Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 7, No. 3, September 2019, pp.1138-1151
Available online at: *http://pen.ius.edu.ba*

Soft-hard data fusion using uncertainty balance principle –corporate credit risk in commercial banking

Sabina Brkić[1], Migdat Hodžić[1], Enis Džanić[2]
[1]School of Engineering, American University in Bosnia and Herzegovina, Sarajevo, B and H
[1]American University in Bosnia and Herzegovina, Sarajevo, B and H
[2]Univerzitet u Bihaću, Bihać, B and H

Article Info

Received May 10, 2019

Keyword:

Soft Data
Hard Data
Credit Default
Uncertainty Balance Principle
Expert Opinion
Data Fusion

ABSTRACT

This study introduces Uncertainty Balance Principle (UBP) as a new concept/method for incorporating additional soft data into probabilistic credit risk assessment models. It shows that soft banking data, used for credit risk assessment, can be expressed and decomposed using UBP and thus enabling more uncertainty to be handled with a precise mathematical methodology. The results show that this approach has relevance to credit risk assessment models in the sense that it proved its usefulness for the purpose of soft-hard data fusion, it modified Probability of Default with soft data modeled using possibilistic (fuzzy) distributions and fused with hard probabilistic data via UBP and it obtained better classification prediction results on the overall sample. This was demonstrated on a simple example of one soft variable, two experts and a small sample and thus this is an approach/method that requires further research, enhancements and rigorous statistical testing for the application to a complete scoring and/or rating system

Corresponding Author:

Migdat Hodzic
School of Engineering, American University in Bosnia and Herzegovina
Trg Solidarnosti 10, Sarajevo, Bosnia and Herzegovina
Email: mhodzic@aubih.edu, migdathodzicmail.com

1. Introduction

In general, any risk analysis can be qualitative or quantitative, depending on the information available and the level of detail that is required [6]. Quantitative techniques rely heavily on statistical approaches while qualitative techniques rely more on judgment than on statistical calculations. Traditional risk models are based on probability and classical set theory which are widely used for assessing market, credit, insurance and trading risk. In case of credit risk, such statistical models use the borrower's characteristic indicators (for corporates usually data from financial statements) and (if possible) macroeconomic variables which were collected historically and are available for defaulting (or troubled) and non-defaulting borrowers. Banks usually implement various scoring and rating tools to build a forecasting model based on correlating default information from the past with the borrowers' characteristics and to use the output of such model for estimated credit worthiness of borrowers with unknown performance, which is done by inputting characteristics into the model. The Internal Rating Based Approach (IRBA) of the Basel Capital Accord allows banks to use their own rating models for the estimation of probabilities of default (PD) as long as the systems meet specified minimum requirements. Most common statistical methods for building and estimation of such models are Regression Analysis, Discriminant Analysis, Logit and Probit Models, Panel Models, Hazard Models, Neural Networks, Decision Trees [27]. These models range from simple to very complex, however, many risks still cannot be analyzed sufficiently by applying classical probability models because of

lack of sufficient experience/historical data, lack of knowledge and vagueness, as well as complex cause-and-effect relationships inherent in certain risk categories. Different to traditional methods for credit risk assessment, fuzzy logic and Theory of possibility can easily integrate linguistic terms and expert opinions (human reasoning) into the assessment enabling more information to be included in case of modeling risks that are not fully understood, as well as cases with insufficient and imprecise hard data. Fuzzy logic systems help simplify comprehensive risk management frameworks and it can provide assistance and/or solutions considering both the available data and experts' opinions for cases where proper quantitative probability model is nonexistent [46]. Fuzzy logic was introduced by Zadeh in 1965 [51] and it has been applied in various industry areas since then such as, in artificial intelligence, computer science, control engineering, decision theory, expert systems, logic, management science, operations research, pattern recognition and robotics [53]. This study is focused on investigating a new method for incorporating additional soft fuzzy information into traditional probabilistic credit risk assessment for corporates, based on soft-hard data fusion via Uncertainty Balance Principle (UBP), for the purpose of improving the predictive power of credit risk assessment model. This is investigated based on data and expert experience of corporate lending of a small commercial bank in Bosnia and Herzegovina (B&H). Market in B&H is very small and it behaves irrationally and often erratically and therefore makes the risk assessment and management decision making process very complex and uncertain which requires new methods for risk modeling to be evaluated. The problem of credit risk assessment in this country is truly insufficiently researched and there are not many scientific publication in this area. One of the major problems for credit risk assessment of corporates from this country lies in the quality of their financial statements, especially for small and medium-sized enterprises. Although there is a legislative framework for accounting and financial reporting, in practice, financial reports are often misleading and prepared in a manner to deceive users mostly with manipulation, falsification or alteration of the accounting documents on the basis of which the financial statements are prepared, deliberate misrepresentation and/or omission of transactions or disclosure of information and the like. All this makes traditional credit risk assessment for corporates very complex and uncertain. This study deals with the use of fuzzy logic as a support tool for traditional evaluation of corporate credit risk in a commercial banking environment, as well as a new approach for soft-hard data fusion using UBP. The main purpose of this study is to investigate and test whether the accuracy of probabilistic credit risk assessment of corporates, evaluated with logistic regression, can be improved using soft and hard data modeling, followed by soft-hard data fusion, in particular using Uncertainty Balance Principle. In literature one can find various methods on how to integrate fuzzy and random data in meaningful ways [34-36, 53], as well as about the area of "random fuzzy sets" and "fuzzy random variables", as well as various "fuzzy" applications [e.g., 1-2, 5, 11-20, 22-24, 33, 37, 39-45, 47-52]. However, this study does not deal with aforementioned. This study is focused on uncertainty transformation of fuzzy to random data by using basic properties of fuzzy and random distributions and the UBP. Hodžić presented a new mathematical approach, the UBP, to deal with uncertainty alignment between fuzzy and random data. He presented a method to describe fuzzy (possibility) distribution in terms of a pair (or more) of related random (probabilistic) events, both fixed and variable with the use of basic properties of both fuzzy and random distributions [28]. This was further expanded to the Uncertainty Balance Principle which was defined to express uncertain data vagueness as represented by a fuzzy data model, with a non-uniqueness of related random data distributions [29]. His method transfers the fuzzy distribution in equivalent random (hard) distribution which can further be combined with the original hard probability distribution using the process of soft-hard data fusion. Fuzzy to random transformations that put probability between necessity and possibility are known from previous research while UBP enables the choice of various cumulative probabilities based on several methods. The result of his approach is that any fuzzy distribution can be thought of as an interplay of two or more probabilistic events and vice versa. UBP can be employed effectively in a variety of data fusion and decision problems where both objective or hard data are to be fused with subjective or soft data [28-31]. Most represented criteria for assessing the quality of a scoring/rating model is in its power to discriminate credit risk (between "non-defaults" and "defaults"). There are many ways to measure such discriminatory power but practitioners usually choose the Cumulative Accuracy profile (CAP) and its summary statistic the Accuracy Ratio (AR), Gini coefficient or the area under the Receiver Operating Characteristic (ROC) which is a similar concept to the CAP. Engelmann et al. [21] show that the AR is just a linear transformation of the area below the ROC curve. However, one should be careful in the use and interpretation of rating/scoring discriminatory measures because they may mislead when it comes to the assessment of the quality of a rating/scoring system [7]. Considering that this study is focused on a small portfolio, and that it will generate pre-fused and fused results from the same portfolio and same time period, it

