DOI: 10.2507/33rd.daaam.proceedings.066

SELECTION OF AN APPROPRIATE PRIOR DISTRIBUTION IN RISK ASSESSMENT

Dubravka Božić & Biserka Runje

This Publication has to be referred as: Bozic, D[ubravka] & Runje, B[iserka] (2022). Selection of an Appropriate Prior Distribution in Risk Assessment, Proceedings of the 33rd DAAAM International Symposium, pp.0471-0479, B. Katalinic (Ed.), Published by DAAAM International, ISBN 978-3-902734-36-5, ISSN 1726-9679, Vienna, Austria
DOI: 10.2507/33rd.daaam.proceedings.066

Abstract

The Bayesian approach, which combines prior information about the quantity to be measured, available before the measurement, and additional information obtained from the measurement, is used in risk assessment in metrology. According to the binary decision rule in risk assessment, there are four outputs: the number of accepted and rejected measurements and the number of falsely accepted and falsely rejected measurements. A falsely rejected measurement represents the producer's risk, while a falsely accepted measurement represents the consumer's risk. These four cases in risk assessment lead us to confusion matrix. In this paper, we evaluate the most suitable prior distribution for modelling the risk for roundness deviation of the inner ring of the bearing. This quantity is always positive; therefore, the choice of prior is limited to those distributions that take only the positive value of the argument. The assessment of the most appropriate distribution was performed by measures derived from confusion matrix and ROC - AUC analysis.

Keywords: consumer's risk; producer's risk; binary decision rule; confusion matrix; ROC curve

1. Introduction

There is no perfect measurement. Different sources of variability affect measurement results and measurement uncertainty. The measured values for the item of interest may or may not be within the interval of allowed values provided by the specifications for the specific product. The verification process of accepting or rejecting inappropriate measurement is the so-called conformity assessment rule. In this decision-making process, wrong decisions may occur that must be considered. There is a producer's risk of rejecting of conforms measurement and consumer's risk of accepting of non-conforms measurement [1]. The reference document [2], adopted by the Joint Committee for Guides in Metrology prescribes the procedure for calculating the producer's risk and the consumer's risk. When calculating the risk, a Bayesian approach is used. This approach combines information obtained by measurement and assumptions about the distribution of parameters that describe the measured data. Measurement data are usually assumed to belong to a normal distribution and are modelled by using a likelihood function. The risk assessment depends on the a priori distribution of the parameters. Different numbers of accepted, rejected, falsely accepted (consumer's risk) and falsely rejected products (producer's risk) are obtained for different priors. Based on these four outputs, it is possible to form a confusion matrix, a well-known term from machine learning. The aim of this paper is to use metrics associated to confusion matrix to estimate the most suitable prior distribution for risk assessment. The prior distribution selection was carried out for a real study case of risk modelling for roundness deviation of the inner ring of the bearing.

2. Background of risk calculation

The Bayesian approach to risk assessment in metrology combines two sources of information. One type of information about the measured variable Y comes from the a priori experience of the measurer himself. The item of interest, measured quantity Y, can take the values denoted by η, and is treated as a random variable with a probability distribution function (PDF), denoted by $g_0(\eta)$. Two parameters are associated to the measured value Y: best estimate $\bar{y} = \mu$, and standard uncertainty $u_0 = s$. This information can be evaluated before measurement performing. Another source of information are the measurement results given by the random variable Y_m. According to [3], the standard measurement uncertainty u_m, and the value of the measured quantity η_m, are associated to the measurement results. Measurement data are modelled via likelihood function $h(\eta_m|\eta)$, by normal PDF. This function, for given η_m, depends on η, and it can be calculated from:

$$h(\eta_m|\eta)=\frac{1}{u_m\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\eta_m-\eta}{u_m}\right)^2\right] \tag{1}$$

As a result of Bayes' rule, the posterior distribution is written as

$$g(\eta|\eta_m) = C\,g_0(\eta)h(\eta_m|\eta), \tag{2}$$

where C is the normalization constant chosen such that $\int_{-\infty}^{\infty} g(\eta|\,\eta_m)d\eta = 1$.

There are two types of risk: specific and global risk. The specific producer's risk is related with conformance probability p_c, the probability that the item of interest is within the interval of permissible value, i.e., within the tolerance interval $[T_L, T_U]$. Symbols T_L and T_U are for the lower and upper limit of the tolerance interval, respectively. Conformance probability is given by

$$p_c = \int_{T_L}^{T_U} g(\eta|\eta_m)d\eta \tag{3}$$

Probability $\bar{p_c}$ that an item of interest is non-confirmed can be calculated as

$$\bar{p_c} = 1 - p_c \tag{4}$$

Another important interval is the acceptance interval $[A_L, A_U]$, that is, the interval of permissible values for the measured quantity. Symbols A_L and A_U stands for the lower and upper limit of the acceptance interval, respectively. Depending on the problem under consideration, the tolerance interval and the acceptance interval can be in several different relationships [4]. The acceptance interval may be entirely within the tolerance interval, and separated from it, from both sides, by a guard band of width w. On this way, the consumer's risk is minimized. If the tolerance interval is within the acceptance interval, the producer's risk is minimized. By placing a guard band between the acceptance interval and the tolerance interval, the probability of making a wrong decision is reduced. From the natural requirements set for the measured quantity, it is possible to define one side tolerance interval when the measured values are limited from below with lower tolerance limit T_L, or from above with upper tolerance limit T_U (Figure 1).

