Data Mining:
A potential detector to find failure in complex components

Authors: E. M. Scheideler, A. Ahlemeyer-Stubbe: Data Mining

in: Elio Padoano, Franz-Josef Villmer, Proceedings 5 th International Conference October 1 and 2, 2015 Trieste, Italy Production Engineering and Management, Volume 11/2015 Publication Series in Logistics Department of Production Engineering and Management Ostwestfalen-Lippe University of Applied Sciences, Lemgo (Germany)

https://www.hs-owl.de/fileadmin/diman/Veroeffentlichungen/PEM_Tagung_zusammen2015.pdf

ABSTRACT:

This paper is aimed to discuss current research using data mining techniques and industry statistics in production environments. The general research approach is based on the idea of using data mining processes and techniques of industry statistics to find rare and hidden patterns behind failures of complex components.

A case study will be applied to illustrate how the technique is carried out and where the limits of this approach occur. The case study deals with a component supplier of printing machines, which received an increasing number of client complaints, all related to one distinct problem. The observed failures seem to occur only among clients with very high quality standards. The affected component undergoes a very complex production process with several steps in different departments. Every single production unit records data information from multiple process variables and at different points in time. In the beginning there was no understanding of the failure causes in production at all. Therefore a huge amount of production data had to be analyzed to find the pattern that discloses the failure.

The data mining process starts with a first step in which the given data sets are prepared and then cleaned, followed up by building a prediction model. The aim is to detect the root causes for failures and to predict potential failures in affected components.

This paper shows how to use data mining to get the answer on pressing production failures.

KEYWORDS:

Data mining, production failure, multi-variant analysis, multivariate process control, predictive modelling, case study

1 PROBLEM BACKGROUND

A product of a printing components manufacturer shows an increasing number of client complaints in recent years, all related to one distinct problem. The nonconforming component leads to more insufficient print quality. Depending on the client´s individual quality standard the failure causes complaints. The different quality standards result in the

fact that the peculiar problem occurs among clients who need a very high print quality for their products or respectively these clients notice the reduced output due to their sensitive products much earlier.

The occurring reduced output causes subsequently, extreme abbreviated service time of the component in the machine, and consequently induce down time for the clients.

At the moment the insufficient components are sent to the manufacturer for amendment. Up to the present claimed parts were completely revised and sent back to the client. This shows that some components are fully operational and some again hold the reduced output, which leads to another client's complaints.

On the part of the manufacturer it is not clear yet which work steps and production parameters in his fabrication release the reduced output and thus the client complaints. During manufacturing, the affected component undergoes a very complex production process with several steps and several production parameters. Another aggravating factor is that the products are manufactured in different departments in shift operation.

So far, the manufacturer has not detected any assessment criteria for avoiding the failure and recognizing the reduced output before delivery in time.

The aim of this study is to detect hints regarding these criteria with the help of statistical analysis, especially Data Mining methods.

Data Mining methods have a long success story in areas of marketing, financial service, fraud detection and health. Health monitoring during operation, anomaly detection on running machines (for example aircrafts) are often done with Machine learning method [1].

In production environments Data Mining is still not a common tool to detect production failure. Using prediction methods of Data Mining to find assessment criteria for avoiding the failure is also not widely distributed. But the needed theories and methods are well known, and solutions can be adapted from the named areas above [2] [3] [4].

This work deals with an existing production problem, how Data Mining can help to solve it.

The overall project is divided into two phases:

1. phase: analysis of the components geometry
2. phase: analysis of the production unit records of the component and data collected during the application of the component in the production environment until the failure occurs

From the manufacturer´s point of view it is important if there is a special significance to find out and to solve possible in-house reasons for the reduced output of the components.

The present paper is dealing solely with the first phase. It is structured by the following topics:

- data
- analysis
- findings
- future prospects

2 INPUT DATA

2.1 Required data

The company is recording a lot of data during the production process. These datasets deal with the geometry, material and information of used semi finished parts of the components. We have also access to some data out of the manufacturing process itself, like information on which production line the component is built, how long it takes, at which time in the year and which person has done special manufacturing processes. Actually there is no access to production process variables like grinding speed.

For this survey we got access to a dataset that covers two years of production and repair details. All build components of the concerned type are listed and the data contain the information, if the problem (failure) occurs or not. The file contains about 100521 Data sets and 65 variables. It includes components with no failures and components with failures as well. The data were delivered in a format that needs some preparation to get ready for analytics.

