Raport Badawczy

Research Report

RB/46/2013

Process control using predicted quality data

O. Hryniewicz, J. Karpiński

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences

POLSKA AKADEMIA NAUK

Instytut Badań Systemowych

ul. Newelska 6
01-447 Warszawa
tel.: (+48) (22) 3810100
fax: (+48) (22) 3810105

Kierownik Zakładu zgłaszający pracę:
Prof. dr hab. inż. Olgierd Hryniewicz

Warszawa 2013

Process control using predicted quality data

Olgierd Hryniewicz and Janusz Karpinski

Abstract SPC procedures are usually designed to control stability of directly observed parameters of a process. However, when quality parameters of interest are related to reliability characteristics it is practically hardly possible to monitor such characteristics directly. Instead, we use some training data in order to build a model that is used for the prediction of the value of an unobservable variable of interest basing on the values of observed explanatory variables. Such prediction models have been developed for normally distributed characteristics, both observable and unobservable. However, when reliability is concerned the random variables of interest are usually described by non-normal distributions, and their mutual dependence may be quite complicated. In the paper we consider the model of a process when traditionally applied assumptions are violated. We show that in such a case some non-statistical prediction models proposed in the area of data-mining, such as Quinlan's C4.5 decision tree, perform better than popular linear prediction models. However, new problems have to be considered when shifts in the levels of process parameters may influence the performance of applied classification algorithms.

1 Introduction

Statistical decision procedures of Statistical Quality Control (SQC) are mainly designed for the analysis of independent, and usually normally distributed, quality characteristics. For modern production processes new measurement techniques allow to describe processes using many characteristics and these characteristics are often assumed to be independent. For many years the T^2 control chart, introduced by Hotelling in the 1947, was the only SPC tool used for SPC of processes described by such multivariate data, see Montgomery (2011). However, during the last

Systems Research Institute, Newelska 6, 01-447 Warsaw, Poland, e-mail: hryniewi@ibspan.waw.pl

twenty years some new techniques have been proposed for dealing with interdependent statistical quality data. For example, control charts for the parameters of the so called profiles have been introduced in order to control not only numerical values of quality characteristics, but the structure of their mutual dependence as well, see the paper by Woodall et al. (2004), the book by Noorsana et al. (2011), and recent papers by Xu et al. (2012), and by Wang and Huwang (2012) for more information. These methods can be used for the analysis of different dependencies of a regression type, both linear and non-linear. However, in practically all cases the proposed models have been obtained under the assumption of normality of measured characteristics. Moreover, it is assumed that all important quality characteristics of interest are *directly* measurable.

In many production processes parameters of produced objects can be measured for all produced items. Thus, one can say that a 100% quality inspection can be implemented for such processes. However, the parameters that can be measured during the production process may not be necessarily the same as the actual quality characteristics that determine the quality of produced items. When reliability is a key quality parameter, the important reliability characteristic such as the life-time cannot be directly measured during a production process. We face the similar situation when the measurement of quality characteristics may have a negative impact on the quality of inspected items. Similarly, the same problem is when the measurements of quality characteristics are costly (e.g. when the time of measurement is too long for a production process), and thus infeasible. In all such cases there are attempts to measure these characteristics indirectly by the measurements of other characteristics.

The problem of an indirect inspection of important quality characteristics was noticed for the first time more than fifty years ago, but since that time it has attracted the attention of relatively few authors. There exist two general approaches to cope with this problem. In the first approach, introduced by Owen and collaborators, see Owen and Su (1977), a multivariate probability distribution of the random vector $(Z, X_1, \dots, X_k)$ is built, where Z is the quality characteristic of interest, and $X_1, \dots, X_k$ are the characteristics that are directly measurable in a production process. The procedures obtained using this approach are acceptable only in the case of the multivariate (usually bivariate) normal (Gaussian) distribution describing $(Z, X_1, \dots, X_k)$. Another approach is based on the assumption that that the relation between the random variable Z and the variables $X_1, \dots, X_k$ is described by a certain (usually linear) regression model. Also in this case the normality assumption about Z is usually used in practice. In both cases there is a certain link of the proposed methods to the multivariate SPC tools mentioned in the first paragraph of this section. However, it has to be assumed that at certain moments of time both predicted and explanatory values have to be observed.

Unfortunately, for many processes the actual multivariate probability distribution of $(Z, X_1, \dots, X_k)$ is different from the normal (Gaussian) one. Moreover, the number of predictors (explanatory variables) $X_1, \dots, X_k$ is not small, and this creates additional problems with the usage of classical statistical procedures. In such cases building of a well-established probabilistic model is rather infeasible. Instead, Hryniewicz (2013) has proposed to use the data mining methodology for a simple classification of inspected items. In the first step used in his approach some (usually two: conforming and nonconforming) classes of inspected items are defined in relation to the possible values of Z. Then, a classifier (e.g. linear classifier, decision tree or artificial neural network) is built using a training data consisted of the limited number of observations $(Z,X_1,\ldots,X_k)$. Finally, the classifier is used in the inspection process for labeling the produced items.

Classifiers used in the inspection process are usually built using small amount of data, named training data. Thus, the results of classification are not error-free. What is more important, however, that the relation between the results of classification and the actual level of the quality characteristic of interest may be quite complicated. Therefore, as it has been noted in Hryniewicz (2013), there is a need to investigate the impact the quality of the classification procedures on the efficiency of SPC procedures used in production processes. Because of the complexity of possible models that describe interdependent and non-normal processes the evaluation mentioned above can be only done using computer simulations

The remaining part of this paper is organized as follows. In Section 2 we describe the simulation experiment proposed in Hryniewicz (2013) for the evaluation of different prediction algorithms. In the next section we use the simulation results for the performance analysis of of four algorithms used for the prediction purposes. We consider the simple linear regression model with a binary output (RegBin), two versions of the Linear Discrimination Analysis algorithms (LDA-s with a symmetric decision criterion, and LDA-as with an asymmetric decision criterion), and the Classification Decision Tree Quinlan's C4.5 algorithm. We evaluate these algorithms in terms of prediction errors for both non-shifted and shifted process levels. In Section 4 we consider the usage of the proposed prediction algorithms for the monitoring production processes with 100% inspection. We consider two cases: when decisions are made using the approach based on the classical Shewhart control chart methodology, and when decisions are made using the approach based on the Moving Average (MAV) control chart methodology. Some conclusions from the performed experiments, and indications for a future work are presented in the last section of the paper.

