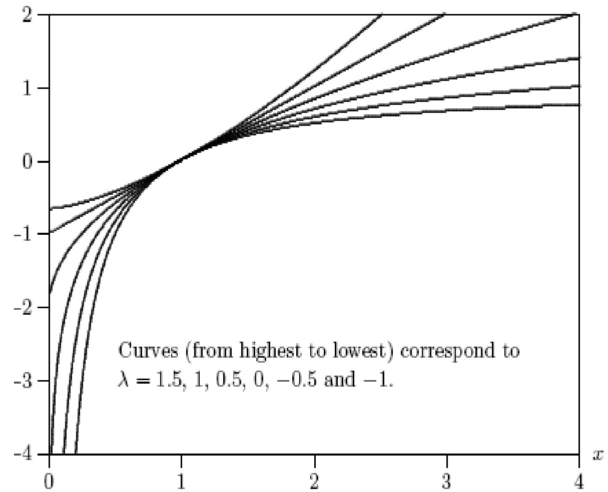
Continuous in power λ

Sharing point (0,1)

$$\lim_{\lambda \to 0} h(y) = \frac{d(y^{\lambda} - 1)/d\lambda}{d(\lambda)/d\lambda} = \frac{y^{\lambda}\ln y}{1} = \ln y$$

Influential points reduction

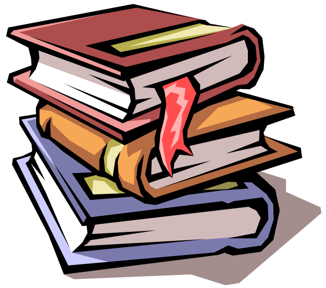




Estimation of λ:

Variance of transformed data minimization.

Problems

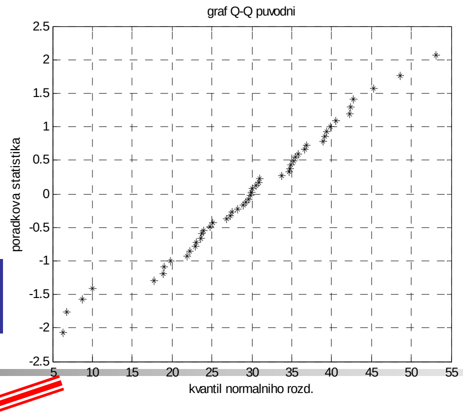


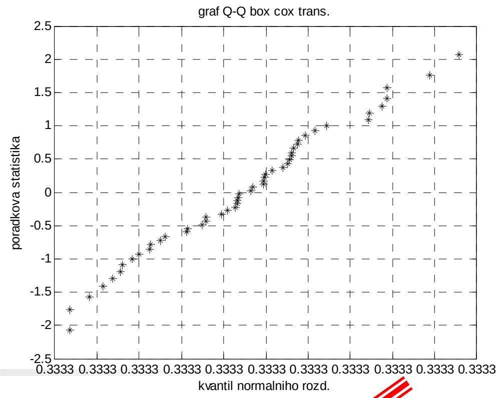
- From Taylor expansion of *h(x)* for fixed λ results

$$D\left(\frac{x^{\lambda}-1}{\lambda}\right)=\frac{1}{\lambda^{2}}D(x^{\lambda})\approx E(x)^{2\lambda-2}D(x)=E(x)^{2\lambda}\delta^{2}$$

where δ is coefficient of variation. For fixed λ will be variance of transformed data higher for higher original data variability. For data with low spread will be therefore variance in transformation low and identification of maximum will be imprecise (high confidence interval covering 1 and 0).
For D(x) $\rightarrow$ 0 is variance D(λ)$\rightarrow \infty$ and therefore the identification of proper transform is difficult.

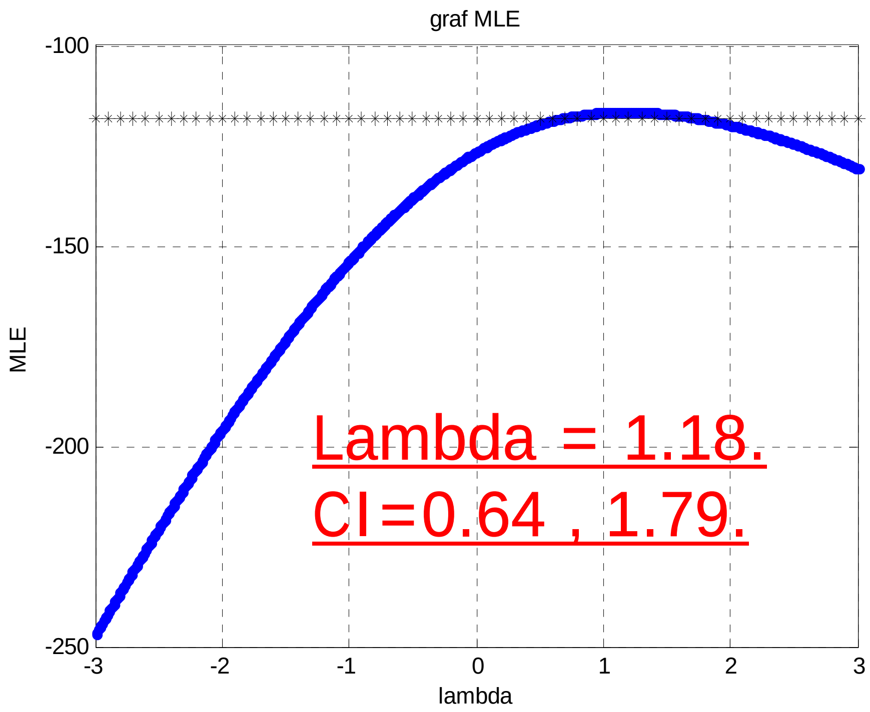
Simulation I

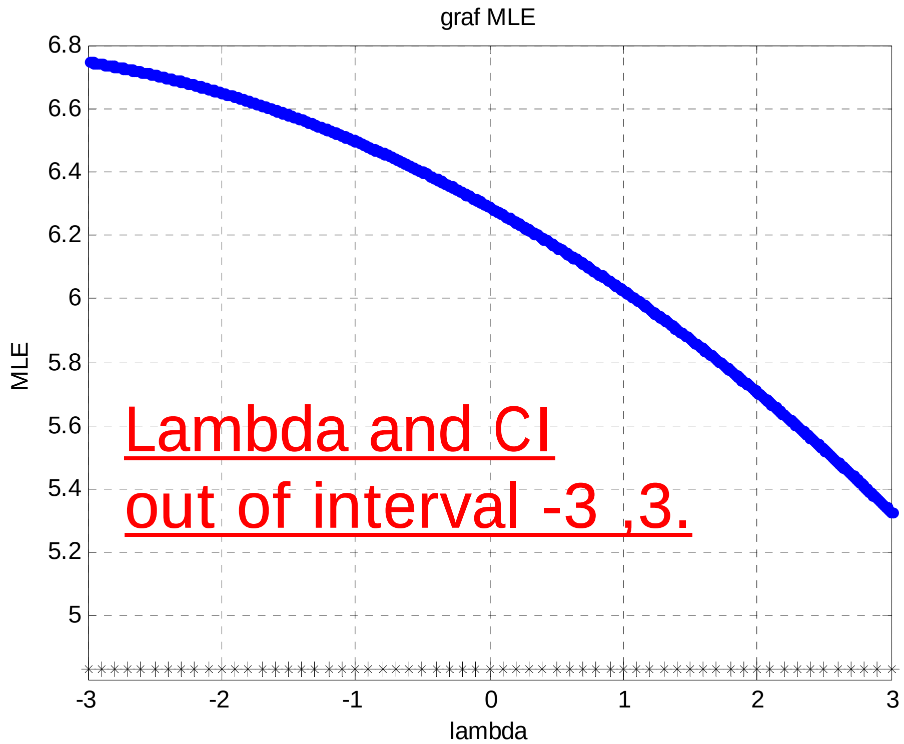




Normal distribution N(30,10)

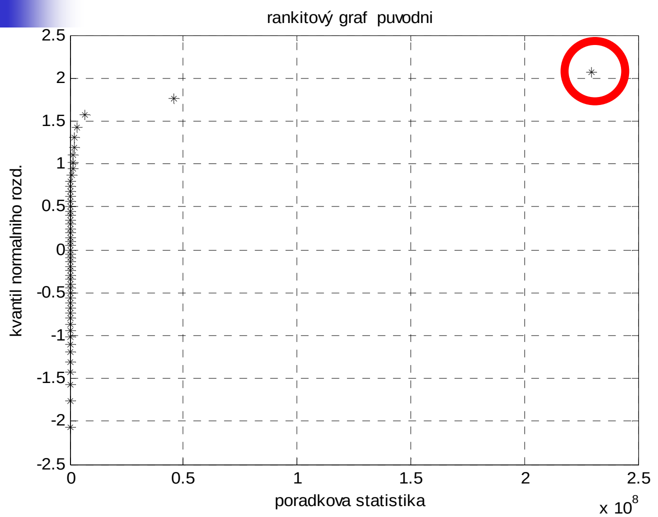
Normal distribution N(30,1)