2.2 Data preparation

Having obtained useful data, it now needs to be prepared for analysis. It is not useful to have the data stored at quite a detailed level in the data mart. But to get relevant results out of the analyses, transformation and aggregation of the data is necessary. It is unlikely that Data Mining algorithms will find hidden patterns without prior data preparation.

Preparation time is expected to be at least 70% of the total time required to do the Data Mining.

In every data preparation step the time constraints during the production have to be considered. This is essential in order to use the results to predict potential failure for the individual component before this is delivered to the client. Therefore, only data that are recorded before the component leaves the company can be used in any kind of prediction. This sounds obvious, but it is based on the fact that the file also contains repairing data: it is important to pay attention to this.

If we the study is not limited to descriptive analyze and it is planned to develop a prediction model, a target variable is needed. In our case the target variable will be created as a binary variable. If the failure comes up or not this has to be transformed in just two values such 0/1. “1” indicates the component has a failure “0” means everything is fine with the component.

In general: the data must be screened for empty areas and it must be decided how to handle them. Depending on the individual situation we use one out of three options:

- to exclude that dataset;
- input a potential good estimator for the missing value;
- do nothing because empty means nothing happened.

Several further data cleaning steps have to happen. Columns in which text is written have to be cleaned such that similar meaning is shown in one equal string. For example, an additional blank changes the string: for the analytical tool the strings then are different.

After all steps of transformation and aggregation, one row for every built component with all the variables inside is needed. Our data file include now “7966” datasets and 13 Variables.

3 ANALYTICS

3.1 Pre-analytics

As a first step, a descriptive analysis was done over all variables. Examining those results, first valuable hints on the data are collected. Fig. 1 gives a typical picture and it gives also good impression on the data we like to use for further analytical steps. Comparing the distribution given in Fig. 1 and 2, it is obvious that it might be more likely that the plotted “ABBALLENLAENGE” might have any impact compared with the “APBAENDER” plotted in Fig. 2 that shows no variance.

