International Research Journal of Applied and Basic Sciences
© 2014 Available online at www.irjabs.com
ISSN 2251-838X / Vol, 8 (7): 963-971
Science Explorer Publications

IRJABS

An integrated fuzzy Data Envelopment Analysis and Data Mining for Performance assessment of Insurance branches

Somayeh Shafaghizadeh[1,*] , Mahdi Yousefi Nejad Attari[2]

1*. Department of Industrial Engineering, Science & Research Branch, Islamic Azad University, Qazvin, Iran
2. Faculty of Industrial and Mechanical Engineering, Bonab branch, Islamic Azad University, Bonab, Iran

corresponding Author email: s_shafaghizadeh@yahoo.com

ABSTRACT: This paper presents a Fuzzy Data Envelopment Analysis (FDEA) model combined with Bootstrapping to assess performance of one of the Data mining Algorithms. We used a two-step process for performance productivity analysis of Insurance Branches in Iran. First, using a Fuzzy Data Envelopment Analysis (FDEA) model, the study analyzes the productivity of eighteen Decision making units (DMUs). Using a Malmquist Index, FDEA determines the productivity scores but cannot give details of factors depend on Regress and Progress Productivity. FDEA model uses a new Latent Variable radial input-oriented technology and simultaneously reduces inputs and undesirable outputs in a single multiple objective linear program.On the other hand, the classification and regression tree (C&R) efficiency model was then utilized to extract rules for exploring and discovering meaningful and hidden information from the vast data bases. The conclusion of the combined model is a set of rules that can be used by policy makers to explore reasons behind the progress and regress productivities of DMUs.
Keywords: Fuzzy Data envelopment Analysis, Classification and Regression, Bootstrapping, productivity, Malmquist Index

INTRODUCTION

Using multiple inputs and outputs, Evaluation of efficiency and productivity of decision-making units (DMUS) such as banks, insurance companies, universities etc. is wrapped. There are some researches showing the importance of process in assessing the performance of a firm. Charnes et al. 1978 recommend a non-parametric approach for measuring the technical efficiency of a set of comparable DMUs.

DEA uses inputs/outputs variables to create an efficient boundary from a series of considered DMUs. A fuzzy set [Zadeh,1965] is usually identified by its characteristic function or membership function. (FDEA) model can be used to develop crisp data envelopment analysis models where uncertainty in classification problems is learned in the form of fuzzy membership function. The efficiency of each DMU is computed by measuring the distance of the DMU from the efficient frontier. Similarly Malmquist productivity Index has been used to evaluate technology change and its effect on inputs and outputs. It is defined as the maximum factor by which inputs in one period could be reduced to produce the same output in a second period.

On the other hand, data mining techniques extracting patterns from large databases have become prevalent. DT is a method usually used to find out meaningful communication and rules by systematically breaking down and subdividing the information in the data. A classification and regression tree (C&R) algorithm innovated by Breiman et al.(1984) is a hierarchical sequence of decision nodes. Each node in a tree strikes one decision at a time until a final node is achieved. Various variables are utilized and a special variable enters the computation only when it is required at a special decision node, and only one variable is utilized at each decision node.

That have applied these methodologies. Sohn & Moon (2004) proposed an approach that can be effectively used for foreshadowing the scale of new technology commercialization projects using the Decision Tree (DT) of data envelopment analysis (DEA) results. Similarly, Seol (2007) proposed and approach that enables firm' s manager to