is considered as adequate to perform comparison of pre-fused and fused credit risk assessment results based on discriminatory power of pre-fused and fused model [7, 21]. For the purpose of this study AR is used to test the approach and Confusion matrix to demonstrate false predictions of the models. Intention of this study is not to produce a whole scoring or rating system used for credit risk assessment because here the focus is to explore if and how UBP can be used to incorporate additional soft information into credit risk assessment models/systems. Our final aim is to be able to improve bank credit risk assessments and other relevant and highly needed information, where various assessment and modeling approach is suggested and/or anticipated, by using exact and more precise mathematical methodology.

2. Soft-Hard Data Fusion using Uncertainty Balance Principle - Credit Risk Assessment in Commercial Banking

2.1 Hard data modeling

In this section we predict corporate probability of default based on hard data, represented through financial statements and financial ratios, which are used for statistical credit risk assessment via assessing the probability of default occurrence in one commercial bank in B&H. In Introduction several statistical methods which are used for risk assessment and default prediction/credit scoring are presented. In the academic literature, as well as in banking practice, the most popular method is the logit model. Logistic regression measures the relationship between the categorical dependent variable and one or more independent or explanatory variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. It predicts the odds of being a case based on the predictor values which are represented through independent variables. The main reason for its popularity in the area of credit risk assessment is that the output from the logit model can be directly interpreted as default probability. Secondly, the resulted logit model enables a check whether the empirical dependence between the potential explanatory variables and default risk is economically meaningful. Thus, the logit method is chosen for the purpose of demonstrating the estimation of default prediction model for corporate exposures. All calculations in this section are performed in RStudio Version 1.1.442. Optimal PD logistic regression model is further used for soft-hard data fusion. The original data, which was supplied by a small commercial bank from B&H, consist of 304 individual companies financed by the bank during five consecutive years (up to year end 2016) and their default/non-default status over the mentioned period. For all individual companies official financial reports, such as the Balance Sheet and Profit and Loss Statement and similar, were obtained from the FIA[1]. The financial data was checked from the perspective of apparent mistakes, homogeneity of all observations, as well as availability of default information for all companies within the data set. Next step was to define the dependent binomial variable. A dichotomous variable is used to represent the dependent variable based on which companies are separated on defaulted and healthy/non-defaulted companies. The value of the dichotomous dependent variable is assigned as 1 in cases which represent a defaulted company, while the value of the dependent variable of 0 is assigned to a healthy/non-default company. Further, the independent or explanatory variables are defined, which are commonly expressed via financial ratios. The financial ratios usually represent the most significant credit risk factors such as leverage, liquidity, productivity, turnover, activity, profitability, firm size, growth rates and leverage development [26]. For the purpose of this research a total of 33 financial ratios are calculated representing aforementioned most important credit risk factors. For three companies official financial statement were not available so these companies were excluded from the final data set. Also, some financial information were missing for a couple of companies. The easiest way to manage those cases is to complete eliminate/exclude the respective observations from the data set and further modeling. However, in this case where the total sample consists of relatively small number of companies this would result in relatively too many observations being lost. Thus, the option of substituting the missing values with the corresponding mean values are chosen, calculated over all observations for the respective time period which basically creates "neutral" values that enable an undistorted assessment by using the remaining input factors. Outliers in the calculated financial ratios are also eliminated by replacing the extreme data points by the 1% respectively the 99% percentile of the according ratio which means the replacement of the values of outliers with the largest or second smallest value in observations excluding outliers. After all data processing, analysis,

[1] Financial – Intelligence Agency which is a state agency for financial, information and intermediation services (FIA, Finansijsko-informatička agencija).

validation and imputation was conducted the final data set was completed. Table 1 shows the final number of observed companies per year, as well as the split of the companies into defaulting and non-defaulting. Shown default ratio is rather high but it moved in line with the total NPL market during observed period, as presented in the Financial Stability Report of CBBH [10].

Table 1. Number of observations and defaults per year

Year	Total	Default companies	Non default companies	Default Ratio
2012	101	11	90	10,9%
2013	55	9	46	16,4%
2014	56	7	49	12,5%
2015	49	10	39	20,4%
2016	39	6	33	15,4%
Total	300	43	257	14,3%

The final data set contains 300 companies and their relevant hard data from Balance sheet and Profit and Loss Statements (total 93 positions), financial ratios (total 33) and relevant default data. For further analysis 36 variables in total are used, of which 33 different financial ratios, Total assets, Total income and Average number of employees per company. The number of chosen variables (36) is still high as the optimal model should contain only a few potential explanatory variables to avoid over-fitting. In order to reduce the number of variables for the logit regression the factor analysis is used. Cross-correlations among 36 variables used in this study were computed. Each variable having correlation with another variable higher than +50% and lower than -50% was removed from further analysis. Remaining variables were first checked if suited for any factor analysis. This is done via Bartlett test of Sphericity and the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (MSA). KMO measured is 0.59, while Bartlett's Test of Sphericity is 1302,901 and is a statistically significant at the 0.000 level. The required sample size recommended in theory for the factor analysis was obtained as the sample used includes 300 observations, or 18 per variable. Hence, factor analysis is considered as an appropriate technique for further analysis of the data. We used Principal Component Analysis (PCA) to achieve further data reduction, as well as to recognize one representative variable from each factor. For further analysis, variables with the highest factor loadings (correlation between the original variables and the factors) are used. All other variables showing high factor loadings within one factor were removed as they show high cross-correlations. PCA provides a variable-by-variable correlation matrix which is used to extract variables that represent a linear combination of the original variables, with an aim to achieve reduction in variables. The coefficients in each linear combination are known as factor loadings. Varimax rotation is used in order to redistribute the variance from earlier factors to later ones in order to achieve simpler, theoretically more meaningful factor patterns, as suggested by Hair, Black, Babin and Andersen [25]. Examining the eigenvalues of factors 14 different factors, explaining the 88 % of the total variance, were recognized which had eigenvalues greater than 1.The next step involved selecting one variable from each of the 14 factors to use further in the logistic regression analysis. A variable with the highest weight from each factor was chosen for the following logistic regression analysis. The Variance Inflation Factor (ViF) of chosen variables for Logistic regression is also tested to identify collinearity among explanatory variables. ViF were in the range of 1.024 to 1.567 for all 14 variables and it is concluded that they can be used further for the identification of an optimal set of explanatory variables of a logit model. We created two logistic regression models, with the finally chosen 14 predictor variables, based on different methods for variable selection. Variable method selection specify how independent variables are entered into the analysis and by using different methods, we can construct a variety of regression models from the same set of variables. Although there are several selection methods, this study uses Enter method (as Model 1) and Backward stepwise method (as Model 2). Enter method represents a variable selection in which all variables in a block are entered in a single step while Backward selection or elimination starts with all predictors in the model and iteratively removes the least contributive predictors. The removal stops when no further variables can be deleted without a statistically significant loss of fit. Training and testing sample was created in R Studio in the split ratio of 80/20. The 80% of the data set was used for training the data and fitting the model, while 20% of