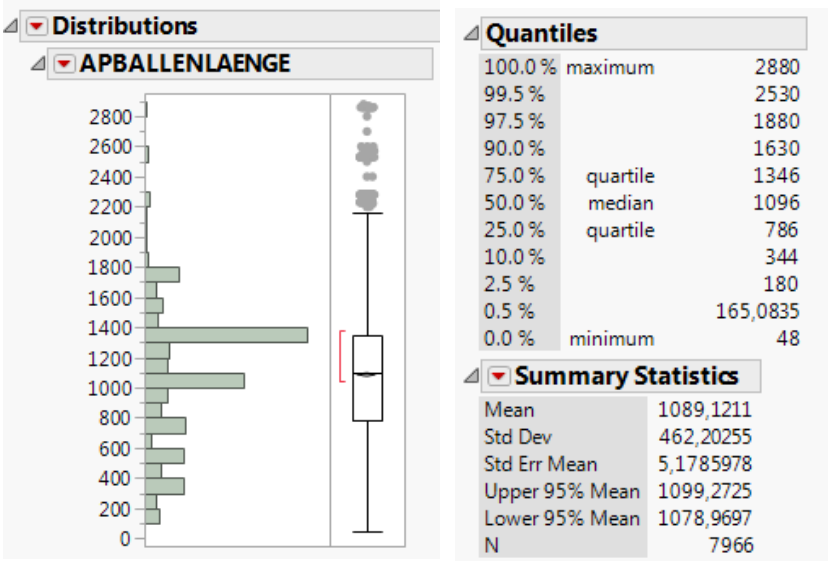


Figure 1: Histogram of geometry parameter length

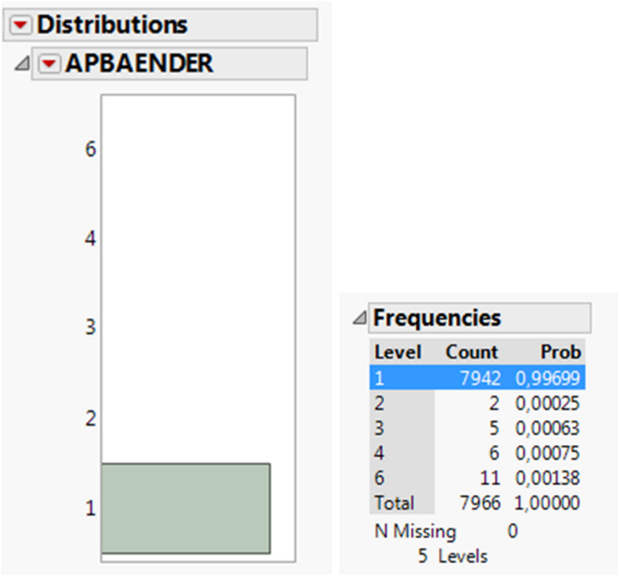


Figure 2: Histogram of geometry parameter “APBAENDER”

One of the most important things is to find out how many failures occur in the data base. In this case 7966 data sets and 84 targets (FEHLERBILD REKLA-INDEX) with 1 mean that the failure are inside. Nearly one percent of the produced components are defective.

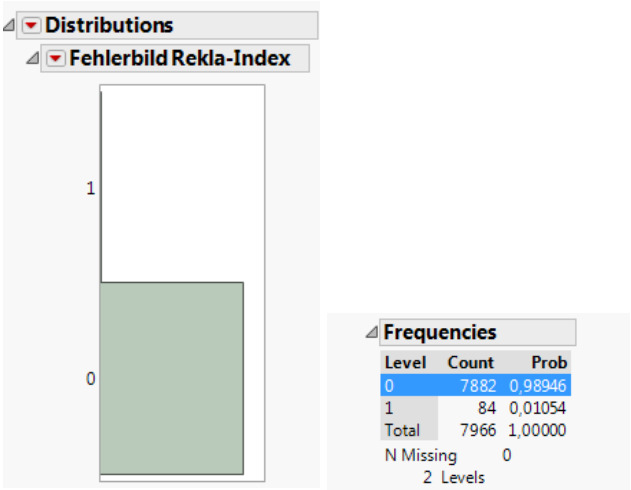


Figure 3: Histogram of target variable “REKLA-INDEX”

From a production point of view, 1% is much too high. But under analytical aspect it might be too low. How we will handle this will be described in chapter 4.

The pre analysis of the failure distribution to the clients, shows that only a small number of clients reclaims failure. Fig. 4 shows the distribution of FEHLERBILD REKLA-INDEX 1 over clients.

The client 11611.001 bought 15 components and made claims for all components (100% error rate). This overview indicates the risk that the target variable „FEHLERBILD REKLA-INDEX” failure or no failure was not answered from all clients with the same quality

standards. It seems that the claims are strongly depending on the individual quality standard to the client itself. That means that the estimated number
of unreported cases of the failure is much higher.

An error rate of 100% leads to suspicion that the detailed terms of use at the clients affects the specific problem. These will be investigated in the second phase of the project.

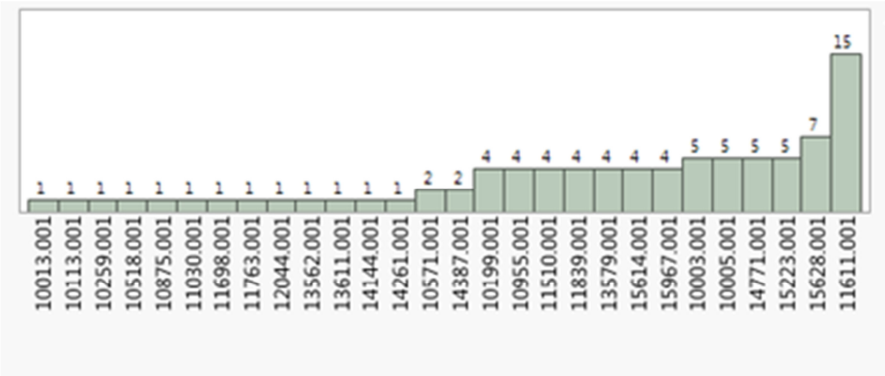


Figure 4: Histogram customer distribution to “FEHLERBILD REKLA-INDEX” 1 (failure)

Beside the fact that there is one client with an error rate of 100%, there are several customers (with several components bought) who have not reported any failures.