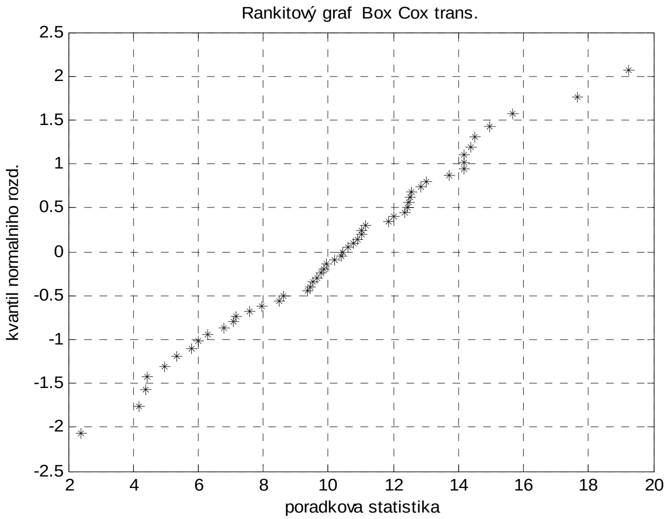


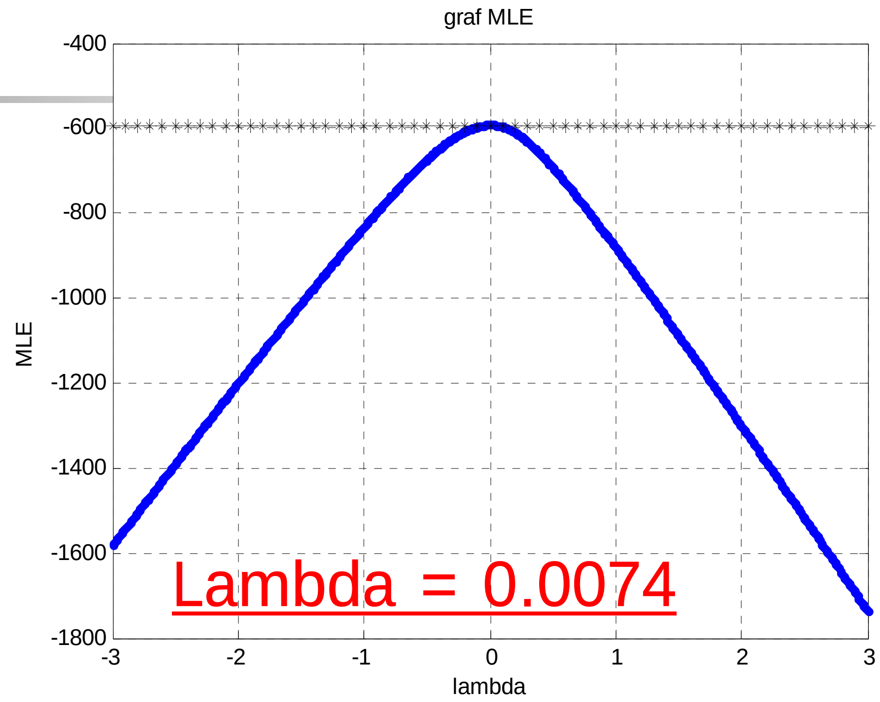


- Exponential transform of N(10, 3.5) + „outlier“ 4x higher

Simulation II

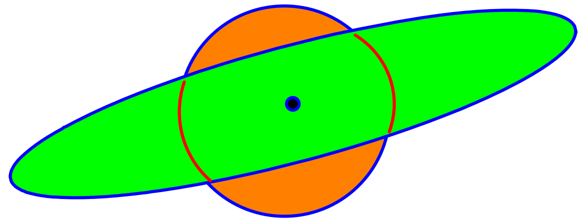






1. Optimal is logarithmic transform ln(x)
2. Extreme outlier influence is supressed

Extension to the multivariate data



- Marginal Box Cox transform (MBC), when individual variables ($\mathbf{x}_1$, … $\mathbf{x}_m$) are transformed independently. Vector $\mathbf{x}_i$ has *n* components.
- Joint Box Cox transform (JBC), when the vector of power transform $\boldsymbol{\lambda} = (\lambda_1 \ldots . \lambda_m)$ is estimated as maximum of condensed likelihood function PL

$$PL = -\frac{\ln(\det(\mathbf{S}(\boldsymbol{\lambda})))}{n} + \sum_{j=1}^{m}\left((\lambda_j - 1)\sum_{i=1}^{n}\ln(x_{ij})\right)$$

Covariance matrix $\mathbf{S}(\lambda)$ for selected λ is computed by the standard manner.

The numerical values of λ computed by MBC and JBC are nearly the same.

Use of Box Cox for cotton fiber quality index

Cotton fiber properties:

- Fiber strength
- Fiber fineness
- Staple diagram parameters (UHM)
- Fiber maturity
- Impurities
- Color

Optimal cotton properties for maximum yarn strength

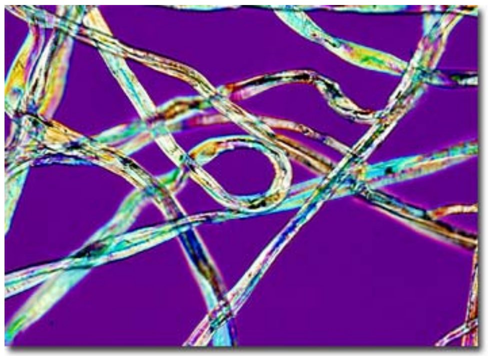
Technology:

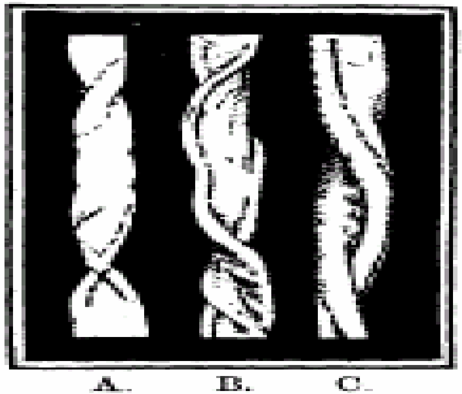
Ring spinning

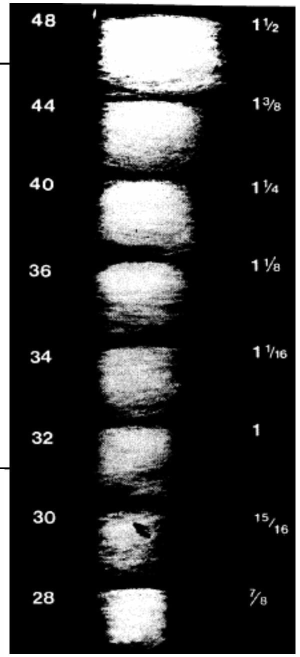
(length > strength > fineness)

Rotor spinning

(strength > fineness > length)







HVI results

Strengh **STR [cN/tex]**

Elongation ant break **EL [%]**

Upper half mean **UHM [mm]**

Uniformity index **UI [%]**

Short fiber content **SF [%]**

Thrash content **TR [%]**

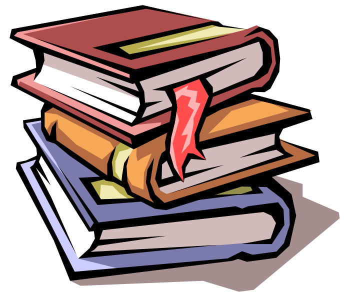
Micronare **MIC [-]**

Fineness **FI [tex]** Maturity **MAT [-]**

Uniformity ratios and descriptive designation	
Below 77	Very low
77-79	Low
80-82	Average
83-85	High
Above 85	Very high

Degree of Strength	HVI Strength (grams per tex)
Very Strong	31 & above
Strong	29 -3 0
Average	26 - 28
Intermediate	24 - 25
Weak	23 & below

Some quality indices



<u>Fiber quality index</u> (FQI)

$$FQI = \frac{UHM * UI * STR}{MIC}$$

<u>Spinning consistency index</u> (SCI)

$$SCI = -414.67 + 2.9 * STR + 49.1 * UHM + 4.74 * UI - 9.32 * MIC + .95 * Rd + 0.36 * b$$

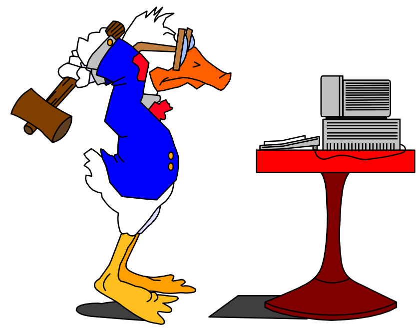
<u>Premium discount index</u> (PDI)

$$PDI = 22.15 * STR^{*} - 4.75 * EL^{*} - 4.37 UHM^{*} + 11.9 UI^{*} - 20.78 * SFC^{*} - 7.8 * MIC^{*}$$

<u>Multiplicative analytic hierarchy process criterion</u> (MIA)

$$MIA = \frac{STR^{0.27} * EL^{0.039} * UHM^{0.291} * UI^{0.145}}{MIC^{0.11} * SFC^{0.145}}$$

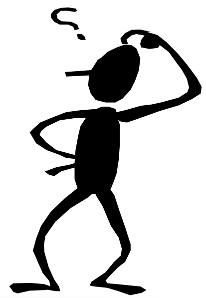
Disadvantages of indices



The main problem with all above mentioned characteristics of cotton fiber quality are:

- strong dependence on the units for individual cotton properties and methods for their evaluation,
- utilization of dimensional parameters based on the limited amount of experimental data (from the past crops),
- no inclusion of individual fiber properties importance for individual spinning technologies.
- no possibility to change parameters for new crops without tedious experimentation
- no defined ranges (limits) for quality indices.
- no possibility to include the direction of some properties influence to quality indices dependent on their real values (case of micronaire).