find inefficient service units in a firm-level and inefficient process in a service unit-level by using an integrated form of DEA and DT. Using A Combination of (DT) and CCR DEA models, Emrouznejad (2010) evaluated the performance of Arabic banks and finally, Samoilenko (2013) tested a DEA-centric Decision Support System (DSS) in order to propose how to assess and manage the relative performances. Some researchers have proposed various fuzzy methods for dealing with the impreciseness and ambiguity in DEA. Fuzzy set algebra developed by Zadeh (1965) is the formal body of theory that allows the treatment of imprecise estimates in uncertain environments. Moheb-Alizadeh et al. (2011) defined the approach of efficiency by DEA into location–allocation models in a fuzzy environment and they showed how this combination can influence the pattern of facility location and the assignment of demands. Due to two main shortcomings, called low discrimination power and unrealistic weight distribution associated with classic form of DEA, a multi-criteria DEA is applied. Fuzzy classification using the data envelopment analysis, pendharkar (2012) developed a fuzzy classification system using data envelopment analysis(DEA) and illustrated application using a simple graduate admission decision-making problem.

In our study, we used FDEA and decision tree (DT) as our main methodologies. Indeed, we propose a two-stage performance evaluation applying FDEA, a non-parametric method by fuzzy model by fuzzy inputs and outputs and crisp efficiency for efficiency evaluation, and a C&R tree, a non-parametric data mining method for classification and regression. In doing so, we use this methodology to evaluate the performance of 18 insurance Branches of Iran insurance cooperation. Productivity scores provide valuable data for the performance assessment of insurance branches while the C&R tree determines further facts that have not been recognized in prior studies.

The proposed fuzzy model, non-parametric method of undesirable outputs with weak disposable inputs technology

Since the technology included the undesirable outputs, Bretholt & Pan (2013) introduced a new model Latent Variable(LV) radial input-oriented technology that is closely associated with a Koopmans Efficient Slacks Based Model. The Latent Variable technology simultaneously reduces inputs and undesirable outputs in a single Multiple Objective Linear Program.

could be built on the following principles:

Using p inputs $(x_{pj}, p=1,2,..,P)$ and producing Q desirable outputs $(y_{qj}, q=1,2,..,Q)$ and R undesirable outputs (z_{rj},r=1,2,…,R), assume that there are J branches of an insurance corporation.

Latent Variable technology uses a Radial Input Model in association with weak disposability applied to aggregate inputs. The weak disposability of inputs aggregate inputs, X_{pj} are reduced by the direct input reduction objective, α as follows:

$$\frac{\sum_{j-=1}^{J} Z_j X_{pj}}{\alpha_0 X_0} = 1 \qquad (1)$$

In this study, by using fuzzy decision rules with fuzzy data can be solved model(2):

$$\begin{aligned}
&VRS\ LV\ Min\alpha: \qquad (2)\\
&\{\forall DMU \parallel j=1,2,...,J : t=K,L\}\\
&s.t.\sum_{j=1}^{J} z_j \tilde{x}_{pj}^{t} = \alpha \tilde{x}_{p0}^{t}, \quad p=1,2,...,P\\
&\sum_{j=1}^{J} z_j \tilde{y}_{qj}^{t} \geq \tilde{y}_{q0}^{t}, \quad q=1,2,...,Q\\
&\sum_{j=1}^{J} z_j \tilde{u}_{rj}^{t} \leq \lambda \tilde{u}_{r0}^{t}, \quad r=1,2,..,R\\
&\sum_{j=1}^{J} z_j = 1\\
&Z\\
&z_j \geq 0\\
&Latent Vaiable \quad 0 \leq \lambda = \frac{\sum_{j=1}^{J} z_j \tilde{u}_{rj}^{t}}{u_{p0}^{t}} \leq 1\\
&\{\forall DMU \parallel r=1,2,...,R : p=1,2,...,P\}
\end{aligned}$$

In this article, a new model for evaluating the efficiency of outputs as inputs is proposed. Considering a shrinkage factor for the undesirable outputs, the dispersion between periods is studied using the Malmquist Productivity Index (MPI). After determination of the DMUs efficiencies, their productivities will be specified for the periods of 2008-2009 and 2009-2010 based on the following formulae.