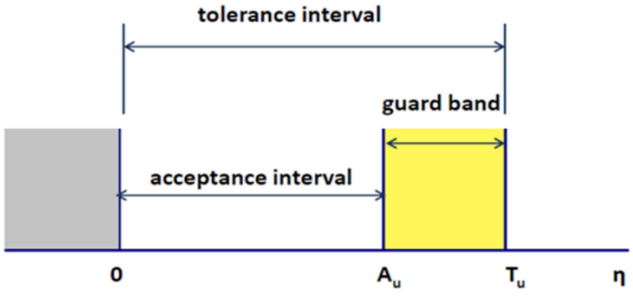

Fig. 1. One-sided tolerance interval bounded with upper limit [4]

The situation shown in Figure 1 corresponds to the case of risk assessment for roundness deviation of the inner ring of the bearing. For the given upper limit of the tolerance interval T_U, the upper limit of the acceptance interval A_U is determined. The width of guard band is equal to

$$w = T_U - A_U = 2ru_m, \tag{5}$$

where the multiplier r is in the range from -1 to 1.

3. Global consumer's and producer's risk

If the true value Y of the item of interest is outside the tolerance interval, and the measured value Y_m is within the acceptance interval, the global consumer's risk is given by

$$R_C = \int_{-\infty}^{T_L}\int_{A_L}^{A_U} g_0(\eta)h(\eta_m|\eta)d\eta_m d\eta + \int_{T_U}^{\infty}\int_{A_L}^{A_U} g_0(\eta)h(\eta_m|\eta)d\eta_m d\eta \quad (6)$$

If the true value Y of item of the interest is within the tolerance interval, and the measured value Y_m is outside the acceptance interval, producer's risk is given by

$$R_P = \int_{-\infty}^{A_L}\int_{T_L}^{T_U} g_0(\eta)h(\eta_m|\eta)d\eta_m d\eta + \int_{A_U}^{\infty}\int_{T_L}^{T_U} g_0(\eta)h(\eta_m|\eta)d\eta_m d\eta \quad (7)$$

Consumer's risk is associated with false accepted or false positive (FP) measurements while producer's risk is associated with false rejected or false negative (FN) measurements. For labelling are used the notation FP and FN, the standard notation in machine learning. The result of the consumer's and producer's risk calculation is most often expressed in percentages. The number of false positive measurements or products could be found by multiplying the R_C value by 100, 1000 or 10000, depending on how many products have to be checked. In this paper, the calculation was performed for 10000 products. In this case, the number of false positive products is equal to $FP = 10000 \cdot R_C$. The number of falsely rejected products is counted as $FN = 10000 \cdot R_p$. The total number of conformed products is equal to $10000 \cdot p_c$. From here, the number of accepted, or true positive (TP) products is equal to $TP = 10000 \cdot p_c - FN$, while the number of rejected, i.e., true negative products (TN) is equal to $TN = 10000 \cdot \bar{p_c} - FP$. According to this binary decision rule, four values that form the confusion matrix are obtained (Figure 2).

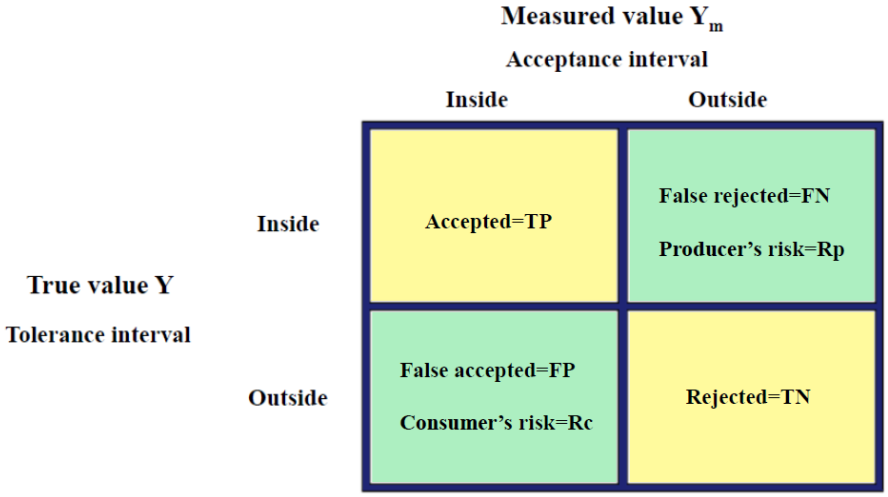

Fig. 2. Confusion matrix for risks calculation

4. Basic assumptions

The roundness deviation of the inner ring of the bearing, with a diameter of $D = 80$ mm, was measured. According to the specifications for these rings, the tolerance for roundness deviation is in the range from $0 - 25$ μm. The determined deviation from roundness amounts to $\bar{y} = 16$ μm, with a standard deviation equal to $u_0 = 5$ μm. Measurement of the deviation from roundness was carried out by an automated system for contact dimensional measurement of the bearing ring. The system enables 100 % control, storage and analysis of data, and separation of bad bearing rings from good ones. By monitoring the mean values and ranges with a control chart, it was determined that the process is stable (under control). The standard uncertainty of measurement results for roundness deviation is $u_m = 1$ μm.