2 Simulation of indirectly observed processes - description of the simulation model

The problem of process control when quality characteristics of interest are not directly observable, but only assessed on the basis of observations of other, possibly related, variables is, as it has bin described in Hryniewicz (2013), much more complicated than the classical one when when all variables of interest are directly observable. The variability of an indirectly quality characteristic of interest consists of two parts. One is related to the inherent variability of the process itself, and the second one is related to unavoidable uncertainty of classification (prediction) procedures. When we have at our disposal only the predicted values of the quality characteristic, these two types of variability are practically inseparable.

As it has been noted in the previous section, the majority of statistical procedures used for the prediction of unobservable quality characteristics is based on the assumption of multivariate normality. Usually this assumption is reduced to the case that the quality characteristic of interest and its observable predictor are jointly distributed according to a bivariate normal distribution. When several possible predictors are available it is also often assumed that these predictors are statistically independent. Under such assumptions simple regression models (usually linear) are built and used for the purpose of quality evaluation. Unfortunately, when quality characteristics of interest describe reliability these simple assumptions are hardly acceptable. First of all, reliability characteristics, such as the life time, are usually described by strongly skewed distributions. Moreover, predictors are frequently modeled by random variables defined on subsets on positive real numbers, and their distributions can be quite far from the normal distribution. Finally, behind the values of observed predictors there are some common physical and chemical phenomena which often make them strongly statistically dependent. One can also add another dimension by assuming strong non-linearity of the relations describing physical phenomena with the observed life times, described e.g. by models of so called competitive risks. Thus, the real models describing the process of e.g. reliability prediction may be very complicated, and usually extremely difficult to identify.

In order to investigate the impact of some of the problems mentioned above on the efficiency of prediction (classification) process a simulation model that consists of three levels has been built. On the first level we have four random variables, denoted by A,B,C,D, respectively, which describe observable characteristics. These variables may be described by several probability distributions (normal, uniform, exponential, Weibull, log-normal) chosen by an experimenter. Observed variables may be pairwise dependent, and their dependence may be described by several copulas (normal, Clayton, Gumbel, Frank) chosen by an experimenter. The strength of dependence is defined by the value of Kendall's coefficient of association τ. Detailed information about the usage of copulas for the description of complex dependence structures can be found in the monograph by Nelsen (2006). These assumptions allow to simulate quite complicated structures of interdependent predictors. On the second level we have four hidden (unobservable) random variables H_A,H_B,H_C,H_D defined on a positive part of the real line. Their probability distributions may be chosen from the set of distributions used in the theory of reliability (exponential, Weibull, log-normal). Each of the hidden random variables is related to its respective observed variable, i.e. H_A to A, H_B to B, etc., and this relation is described by a chosen copula (with a given value of Kendall's τ) describing their joint probability distribution, and a certain linear dependence between their expected values. Finally, on the third level, hidden random variables are transformed to the final random variable T that describes the life time that can be observed only in specially designed experiments. The relation between H_A,H_B,H_C,H_D and T is strongly non-linear, and

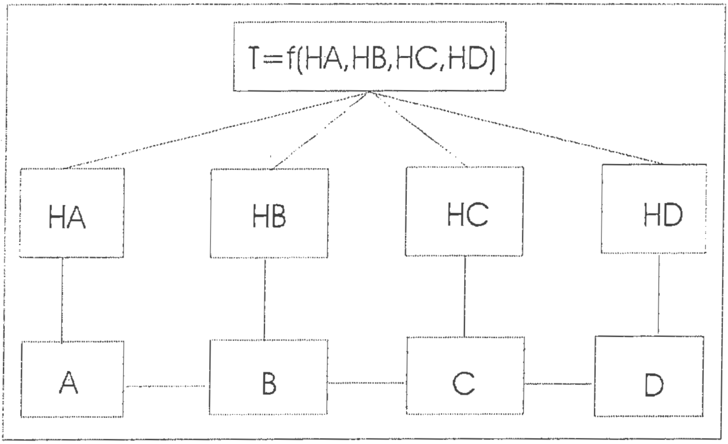

Fig. 1 Structure of the simulation model

is described by operators of a "min-max" type. The structure of the simulation model described above is presented on Figure 1.

The simulation system described above allows to simulate sets of data with a very complex, and practically impossible to be predicted in advance, structure. In this paper we show the results of experiments using the model described in Hryniewicz (2013). In this model A is distributed according to the normal distribution $N(5;0.5)$, B has the exponential distribution with the expected value equal to 10, C is distributed according to the log-normal distribution such that log of C is distributed according to the $N(5;1)$, and D has the Weibull distribution with the scale parameter equal to 10, and the shape parameter equal to 2.

The dependence between A and B has been described by the Clayton copula with $\tau = 0,8$. The joint distribution of B and C is described by the normal (gaussian) copula with $\tau = -0,8$ (Notice that this is bivariate "normal" distribution, but with non-normal marginals!), and the joint distribution of C and D has been described by the Frank copula with $\tau = 0,8$.

The hidden variable H_A is described by the log-normal distribution such that log of H_A is distributed according to the $N(3;1)$, and its joint probability distribution with A has been described by the normal copula with $\tau = -0,8$. The joint distribution of H_B and B has been described by the Frank copula with $\tau = 0,9$, and the marginal distribution of H_B is assumed to be the exponential with the expected value equal to 10. The joint model of H_C and C is similar, but the copula describing the dependence in this case is the Gumbel copula, and the marginal distribution of H_C is also assumed to be the exponential, but with the expected value

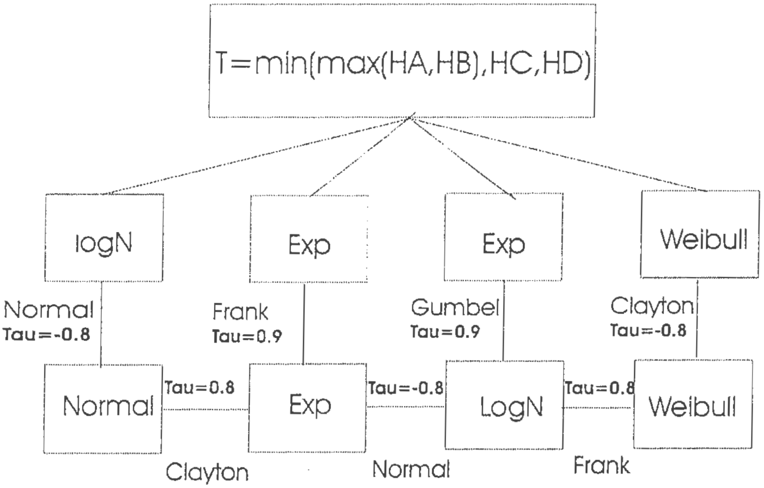

Fig. 2 Structure of the simulation experiment

equal to 20. Finally, the hidden variable H_D is described by the Weibull distribution with the scale parameter equal to 50, and the shape parameter equal to $1,5$, and its joint probability distribution with D has been described by the Clayton copula with $\tau = -0,8$. The random variable T that describes the life time has been defined as $T = min[max(H_A, H_B), H_C, H_D]$.