The MPI is used to assess technology changes and change effect on the inputs and outputs

The largest factor MPI is defined by the inputs that can be reduced in one period and determine the same output's production in a second period. Suppose the production technology in the K period when the main reduction coefficient is as follows and the target values are in the L period.

$$\lambda_j^k(X_j^L, Y_j^L, U_j^L) \quad (3)$$

In general, the introduction of the DEA hidden variable technology is a first step towards the analysis of undesirable outputs and the consideration of external effects on the company and the society. Using the dense hidden variable reduction model, this model theoretically presents the production of simultaneous reduction of undesirable outputs and inputs through causal relationships. Equation (4) shows DMUs productivity results; accordingly, $MPI > 1$ indicates progress, $MPI = 1$ shows no change, and $MPI < 1$ is indicative of regress during the period.

METHODS

Combining FDEA with C&R tree

The proposed C&R tree in this study includes four main components

The first component, is the output (dependent) variable. Based on the independent (predictive) variables, this variable is used to predict.

In this study, the output variable is the obtained productivity scores that have been divided into three groups of progressive productivity (target>1), regressive productivity (target<1), and without change productivity (target=1).

The second component is the independent (predictive) variables. The number of independent variables is related to the purpose of investigation.

In this case, the independent variables are external and internal factors (Table 1).

The third component is the set of training data, which includes both output and independent variables values coming from a group of FDMUs we want to predict.

The fourth component is the test or the set of additional data coming from specific DMUs that require more precise prediction. This test data set may not exist in practice. It is normally believed that a test data set is required to enforce the decision laws; however, it is not always necessary to determine the efficiency of the decision rules.

Using DEA/C&R, the evaluation process of efficiency and productivity of insurance branches is presented in Figure(1). As shown in figure, first FDEA is applied to measure the efficiency and productivity of each branch with three inputs (administrative costs, insurance costs, and the number of branches) and three outputs (revenue from insurance premiums, the loan payments, the compensation payments). According to these results, the branches will be divided into three groups of efficient, inefficient, and without change branches.

In the next stage, insurance-related environmental factors such as age of the insurance branches, their ratings, and the number of issued insurance policies are considered as inputs to the C&R tree analysis while productivity scores, obtained in the first phase, are regarded as the outputs (Table 1). Clearly this is a general framework applicable in conducting all types of analyses in every organization including insurance companies and banks.

If this method is used for other purposes, both input and output variables in the first stage (FDEA) can be appropriately adjusted for the evaluation model. Therefore, the inputs to the second stage are supposed to be chosen according to the expectations of insurance experts and policy makers. The end results are usually a set of rules related to both input factors and FDEA productivity scores.

Table 1. Input factors in the C&R tree

Variable	Variable type	Min	Max	Mean	Std
Age of the insurance branches	Numerical	1386	1324	1372	12.93
Level of the branch[a]	Categorical	1	3		
Geographical branch[b]	Categorical	1	5		
No. of staff	Numerical	10	159	48.6	39.06
Qualifications of staff[c]	Categorical	1	5		
No. of loan	Numerical	2	1007	354.9	325.1
No. of insurance policies	Numerical	12	6894	1251	1780.5
No. of claim paid	Numerical	6	779	263.4	247.7