For the given data, nine prior distributions were chosen: four non-parametric distributions, two one-parameter distributions and three two-parameter distributions. The parameters of the selected distributions were determined by using the arithmetic mean of the roundness deviation $\bar{y}$ and the standard deviation u_0. These two quantities are ignored with non-parametric distributions. With one-parameter distributions, the standard deviation is ignored. The estimation of parameters, for one-parameter and two-parameter distributions is determined according to the formulas that can be found in [5]. Parameters calculated for the selected distributions are shown in Table 1

Prior	Parameters	Arguments	p_c	IR_T	AUC
Uniform, U[0, 26.3213]	−	$\eta \geq 0$	0.9498	0.0529	0.9847
Uniform, U[0, 50]	−	$\eta \geq 0$	0.50	1	0.9988
Cauchy	−	$\eta \geq 0$	0.9873	0.0129	0.9999
Gibrat	−	$\eta > 0$	0.9994	0.0006	1
Rayleigh, R(α)	$\alpha = 12.8$	$\eta \geq 0$	0.8520	0.0174	0.9972
Maxwell, M(α)	$\alpha = 10.025$	$\eta \geq 0$	0.8986	0.1128	0.9963
Gamma, Γ(α, λ)	$\alpha = 10.24\,, \lambda = 0.64$	$\eta > 0$	0.9498	0.0529	0.9964
Beta, B(α, β)	$\alpha = 10.24, \beta = 639978.52$	$\eta > 0$	0.9498	0.0529	0.9964
Truncated normal, TN(μ, σ)	$\mu = 16.0119\,, \sigma = 4.9809$	$\eta \geq 0$	0.9643	0.0370	0.9949

Table 1. Prior information

Given that the risk of roundness deviation is assessing, chosen are distributions that take positive values of the argument, and that allows a small amount of measured values equal to zero so that $\eta \geq 0$. Also are selected prior distributions where the probability that the measured values are equal to zero is negligible, and $\eta > 0$.

Next step in risk assessment is the equidistant subdivision of the $[-1, 1]$ interval, which is the domain of the multiplicative factor *r*. The subdivision with the threshold value equal to 0.1 gives the 21 nodes. For each node, the upper limit of the acceptance interval is found out. According to the (5), the step of changing the value for the upper limit of the acceptance interval is equal to 0.2 μm. The method allows estimation of the limit of the acceptance interval, which is lower than the upper limit of the tolerance interval, as well as the determination of the limit of the acceptance interval, which is higher than the upper limit of the tolerance interval. In this case, the range for the acceptance limit goes from 23 μm, in the case when $r = 1$, to 27 μm, for $r = -1$. Considering the given tolerance interval for risk calculation of roundness deviation, allowed values are in the range from 23 μm to 25 μm. All data required for risk calculation according to the (6) and (7) were determined on this way. For the selected prior distribution, and for each of the 21st subdivision nodes, one confusion matrix is generated.

5. Comparison with machine learning

The creating a confusion matrix in risk assessment does not need a large amount of data divided into a training set and a test set, as in machine learning. Only four data are using for the creating confusion matrix: best estimate $\bar{y}$, and standard uncertainty u_0 for the item of interest Y (which can be set up before measurement according to our prior beliefs), the standard measurement uncertainty u_m, associated to measurement data, and upper tolerance limit T_U. In comparison with machine learning, can be said that in risk assessment, these four data are formed a training set. Based on these four data and given threshold of the upper limit of the acceptance interval, by using (6) and (7), confusion matrixes are generated. In this paper, each confusion matrix contains 10000 data. In risk assessment, confusion matrix gives a comparison between true and measured values. Equation (6) and (7) for the consumer's and the producer's risk estimating are classifiers used to orders data into classes. And these are also the rules that enable the prediction of the number of TP, TN, FP and FN products for a given size of the output data set. A confusion matrix of type 2×2 has two classes. In risk assessment, we can talk about the classes "inside" and "outside" of a specific interval. The difference compared to machine learning is that when assessing risk, we distinguish between the intervals to which true and measured values belong. We are talking about measurements that are "inside tolerance interval" and "outside tolerance interval" for true values and "inside acceptance interval" and "outside acceptance interval" for measured values. Several different algorithms are used for binary classification of data in machine learning [6]. Some of the most famous are: Naive Bayes algorithm, logistic regression, k-nearest neighbours, support vector machine, decision tree and neural network. Equation (6) and (7) for ordering into classes, are most similar in origin to the Gaussian Naive Bayes classifier. Both, machine learning and risk assessment combine prior distribution, and the likelihood function for the normal distribution.

Depending on the prior distribution, confusion matrix in risk assessment can contain balanced or imbalanced data. In machine learning, imbalanced data represent a problem. Almost all standard machine learning algorithms give a good and reasonable result only for the balanced data [7]. In metrology, imbalanced data are not only allowed, but they are also desirable. With a well-performed measurement, there is always a disproportion between the number of accepted and the number of rejected products. In the production process, it is required that the number of accepted products that meet the specifications greatly exceeds the number of products that do not meet the required standards. We will define several different ways in which we can verify that data are imbalanced. In metrology, we are interested in conformed measurement or products. In this paper, we take the "inside tolerance interval" class for the majority class, and the "outside tolerance interval" class for the minority class. In that case, the size of the majority class is the same as the number of confirmed products, and the size of the minority class is equivalent to the number of non-conformed products. We define imbalanced ratio, or skew, as the ratio of the number of products in the minority class to the number of products belonging to the majority class [8]. The ratio refers to data that is inside or outside the tolerance interval and is denoted with IR_T. This ratio provides data on the balance of data in risk assessment. If the IR_T is equal to 1, data are balanced, and if the IR_T is close to zero then data are imbalanced. The results for IR_T, for selected priors, are presented in Table 1.