The parameters of the aforementioned distributions have been found experimentally in such a way, that unreliable items have their lifetimes T smaller than 5. Moreover, the relation between the observed variables A, B, C, D and their hidden counterparts H_A, H_B, H_C, H_D is such that a shift in the expected value of each observed variable, measured in terms of its standard deviation, results with the similar shift of the expected value of its hidden counterpart, measured in terms of its own standard deviation. The description of this experiment is presented on Figure 2.

3 Quality prediction of indirectly observed processes - evaluation of simulation experiments

3.1 Classification methods used for the prediction of quality characteristics

In the experiment described in Section 2 we generate the values of the random variable T that has the interpretation of the life time. For the purposes of quality inspection it is important if this value is greater than a certain critical value T_{crit}. Therefore, we need a classification method for the appropriate labeling produced item basing of the results of the measurements of some explanatory variables. In other words, we have to classify measure items into two classes: conforming (with $T > T_{crit}$) and nonconforming (if otherwise) items. There exist dozens of methods used for solving such classification problems. Some o them, based on some statistical assumptions, have certain optimal properties. The properties of other methods, mainly based on a data mining approach, can be only assessed experimentally. In our research we have considered the performance of several classification methods in the analysis of data generated by our simulation system.

First considered classification method is a naive linear regression. Let us label the class of unreliable items (i.e. those whose life-time is shorter than 5) by 1, and the class of remaining items by 2. Then, let us consider these labels as real numbers, treating them as observations of real dependent variable in the regression analysis of the following form:

$$CL = x_A * A + x_B * B + x_C * C + x_D * D + x_F, \tag{1}$$

where CL is the predicted class of an item described by explanatory variables A, B, C, D, and x_A, x_B, x_C, x_D, x_F are respective coefficients of regression equation estimated from a training set of n elements. The value of CL estimated from (1) is a real number, so an additional requirement is needed for the final classification (e.g. if $CL < 1,5$ an item is classified as unreliable, and otherwise, as a reliable one). The only advantage of this naive method is its simplicity. It can be easily implemented using statistical tools available in spreadsheets. We have also used the linear regression model for the prediction of T and further classification, but the results of classification were similar to the results of classification using this naive binary regression approach and are not presented in this paper.

The second considered classifier is based on classical statistical results of Fisher. It is known as the Linear Discrimination Analysis (LDA), and is described in many textbooks on multivariate statistical analysis, and data mining (see, e.g. Hastie et al. (2008). In this method statistical data are projected on a certain hyperplane estimated from the training data. Those data points who are closer to the mean value of the projected on this hyperplane training data representing the class 1 than to the mean value of training data representing the remaining class 2 are classified to the class 1. Otherwise, they are classified to the class 2. The equation of the hyperplane is given by the following formula:

$$L = y_A * A + y_B * B + y_C * C + y_D * D + y_F, \tag{2}$$

where L is the value of the transformed data point calculated using the values of explanatory variables A, B, C, D, and y_A, y_B, y_C, y_D, y_F are respective coefficients of the LDA equation estimated from a training set of n elements.The calculation of the LDA equation (2) is not so simple. However, it can be done using basic versions of many statistical packages such as SPSS, STATISTICA, etc.

If Z_L denote the decision point, a new item is classified to the class 1 if $L \geq Z_L$, and to the class 2, otherwise. In our research we have considered two methods of the calculation of Z_L, and thus two method of classification. In the first method, called in this paper as the LDA-s, this point, denoted by Z_{Ls}, is just the average of the mean values of the transformed data points from the training set that belonged to the class 1 and the class 2, respectively. In the second method, called in this paper as the LDA-as, the decision point, denoted by Z_{Las}, is - as suggested in Murtagh (1986) - the weighted average of the center points of data sets representing the class 1 and the class 2, with weights proportional to the sizes of these classes in the training data.

The fourth considered classification method is based on the implementation of the one of the most popular data mining classification algorithms, namely the classification decision tree (CDT) algorithm C4.5 introduced by Quinlan (1993), and described in many textbooks on data mining, such as Witten et al. (2011). In our experiments we used its version (known as J48) implemented in the WEKA software, available from the University of Waikato, Hamilton, New Zealand, under the GNU license. The decision tree is constructed using "IF..THEN..ELSE" rules, deducted from the training data. In the description of this classification method in this paper we use the notation of the MS Excel function $IF(lt, t, f)$, where lt is a logical condition (e.g. $C < 50$), t is the action when $lt = true$, and f is the action when $lt = false$. The actions t and f can be implementations of other IF functions, or - finally - the assignments of classes to the considered items.

3.2 Description of classifiers used in the simulation experiments

The classification rules (binary regression models, LDA linear equations with respective decision rules, and decision tree "IF.. THEN..ELSE" rules) are built using certain training data sets consisted of the values of all explanatory variables and the values of the quality variable of interest. In our experiment these training data sets have been generated by our simulation program using the model described in the previous section. In the artificial intelligence community it is assumed that good training data sets should consist of several hundreds of items. In our reliability prediction problem such large data sets are absolutely infeasible. In our simulation experiments for the generation of training data sets we have taken an upper feasible limit on the number n of observed items, namely $n = 100$. In order to estimate the effect of the randomness of the training data sets on the classification decision rules, and finally on the results of classification during a production process, we have generated several different data sets. For each of these data sets we have built all considered classification rules. Similarly to Hryniewicz (2013) we present the results for only 10 such sets.

The coefficients of the regression equation (1) are presented for the considered ten data sets in Table 1. Those coefficients that have been indicated by the regression statistical tool as statistically non-significant have been printed in this Table in *italics*. However, we have to remember that in the calculation of the significance of regression coefficient it is assumed that observed data are distributed according to the normal probability distribution. In our case it is is obviously not true, so in our classification experiment we have used full regression equations.

Table 1 Regression model - different sets of training data

Dataset	x_A	x_B	x_C	x_D	x_F
Set 1	*-0,133*	-0,034	*-0,0001*	-0,026	3,036
Set 2	*0,010*	-0,039	*0,0001*	-0,033	2,321
Set 3	*0,144*	-0,033	*<0,0001*	*0,0002*	*1,359*
Set 4	*-0,194*	-0,034	*-0,0005*	*-0,019*	3,396
Set 5	-0,346	-0,021	-0,0006	*0,002*	3,763
Set 6	*0,073*	-0,047	*<0,0001*	*-0,023*	2,058
Set 7	*-0,142*	-0,040	-0,0004	*-0,0008*	3,019
Set 8	*0,109*	-0,038	*<0,0001*	*-0,001*	1,611
Set 9	-0,346	-0,039	*0,0002*	-0,034	4,054
Set 10	*-0,002*	-0,018	-0,0005	0,042	1,745

Just a first look at Table 1 reveals that the estimated regression equation (1) may be completely different, depending on the chosen training data set. However, some general pattern is visible: only explanatory variable B is significant for all regression lines. On the other hand, explanatory variable C seems to be of no practical importance in the classification process.