the data set was used to test and validate the final model. Logistic regression result of Model 1 is shown in Table 2, while for Model 2 in Table 3.

Table 2. Model 1 for default prediction of corporate portfolio in the respective bank

	Estimate (B)	Std. Error	z value	Pr(>|z|)
(Intercept)	-2,45E+00	1,09E+00	-2	0,02408 *
Net profit or loss	-1,39E-07	1,21E-07	-1	0,25362
Average number of employees	-8,28E-04	1,54E-03	-0,537	0,59145
Long term assets to total assets	3,30E+00	1,30E+00	3	0,01135 *
Financial investments to total asset	4,79E+00	1,80E+00	3	0,00775 **
Shareholder equity ratio	-3,80E+00	1,38E+00	-3	0,00581 **
Financial leverage	8,14E-03	1,34E-02	0,606	0,54449
Current liquidity	-1,62E-01	3,81E-01	-0,424	0,67127
Credit exposure from business	3,01E-01	2,98E-01	1	0,31142
Short-term assets turnover ratio	-1,17E-01	1,59E-01	-0,737	0,46122
Inventories turnover ratio	-8,04E-03	7,63E-03	-1	0,29212
Inventory days	2,52E-03	1,38E-03	2	0,06788 .
ROE	-8,44E-01	1,33E+00	-0,635	0,52513
Operating margin	-2,30E+00	9,43E-01	-2	0,01454 *
Added value per employee	-3,52E-06	4,02E-06	-0,876	0,38115

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 3. Model 2 for default prediction of corporate portfolio in the respective bank

	Estimate (B)	Std. Error	z value	Pr(>|z|)
(Intercept)	-1,792129	0,730099	-2,455	0,014103 *
Long term assets to total assets	2,728684	1,134073	2,406	0,016124 *
Financial investments to total asset	3,082175	1,519427	2,029	0,042508 *
Shareholder equity ratio	-3,852924	1,155605	-3,334	0,000856 ***
Short-term assets turnover ratio	-0,276585	0,136526	-2,026	0,042777 *
Inventory days	0,001758	0,001310	1,342	0,179473
Operating margin	-2,734262	0,741854	-3,686	0,000228 ***

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 4. Logit models validation results

Model	Type of method	AIC	McFadden	Cox and Snell (ML)	Nagelkerke (Cragg and Uhler)	Df.diff	LogLik.diff	Chisq	p.value
M1	Enter	176,36	0,279	0,209898	0,367841	-14	-28,271	56,543	0,0046942
M2	Backward	142,53	0,288	0,215861	0,378754	-6	-26,019	52,038	0,0000183

Finally, the derived logit models had to be adequately validated and tested. The first step in this process was to conduct relevant statistical tests in order to verify the model's robustness and goodness of fit, as shown in Table 4. We calculated the coefficient of determination, denoted as R^2, which do not assess the goodness-of-fit but is rather a measure based on various comparisons of the predicted values from the fitted model to those from model 0, the no data or intercept only model [32]. While R^2 statistics can be suggestive, it is most useful when comparing competing models for the same data. R^2 summarizes the proportion of variance in the dependent variable associated with the independent variables. This value ranges from 0 to 1 and larger R^2 values indicate that more of the variation is explained by the model. For regression models with a categorical

dependent variable, it is not possible to compute a single R^2 statistic that has all of the characteristics of R^2 in the linear regression model. Thus, several pseudo R^2 are used in such cases. For both models Cox and Snell's R^2, Nagelkerke's R^2 and McFadden's R^2 (please see Table 4) is calculated. Cox and Snell's R^2 (M1=0,210; M2=0,216) is based on the log likelihood for the model compared to the log likelihood for a baseline model. Nagelkerke's R^2 (M1=0,368; M2=0,379) is an adjusted version of the Cox & Snell R^2 that adjusts the scale of the statistic to cover the full range from 0 to 1 because Cox and Snell's R^2 has a theoretical maximum value of less than 1, even for a "perfect" model. McFadden's R^2 (M1=0,279; M2=0,288) is based on the log-likelihood kernels for the intercept-only model and the full estimated model. We used the log likelihood ratio test to compare the goodness of fit of two statistical models, a null (small) model against an alternative model (more complex model). Presented models show following goodness-of-fit indicators: the LogLikelihood shows a value of -28,271 for Model 1 and -26,019 for Model 2, Chi-square of Model 1 is 56,543 and Model 2 is 52,038, while Model 1 being statistically significant at $p < 0,005$ and Model 2 at $p < 0,001$. Based on provided results in Table 4 we conclude that Model 2 is a better fit for used data because it has higher R^2 and is statistically significant at a lower p value. Thus, this model is chosen as the final model for default prediction of corporate portfolio in respective bank, based on provided hard data set. According to final logit model (M2) an increase in "Shareholder equity ratio", "Short-term assets turnover ratio" and "Operating margin" are associated with a decrease in probability of default occurrence. On the other hand, increase in other explanatory variables shown in Table 3 are associated with an increase in probability of default occurrence. Thus, the empirical dependence between the potential explanatory variables and default risk is economically meaningful. As a second validation test, the estimated final model should be applied to a validation/testing sample to test how the model produces out-of-sample forecasts. For this step assessing the quality of the model based on the accuracy ratio concept is chosen. The testing was considered on the initially created testing sample and by assessing three different hit indicators:

a) Accuracy Ratio (AR) of non-defaulted/healthy companies' correct prediction,
b) AR of defaulted companies correct prediction and
c) AR of total model prediction (also referred to as the total model hit ratio).