Utility Value Concept I



Let we have K utility properties $R_1, \ldots, R_K$ (cotton fiber properties measured by HVI). Based on the direct or indirect measurements it is possible to obtain some **quality characteristics** $x_1, \ldots, x_K$ (mean value, variance, quantiles etc.). These characteristics represent utility properties. Functional transformation of quality characteristics (based often on the psycho physical laws) lead to partial utility functions

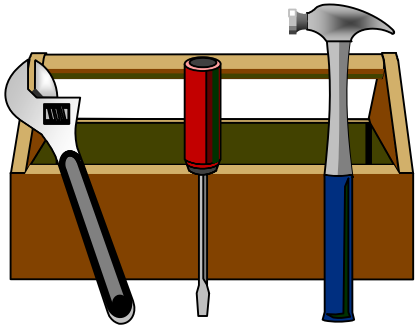
$$u_i = f(x_i, L, H)$$

L is value of characteristic for just non acceptable cotton ($u_i = 0.1$) and *H* is value of characteristic for just fully acceptable product ($u_i = 1$)

Utility value *U* (quality index) is weighted average of u_i with weights β_i

$$U = ave(u_i, \beta_i)$$

Utility Value concept II



Weights β_i correspond to the importance of given utility property and are closely connected with **area of cotton application**.

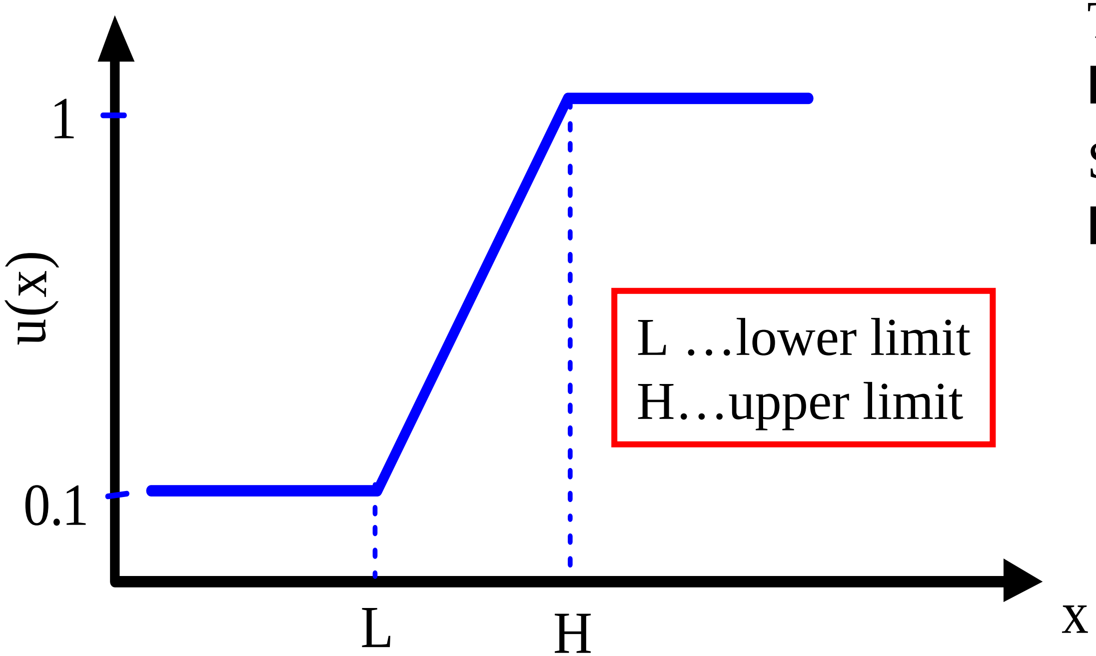
The weighted **geometric mean** used as average has following advantages:

- For zero value of u_i is also U = 0. This means that non acceptable utility property cannot be replaced by combinations of other utility properties.
- Geometric mean is for not constant u_i always lower that arithmetic mean. This reflects evaluation based on the concept that the values of utility properties close to unsatisfactory cottons are more important for expressing the quality than those close to optimum cotton.

$$U = exp\left(\sum_{j=1}^{m} \beta_j \ln(u_j)\right)$$

One side bounded properties (cotton quality index)

- Degree of quality is monotone increasing or decreasing function of quality characteristic x.



$$u(x) = \frac{0.9}{H-L}(x-H)+1$$

LB (lower is better) properties

Thrash content **TR [%]**

L=6 H=2

Short fiber content **SF [%]**

L=18 H=6

UB (upper is better) properties

Strength HVI **STR** [g/tex]

L=23 H=31

Length **UHM** [mm]

L=25 H=32

Uniformity index **UI [%]**

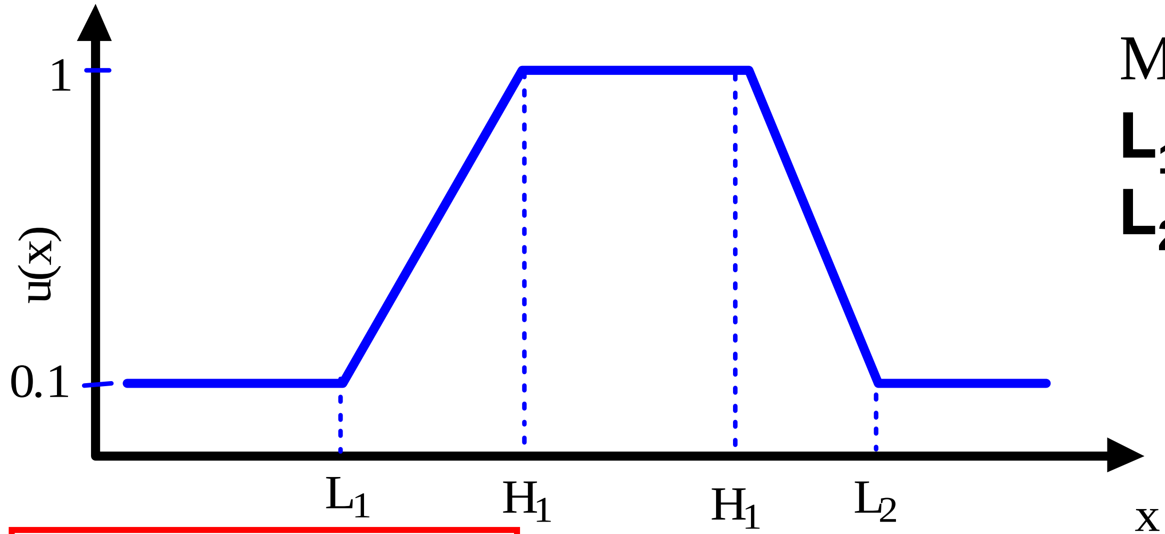
L=77 H=85

Elongation **EL [%]**

L=5 H=7.7

Two side bounded properties (cotton quality index)

- Degree of quality is monotone decreasing function of property value x on both sides from optimal (constant) region.



Micronaire **MIC** [-]