A comparison of the decision model parameters for different training data sets in the LDA case is presented in Table 2. In this case we cannot say about statistical significance of the parameters of the decision rule. However, the general impression is that some predictors are of limited importance for the classification purposes. The particular models look completely different depending on the training data set. However, in all the cases the explanatory variable C seems to have no effect (very low values of the coefficient describing this variable) on the classification.

Now, let us considered different decision rules estimated for the CDT algorithm. Because of a completely different structure of decision rules presented in Table 3 we cannot compare directly these rules with the rules described by the equations (1-2). They also look completely different for different training data sets, but in nearly all

Table 2 Linear discrimination analysis - different sets of training data

Dataset	y_A	y_B	y_C	y_D	y_F	Z_{Ls}	Z_{Las}
Set 1	0,687	0,174	0,001	0,133	-6,338	0,628	1,255
Set 2	-0,045	0,178	-0,001	0,151	-2,710	-0,014	-0,028
Set 3	-0,646	0,148	<0,0005	-0,001	1,663	0,464	0,927
Set 4	0,880	0,152	0,002	0,087	-7,499	0,585	1,171
Set 5	1,500	0,091	0,003	-0,008	-9,121	0,254	0,508
Set 6	-0,342	0,219	<0,0005	0,107	-1.399	0,706	1,412
Set 7	0,703	0,196	0,002	0,037	-6,044	0,784	1,568
Set 8	-0,501	0,173	<0,0005	0,006	0,636	0,629	1,258
Set 9	1,458	0,127	0,001	0,143	-10,048	0,344	0,687
Set 10	0,008	0,087	0,002	-0,206	0,272	0,771	1,542

cases (except for the Set 9) decision are predominantly (and in one case exclusively) based on the value of the explanatory variable C.

Table 3 Decision trees - different sets of training data

Dataset	Decision rule
Set 1	$IF(C <= 70,0181; IF(C <= 56,1124; 1; IF(C <= 63,2962; 2; 1)); IF(D <= 16,4381; 2; IF(A <= 4,3509; 2; 1)))$
Set 2	$IF(C <= 56,4865; 1; IF(D <= 17,3301; 2; IF(A <= 4,0217; 2; 1)))$
Set 3	$IF(C <= 73,6148; IF(C <= 57,1355; 1; IF(A <= 5,0876; 2; IF(D <= 4,497; 2; 1))); IF(D <= 17,3499; 2; 1))$
Set 4	$IF(C <= 70,2191; 1; IF(D <= 15,9098; 2; 1))$
Set 5	$IF(C <= 73,1584; 1; (IF(D <= 17,0516; (IF(C <= 87,8921; (IF(D <= 5,0679; 2; 1)); 2)); 1)))$
Set 6	$IF(C <= 60,3912; 1; IF(D <= 16,3504; 2; 1))$
Set 7	$IF(C <= 71,8184; 1; 2)$
Set 8	$IF(C <= 71,4456; 1; (IF(C <= 983,0929; 2; (IF(D <= 18,8213; 2; 1)))))$
Set 9	$IF(B <= 14,7339; (IF(D <= 16,7482; (IF(D <= 4,527; 1; 2)); 1)); 1)$
Set 10	$IF(C <= 60,5044; 1; 2)$

One can notice that the weight assigned to the explanatory variables in the CDT algorithm is nearly exactly opposite to the weights assigned in the RegBin (1) and LDA (2) classification models. In order to explain this shocking difference one should look at Figure 3. The dependence between the life time E and the explanatory variables C and D in not only non-linear, but non-monotonic as well. This dependence cannot be captured by the measures of linear correlation in the linear models (1)-(2). However, it seems that the explanatory potential of these two variables is much greater than the potential of the variables A and B. We have investigated this problem in our further simulation experiments.

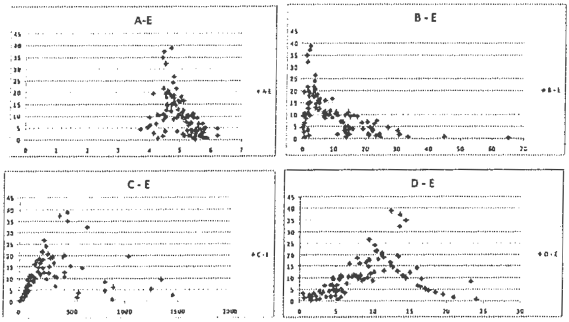

Fig. 3 Dependencies between the life-time E and explanatory variables A, B, C, D [Hryniewicz (2013)]

3.3 Evaluation of classifiers - constant process levels

In order to evaluate the quality of different classification methods used for the prediction of quality characteristics generated by our simulation model we have performed several tests. Each of the performed experiments consisted of 1000 simulation runs. For each run a set of classifiers that were built using the same training set was chosen randomly from among the available 10 sets. Then, a sample of n items described by vectors of variables (T, A, B, C, D) was generated using the already defined model. The values of (A, B, C, D) were used for the prediction of the quality of the simulated items. The actual (generated) value of $T$was used for the calculation of the actual fraction nonconforming, and the labels obtained by the classifiers were used for the calculation of the predicted fraction nonconforming. Thus, obtained for the each type of classifier values of the reported fraction nonconforming are averaged over 10 particular classifiers used for the prediction of the quality. In Table 4 we present the results of the estimation of actual and reported fractions nonconforming (process levels) using the samples of $n = 5000$ items. Together with the average values (Avg) we present the values of the coefficients of variation (CVar) which represent the variability of the estimated process levels, and the coefficient of skewness (Skew).

Table 4 Actual and reported values of the fraction nonconforming

	Actual	RegBin	LDA-s	LDA-as	C4.5
Avg	0,253	0,161	0,250	0,183	0,255
CVar	0,024	0,181	0,180	0,356	0,193
Skew	0,052	0,624	0,231	1,133	0,349

From the analysis of the data presented in Table 4 one can easily see that the reported level of the fraction nonconforming may be significantly different from the actual one. This is the result of wrong classifications made by the classifiers. There are two types of misclassified ides: false positives (FP), i.e. nonconforming items classified as conforming ones, and false negatives (FN), i.e. conforming items falsely classified as nonconforming. This problem will be discussed later on, but now we have to notice its most important consequences. When the fraction of FP's is greater than the fraction of FN's the reported fraction nonconforming is lower than the actual one, as it is the case for RegBin, LDA-s, and LDA-as classifiers. When the reported fraction nonconforming is greater than the actual one, as it is the case for the C4.5 classifier, the fraction of FN's is greater than the fraction of FP's. Usually non-detection of a nonconforming item leads to worse consequences than a false detection of a conforming item, and in such a case RegBin and LDA-as classifiers seem to be inferior to their remaining two competitors.