The model was also tested with different classification cut off points. After the comparison between the various classification cut off points, as shown in Table 5, the 0,25 was retained since it predicts classification of companies significantly better in two indicators compared to other classification cut off point. This means that PDs, calculated by the model, which are below 0,25 classification cut off are to be considered as non-defaulted companies, while PDs above this point are to be considered as defaulted. It is, however, not considered that one always chose the cut off that gives the highest classification rate. Depending on the relative costs of false positives and false negatives, one can also choose to use a cutoff that gives a slightly lower correct classification in order to minimize cost. For the purpose of this study the cut-off that gives the highest classification rate is assumed without further analysis of the relative costs.

Table 5. Various classification cut off results

Model	Type of method	Classification Cut-Off	AR all	AR default	AR healty
M2	Backward	0,5	0,87	0,27	0,96
		0,3	0,87	0,55	0,92
		0,25	0,88	0,64	0,92

With applied 0,25 classification cut off on the testing sample, the model resulted in AR of 92% for non-defaulted/healthy companies correct prediction, AR of 64% for defaulted companies correct prediction and total model hit ratio of 88%. Thus, PD logit model with 0,25 classification cut off will be further used for the soft hard data fusion. The Bank provided different corporate categories based on Total Asset volume, of which Extremely small (up to BAM 100k), Small (from BAM 100k to BAM 500k), Medium (from BAM 500k to BAM 5m), Large (from BAM 5m to BAM 20m) and Extremely large (more than BAM 20m). Based on provided threshold for corporate segmentation Extremely small companies make up 4% share, Small corporates 14,3%, Medium corporates 41,3%, Large corporates 20,7% and finally Extremely large corporates 19,7% of the share. The most representative corporate category, in terms of number of companies within the

category, is the Medium sized corporates and thus soft-hard data fusion will be tested on this segment. In Figure 1 the Cumulative distribution function (cdf) and Probability density function (pdf) of a Medium size company defaulting versus predicted PD score per Logistic Regression output are shown.

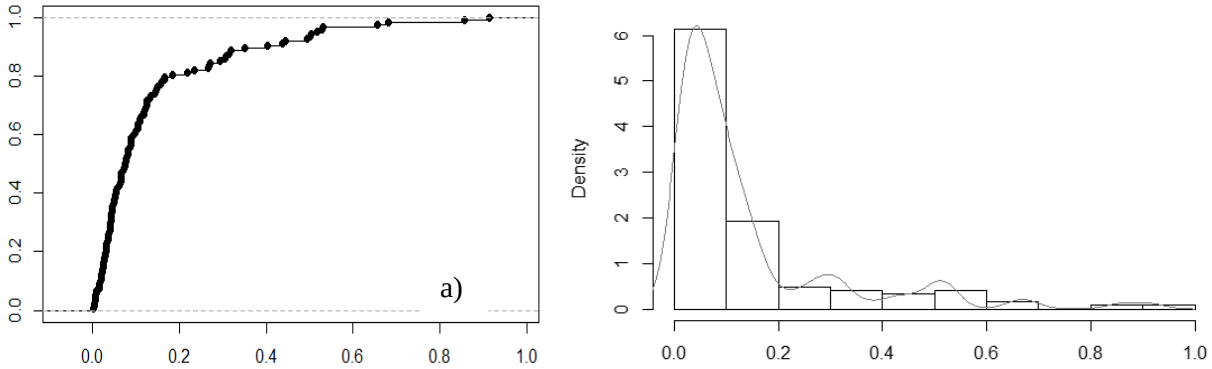


Fig.1. a) Logistic regression curve showing probability of a Medium size company defaulting versus predicted PD score and its b) Probability density function of predicted PD score per Logistic regression

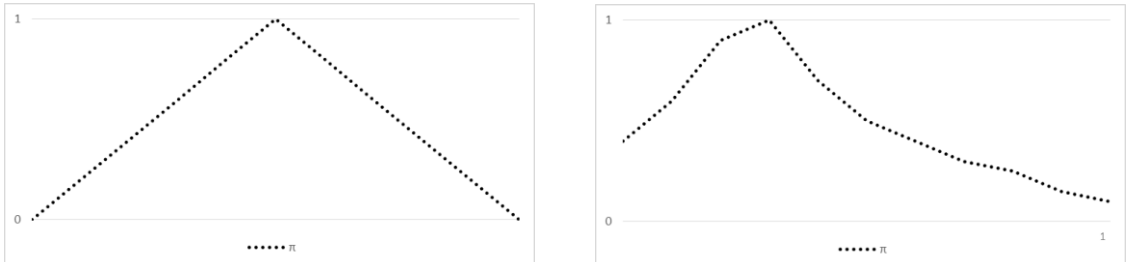
2.2 Soft data modeling

In previous research studies fuzzy distributions of soft data/variables used for corporate credit risk assessment in commercial banking were designed and developed by applying Type-1 [8] and Type-2 fuzzy logic [9]. For this purpose expert sample was created ad hoc with a commercial bank in Bosnia and Herzegovina that was willing to take part in this project at this initial phase (the same bank that supplied the hard data described in previous section). Various experts were interviewed for the types of soft variables used for conducting credit risk assessment of corporates. They provided all information about the process, data processing and various inputs used for credit risk assessment. Mentioned experts and relevant literature were consulted in the definition of membership functions. Experts have provided inputs for generating universe of discourse and the number and description of membership functions related to each soft variable. Membership functions were generated completely unmotivated and are expressing the interviewed experts own opinion and experience. Data processing was done by listing all identified soft variables and by recording and mapping their membership values into membership functions based on inputs from interviewed experts. The final step in that study was to generate a list of the most significant soft variables and their descriptions, as well as to create graphical illustrations of possibility distribution of each soft variable. Design and development of fuzzy/possibilistic distributions (the term fuzzy and possibilistic distributions are used interchangeably) of soft data/variables used for corporate credit risk assessment served as first step in the process of creating a new credit risk assessment model based on soft-hard data fusion via UBP. Next step in this process is to transform identified soft data into hard and prepare the data for soft-hard fusion. Thus, in order to illustrate this approach the research objective within this section is to model the obtained soft variables into hard data using UBP. With this we aim to present the linguistic and intuitive (soft) information about bank credit risk data, expressed through a series of mathematical fuzzy (possibilistic) distributions, which can be handled quantitatively and combined (fused) with related probabilistic data. For the purpose of this study the testing of our approach is based on using one soft variable provided by two different interviewed experts. All other identified soft variables, as well as inclusion of inputs from all interviewed experts, will be used/extended in future research. In order to achieve the aforementioned objective, soft to hard modeling via UBP is envisaged through two main activities, which consists of:

a) Transformation of fuzzy to ProCDs, ProPDs and association of specific probabilities.
b) Definition of 5 possible Ps1-5 for testing the area of interest and its impact on hard data.