[a] 1, Assembled; 2, Super; 3, Level 1

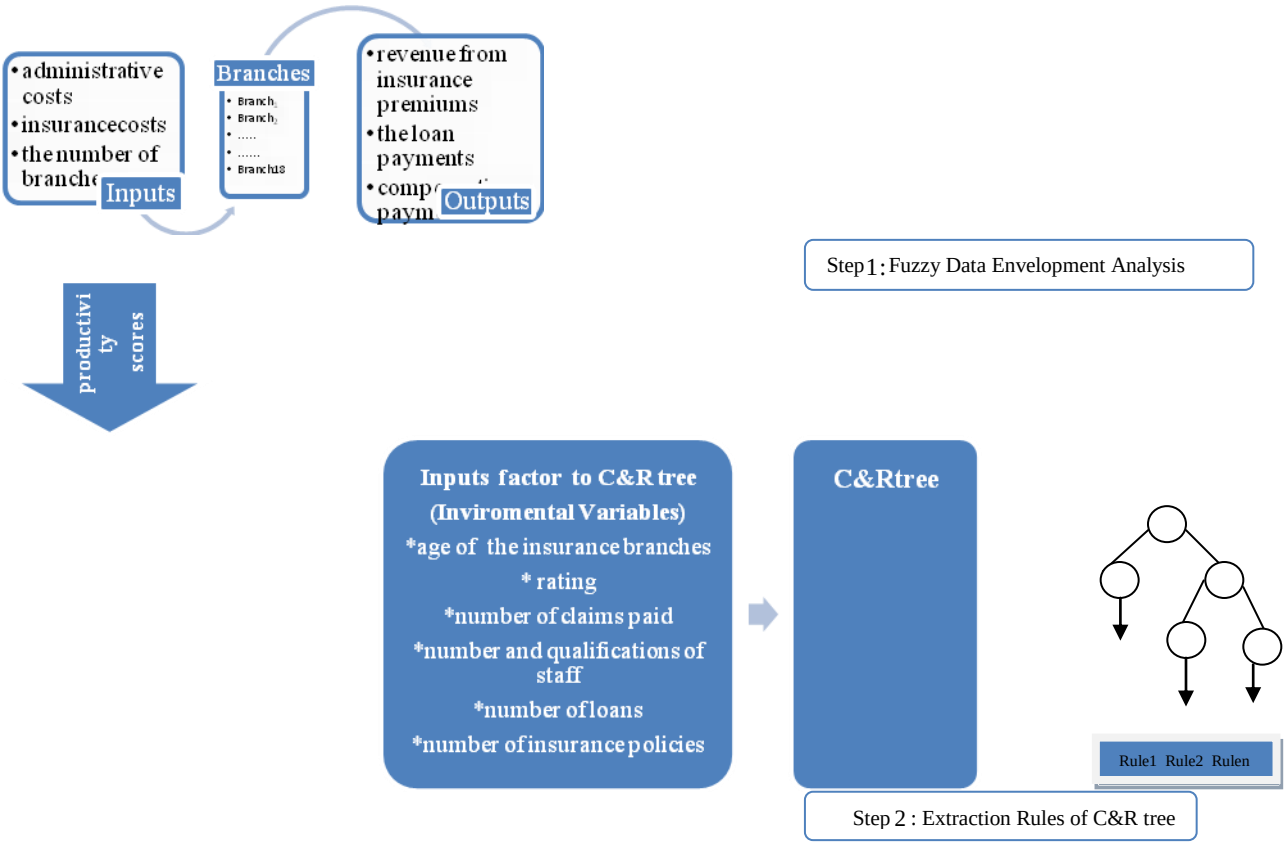

Figure 1. FDEA/C&R methodology for assessing Iran Insurance

DEA/C&R bootstrapping for evaluation of insurance branches

One of the problems of using DEA/C&R is that in many DEA studies, there are not sufficient data to generate a decision tree. In view of that, the Bootstrapping technique has been proposed to increase the number of DMUs before generation of a decision tree (Emrouznejad & Anouze, 2010). This method consists of three steps. First, the values of efficiency and productivity of each branch are calculated. Then, according to the obtained efficiency and productivity values, branches will be grouped into three classes of progressive productivity(target >1,MPI >1), regressive productivity (target<1, MPI<1), and without change productivity (target=1, MPI=1).

Producing an accurate C&R tree requires a large database. In case of the present study, only 18 insurance branches have been investigated; thus, by 100 times application of re-sampling bootstrapping technique (described by Efron and Tibshirani, 1993), the database is enlarged sufficiently.

Consequently, in the second step, 18 units (with replacement) are randomly chosen and the resampling bootstrapping technique is applied for 100 times to obtain 1800 units. After 100 times re-sampling, the data base is divided into training and testing groups with ratio of 7 to 3 (Zhou & Jiang, 2003).