From Table 4 one can also see that the coefficient of variation for the reported fraction nonconforming (process levels) is significantly greater than for the case of the actual one. This is the result of additional variability introduced by the random choice of classifiers. In Table 5 we show how the reported fraction nonconforming vary for different sets of training data used for building classifiers. The results presented in this table are based on the simulation of samples of $n = 1000$ items. The total number of simulation runs was 1000. Hence, for each of the considered sets of training data (and the classifiers built using these data) there were, on average, 100 data points. Therefore, the particular values of the fraction conforming displayed in this table are not very accurate.

Table 5 Actual and reported values of the fraction nonconforming - different sets of training data

Set	Actual	RegBin	LDA-s	LDA-as	C4.5
Set 1	0,252	0,137	0,262	0,170	0,220
Set 2	0,255	0,155	0,328	0,332	0,205
Set 3	0,254	0,141	0,212	1,148	0,259
Set 4	0,252	0,131	0,179	0,129	0,304
Set 5	0,254	0,207	0,303	0,250	0,337
Set 6	0,252	0,164	0,210	0,145	0,254
Set 7	0,253	0,184	0,263	0,176	0,234
Set 8	0,254	0,157	0,213	0,144	0,244
Set 9	0,252	0,214	0,289	0,221	0,323
Set 10	0,252	0,129	0,232	0,118	0,185

The information contained in Table 5 is more readable when presented graphically, as on Figure 4. From this picture one can see that the RegBin classifier consistently reports lower fraction nonconforming, and its shows that the fraction of FP's is for this classifier excessively high. The LDA-as classifier behaves similarly, except for the cases of Set 2 and Set 5 where it reports, respectively, fraction nonconforming higher or equal to the actual one. This is also the reason for the large coefficient

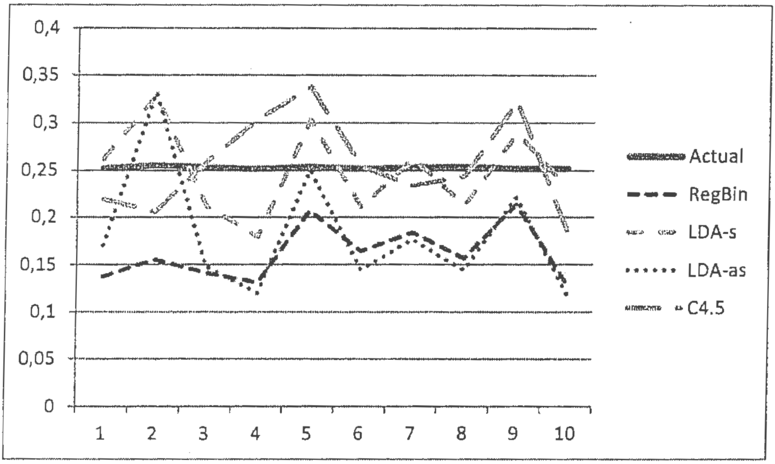

Fig. 4 Actual and reported values of the fraction nonconforming - different sets of training data

of variability of this classifier, as it can be read from Table 4. The behavior of the LDA-s and C4.5 classifiers is similar. The reported fraction nonconforming varies significantly from set to set of training data for both these classifiers. However, the coefficient of variation for the LDA-s classifier is slightly smaller.

As it has been already noted, the difference between the reported fraction nonconforming and the actual one shows if the fraction of FP's exceeds that of FN's or not. However, it is much more important to know what are the actual values of the fractions of misclassified items. On Figure 5 we show the results of another simulation experiment (for the sample size $n = 500$) where the actual fractions of FP's are presented for different classifiers and different training data sets. From this figure we see that the C4.5 classifier outperforms its competitors, as for nearly all training data sets it gives the lowest fraction of false classifications of the FP type. The RegBin and LDA-asym classifiers are definitely bad with respect to the fraction of FP's.

The similar picture that presents the actual fractions of FN's for different classifiers and different training data sets is shown on Figure 6. In this case the classifiers that have worse behavior with respect to FP's look better. However, the C4.5 classifier which outperforms the other ones with respect to FP's does not perform very bad

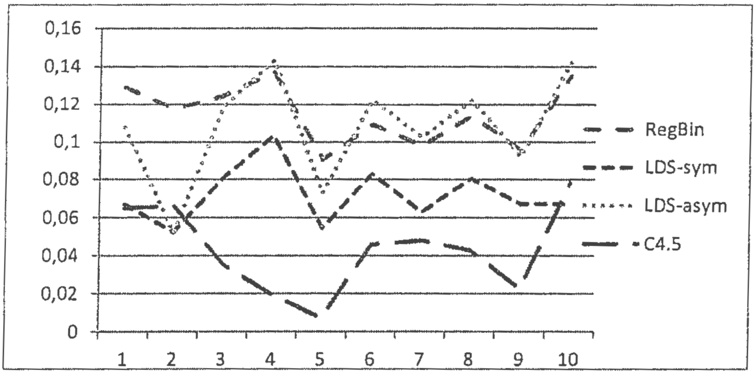

Fig. 5 Fraction of FP's for the process under control - different sets of training data

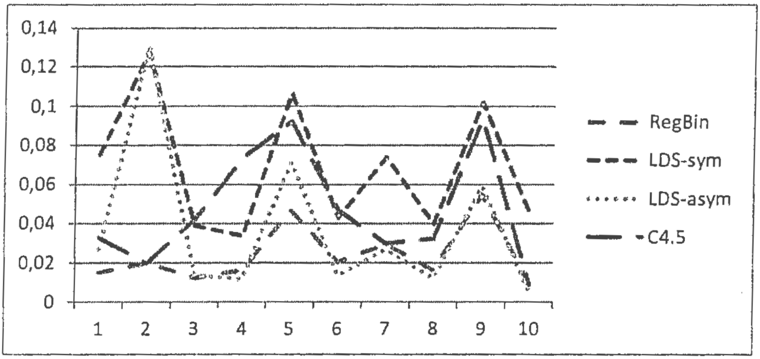

Fig. 6 Fraction of FN's for the process under control - different sets of training data

with respect to FN's. So one can say that for the process under control this classifier is the best from among those considered in this paper.