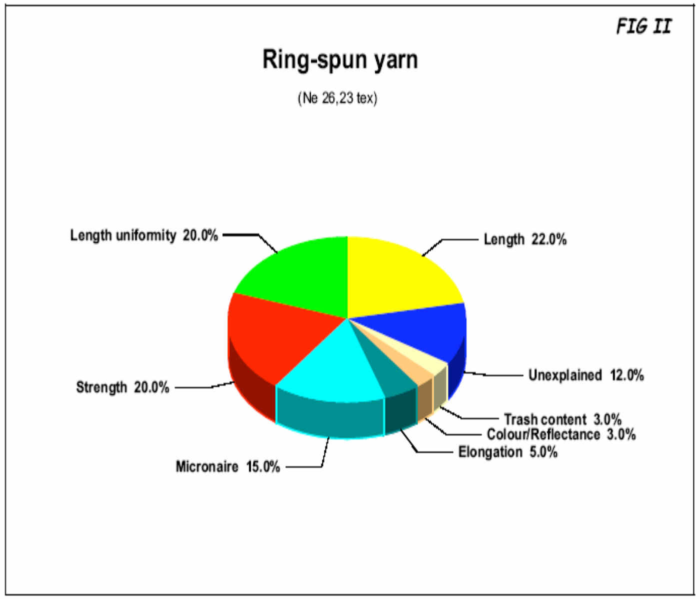
L_1=3.4 H_1=3.7
L_2=5 H_2=4.2

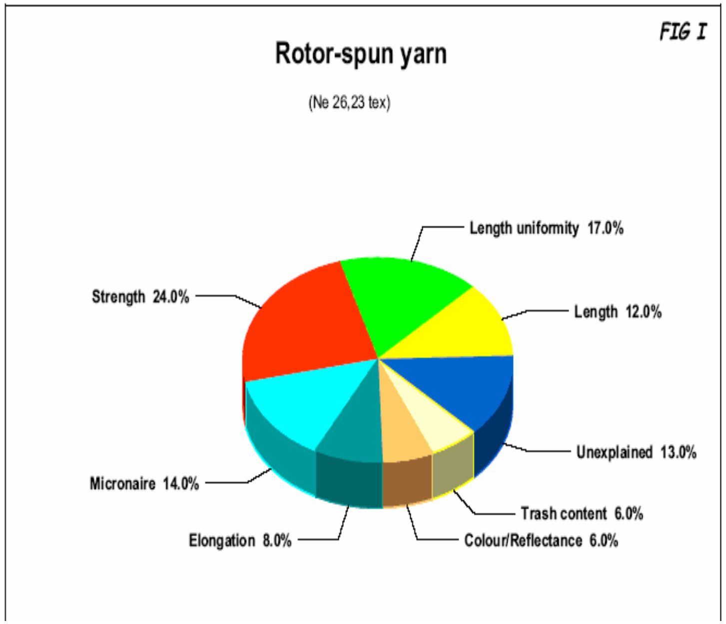
L_1, L_2 …lower limits
H_1, H_2 …upper limits

Weights for cotton quality index

- Contribution of HVI characteristics to yarn strength
- Weight b of characteristics is percentages divided by 100 and then standardized (sum of weights should be one).

Property weight	Rotor	Ring
UI [%]	0.20	0.22
MIC [-]	0.16	0.17
UHM [mm	0.14	0.24
STR [g/tex]	0.28	0.22
EL [%]	0.09	0.06
SF [%]	0.06	0.06
TR [%]	0.07	0.03





Utility function computation

In program COMPLEX written in MATLAB the simulation technique is applied. It is based on the assumption that for each utility property R_j the mean value x_j and variance s^2_j are determined from the measured data.

Utility value *U* determination

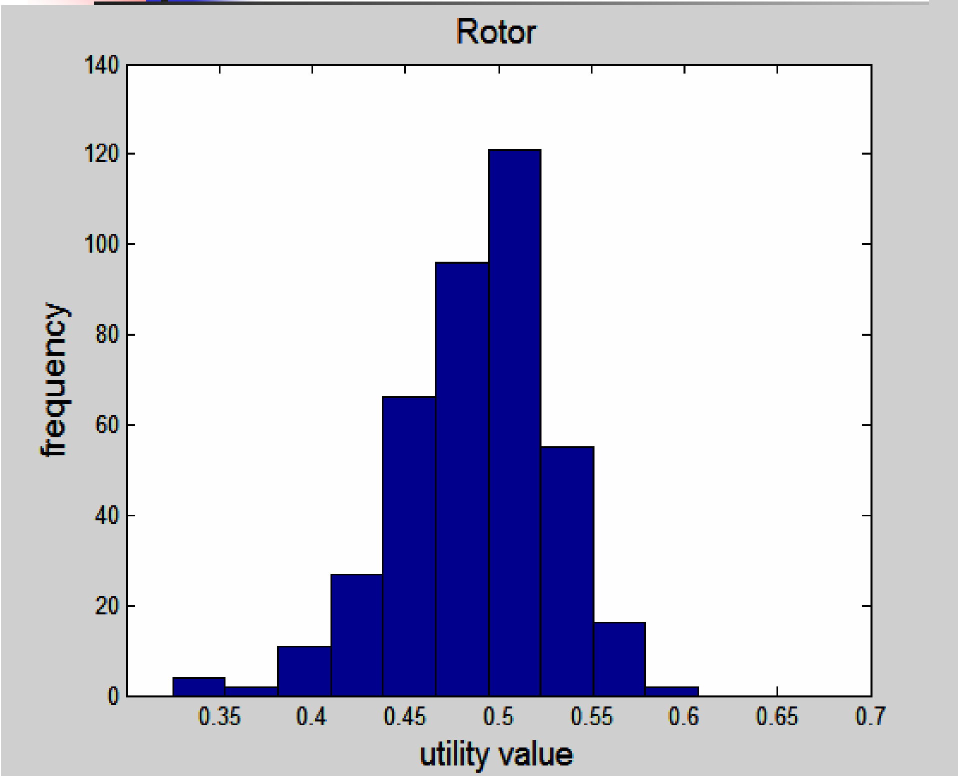
I. Generation of $x(k)_j$ (j=1,.....m) values having normal distribution with mean values x_j and variances s^2_j. The pseudorandom number generator built in MATLAB is used.

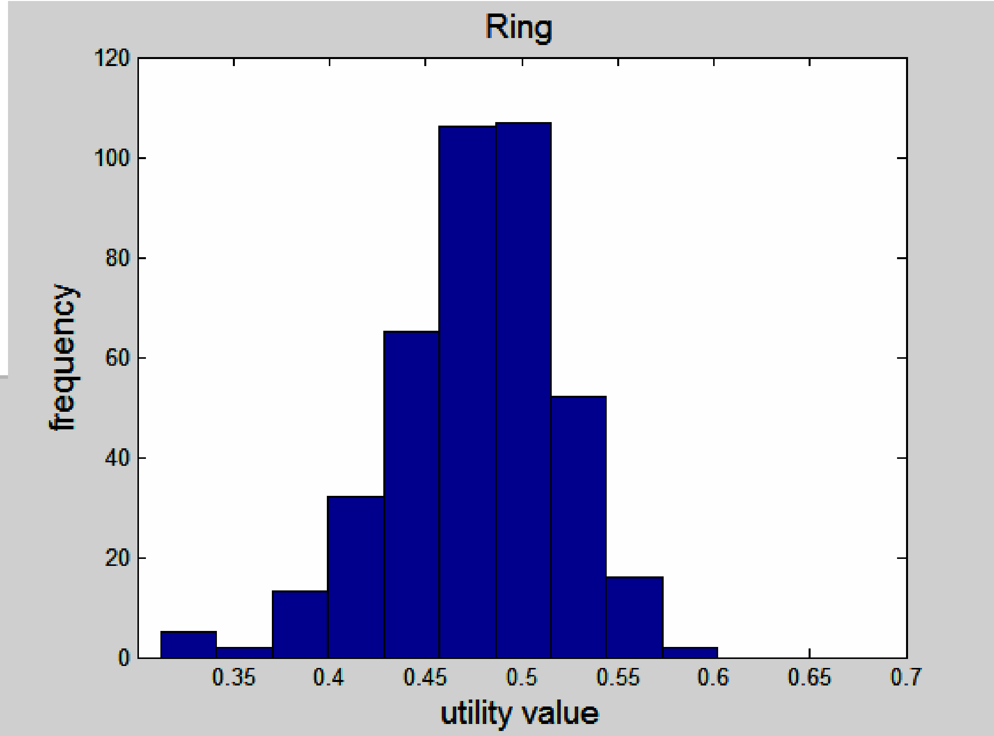
II. Calculation of the utility value $U(k)$

III. The steps I and II are repeated for $k = 1,.....n$ (usually $n = 600$ is chosen).

IV. Construction of a non-parametric estimator of probability density function and histogram from the values $U(k)$ (k=1,.....n) and computation of the $E(U)$, $D(U)$ estimates.

Simulation results





Complex criterion rotor :

Mean	lower limit	upper limit
4.90e-001	4.86e-001	4.94e-001

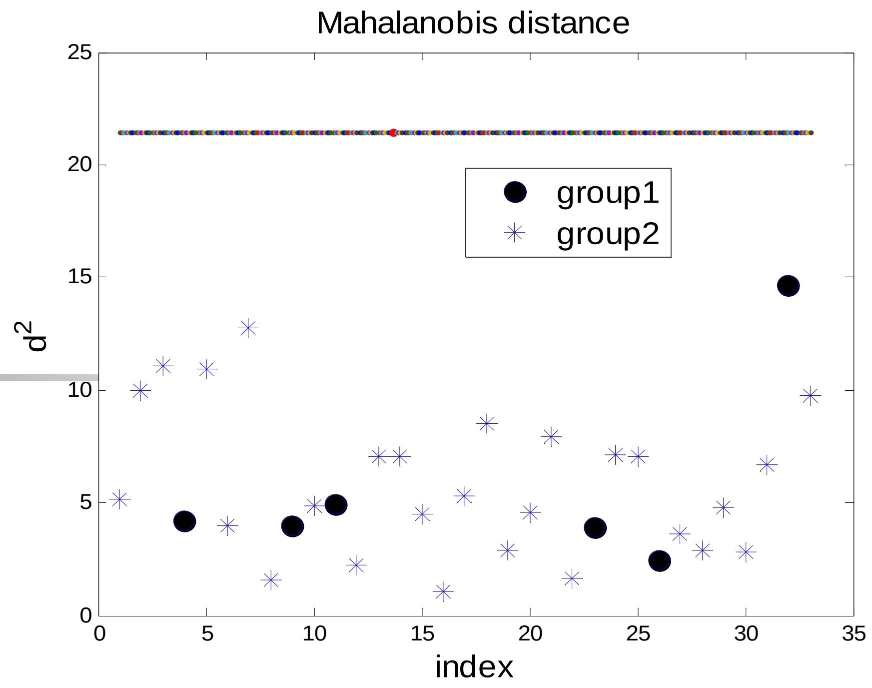
Complex criterion ring :