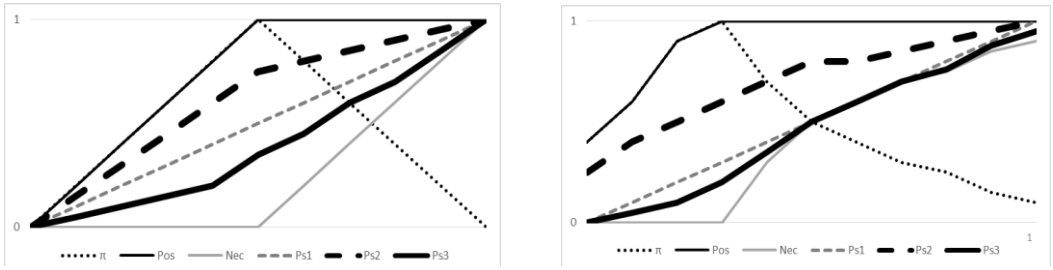
We chose one soft Type-1 fuzzy distribution and inputs from two experts, hereinafter referred to as Expert 1 and Expert 2, for the objective of soft to hard transformation and finally the soft-hard fusion. The chosen soft variable is presented in Figure 2 in the form of a Type-1 fuzzy distribution. Due to confidentiality estimation results that were given by the bank experts are not disclosed but instead a graphical illustration of the fuzzy distribution results is shown. The soft to hard uncertainty transformation via UBP is based on a three step

methodology [30]. The first step is to decompose fuzzy distribution via cumulative probability distribution. This is demonstrated in Figure 3 where π is the original fuzzy distribution, *Pos* is the possibility distribution while *Nec* represents the necessity.



a) Expert 1 b) Expert 2

Fig.2. Type-1 fuzzy distribution of chosen soft variable and experts



a) Expert 1 b) Expert 2

Fig. 3. Decomposition of fuzzy distribution via cumulative probability distribution and soft to hard alignment

UBP enables the choice of various cumulative probabilities based on several methods and in the next steps of UBP specific probabilities are determined using related probability density functions (ProPD), as well as associate specific probabilities with probabilistic events, as demonstrated in Figure 3. With the application of UBP soft to hard alignment process five different test cases of Soft default probability (hereinafter referred to as *Ps*) are defined, where Ps1 is a neutral case, Ps2 lies between Ps1 and Possibility distribution (Pos), Ps3 is between Ps1 and Necessity (Nec), while two extreme cases are represented by Pos in case of Ps4 and Nec in case of Ps5. In Figure 4 probability density functions of all defined *Ps* cases for both experts is shown.

2.3 Soft-Hard data fusion

This section contains fusion of data obtained from soft and hard data modeling. This study is mainly interested in pre-fused and fused PD of observed companies and its implication on the model accuracy of Default and Non-Default predictions. Thus, hard data represent the PD demonstrated in Figure 1, while the soft data is based on inputs from two experts represented via possibility distributions demonstrated in Figure 3 in previous section. There are various methods for the soft-hard data fusion [e.g. 38]. In this study two methods are used of which first method is an ad-hoc, intuitive and simple method of averaging hard and soft distributions (1). Assumption underlying this method is that both hard and soft data are from the same source only different detections were used to collect them and so averaging is used as an ad-hoc method of combining them. We have:

$$\text{PDF (or PMF) fused} = (\text{PDF (or PMF) hard} + \text{PDF (or PMF) soft}) / 2 \quad (1)$$

For the second fusion method a normalized independent fusion is used, which assumes independence of the source for hard and soft data and thus resulting value is the product of soft and hard pdf/pmf normalized with their total sum:.

$$\text{PDF (or PMF) fused} = (\text{PDF (or PMF) hard} * \text{PDF (or PMF) soft}) /$$

SUM(PDF (or PMF) hard * PDF (or PMF) soft) (2)

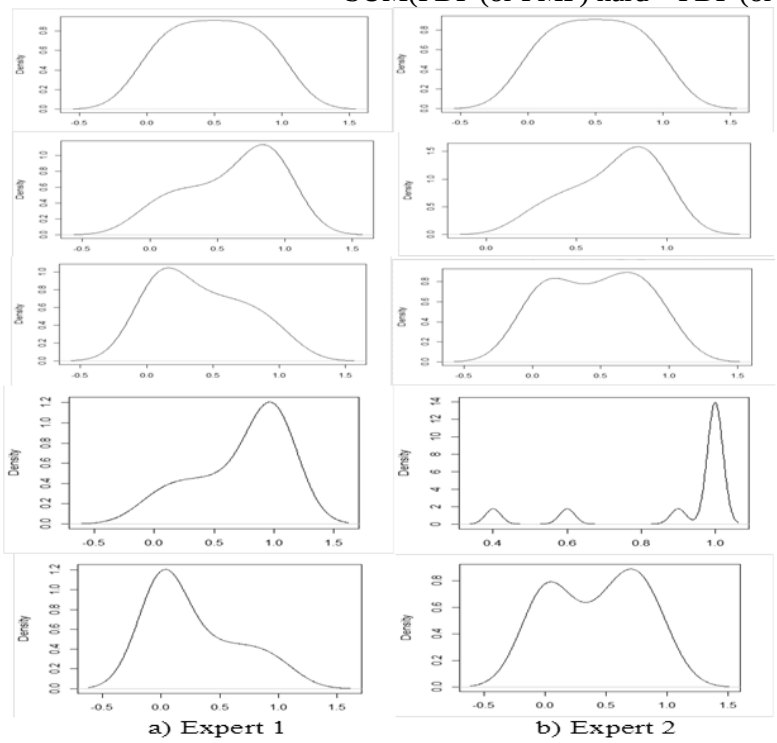


Fig.4. Probability density functions of related ProPD
(Ps1 on the top, Ps5 on the bottom)

The validity of each of the fusion methods is tested on the banking data base we used. No attempt was made to optimize this choice and this is left for a follow up research.
The fusion is considered for following cases:

- Soft-hard data fusion based on first fusion method and inputs from Expert 1,
- Soft-hard data fusion based on first fusion method and inputs from Expert 2,
- Soft-hard data fusion based on second fusion method and inputs from Expert 1
- Soft-hard data fusion based on second fusion method and inputs from Expert 2.