3.2 Multi variate Analysis

After finishing the descriptive analysis, a multi variate analysis is done with the main focus on variables containing geometry information. It is expected to find any kind of relationship (correlation, collinearity) between the different parameters. Some of these relations are pretty obvious and easy to detect by the use of domain knowledge, but others are quite astonishing. To see not only the relations between the variables, the statistics are also grouped by the target Variable (FEHLERBILD REKLA-INDEX).

The scattering matrix of a multi way analysis contains scatter diagrams for each combination of the variables. On the diagonal the histogram of each variable is shown. The left part of Fig. 5 shows the scattering matrix for the “FEHLERBILD REKLA-INDEX” value 0 (no failure), the right part of Fig. 5 “FEHLERBILD REKLA-INDEX” value 1 (failure).

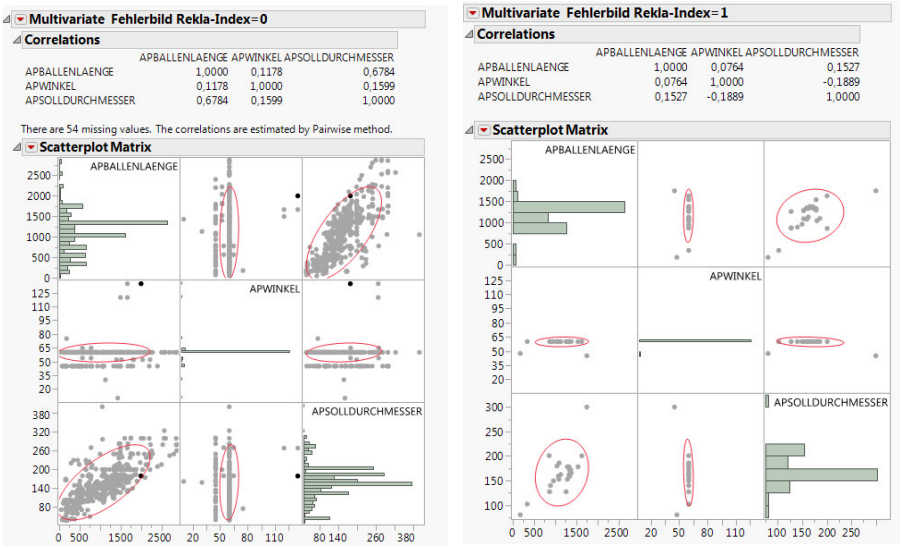


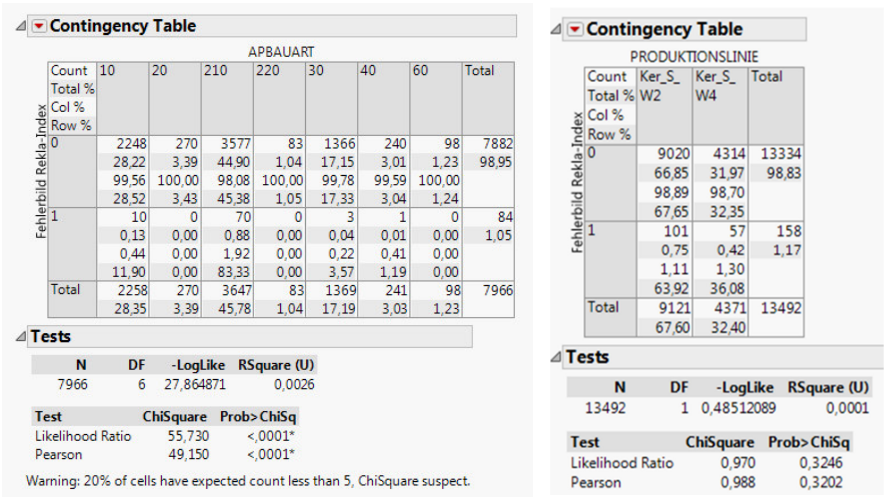
Figure 5: Multivariate Correlations geometry for Target value 0 / 1

Comparing both parts of the figure, it is obvious that there are different patterns for “FEHLERBILD REKLA-INDEX” 1 and “FEHLERBILD REKLA-INDEX” 0. For example, for the combination “APBALLENLAENGE” / “APSOLLDURCHMESSER”, it can be seen that the failure are likely to appear on components with 130-225mm “APSOLLDURCHMESSER” and a APBALLENLAENGE” of 750mm to 1750mm. For the no failure case a very different pattern is shown.