3.4 Evaluation of classifiers - variable process levels

In the section 3.3 we have considered the case when the probability distributions of the explanatory (predictive) variables *A*, *B*, *C*, and *D* are the same for the training data used for the construction of classifiers and the process data. Thus, the actual and reported fraction nonconforming for the training and process data are governed by the same probability distribution. However, processes where the considered classifiers are used may vary in time in many different ways. In this paper we consider only the simplest case when the change of the process parameters is described by the shift (up or down) of the expected value of only one of explanatory variables. We will report these shifts as the multiples of the standard deviations of the respective explanatory variables.

First of all, we have to note that the shift of the expected value of an explanatory variable may result in two mechanisms of the change of the observed process level. First, it changes, but to usually unknown extent (because of very complicated relations between the values of considered variables), the actual fraction of nonconforming items. Second, it results in changing the efficiency of the used classifiers. After such a shift has occurred the existing classification rules do not fit to the actual data, and the probabilities of false classification, both for FP's and FN's, may change quite dramatically.

Let us consider the shifts of the expected values of the explanatory variables that results in the deterioration of the actual process level described by the actual fraction nonconforming. In our model we observe such deterioration for the positive (upwards) shifts of the expected value of A, and the negative (downwards) shifts of the expected values of B, C, and D. In Table 6 we present the values of actual and reported fraction nonconforming when the magnitude of shifts of the expected values of the explanatory variables is small, and is equal to the $0,5\sigma_X$, where σ_X is the standard deviation of an explanatory variable X. We have chosen such small shifts in order to show the effect of the worsening of the efficiency of the considered classifiers.

Table 6 Actual and reported values of the fraction nonconforming for a deteriorated process

Shift	Actual	RegBin	LDA-s	LDA-as	C4.5
No shift	0,254	0,162	0,248	0,181	0,257
A (up)	0,252	0,180	0,277	0,204	0,260
B (down)	0,253	0,117	0,192	0,134	0,255
C (down)	0,303	0,152	0,235	0,173	0,608
D (down)	0,270	0,155	0,238	0,176	0,219

The results presented in Table 6 show a really strange behavior of the considered classifier. When the expected value of A is shifted the actual fraction nonconforming remains practically the same (the observed differences may be explained as the effect of the randomness of the simulation process), but all classifiers show significant (except for the C4.5) deterioration of the process. When the expected value of B is shifted the situation with the actual fraction nonconforming is similar, but the classifiers behave in a completely different way. They show significant (except for the C4.5) improvement of the process (lower value of the fraction nonconforming). The situation becomes dramatically bad in the case of the shift of the expected value of C. In this case the actual fraction nonconforming increases significantly, but this is reported only by the classifier C4.5. All the remaining classifiers show the improvement of the actually deteriorated process! When the expected value of D is shifted the actual fraction nonconforming increases quite significantly, but this has not been reported only by any of the considered classifiers!

This strange behavior may be explained when we look at the definitions of classifiers presented in Tables 1 - 3. The classification rules for the RegBin, LDA-s, and LDA-as classifiers depend practically exclusively on the values of A and $B4$. On the other hand, the decision rules of C4.5 depend predominantly on the values of C (only in one case the decision rule does not depend on the value of C) and D.

Now, let us look at the case when shifts in the expected values of the explanatory variables result in the improving of the process. Similarly to the case discussed above assume that the magnitude of the shift is equal to $0,5\sigma_X$, but in this case the expected value of A decreases, and the expected values of B, C, and D increase. The actual and reported fraction nonconforming for such a case are presented in Table 7.

Table 7 Actual and reported values of the fraction nonconforming for a improved process

Shift	Actual	RegBin	LDA-s	LDA-as	C4.5
No shift	0,254	0,162	0,248	0,181	0,257
A (down)	0,252	0,151	0,232	0,170	0,252
B (up)	0,253	0,209	0,303	0,230	0,264
C (up)	0,234	0,179	0,278	0,197	0,151
D (up)	0,244	0,175	0,282	0,207	0,333

Just as in the previously considered case of a deteriorated process the behavior of the considered classifiers is often different than expected. Only in the case of the shift in A the classifiers behave as expected (they show the improvement). In contrast, in the case of the shift in D, and of the shift in B, the classifiers behave contrary to the expectation (they show the deterioration). In the case of the shift of the expected value of C only the C4.5 classifier behaves as expected.

On Figure 7 we show how the fraction of FP's changes for different classifiers when we use different training sets, and the process is deteriorated due to the negative shift of the expected value of the explanatory variable D. The impact of the same shift on the fraction of FN's is presented on Figure 8.

From Figure 7 and Figure 8 we can see that the classifier C4.5 is consistently better than its competitors when the fraction of FP's is considered. Its behavior is nearly as good if we consider the fraction of FN's (only RegBin is better).

The similar situation takes place when we consider the deterioration of the process due to a negative shift of the expected value of the explanatory variable C. On Figure 9 we show how the fraction of FP's changes in this for different classifiers and different training sets.

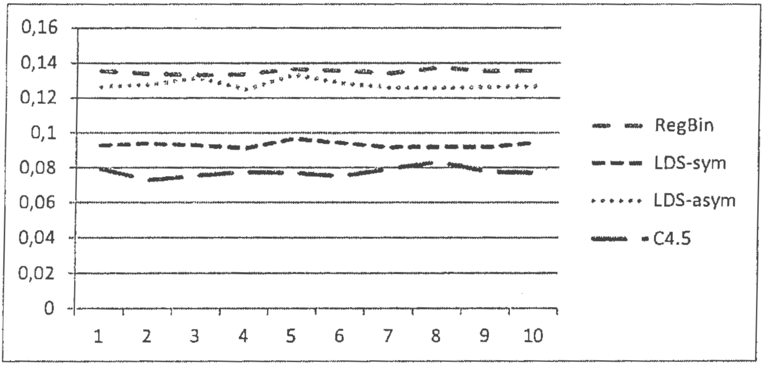

Fig. 7 Fraction of FP's for the negative shift of *D* - different sets of training data

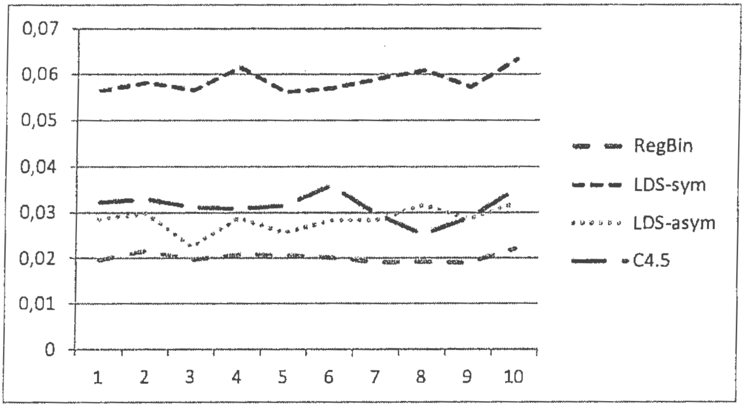

Fig. 8 Fraction of FN's for the negative shift of *D* - different sets of training data

The behavior of the classifier C4.5 in the case presented on Figure 9 is nearly perfect. For several sets of training data the classifier built on this training data shows ideal performance - no FP's have been observed. In the remaining cases the behavior of the C4.5 is significantly better than the behavior of its competitors. Unfortunately, the behavior of this classifier with respect to the fraction of FN's is dramatically worse, as it is seen on Figure 10. Due to a relatively small shift in *C* the C4.5 classifier begins to report the majority of all classified items as nonconforming.