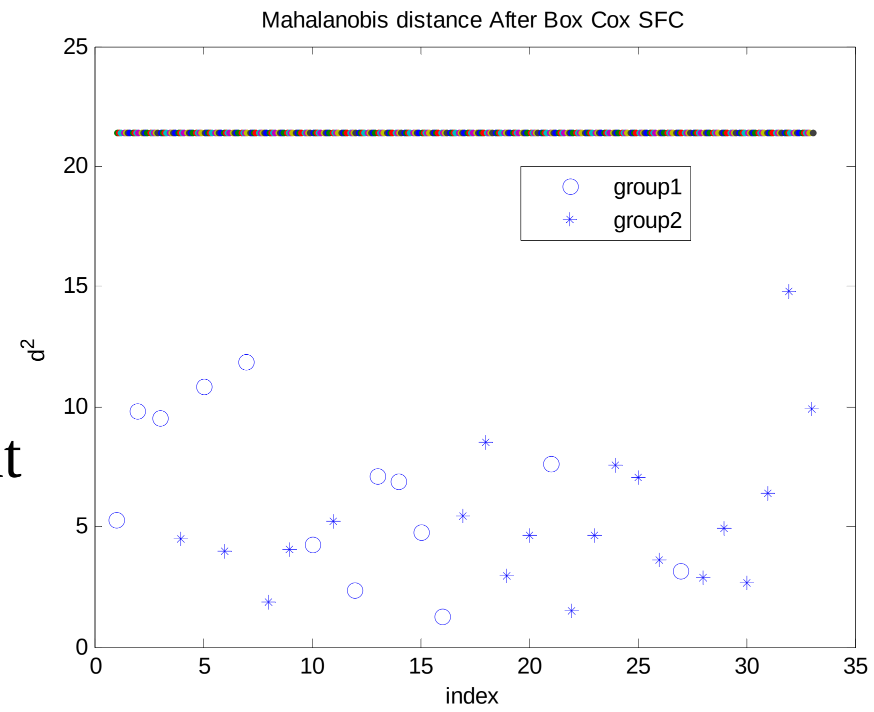
Mean	lower limit	upper limit
4.79e-001	4.75e-001	4.84e-001

Coefficient of variation = 3 %

Experimental part

- The results of the crop study of 1997 and 1998, which includes 33 sets of cottons were used for comparison of *U* with some other cotton quality indices.
- For different cotton varieties the International Textile Center (USA) evaluated the all characteristics required to computation of cotton quality index U excluding thrash content (in evaluation of U the value *TR = 2* was selected).



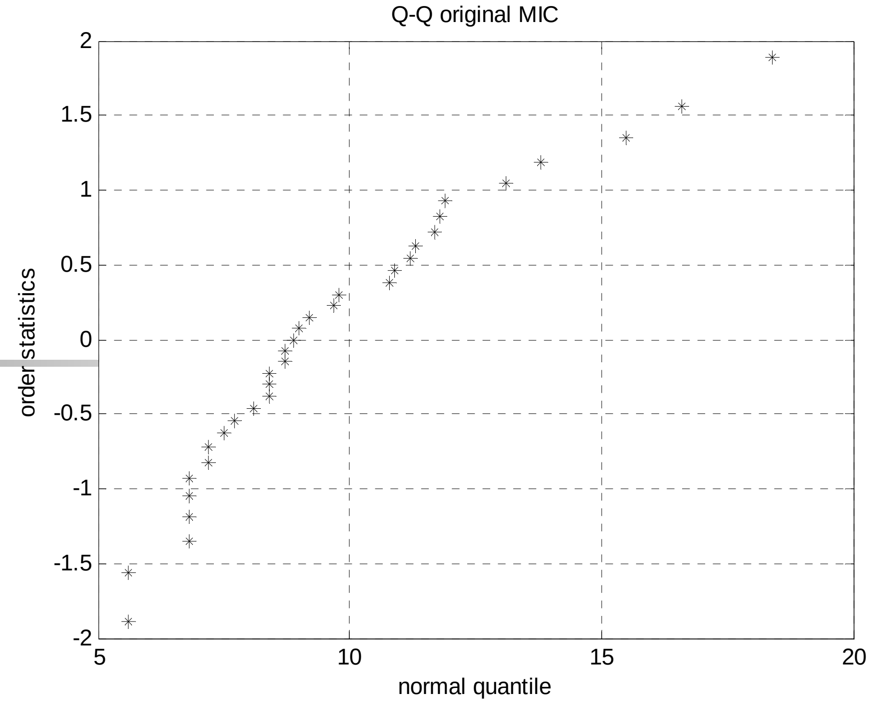


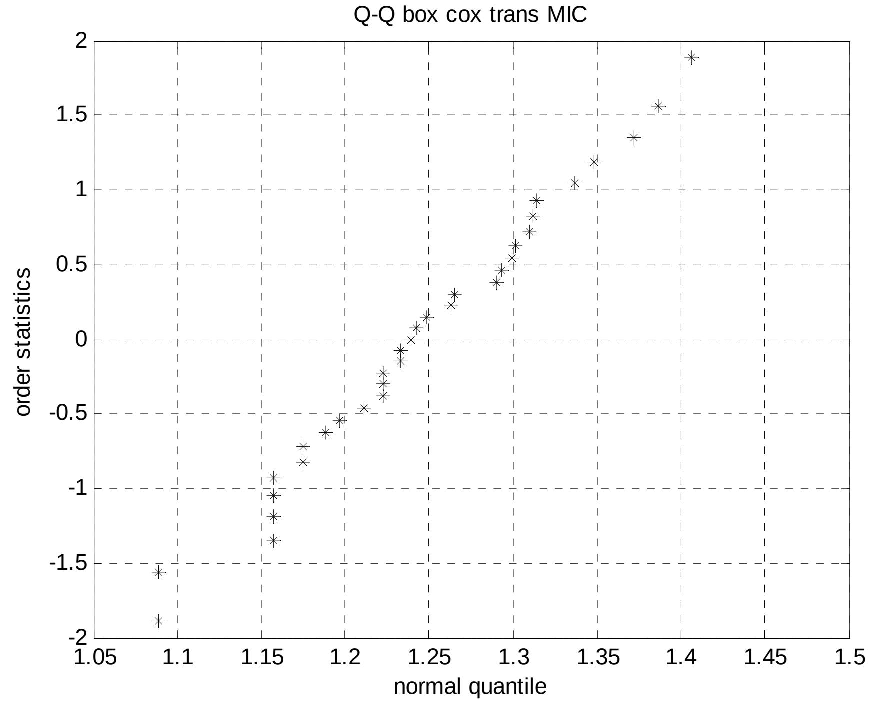
Box Cox transform

- For all variable excluding SFC is Box Cox transform inefficient.
- For SFC is optimal lambda = -0.5800

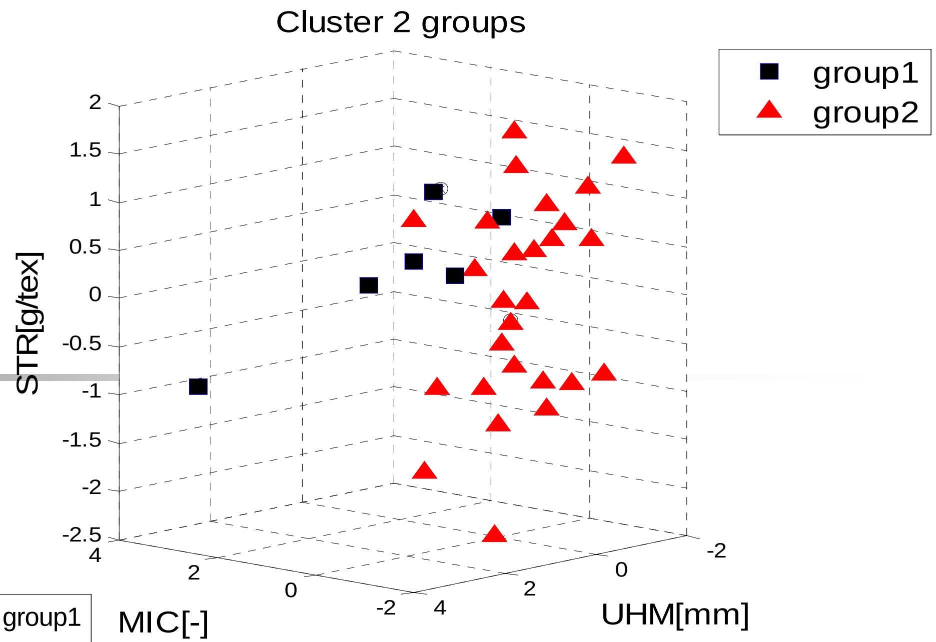
CI =[-1.6900   0.5].