With the help of contingency tables and Chi-Square testing the individual power of a variable to explain the target variable is detected. An example is the contingency table of “APBAUART” versus “FEHLERBILD REKLA-INDEX” (Fig. 6, left part) and “PRODUKTIONSLINIE” versus “FEHLERBILD REKLA-INDEX” (Fig. 6, right part). The variable “APBAUART” is an example of a potential explanatory variable and the variable “PRODUKTIONSLINIE” (production line) is an example of a variable with no single impact on the failure.

It can be easily seen that both contingency table and the result of the chi-square testing give clear hints about which variables are potentially good explanation factors to detect failure. Although the software indicates invalidity of the chi square test for “APBAUART”, due to many cells occupied too less, these overall results are very obvious.

“APBAUART” 210 is a clear candidate to indicate potential failure.



Contingency Table

APBAUART

Fehlerbild Rekla-Index	Count / Total % / Col % / Row %	10	20	210	220	30	40	60	Total
0		2248	270	3577	83	1366	240	98	7882
		28,22	3,39	44,90	1,04	17,15	3,01	1,23	98,95
		99,56	100,00	98,08	100,00	99,78	99,59	100,00	
		28,52	3,43	45,38	1,05	17,33	3,04	1,24	
1		10	0	70	0	3	1	0	84
		0,13	0,00	0,88	0,00	0,04	0,01	0,00	1,05
		0,44	0,00	1,92	0,00	0,22	0,41	0,00	
		11,90	0,00	83,33	0,00	3,57	1,19	0,00	
Total		2258	270	3647	83	1369	241	98	7966
		28,35	3,39	45,78	1,04	17,19	3,03	1,23	

Tests

N	DF	-LogLike	RSquare (U)
7966	6	27,864871	0,0026

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	55,730	<,0001*
Pearson	49,150	<,0001*

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

Contingency Table

PRODUKTIONSLINIE

Fehlerbild Rekla-Index	Count / Total % / Col % / Row %	Ker_S_W2	Ker_S_W4	Total
0		9020	4314	13334
		66,85	31,97	98,83
		98,89	98,70	
		67,65	32,35	
1		101	57	158
		0,75	0,42	1,17
		1,11	1,30	
		63,92	36,08	
Total		9121	4371	13492
		67,60	32,40	

Tests

N	DF	-LogLike	RSquare (U)
13492	1	0,48512089	0,0001

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	0,970	0,3246
Pearson	0,988	0,3202

Figure 6: Contingency table APBAUART / FEHLERBILD REKLA-INDEX

As a result of the analytics done in this chapter we are able to do some feature reduction and to find potential good interaction to be used in a predictive model (Section 4).

4 PREDICTIVE MODEL

Based on the pre-analytics results above, it is planned to develop a predictive model that can be used to forecast whether a module is likely to fail or not. As a side aspect the model also generates clear hints of how to avoid future failure during the ongoing production.

The modelling process follows a general data mining process (Fig. 7).

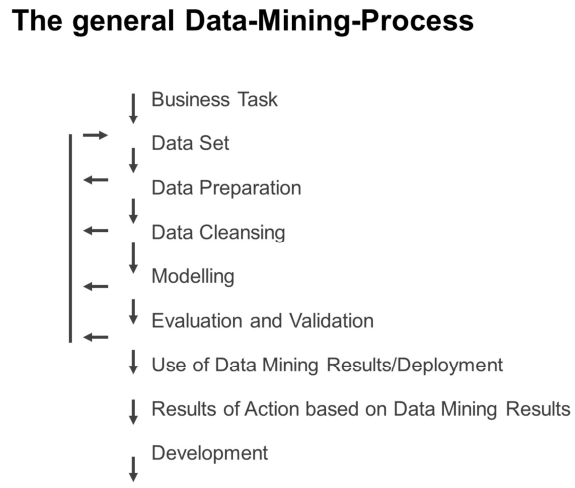


Figure 7: general Data-Mining- Process [2]

As described above, all steps until modelling are already done. We are now focusing on the modelling part.

To ensure good and reliable model quality to work with a train and test approach on one hand or to use cross validation on the other hand is indicated. This ensures overfitting of the model as well. Based on the fact that just 1% of all datasets are “FEHLERBILD REKLA-INDEX” =1. It was decided to use cross validation as evaluation and validation technique. Details can be found in [3].