In addition to aforementioned, two options of fused PD calculations are explored, the Complex and Simple method. The Complex fused PD calculation method consists of group pdf balancing, while the Simple fused PD calculation method consists of individual pdf balancing. Thus, the fusion output is presented in eight different testing cases. Figure 5 demonstrates soft-hard fusion results for the case of Expert 1. Starting from the left side of Figure 5, all test cases of soft default probability transformed using UBP, which are used for the fusion with hard data, are shown. Next a comparison of cdf of pre-fused (hard) PD and fused PD is shown for both first and second fusion method for observed companies, as well as for all test cases of Ps. On the right side the Figure 5 a comparison of pre-fused (hard) PD and fused PD, calculated based on Simple and Complex PD method, is shown per company for both first and second fusion methods, as well as for all test cases of possibility functions/distributions. The same was calculated for the case of Expert 2. Performance of fused Default vs Non-Default classification model is evaluated based on the AR (true predictions) and error or confusion matrix (false predictions). Such matrix gives an overview of prediction results on a classification problem. Results from all testing cases show that the minimum false rate achieved, without worsening results of the pre-fused model, is:

- **13,71%** for False Total predictions (vs **15,32%** in pre-fused model),
- **9,35%** for False Default predictions (vs **11,21%** in pre-fused model) and
- **41,18%** (same as is pre-fused model).

This is achieved in several testing cases under different fusion methods, PD calculation and Ps, which further implies that several types of fusion methods, fused PD calculations, as well as different options for *Ps* can be used for further PD optimization, validation, calibration and etc.

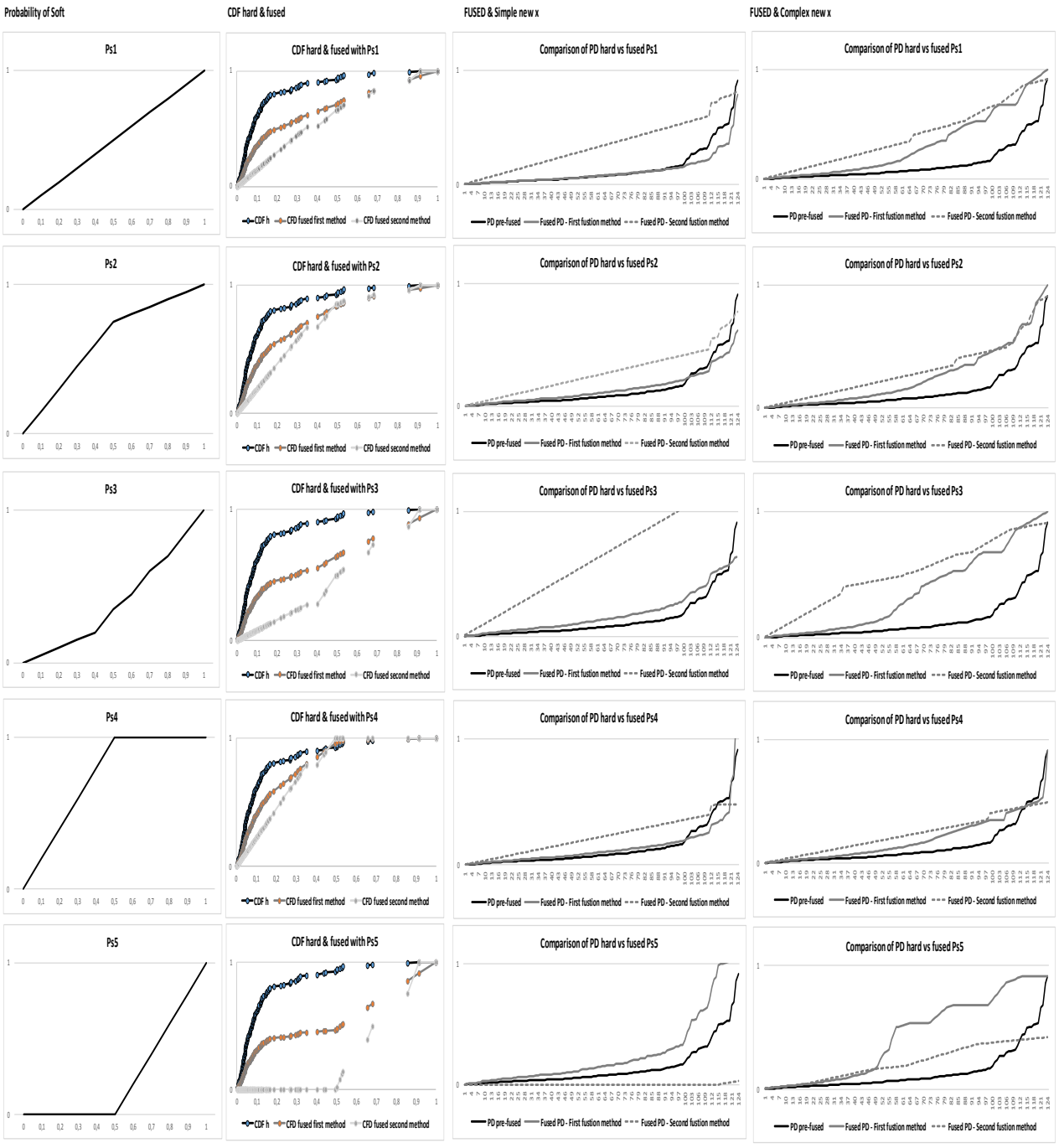


Fig.5. Soft-hard fusion results for the case of Expert 1

3. Future Research

3.1. Limitations of the research

There are several limitations that apply to this research, but four of them are of particular importance. Firstly, using only one number, such as the AR in this case, to compare different models that are generated on the same data contains little information from a statistical point of view. When assessing the quality of a complete scoring and/or rating systems it is desirable to perform rigorous statistical test which are the only way to

obtain a sound decision about the superiority of one rating model over the other [21]. Nevertheless, this study did not aim to produce full scoring/rating models but rather to investigate the usage of UBP on a corporate data set in order to set foundations for further research. Secondly, approach was tested only on the corporate portfolio of one bank, using only one corporate segment, one soft variable and inputs of only two experts. Aim was to test the application/utilization and implications of UBP and for this it was enough to create several testing cases based on aforementioned sample. However, many soft variables impact performance of a company, as well as different expert opinions/perceptions about soft variables impact the decision making process, hence for a complete expert system or scoring model all of them should be included in the modeling of the predictions. Third limitation is reflected in the ability of experts to properly/correctly transform their opinions, perceptions and experiences into relevant fuzzy distributions. If soft data is not correctly modeled the fusion output will not result in improvements of the model's predictive power and could also significantly diminish its predictive power. Finally, the forecasting or predictive power of any statistical model is dependent on assuming unchanged relationship between the model's variables and the default event which considers application of historical relationship between the two for future development. Given the wide range of possible events which can influence companies operations this assumption cannot be guaranteed over longer time periods. In order to overcome this limitation it is necessary to regularly revalidate the model, as well as recalibrate the model based on revalidation results, in order to ensure model's targeted predictive power. All mentioned limitations, as well as other such as limitations of AR, necessity for PD calibration, inherent limitations of logistic regression and Type-1 fuzzy distributions, will be addressed in further research.