In the third step, based on classified efficiency scores (<1 ،=1 ،1 >) as target variables and other uncontrollable variables (branches' rating, location, number of employees, etc.) as inputs to the C&R tree, the logical decisions are extracted.

EXPERIMENTAL RESULTS

FDEA (first stage)

In efficiency and productivity literature, the key factors to identify input and output variables of each insurance branch are its financial balance sheet and amounts of income, profits, and losses.

Indices used in this thesis were collected over a long period of time, with reference to every branch, and based on the managers' point of views (Table2). Then using the Latent Variable Model (LVM), efficiency and productivity of the insurance branches in the years 2008-2010were measured.

During the review process of productivity in the years 2008-2009, five branches displayed progression and 13 branches showed regressive trends. Similarly, in the years 2009-2010, 8 branches were productive and 10 branches were not. On average, 36.2% of the branches were productive and 66.2% were not. However, due to the high dispersion, all values of the input data were normalized before entering the tree for not reducing the prediction accuracy.

Table2. Input/output variables in DEA

Variable($)	Min	Max	Mean	Std
Inputs				
Administrative costs	11.94	1817808	179091	325464
Insurance cost	23.8	8861261	714697055	1480265
No. of branches	2	271	92.129	54.28
Outputs desirable	1085.3	12356881	1539634	2136065
Revenue from insurance premiums				
Loan payments	6519.033	6168493.033	1941592.7	1775229.3
Output undesirable	2756.8	6800070.1	1163839.4	1216054.7
Compensation payment				

Bootstrapping (second stage)

As mentioned before, 18 units (with replacement) were randomly chosen and the resampling bootstrapping technique was applied for 100 times to obtain 1800 units. This process led to a greater accuracy in the prediction of the C&R tree.

C&R analysis (third stage)

According to the FDEA, the insurance branches were divided into three groups of progressive productivity (1<MI), regressive productivity (1>MI), and without change productivity (1=MI). These groups are used as the target variable in the C&R tree.

Table(3)shows efficiency and productivity scores of the insurance branches in the C&R tree.

DISCUSSION AND CONCLUSION

For all attributes, impurity levels before and after the pruning are measured and the feature that further reduced the impurities is selected. The purity index is based on the least amount of impurities in each node. In consequence, multiple regression decision trees are plotted for each period.

Regression tree analysis in the years 2009-2010

Table (4) shows the predictive accuracy of the generated tree.

As stated by the prediction, in the years 2009-2010, out of the whole 1800 cases 997 cases had MPI<1 and 803 cases had MPI>1. Out of 1246 training data, 1246 cases were predicted correctly with the accuracy of 100%.

The overall accuracy of the prediction C&R tree was 100% indicating a high level of confidence.

In Figure(3), the generated C&R tree with 579 cases of progressive productivity, 667 cases of regressive productivity, and 8 nodes is presented.

In Figure(2), it can be observed that the number of paid losses in the year 2010 is the most important factor (67%)The age of the branches was the second important factor (30%) and the other variables (the number of employees with MA or PhD degrees, branch's rating, and the number of issued policies in the year 2010) were of equal importance (6%)

Table 3. Productivity scores LVM model by Malmquist Index(2009-2010)

DMUs	L	m	u	MIP
DMU1	1.79823	5.88022	8.90126	0.2675
DMU2	0.63973	2.09193	3.16668	0.3715
DMU3	0.6656	2.17651	3.29472	0.272
DMU4	0.5919	1.93551	2.92991	0.4712
DMU5	0.69093	2.25935	3.42012	0.2427
DMU6	0.7246	2.36944	3.58677	1.6227
DMU7	0.49887	1.63129	2.46939	0.2523
DMU8	0.4569	1.49406	2.26166	1.7953
DMU9	0.41523	1.35781	2.05541	0.5097
DMU10	0.6523	2.13302	3.22889	0.4258
DMU11	0.5745	1.87862	2.84378	0.2666
DMU12	0.79193	2.58962	3.92007	1.7225
DMU13	0.6422	2.09999	3.17889	1.4417
DMU14	0.95333	3.1174	4.719	1.6023
DMU15	0.66047	2.15973	3.26931	0.4574
DMU16	4.15117	13.574320	20.5483	8.0377
DMU17	2.5359	8.29239	12.5527	6.3628
DMU18	2.53837	8.30046	12.5649	1.1935