The IR_T determined in this way is unique for all confusion matrix generated for different thresholds of the upper limit of the acceptance interval. Data balance can also be checked by determining the ratio of the number of products that are "outside the acceptance interval" to the number of products or measurements that are "within the acceptance interval". This ratio is denoted with IR_A. The ratio is not unique and changes with the change in the value of the upper limit of the acceptance interval (Figure 3a). The third way of checking data balance is possible by forming the TN/TP ratio. This ratio is also not unique and its value changes with a change in the value of the upper limit of the acceptance interval (Figure 3b). For imbalanced data, the ratios IR_A and TN/TP take on small values close to zero, and for balanced data they can be greater than 1. In metrology, as a result of the risk assessment, we want to have as many true positive products and as few true negative products as possible. Which means that we are always interested in imbalanced data. On the other hand, a perfect separation of products into those that are TP and those that are TN in metrology is not allowed, in the sense that all measurements are classified into these two groups. This means that it is not allowed that the values of FP and FN are both at the same time equal to zero. It is always assumed that there is a certain measurement uncertainty and a certain number of unacceptable measurements, whether they are of the FP or FN type.

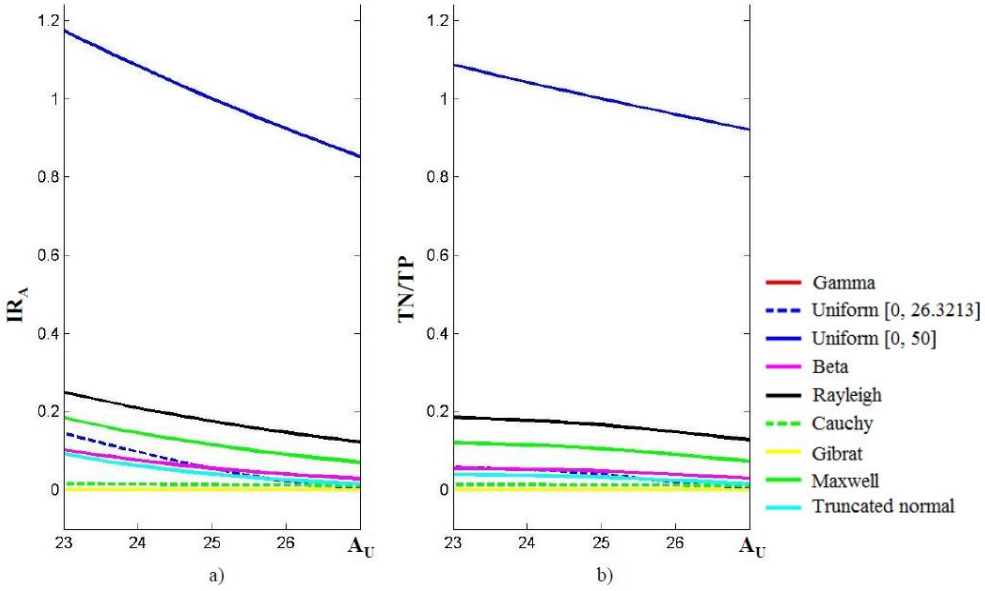

Fig. 3. Imbalanced ratio, a) IR_A, b) TN/TP

6. Results and analysis

The values of TP, TN, FP, and FN in the generated confusion matrix are different for different values of the upper limit of the acceptance interval. These matrices serve for the selection of the appropriate prior distribution. The assessment of the appropriate prior distribution was performed using metrics associated with confusion matrix. Three characteristic metrics: accuracy, precision and recall are calculated as follows:

$$\text{accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FN} + \text{FP} + \text{TN}) \tag{8}$$

$$\text{precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{9}$$

$$\text{recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{10}$$

Accuracy is metrics that tell how many rings are correctly classified into TP and TN categories in relation to the total amount of data. This metric is extremely sensitive to imbalanced data [9]. In the case of imbalanced data desirable in metrology, the number of data classified in the TN category is negligibly small, and it is valid that TP≫TN. According to (8), for such a value relationship between TP and TN, the accuracy actually provides a prediction of the number of rings classified in the TP category in relation to the total amount of data, Figure 4a. This is true for all chosen priors except for the uniform distribution defined on the interval [0, 50]. According to Table 1, Figure 3a and Figure 3b, the uniform distribution U[0, 50] is the only one of the selected distributions that has balanced data. The IR_T value for this distribution is the highest and is equivalent to 1. Compared to all other distributions, this distribution takes the highest values for the ratios IR_A and TN/TP over the entire range for the upper limit of the acceptance interval. According to Figure 4a, the accuracy for U[0, 50] is generally higher than the accuracy of the one-parameter distributions: Rayleigh and Maxwell distributions, but it is lower than the accuracy of the two-parameter distributions: truncated normal, gamma and beta.

Although the accuracy metric has a high value for U[0, 50], according to the requirements set in metrology, we cannot say that this prior is good for risk assessment in the case of roundness deviation of the inner bearing ring. In each of the figures presented in this paper, the graphs for the gamma and beta distributions completely coincide on the interval [24, 26]. On the interval [23, 24], the gamma distribution has higher values for R_c, that is, a larger number of FP rings compared to the beta distribution, and therefore a smaller number of TP rings. On the interval of [26, 27] gamma compared to beta distribution has smaller values for R_P, that is, the number of FR rings. Given that this difference is negligible, of the order of 10^{-4}, we can consider that the gamma and beta distributions have the same behaviour over the entire range for A_U.