3.2. Future research

Further research shall address identified limitations of this research, particularly in terms of including all relevant aspects into a final and complete credit risk assessment model based on fused soft-hard data using UBP. Soft inputs from all experts and all soft variables shall be included and tested with various options for group or individual soft fusion in order to find the optimal combination for soft-hard data fusion. The finalized and complete credit risk assessment model shall include optimization of soft hard PD functions, calibration of PD, as well as other classification performance measures and in and out of sample testing. It shall also be compared with other credit risk assessment models which already have soft information included based on other/different methodologies in order to test strengths and weaknesses of using UBP compared to other methods which include soft data. The research will be extended to incorporate a study on other bank portfolios and other types of borrowers (e.g. retail) in order to see how the approach works for smaller vs larger portfolios. Further research will test the soft-hard data fusion via UBP on other credit risk assessment methods (e.g. hazard), other data fusion methods and other methods for fused PD calculation. Future research shall also explore the impact of Type-2 fuzzy/possibility distributions to address inherent limitations of Type 1 fuzzy distributions so that broader range of banking data uncertainties can be handled and combined with the corresponding hard data with an aim of finally being incorporated into a new (and potentially superior) soft-hard data fusion model for credit risk assessment and other similar risk assessments.

4. Conclusions

This study introduces a new concept/approach, the UBP, which is a new fuzzy to random uncertainty alignment methodology based on which fuzziness can be described as precisely defined non unique randomness [29], and which can be applied to credit risk assessment models. It presents a new methodology to deal with soft and hard data fusion in the credit risk assessment modeling, by using fused data to enhance discriminatory power of the predictions and thus the decision making process. We demonstrated that soft banking data used for credit risk assessment can be expressed and decomposed using UBP and its defined three step methodology. Main contribution of UBP is that it enables the choice of various cumulative probabilities based on several methods and thus enables more uncertainty to be handled with a precise mathematical methodology. We tested two fusion methods and demonstrated that by group and individual pdf balancing one can obtain new fused PD scores which can be further used for various testing, validation, calibration and similar activities required for a complete scoring and/or rating model/system. This study demonstrated the usage of UBP as a new methodology for incorporating additional information into credit risk assessment. Pre-fused and fused PDs are presented for different fusion and new fused PD calculation methods. All test cases of soft possibilistic distributions, generated based on Type-1 fuzzy logic and decomposed via

UBP, demonstrated different shifts of PD compared to pre-fused results. In addition, different results are shown for AR and Confusion matrices in case of different soft possibilistic Type-1 data. The optimal AR obtained for presented data contains improvement in first two categories (total predictions and default predictions) while no change in the last category (non-default predictions) compared to the pre-fused model. This is achieved in several testing cases, which further implies that various types of fusion methods, fused PD calculations, as well as different options for *Ps* can be used for further PD optimization and model testing, validation, PD calibration and etc. The results show that UBP has relevance in the sense that it proved its usefulness for the purpose of soft-hard data fusion, it changed PDs with soft data modeled using possibilistic distributions and fused with hard probabilistic via UBP by getting better prediction results on the overall sample (the total model hit). However, this was demonstrated on a simple example of one soft variable and a very small sample. Real scoring and rating systems are extremely complex and require rigorous statistical tests in the validation process [3-4], and this will be applied and reported in further research.

References

[1] Agarwal, P., & Najal, H.S. (2015). Possibility theory vs possibility theory in fuzzy measure theory, *Int. Journal of Engineering Research and Applications 5*(5), 37–43.

[2] Babashamsi, P., Golzadfar, A., Yusoff, N.I., Ceylan, H., & Nor, N.G. (2016). Integrated fuzzy analytic hierarchy process and VIKOR method in the prioritization of pavement maintenance activities. *Int. J. Pavement Res. Technol.*, *9*(2), 112-120.

[3] Bank for International Settlements (February, 2005a), Working Paper No. 15: Studies on the Validation of Internal Rating Systems, Basel: Basel Committee on Banking Supervision, Available at: www.bis.org.

[4] Bank for International Settlements (July, 2005b), An Explanatory Note on the Basel II IRB Risk Weight Functions, Basel: Basel Committee on Banking Supervision, Available at: www.bis.org.

[5] Bellman, R., & Zadeh, L. (1979). Decision-making in a fuzzy environment. *Management science.*

[6] Bennett, J.C., Bohoris, G.A., Aspinwall, E.M., & Hall, R.C. (1996). Risk analysis techniques and their application to software development. *European Journal of Operational Research, 95*(3), 467-475.

[7] Blochwitz, S., Hamerle, A., Hohl, S., Rauhmeier & R., Rösch, D. (2005). Myth and reality of discriminatory power for rating systems. *Wilmott Magazine*, pp. 2-6.

[8] Brkić, S., Hodžić, M., & Džanić, E. (2017). Fuzzy Logic Model of Soft Data Analysis for Corporate Client Credit Risk Assessment in Commercial Banking. *Fifth Scientific Conference with International Participation "Economy of Integration" ICEI 2017*, Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3079471

[9] Brkić, S., Hodžić, M., & Džanić, E. (2019). Soft Data Modeling via Type 2 Fuzzy Distributions for Corporate Credit Risk Assessment in Commercial Banking, In Avdakovic, S. (Ed) *Advanced Technologies, Systems and Applications III* (pp. 457-469). Springer, Cham.

[10] Central Bank of Bosnia and Herzegovina (2017). *The Financial stability report 2017.* Retrieved October 10, 2018, Available at: www.cbbh.ba

[11] Coolen, F.P.A., et al. (2010). *Imprecise probability*. In: Lovric M. (eds) International Encyclopedia of Statistical Science. Berlin, Heidelberg: Springer.

[12] de Cooman, G. (1996). Possibility theory 1, the measure- and integral-theoretic groundwork. Universiteit Gent, Vakgroep Elektrische Energietechniek.

[13] Dubois, D. (2006). Possibility theory and statistical reasoning. Institut de Recherche en Informatique de Toulouse.

[14] Dubois, D., & Prade, H. (1983). Unfair coins and necessity measures: towards a possibilistic interpretation of histograms. *Fuzzy Sets Syst., 10*(1), 15–20.

[15] Dubois, D., & Prade, H. (1986). Fuzzy sets and statistical data. *Eur. J. Oper. Res., 25*(3), 345–356.

[16] Dubois, D., & Prade, H. (1987). The mean value of a fuzzy number. *Fuzzy Sets Syst., 24*(3), 279–300.