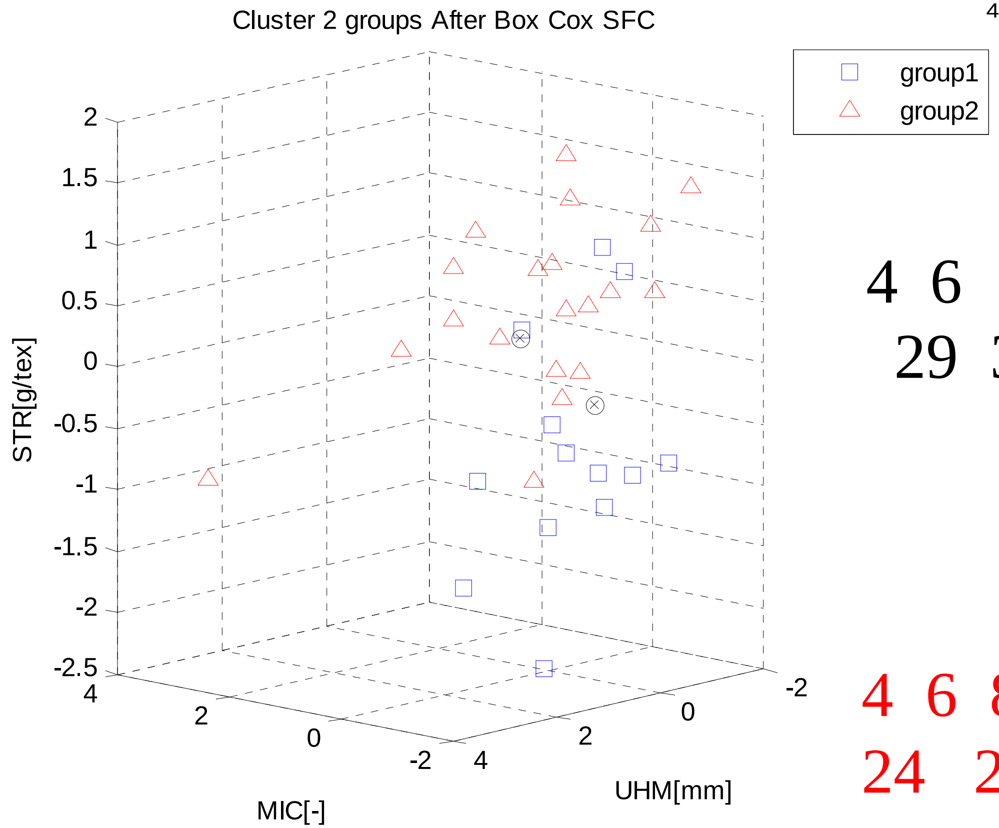
The logarithmic transform is selected





Cottons clustering



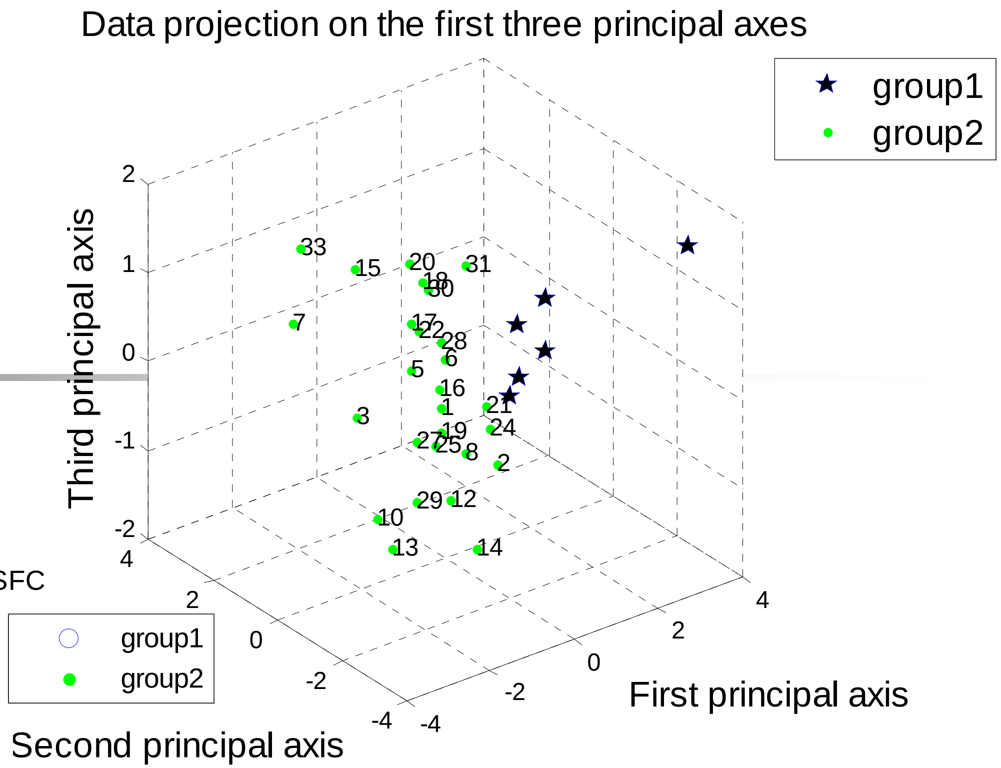


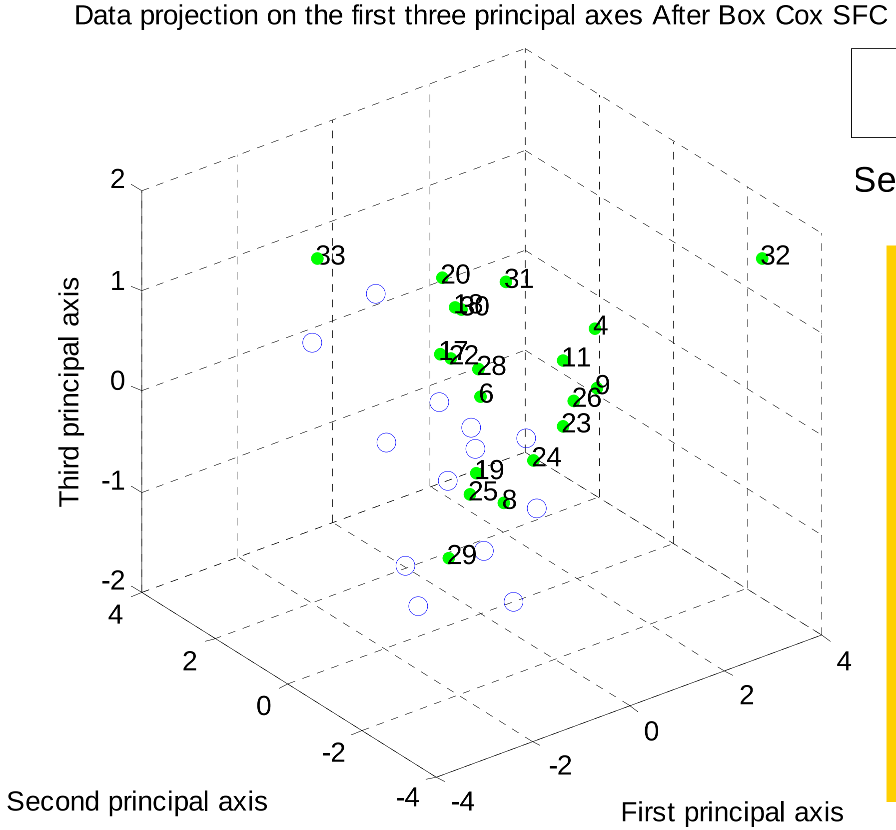
4 6 8 9 11 19 20 23 24 25 26 28 29 30 31 32