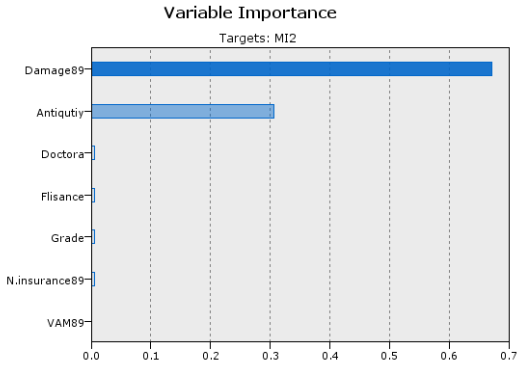

Figure 2. Importance of predictor variables(2009-2010)

Table 4. Predicted accuracy of the tree

Results for output field MI2
Comparing $R-MI2 with MI2

'Partition'	1_Training		2_Testing	
Correct	1,246	100%	554	100%
Wrong	0	0%	0	0%
Total	1,246		554	

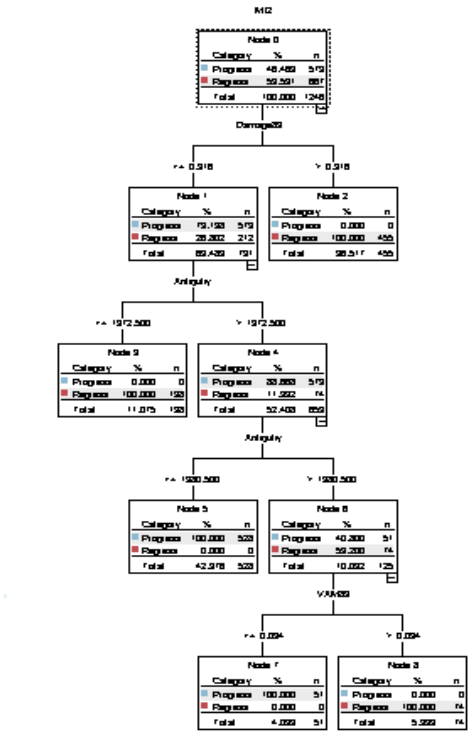

Figure 3 . C&R tree for Iran Insurance

Extracting rules for insurance branches with progressive productivity (579cases)

Rule 1: if the number of paid losses is smaller than or equal to 0.316 and the branch establishment year is between the years 1993and 2001,the branch has progressive productivity (528 cases).

Rule 2 : if the number of paid losses is smaller than or equal to0.316 , the branch establishment year is after the year 2001, and the number of paid loans is less than or equal to 0.034, the branch has progressive productivity (51cases).

Extracting rules for insurance branches with regressive productivity(667cases)

Rule 3 : if the number of paid losses in the year 2010 is smaller than or equal to 0.316 and the branch establishment year is before the year 1993, the branch has regressive productivity (138 cases).

Rule 4 : if the number of paid losses in the year 2010 is smaller than or equal to 0.316, the branch establishment year is after the year 1993, and the number of paid loans is more than 0.034, the branch has regressive productivity (74 cases).

Rule 5 : if the number of paid losses is bigger than 0.316, the branch has regressive productivity (455 cases).

Final evaluation

This thesis tries to introduce a combination of FDEA and C&R tree. In this study, 18 insurance branches in Tehran are examined.

In general, the efficiency and productivity scores can be obtained using FDEA and MPI. However, these methods cannot explain the related factors to inefficiency and unproductivity, especially in case of variables that are not numerical.