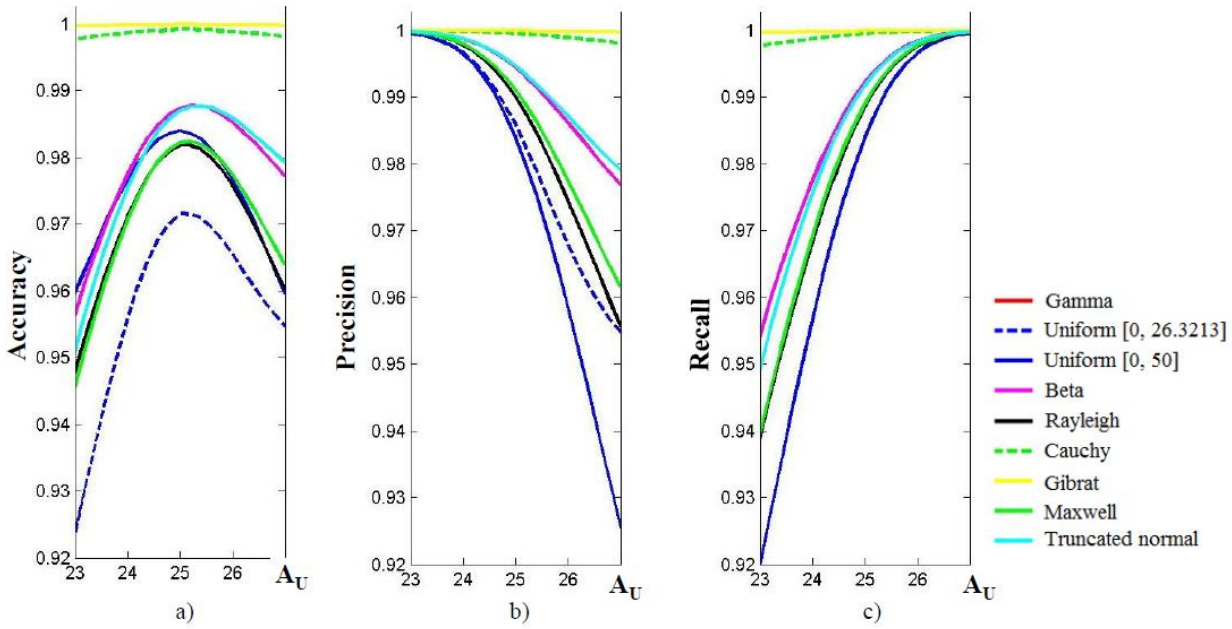

Fig. 4. a) Accuracy, b) Precision, c) Recall

For all distributions, accuracy increases for $A_U \in [23, 25]$ and decreases for $A_U \in [25, 27]$. The Rayleigh's distribution on the interval [23, 24.5] has a slightly higher accuracy than the Maxwell's distribution. Of the two-parameter distributions, beta and gamma distributions have higher accuracy than the truncated normal distribution. The Cauchy's and Gibart's distribution, among the all chosen non-parametric distributions, achieve the highest values for all three metrics, Figure 4. The smallest values for accuracy over the entire range for upper acceptance limit are achieved by the uniform distribution U[0, 26.3213]. This distribution is artificially generated so that the conformance probability, on the interval [0, 25], for this distribution, is equal to 0.9498 and has the same value as the conformance probability for the gamma and beta distributions.

The precision measure, according to (9), gives a prediction of the number of rings classified in the TP category in relation to the total number of rings from the "inside acceptance interval" class. As the upper limit of the acceptance interval increases and the width of the guard band decreases, the number of rings classified in the TP category increases, but the number FP in the denominator of expression (9) also increases, therefore the precision decreases for all distributions on the interval [23, 27]. Given that data for U[0, 50] are balanced, the precision for this distribution is the lowest, Figure 4b. Rayleigh's distribution achieves slightly lower values compared to Maxwell's. Truncated normal distribution achieves slightly higher values for precision, compared to beta and gamma distribution.

The recall measure, according to (10), gives a prediction of the number of rings classified in the TP category in relation to the total number of conformed rings. As the upper limit of the acceptance interval increases, the number of TP increases, but a bearing ring number classified in the FN category decreases, therefore the recall for all priori distribution increases on the interval [23, 27]. According to the recall measure, both uniform distribution U[0, 50] and U[0, 26.3213] show almost the same behaviour. Their graphs overlap, unlike the graphs for accuracy and precision. The difference between the precision metrics values, taken by these two distributions, is negligible, and amounts to 10^{-4}.

The true negative rate (TNR) metric provides a prediction of the number of rings classified as TN in relation to the total number of non-conformed rings and can be calculated as:

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{FP} + \mathrm{TN}) \tag{11}$$

Given that the number of rings classified in the TN category decreases as the upper limit of the acceptance interval increases, and the number of rings from the FP category increases, the TNR decreases on the interval [23, 27], for all distributions, Figure 5a.