First cluster

4 6 8 9 11 17 18 19 20 22 23 24 25 26 28 29 30 31 32 33

Cottons in PCA space



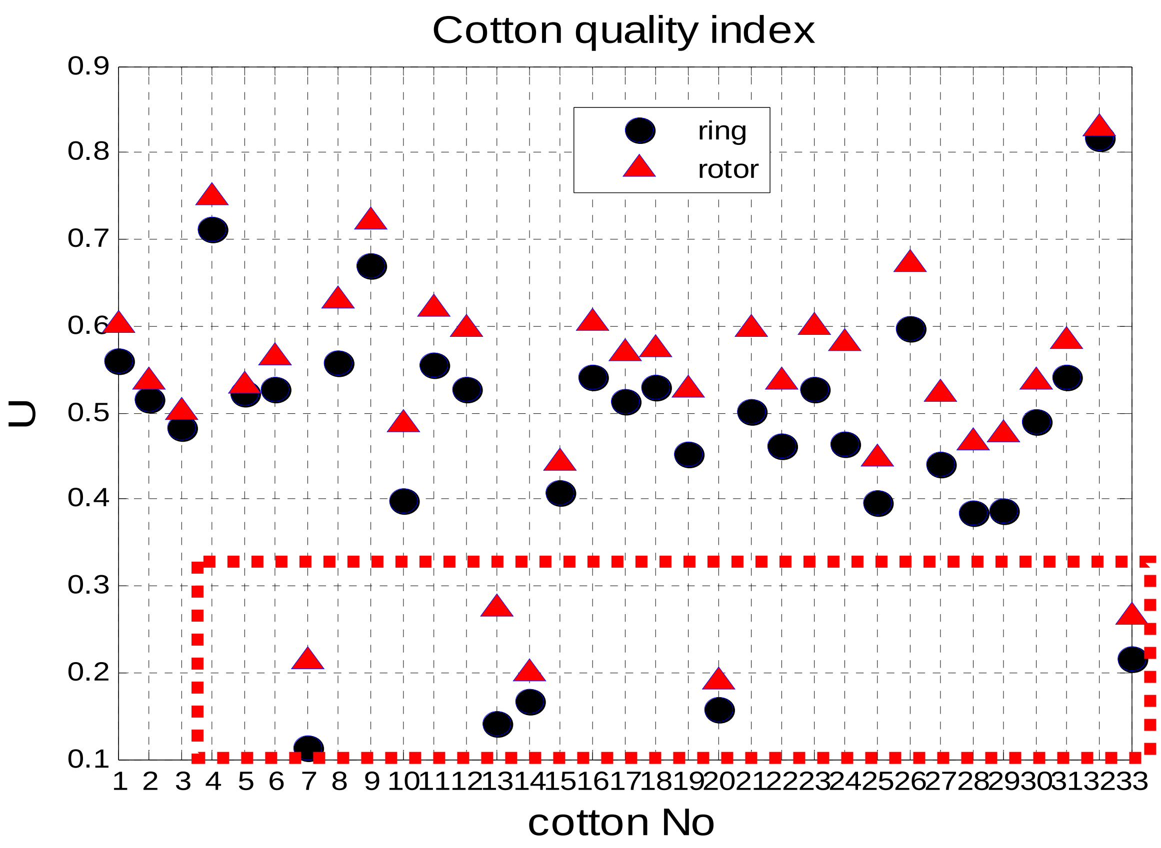


The differences due to transform of SFC of cotton in the principal component spaces are low

- The standardized data were used

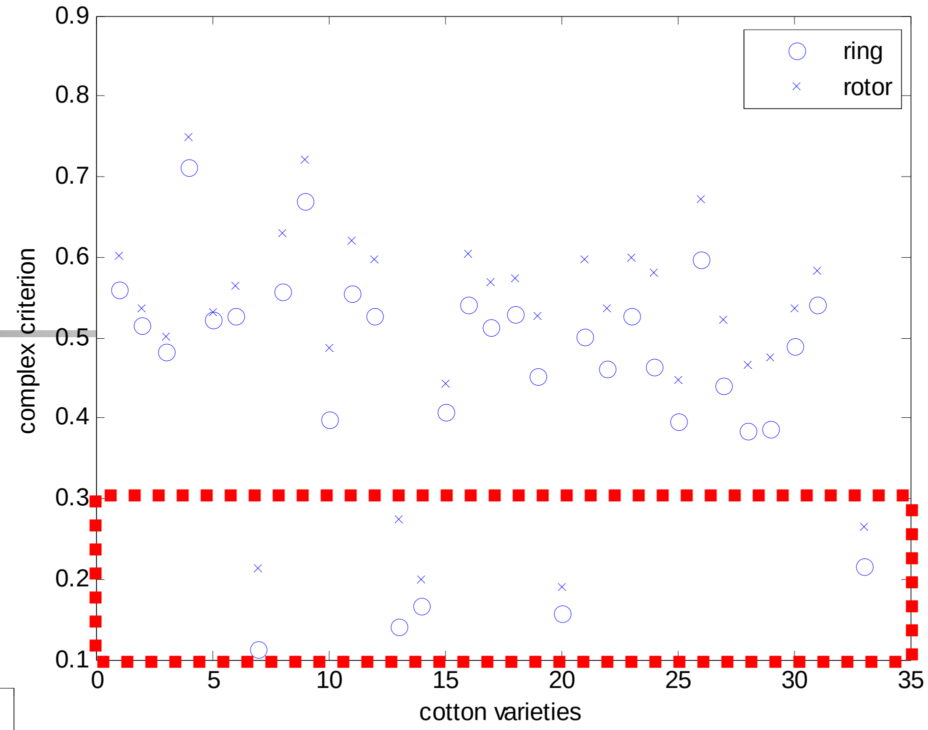
Cotton quality index

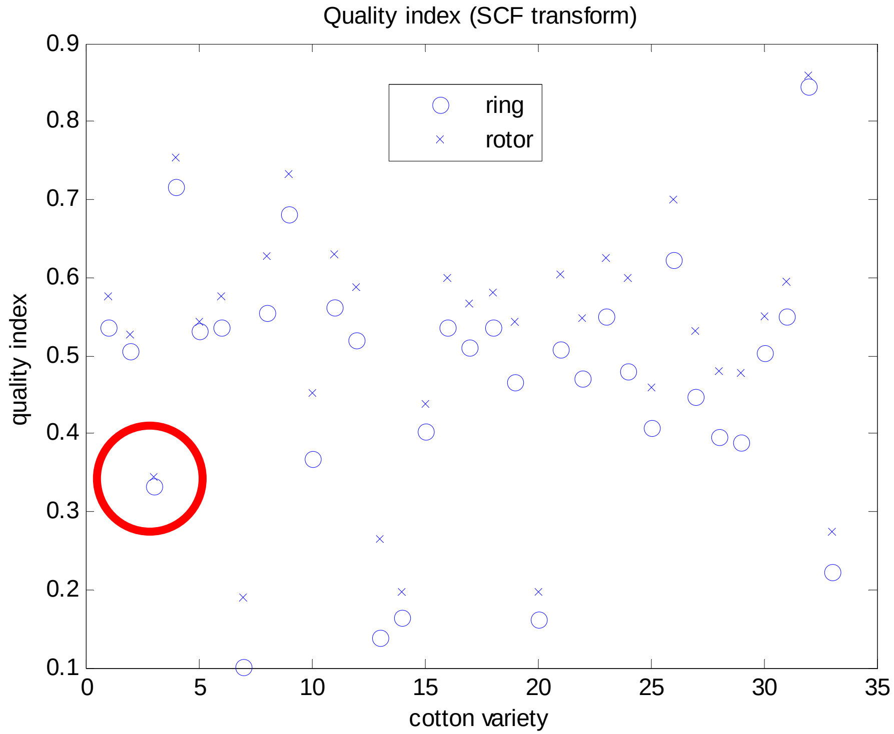
The cotton quality index *U* for all cotton varieties

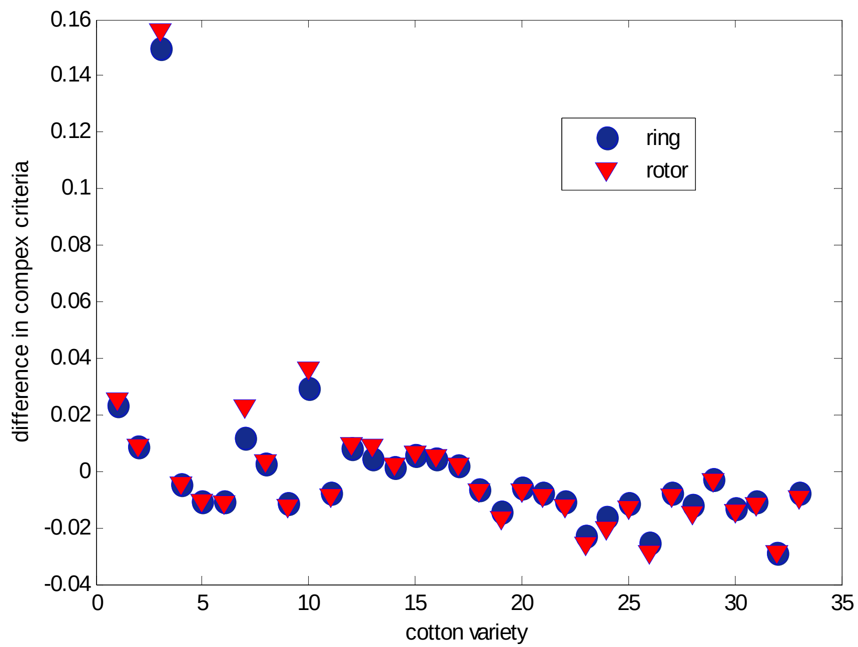


It is visible that according to the cotton quality index are cotton varieties separated into two groups. In the low value of *U* are cottons with very low UHM (No. 7, 13, 33) or very high micronaire (No. 7, 14, 20, 33).