As modelling technique decision trees are chosen. The main drivers behind this decision are the facts that in most cases they come up with robust results, a good graphical representation and the opportunity to translate the model itself in actionable rules [2] [3] [4].

Best results are produced by the C4.5 algorithm. The C4.5 algorithm is a decision tree algorithm that is a further development of the ID3 [5]. The main advantages of C4.5 algorithm are the use of categories and numeric values and an error based pruning. C4.5 is in use and described in detail in several data mining text book for the last 20 years. Apart from the paper by Quinlan [5], an applied description of C4.5 is also given in [6].

Confusion Matrix		To (predicted Target)		
		1	**0**	**all**
from (real Target)	**1**	**35**	49	84
	0	93	**7789**	7882
	all	128	7838	7966

Figure 7: Confusion Matrix

The confusion matrix above (Fig. 7) means, that if the model provides a list of 128 components, 27.3% (35 components out of 128) of them are genuinely positive and 41.6% of genuine positives (35 out of 84 components) do appear on the list. In contrast, randomly choosing 128 components from 7966 would yield only 1.6% of positives on average, and 98.4% of actual positives would be missed.

It is obvious that it is not likely to find a model that predicts 100% of all failures correctly. But the found model is a good starting point to reduce the amount and the percentage of components that failed.

5 RESULTS

In general the project shows the power of Data Mining techniques to solve quality problems in production areas. Differently from traditional quality management techniques, like e.g. Statistical process control (SPC) –Charts [7], a model was worked out under a Data Mining process [8]. This Data Mining model enables the company to optimize their production and to predict potential failures before they are delivered to the client side, even when a certain

overestimation will waste efforts in additional quality control on those that are wrongly estimated as potential failures. But under the bottom line it saves money, because the cost of reclamation, after the component is part of the clients' production line, are much higher. Client's satisfaction and the related increased likelihood of future purchases are another monetary value as well.

Based on some detailed findings in phase 1, the manufacturer also started technical investigation of the problem. The first finding indicates a waviness on the surface with very low amplitudes [9].

6 FUTURE

As indicated in the beginning, based on the experience of this first project phase we will conclude with the second phase. The major drivers of the second phase are:

- to develop a "standard" quality measurement tool kit to ensure that the amount of today's undetected failures will be reduced and to make failure quotes comparable;
- to get more detailed production unit record. Actual there are some hints that the failure pattern can be caused by special dynamic behavior;
- data/detailed information on the environment and the circumstance of how the component is embedded in the customer production line;
- results of the physical findings will be included as well.

At the end of phase 2 we expect an improved and more precise model that helps to detect potential failures as soon as possible and that reduces the amount of reclamation.

REFERENCES

[1] Perner, P. (2013): Machine Learning and Data mining in pattern recognition, 9 th International conference, MLDM 2013, New York, NY, USA July 19-25, 2013 Proceedings
[2] Ahlemeyer-Stubbe, A., Coleman, S. (2014) A Practical Guide to Data Mining for Business and Industry, John Wiley&Sons
[3] Perner, P.(2002) Data Mining on Multimedia Data, Lecture Notes in Artificial Intelligence, Vol. 2558. Springer Verlag,
[4] Perner, P. (2015) Decision Tree Induction Methods and Their Application to Big Data In: Xhafa. F, Barolli L., Barolli A, Papajorgji P. (Eds.), Modeling and Optimization in Science and Technologies, Modeling and Processing for Next-Generation Big-Data Technologies With Applications and Case Studies, Volume 4 2015: 57-88, Springer Verlag.
[5] Quinlan, J. R. (1993) C4.5 : Programm for Machine Learning, Morgan Kaufmann Publishers Inc. San Mateo CA
[6] Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods and Algorithms, IEEE Press, Hoboken
[7] Juran, J.M., De Feo, J. A. (1999) Juran's Quality Handbook, 5 Edition Mcgraw-Hill Education
[8] Geiger, W., Kotte W., (2008) Handbuch Qualität: Grundlagen und Elemente des Qualitätsmanagements: Systeme – Perspektiven, Springer Verlag
[9] WZL - Werkzeugmaschinenlabor der RWTH AACHEN: AiF 15539 Gezielte Prozessführung zur Vermeidung von Kurzwelligkeiten beim Außenrund-Einstechschleifen, http://www.wzl.rwth-aachen.de