[17] Dubois, D., & Prade, H. (1992). When upper probabilities are possibility measures. *Fuzzy Sets Syst., 49*(1), 65–74.

[18] Dubois, D., & Prade, H. (2002). Possibility theory probability theory and multiple valued logics: a clarification. *Annal. Math. Artif. Intell., 32*, 35–66.

[19] Dubois, D., & Prade, H., (1988). *Possibility Theory*. New York: Plenum.

[20] Dubois, D., Prade, H. & Smets, P. (2001). New semantics for quantitative possibility theory. *2nd International Sympoium on Imprecise Probabilities and Their Applications* (pp. 152-161). Ithaca, New York.

[21] Engelmann, B., Hayden, E., & Tasche, D. (2003). Testing for Rating Accuracy, Risk 16, January, 82-86.

[22] Eschenbach, W. (2012). Triangular Fuzzy Numbers and the IPCC. Retrieved October 19, 2018, from https://wattsupwiththat.com/2012/02/07/triangular-fuzzy-numbers-and-the-ipcc/

[23] Feller, W. (1950). *An Introduction to Probability Theory and Its Applications*. New York: Willey.

[24] Garibaldi, J.M., & John. R.I., (2003). Choosing Membership Functions of Linguistic Terms. *Proceedings of the 2003 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2003),* (pp. 578-583). St. Louis, USA.

[25] Hair Jr. J., Black W. C., Babin B. J. & Andersen R. E. (2010). *Multivariate data analysis*, 7th Edition, Prentice Hall.

[26] Hayden E. (2011). Estimation of a Rating Model for Corporate Exposures. In Engelmann, B., & Rauhmeier, R, (eds) *The Basel II Risk Paramenters* (pp.13-24). London: Springer.

[27] Hayden E., Porath D., (2011). Statistical Methods to Develop Rating Models. In Engelmann, B., & Rauhmeier, R, (eds) *The Basel II Risk Paramenters* (pp.1-12). London: Springer.

[28] Hodžić, M. (2016a). Fuzzy to Random Uncertainty Alignment. *Southeast Europe Journal of Soft Computing, 5*(1), 58-66.

[29] Hodzic, M. (2016b). Uncertainty Balance Principle. *IUS Periodicals of engineering and natural sciences, PEN 4(*2), 17-32.

[30] Hodzic, M. (2018). Soft to Hard Data Transformation Using Uncertainty Balance Principle. In Hadzikadic, M. & Avdakovic, S. (Eds) *Advanced Technologies, Systems and Applications II* (pp. 785-809). Springer International Publishing.

[31] Hodzic, M. (2019). A Platform for Human-Machine Information Data Fusion, In Avdakovic, S. (Ed) *Advanced Technologies, Systems and Applications III* (pp. 430-456). Springer, Cham.

[32] Hosmer, D., W., & Lemeshow, S. (2010). Applied Logistic Regrresion, 2nd Ed. NJ, USA. John Wiley & Sons, Inc.

[33] Iancu, I., Mamdani, A. (2012). Type fuzzy logic controller. In Dadios, E. (ed.) *Fuzzy logic—controls, concepts, theories and applications* (pp. 325-350). InTechOpen.

[34] Jenkins, M.P., et al. (2015). Towards context aware data fusion: modeling and integration of situationally qualified human observations to manage uncertainty in a hard-soft fusion process. *Information Fusion, 21*(1), 130–144.

[35] Kandemir, E. et al. (2017). A Comparison of Perturb & Observe and Fuzzy-Logic Based MPPT Methods for Uniform Environment Conditions, *Periodicals of engineering and natural sciences,5(1*), 16-23.

[36] Kaufmann, A., & Gupta, M.M. (1985). *Introduction to Fuzzy Arithmetic, Theory and Applications*. New York: Reinhold, Van Nost.

[37] Kaur, B., Bala, M., & Kumar, M. (2014). Comparitive analysis of fuzzy based wildfire detection techniques. *Int. J. Sci. Eng. Res., 5*(7), 813-818.

[38] Leon-Garcia, A. (2008). *Probability, statistics, and random processes for electrical engineers* (3rd ed). Upper Saddle River, NJ: Pearson Prentice Hall - Pearson Education, Inc.

[39] Liu, B. (2012). Why is there a need for uncertainty theory? *J. Uncertain. Syst., 6*(1), 3–10.

[40] Mauris, G. (2011). Possibility distributions: a unified representation of usual direct-probability-based parameter estimation methods. *Int. J. Approx. Reason.*, *52*, 1232–1242.

[41] Narukawa, Y., Torra, V., & Gakuen, T. (2016). Fuzzy measure and probability distributions: distorted probabilities. Retrieved July 20, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.2292&rep=rep1&type=pdf

[42] Onuwa, O.B. (2014). Fuzzy expert system for malaria diagnosis. *Orient. J. Comput. Sci. Technol.*, *7* (2), 273–284.

[43] Raufaste, E., & Neves, R.D.S. (1998). Empirical evaluation of possibility theory in human radiological diagnosis. In Prade, H. (ed.) *13th European Conference on Artificial Intelligence*. Wiley.

[44] Sanchez, L., Casillas, J., Cord, O., & Jose del Jesus, M. (2002). Some relationships between fuzzy and random set-based classifiers and models. *Int. J. Approx. Reason.*, *29*, 175–213.

[45] Şentürk, S. (2010). Fuzzy regression control chart based on a-cut approximation. *Int. J. Comput.Intell. Syst.*, *3*(1), 123–140.

[46] Shang, K., & Hossen, Z. (2013). Applying Fuzzy Logic to Risk Assessment and Decision-Making. Casualty Actuarial Society, Canadian Institute of Actuaries, Society of Actuaries, 2, 209-218.

[47] Shapiro, A.F. (2009). Fuzzy random variables. *Insur. Math. Econ.*, *44*, 307–314.

[48] van der Helm, R. (2008). Towards a clarification of probability, possibility and plausibility: How semantics could help futures practice to improve. *Foresight*, *8*(3), 17–27.

[49] Vladareanu, V. et al. (2019). *Adaptive Neural Network Fuzzy Inference System for HFC Processes, IUS Periodicals of engineering and natural sciences, PEN 7(1)*, 311-317.

[50] Yang, M.S., & Liu, M.C. (1998). On possibility analysis of fuzzy data. *Fuzzy Sets Syst.*, *94*, 171–183.

[51] Zadeh, L.A. (1965). Fuzzy sets. *Inf. Control*, *8*(3), 338–353.

[52] Zadeh, L.A. (2008). Is there a need for fuzzy logic? Info. Sci., 178(13), 2751–2779.

[53] Zimmermann, H. J. (2001). *Fuzzy Set Theory – and Its Applications* (4th Edition). New York, Kluwer Academic Publishers.