Influence of transform

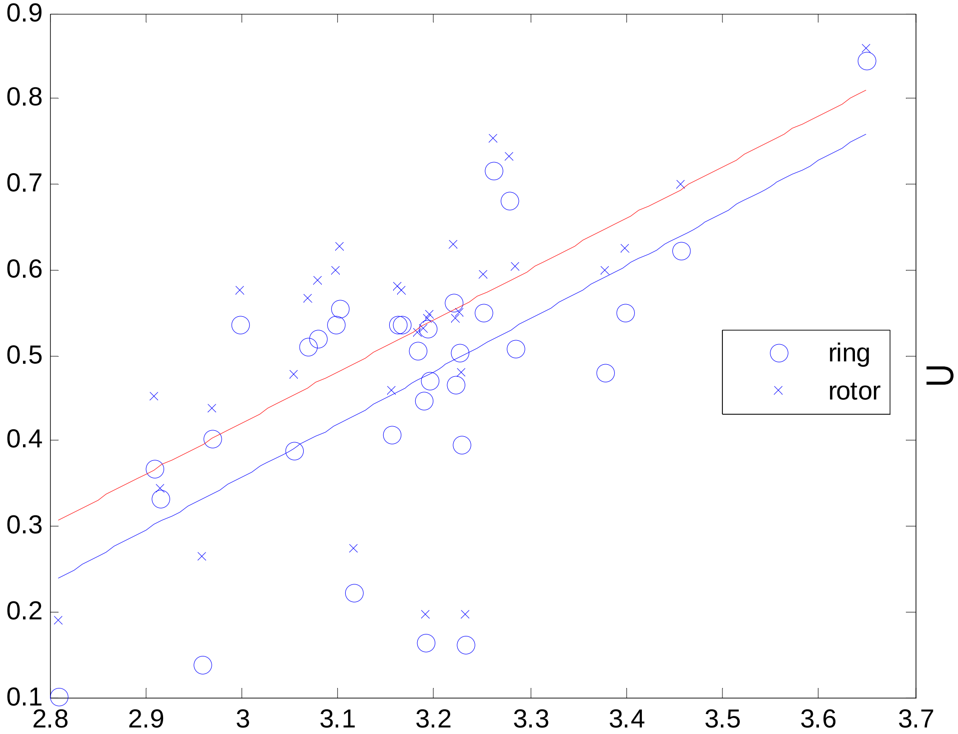




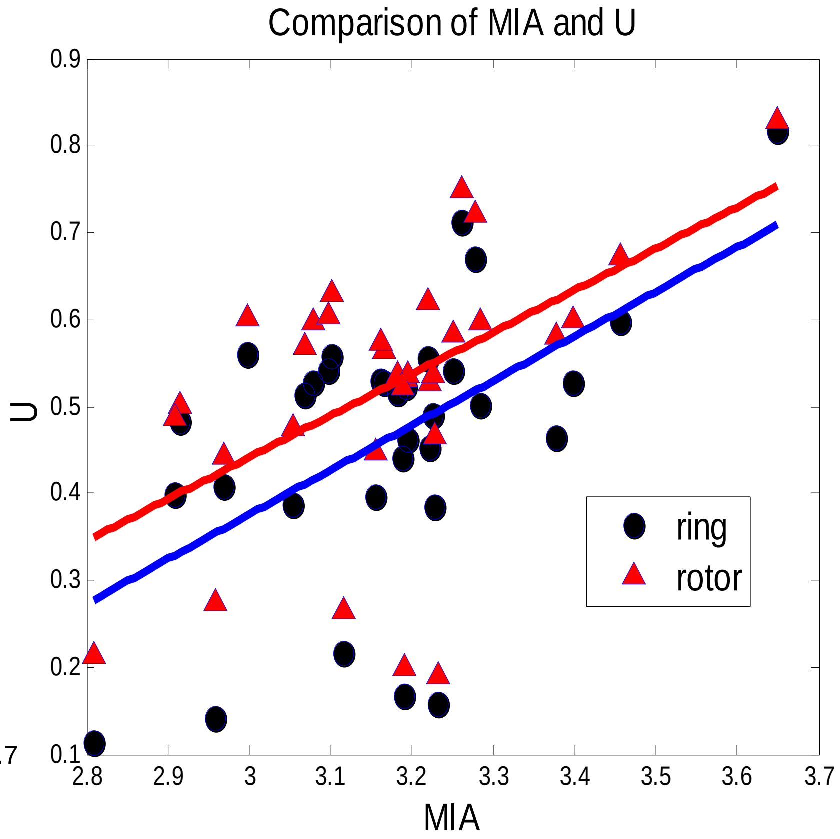


Inter dependencies

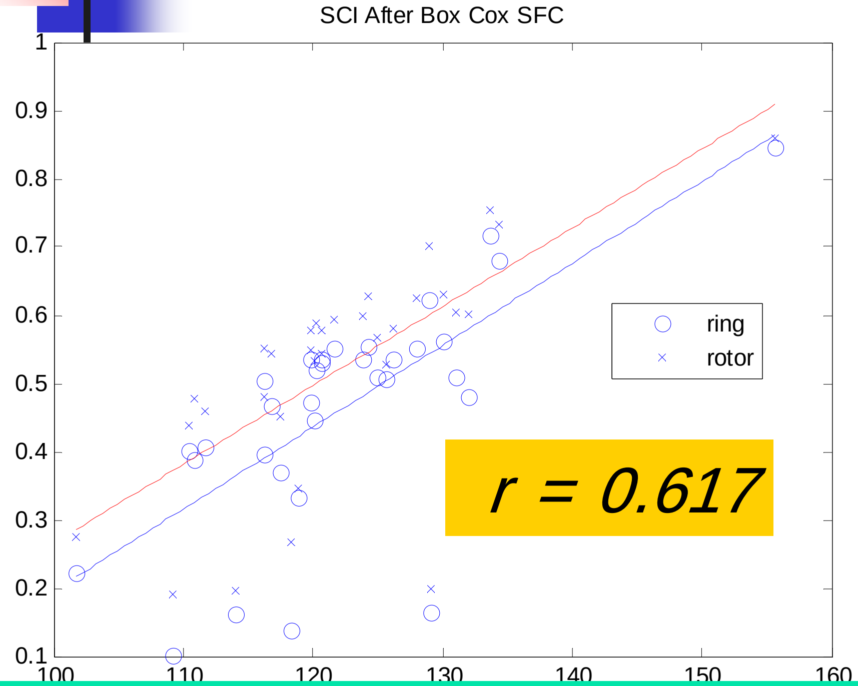
r = 0.624

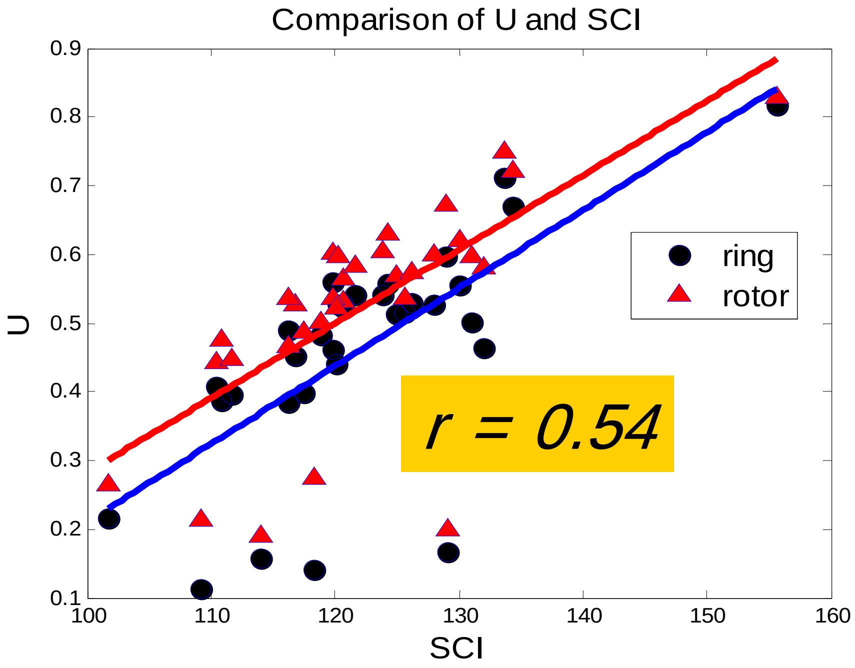


r = 0.53



Inter dependencies





The best correlation exists between U and SCI. In each cases are visible the outlying points with very low values of *U*. Main reason of this deviations are higher low limits for UHM (25 mm) and low higher limit for MIC (5) used for computation of U.

Application of U

The complex criterion U (cotton quality index) can be used for creation of:

1. Control charts (on line monitoring and control of production)
2. Expressing of spinning ability of cotton fibers

The differences between utility values for various cottons can be visualized by comparing of corresponding confidence intervals

Complex criterion of quality i.e. cotton quality index is useful for prediction of usefulness of cotton fibers in textile mills or for characterization of differences between various varieties.

Conclusion

- Described procedure for evaluation of cotton quality index (U) can be very simply modified for other selected properties or other set of weights.
- Based on preliminary results it will be probably necessary to solve problems with some cotton varieties having small micronaire due to fineness and relatively high strength.
- For some cases will be Box Cox transform beneficial (in other case it is necessary to add restriction to the L_1 and H_1)