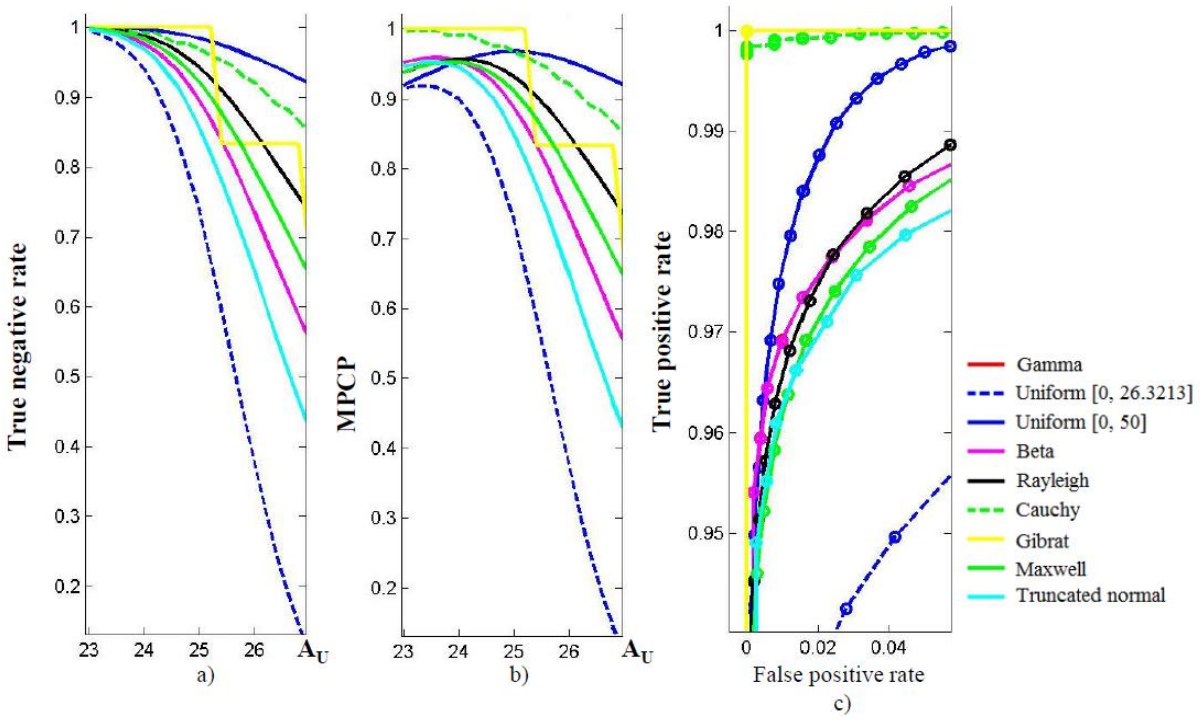

Fig. 5. a) True negative rate, b) Maximum probability of correct prediction, c) ROC curves

On the interval [23, 25], the Gibrat's distribution has the highest values for TNR, compared to all other distributions. The TNR for Gibrat's distribution decreases "stepwise" on the interval [25, 27], which, according to the assumptions, is not an interval of allowed values. Among all selected distributions, according to all selected metrics, the Gibrat's distribution is the distribution that best separates rings into categories TP and TN, with a few rings in categories FP and FN. Unlike other distributions, the number of FP and FN rings, for the Gibrat distribution, is either 0 or 1 or 2. That is why the graph of this distribution is stepwise, Figures 5a, 5b, 6b, 6c. A similar stepwise shape, but to a lesser extent, have the Cauchy's distribution, Figures 5a and 5b.

In order to determine the maximum probability of correct prediction (MPCP) the false negative rate (FNR) and the false positive rate (FPR) are defined as following:

$$\mathrm{FNR} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN}) \tag{12}$$

$$\mathrm{FPR} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN}) \tag{13}$$

The false negative rate gives the number of rings classified in the FN category in relation to the total number of conformed rings. The false positive rate gives the number of rings classified in the FP category in relation to the total number of non-conformed rings. Maximum probability of correct prediction is a measure associated with the confusion matrix that tells us what the maximum probability is that we did not choose the FP and FN ring. It is defined as:

$$\mathrm{MPCP} = (1 - \mathrm{FPR})(1 - \mathrm{TNR}) \tag{14}$$

This measure allows us to determine the maximum upper limit of the acceptance interval for each individual distribution, Figure 5b. The results are as follows: According to the MPCP, the upper limit of the acceptance interval for U[0, 26.3213] is $A_U = 23.4$, the associated risks are $R_C = 0.0864$ %, and $R_P = 6.17$ %. For the uniform distribution, U[0, 50] holds $A_U = 25$, $R_C = 0.8$ %, $R_P = 0.8$ %. It is about balanced data, so this result is expected. For the Cauchy distribution holds $A_U = 23.6$, $R_C = 0.0037$ %, $R_P = 0.16$ %. Gibrat distribution: $A_U = 25$, $R_C = 0.0033$ %, $R_P = 0.0040$ %. In this case, MPCP is equal to 1. Rayleigh distribution: $A_U = 24$, $R_C = 0.18$ %, $R_P = 2.71$ %, Maxwell's distribution: $A_U = 24$, $R_C = 0.17$ %, $R_P = 2.77$ %. The Rayleigh and Maxwell distributions are both one-parameter distributions that reach the maximum value for MPCP at $A_U = 24$, but for that value, the consumer risk for Maxwell distribution is lower compared to Rayleigh distribution. For the gamma distribution holds $A_U = 23.6$, $R_C = 0.0522$ %, $R_P = 2.93$ %, Beta distribution: $A_U = 23.6$, $R_C = 0.0523$ %, $R_P = 2.94$ %, and truncated normal distribution: $A_U = 23.4$, $R_C = 0.0331$ %, $R_P = 3.77$ %.

One of the methods by which we can evaluate the behaviour of the prior is Receiver Operating Characteristic (ROC) analysis, Figure 5b. Area under ROC Curve (AUC) is a measure which tells how TPR changes with increasing of FPR. This measure also isn't sensitive to imbalanced data.