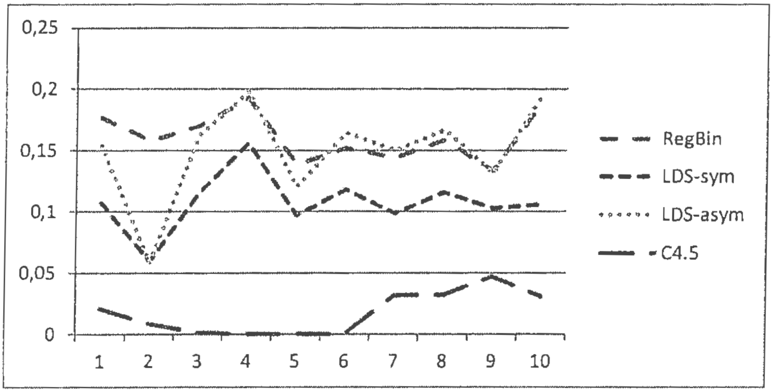

Fig. 9 Fraction of FP's for the negative shift of *C* - different sets of training data

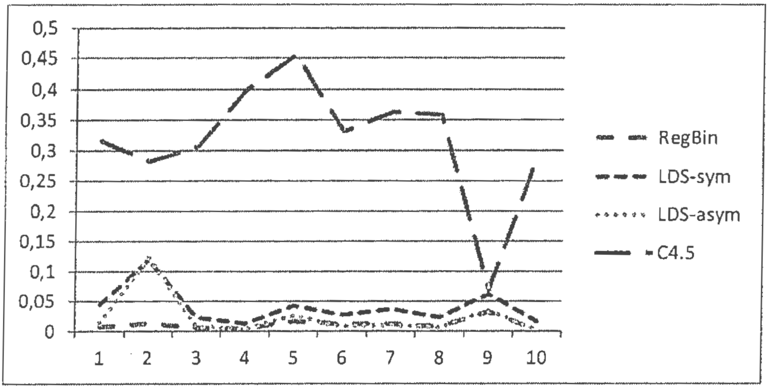

Fig. 10 Fraction of FN's for the negative shift of *C* - different sets of training data

The impact of this phenomenon on the performance of SPC procedures will be discussed in the next section of this paper.

Let us finish this section with a close summary. Neither of the considered classifiers is consistently better than its competitors. However, a certain ranking of these classifiers can be made. The binary regression RegBin classifier is definitely the worse, and this is not unexpected as the considered problem is strongly nonlinear. Also the LDA-as classifier built using Fisher's linear discriminant function with asymmetrically located decision point has rather disqualifying properties. The similar classifier, but with symmetrically located decision point (LDA-s), performs much better, but in some cases its behavior is really bad. The clear winner is Quinlan's C4.5 classifier which outperforms its competitors in the majority of different simulation experiment. Its only visible deficiency is the highly excessive rate of false negative classifications for certain patterns of process deterioration.

4 SPC procedures for monitoring the process with predicted values of quality characteristic

The ultimate goal of any SPC procedure is to keep the process at an acceptable level. Even if we observe all items in a process we can still use SPC procedures for monitoring the process quality. For example, we can divide the entire process into consecutive segments of n elements, and treat these segments as samples for charting purposes. Alternative approaches, such as using a sliding window for monitoring the process, are also possible. In our research we considered two approaches: division of the process into segments of $n = 100$ items considered as samples for charting the Shewhart p-chart, and using a mowing average chart (MAV) with a sliding window of $n = 100$ items.

For the construction of the Shewhart control p-chart one needs a good estimator of the fraction nonconforming p. By a a good estimator we understand a stable estimator characterized by low variability. The estimated value of the process fraction nonconforming p is obtained from the analysis of Phase I process data when the process is under control. In Table 8 we present the coefficients of variation of the estimates of p for actual and reported process levels, and different sample sizes n (i.e. durations of the Phase I).

Table 8 Coefficients of variation of the estimators of p

n	Actual	RegBin	LDA-s	LDA-as	C4.5
200	0,125	0,242	0,218	0,385	0,228
500	0,075	0,210	0,195	0,360	0,206
1000	0,054	0,191	0,187	0,348	0,193
2000	0,039	0,186	0,184	0,352	0,192
5000	0,024	0,181	0,180	0,356	0,193

From Table 8 we can see that stable estimates of p can be obtained from samples of $n = 1000$ items. For this sample size the variability of the reported values of p depends upon the variability induced by the variability of the training data sets used for building classifiers. However, if this sample size is not feasible we can use sample sizes of $n = 500$ items. In our simulation experiments for the construction of SPC procedures we use the sample size n equal to 1000, as this seems to be a good compromise between the stability and the need to built the procedure as quickly as possible.

For the evaluation of the performance of the considered SPC procedures we have used the following simulation experiments:

1. A set of classifiers built using the same training data is chosen randomly from the set of 10 possible options.
2. A Phase I sample of $n = 1000$ items is generated, and actual and reported fraction nonconforming p are estimated.
3. Classical Shewhart p-chart with 3-sigma limits is designed.
4. Consecutive segments of the process of the length of $m = 100$ items are generated.
5. For each of the production segments the estimated values of p are compared with the control lines calculated in Step 3).
6. Steps 4) - 5) are repeated until the out-of-control signal is observed, and a respective run length is determined.
7. Steps 1) - 6) are repeated 100 times, and the respective average rung length's (ARL's) are calculated.

Such experiments in certain cases were repeated several times, so the total number of simulation runs varies from 100 to 500. This limited number of simulation runs does not allow us to evaluate precise values of ARL's, so their values presented in the following tables should be considered as only approximate.