If the ROC curve is closer to the upper left corner, the AUC value is higher. If the AUC is equal to 0.5, it represents chance, while the value 1 corresponds to perfect classification. Values below 0.5 are not considered [10]. For the selected priors, the AUC values are shown in Table 1. The AUC number for the Gibrat distribution is equal to 1. It is followed by the Cauchy distribution with an AUC number equal to 0.9999. The uniform distribution U[0.50] with balanced data has a surprisingly large value for the AUC number, equivalent to 0.9988. As said before, this distribution is not suitable for risk assessment. The behaviour of priors is also analyzed for three other known metrics associated with confusion matrix: F1 score, kappa statistics and Matthew's correlation coefficient (MCC), Figure 6.

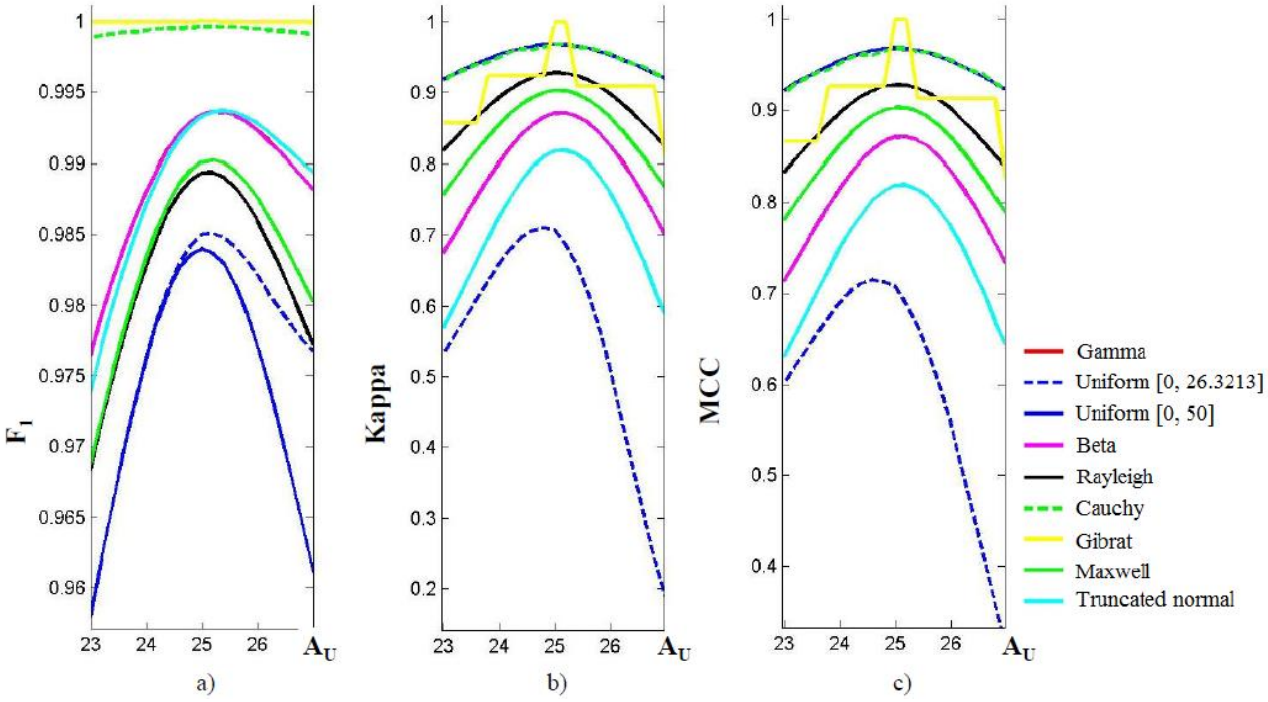

Fig. 6. a) F1 score, b) Kappa, c) Matthew's correlation coefficient

These metrics are calculated as follows:

$$\mathrm{F1} = 2 * (\mathrm{precision} * \mathrm{recall})/(\mathrm{precision} + \mathrm{recall}) \tag{15}$$

$$\mathrm{kappa} = 2 * (\mathrm{TP} * \mathrm{TN} - \mathrm{FN} * \mathrm{FP})/((\mathrm{TP} + \mathrm{FP}) * (\mathrm{FP} + \mathrm{TN}) + (TP + FN) * (\mathrm{FN} + \mathrm{TN})) \tag{16}$$

$$\mathrm{MCC} = ((\mathrm{TP} * \mathrm{TN}) - (\mathrm{FP} * \mathrm{FN}))/\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)} \tag{17}$$

The F1 score is the harmonic mean of the precision and recall metrics. It is a metric that is suitable for unbalanced data such as those encountered in metrology, in the case when the number of rings of the bearing classified in the TP category is much higher compared to the number of rings classified in the TN category. The values for the F1 score are in the range from 0 to 1. If the F1 score is equal to one, we are talking about the perfect separation of each ring into the appropriate class. If the F1 value is equal to 0, the model poorly classifies the measurements (rings) into the appropriate classes. According to the F1 score metric, Gibrat's and Cauchy's distributions best classify the rings into the corresponding classes, Figure 6a. They are followed by two-parameter distributions, and then followed one-parameter distributions. In the end are, as were expected, uniform distributions as poor classifiers.