The most important characteristic of any SPC procedure is its Average Run Length (*ARL*). When an SPC procedure of a Shewhart control chart type is used for monitoring 100% inspected process it is the average number of segments inspected till the moment of an alarm signal. In nearly all practical cases when fraction nonconforming is monitored only charts with upper control lines are used. In our case, however, two-sided control chars are needed, because - as it has been shown in Section 3.4 - actual deterioration of a process may lead, for certain classifiers, to the decrease of the reported fraction nonconforming. In Table 9 we present the values of this characteristic for the process under control (No shift), and processes with shifted expected values of the explanatory variables. In this table we consider only such shifts that lead to the worsening of process levels.

Table 9 Values of the ARL for the Shewhart control chart

Shift	Actual	RegBin	LDA-s	LDA-as	C4.5
No shift	304,6	320,4	323,7	329,7	276,3
A (up)	267,8	297,3	202,9	259,1	237,6
B (down)	298,1	121,9	71,6	142,6	307,9
C (down)	32,2	369,8	280,2	313,2	19,3
D (down)	157,3	337,1	229,8	321,5	168,3

The results presented in Table 9 show undoubtedly, that only the chart based on the values predicted by the C4.5 classifier gives values of the ARL similar to those obtained as if the actual values were observed. The remaining classifiers give acceptable results only for the case of the upward shift in A. In the case of other deteriorations they simply do not signal the worsening of the process quality.

When a process is continuously monitored one can think about a control procedure using the moving average (MAV) approach. Fraction nonconforming is calculated for a sample of the last m, and is moved when a new measurement is available. In contrast to the previously considered monitoring of consecutive segments of a process, the values of the fraction nonconforming evaluated from consecutive samples are highly correlated (with the correlation coefficient equal to $1-1/m$). Therefore, the standard deviation of the monitored statistic cannot be calculated from a well known formula valid for the binomial distribution. We propose to estimate it from the total number of n Phase I observation, i.e. from $n-m+1$ MAV samples. In Table 10 we present the averaged results of this estimation for $n=1000$ and $m=100$ when different classifiers have been used for classification purposes.

Table 10 Standard deviations of the MAV values of p - averaged over different sets of training data

Set	Actual	RegBin	LDA-s	LDA-as	C4.5
Avg σ_p	0,041	0,034	0,040	0,035	0,041

The ARL values for the two-sided MAV control are presented in Table 11.

Table 11 Values of the ARL for the MAV control chart

Shift	Actual	RegBin	LDA-s	LDA-as	C4.5
No shift	13525	7963	6921	10093	10431
A (up)	9735	8093	5738	7783	6992
B (down)	7410	2469	1967	24742	8232
C (down)	825	38994	25326	19892	614
D (down)	3206	65952	6845	87288	5536

The analysis of the results presented in Table 11 confirms the finding obtained in the case of the procedure based on the Shewhart control chart. Only the chart based on the predictions of the C4.5 classifier gives satisfactory results. Interesting is the comparison of these results with those presented in Table 9 when we compare the numbers of items inspected till the alarm signal. In order to do this comparison we have to multiply the cells of the Table 9 by 100. The MAV procedure triggers false alarms (when process is under control) more frequently, but on the other hand the average time to alarm signal (measured in the number of inspected items) is for this procedure visibly smaller when the best classifier (C4.5) is applied. Additional investigation of the MAV control lines is thus needed if we want to decrease the rate of false alarms when the process is under control.

In all comparisons made in this section we used the ARL characteristic for the comparison of different SPC procedures. This value may be very misleading if the probability distribution of the values of p is skewed and has a large variance. Unfortunately, the analysis of data presented in Table 4 shows that the estimates of p made using considered classifiers are highly variable and skewed. Therefore, the probability distribution of run length's is also highly skewed. Therefore, the ARL values obtained from a very limited number of simulation runs may be inaccurate. We have to take this into account if we want to draw from the presented results conclusions of the quantitative character. For example, for the comparison of different procedures we may use the median of the run length. In Table 12 we compare the values of the ARL with the corresponding values of the median.

Table 12 Values of the ARL and the median of RL for the MAV control chart

Shift	Actual	RegBin	LDA-s	LDA-as	C4.5
ARL	13525	7963	6921	10093	10431
Median	2110	1410	2117	1539	1302

This comparison shows the comparison of the ARL's does not lead to the same conclusions as the analysis of the the medians. This problem needs definitely further investigation.

5 Conclusions

The results presented in this paper extend the results presented in Hryniewicz (2013). They show that in the case of non-normal distributions of characteristics of interest, and non-linear dependencies between observable (explanatory) and not directly observable (only predicted!) values of processes the properties of control charts designed using the standard methodology may be not satisfactory. Several popular classifiers used for prediction purposes have been investigated, and only Quinlan's C4.5 decision tree classifier have shown acceptable average performance. Moreover, their performance is difficult to predict in advance, as it has been already shown in Hryniewicz (2013). Further research is needed with the aim to analyze the impact of the size of training sets, and the size of the Phase 1 samples, on the characteristics of control charts. Additional research on the possible application of more complicated classifiers is also needed. The results presented in this paper show that the application of modern data mining techniques for SPC purposes, which is strongly advocated by some specialists, is promising but, as for now, the obtained results are far from beeing satisfactory from a practical point of view.

References

Hastie T, Tibshirani R, Friedman J (2008) *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* 2nd edn. Springer, New York.

Hryniewicz O (2013). Statistical evaluation of the results of inspections based on the data mining methodology. Submitted to *Frontiers in Statistical Quality Control 11.*

Montgomery DC (2011). *Introduction To Statistical Quality Control.* 6th edn. Wiley, New York.

Murtagh F (1986) *Multivariate Analysis.* Kluwer, Dordrecht.

Nelsen RB (2006). *An Introduction To Copulas.* 2nd edn. Springer, New York.

Noorsana R, Saghaei A, Amiri A (2011). *Statistical Analysis of Profile Monitoring.* Wiley, Hoboken NJ.

Owen DN, Su YH (1977). Screening based on Normal Variables. *Technometrics*, 19, 65-68.

Quinlan JR (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann, Los Altos, CA.

Witten IH, Frank E, Hall MA (2011). *Data Mining. Practical Machine Learning Tools and Techniques.* 3rd edn. Elsevier, Amsterdam.

Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004). Using Control Charts to Monitor Process and Product Profiles. *J Qual Techn*, 36, 309-320.

Wang YT, Huwang L (2012). On the Monitoring of Simple Linear Berkson Profiles. *Qual Rel Engin Int*, 28, 949-965.

Xu L, Wang S, Peng Y et al. (2012). The Monitoring of Linear Profiles with a GLR Control Chart. *J Qual Techn*, 44, 348-362.