The kappa statistic is a measure of agreement between predicted and observed data. Both measures: Kappa statistic and MCC take values in the range from -1 to 1. All values of those measures greater than 0.8 are considered to represent strong agreement [11]. The priors with values lower than 0.8 are U[0, 26. 3213], and partially truncated normal distribution, Figures 6a, 6b. The kappa statistics and MCC gives the same arrangement of priors for the ring of bearing classification, only the values of the kappa statistic are slightly smaller compared to the MCC measure. The balanced data associated with U[0, 50] distribution have higher values of the MCC and kappa statistic. The uniform distribution U[0, 50] and the Cauchy's distribution assume almost the same values. The maximum absolute difference between the values taken by these two distributions is of the order 10^{-3} in the favour of the U[0, 50] distribution.

7. Conclusion

Among all selected priors and according to all selected metrics associated with confusion matrix, the Gibrat distribution is imposed as the best choice for prior distribution in risk assessment. This prior is also the best choice among all those priors where it is implied that the argument of the distribution is $\eta > 0$. If a small measurement value equal to zero is allowed, that is, $\eta \geq 0$ is valid, then the Cauchy distribution is the best choice for the prior distribution. The Gibrat's and Cauchy's distribution are also the best choice among all the tested non-parametric priors. In metrology, we choose a non-parametric uniform prior if we do not want to influence the results of the risk assessment with our personal belief or if we believe that each of the measured values is equally likely, and we do not want to favour any one value over another. According to the performed analysis, the uniform distribution U[0, 50] is not a recommended distribution for the risk assessment because it generates balanced data in the confusion matrix. The uniform distribution U[0, 26. 3213] is also not appropriate prior for risk estimation. On all tested metrics, this prior is almost always low-graded. The AUC number for this prior is the lowest one compared to all other priors. One-parameter distributions for risk assessment should only be chosen when data for the standard uncertainty u_0 are not available, and there is no other choice. For all analyzed metrics, the results for the Rayleigh and Maxwell distributions are in between of the results for the chosen two-parametric and one-parametric distributions. If all the necessary data are available, as the best estimate $\bar{y}$ and standard uncertainty u_0, and if it is necessary to include them in the risk assessment, then two-parameter distributions are the best choice. The gamma or beta distribution are appropriate for risk assessment if the argument values are strictly positive values, and the truncated normal distribution is appropriate if the proportion of measured values equal to zero is significant.

It remains to be seen how changing the basic parameters used in risk assessment will affect the appearance of the confusion matrix and the selected metrics. And will changing the default parameters affect the choice of prior. Also, one of the future stages in the research will be the application of machine learning techniques to the data generated in this way. Models trained on imbalanced data will easily recognize data belonging to the majority class, but will not recognize data belonging to the minority class. Therefore, it will be necessary to apply machine learning algorithms that take into account the unbalancing of the data or to apply data balancing techniques such as data class weighting and various over and under-sampling techniques.

8. References

[1] Runje, B.; Horvatić Novak, A.; Razumić, A.; Piljek, P.; Štrbac, B. & Orošnjak, M. (2019). Evaluation of Consumer and Producer Risk in Conformity Assessment Decisions, Proceedings of the 30th DAAAM International Symposium, Zadar, Croatia, ISSN 1726-9679, ISBN 978-3-902734-22-8, Katalinic, B. (Ed.), pp. 0054-0058, Published by DAAAM International, Vienna, DOI: 10.2507/30th.daaam.proceedings.007

[2] https://www.bipm.org/en/committees/jc/jcgm/publications, (2012). JCGM 106:2012 Evaluation of measurement data – The role of measurement uncertainty in conformity assessment, Accessed on: 2022-09-19

[3] https://www.bipm.org/en/committees/jc/jcgm/publications, (2018). JCGM 100:2008 Evaluation of measurement data – Guide to the expression of uncertainty in measurement, Accessed on: 2022-09-19

[4] https://www.eurolab.org/pubs-techreports, (2017). EUROLAB Technical Report No.1/2017 - Decision rules applied to conformity assessment, Accessed on: 2022-09-19

[5] Đorić, D.; Mališić, J.; Jevremović, V. & Nikolić-Đorić, E. (2007). Atlas raspodela, Građevinski fakultet Univerziteta u Beogradu, ISBN: 978-86-7518-077-7, Beograd

[6] Mandić, M. & Kraljević, G. (2020). Two-Layer Architecture of Telco Churn Auto-ML, Proceedings of the 31st DAAAM International Symposium, Mostar, BiH, ISSN 1726-9679, ISBN 978-3-902734-29-7, Katalinic, B. (Ed.), pp. 0788-0792, Published by DAAAM International, Vienna, Austria, DOI: 10.2507/31st.daaam.proceedings.109

[7] He, H. & Garcia, E. A. (2009). Learning from Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, Vol 21, No 9, pp. 1263-1284, DOI: 10.1109/TKDE.2008.239

[8] Jeni, L. A.; Cohn, J. F. & De La Torre, F. (2013). Facing Imbalanced Data-Recommendations for the Use of Performance Metrics, International Conference on Affective Computing and Intelligent Interaction (ACII), Geneva, Switzerland, pp. 245-251, DOI:10.1109/ACII.2013.47

[9] Buda, M., Maki, A. & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks, Vol. 106, pp. 249-259., ISSN 0893-6080, DOI: 10.1016/j.neunet.2018.07.011

[10] Fawcett, T. (2006). An introduction to ROC analysis, Pattern Recognition Letters, Vol. 27, No. 8, pp. 861-874., ISSN 0167-8655, DOI: 10.1016/j.patrec.2005.10.010

[11] McHugh, ML. (2012). Interrater reliability: the kappa statistic, Biochemia Medica, Vol. 22, No. 3, pp. 276-272., DOI: 10.11613/BM.2012.031