Considering factors associated with efficiency and productivity, C&R tree can present a better understanding of the FDEA results.

Despite the proposed method in the present study is examined in the insurance industry, it potentially has much broader applications.

Regarding DMUs' efficiency and productivity evaluation, the proposed DEA/C&R method can be applied as a framework for further research.

The results of this combined method are a set of rules, which can be applied by policy makers to explore the reasons behind DMUs' efficiency and inefficiency.

Creating a good and reliable C&R tree usually requires a large database and many observations; but in most of the reported DEAs, the number of DMUs is not large enough to generate a proper C&R tree. In order to solve this problem, the Bootstrapping method was proposed in this study. Nonetheless, further investigations seem quite necessary for an appropriate application of this method.

CONCLUSION

Fuzzy Data Envelopment Analysis (FDEA) is a management tool for efficiency and productivity assessment. This paper presented a framework for relating FDEA to classification and regression analysis. While the FDEA provides valuable and acceptable results, the C&R analysis reveals additional facts that were unclear in previous studies.

Unlike previous studies in the fields of FDEA and insurance industry that just tried to identify the impacts of different factors on the efficiency with the same impact level, the proposed C&R tree is based on the analysis of impact levels of different factors related to efficiency and productivity of insurance branches.

Exploring the variables' importance and influence on variables' dependence with the least amount of impurities to reach the target node (through the Clementine software), can lead to an in-depth analysis with the lowest amount of error by combining environmental factors with efficiency and productivity scores (obtained from the FDEA).

In previous studies, only the key parameters in the efficiency or inefficiency of the insurance branches have been evaluated and no environmental factor related to progressive/regressive productivity has been addressed yet. For example, the number of losses, paid loans, and age of the branches are not considered as important factors in the efficiency/inefficiency issue; however, according to the extracted rules, they are influential variables in the efficiency/inefficiency evaluation of the insurance branches with different impact levels.

Furthermore, using numerical and categorical variables with different degrees of importance, rules were extracted for each specific DMU and used to identify productivity or unproductivity of the selected insurance branches.

Unlike previous studies on FDEA applications, which focused only on the numerical calculations of efficiency and productivity, this paper studied factors related to efficiency and productivity of insurance branches, using C&R tree. In addition, possible rules were extracted for every DMU, using both numerical and categorical variables by fuzzy value. Obviously these rules are very useful for policy makers and can improve their decision-making processes.

Future studies

There are a number of additional issues of practical importance to those who study C&R trees (independent factors for the insurance sectors, application of various rules and accurate measurement, and improvement of the Bootstrapping method). Despite these issues have not been addressed in the current investigation, their inclusion in other studies can broaden the field for the development of future studies.

In future research, databases with larger sample size can be chosen to avoid using the Bootstrapping method. It must be noted that the use of simulation in this paper was one of the limitations.

Fuzzy decision tree can be used instead of crisp decision tree because it offers beneficial results in case of insurance industry's qualitative data. It is also possible to set the DEA efficiency and productivity results as output variables. Moreover, depending on the type of data and the importance of input variables, other trees such as, and can be used.

REFERENCES

Breiman L, Friedman J, Olshen R, Stone C.1984. Classification and Regression Trees, Pacific Grove, CA: Wadsworth-Monterey.

Bretholt A, Pan J.2013. Evolving the Latant Variable model as an environmental DEA technology, Omega $41, 315-325.$

Charnes A, Cooper WW, Rhodes E.1978. Measuring the efficiency of decision making units. European Journal of Operational Research, $2(6), 429-444.$

Emrouznejad A, Anouze A.2010. Data envelopment analysis with classification and regression tree- a case of banking efficiency, Expert System the Journal of Knowledge Engineering, 231-246.

Moheb-Alizadeh H, Rasouli SM, Tavakkoli-Moghaddam R.2011. The use of multi-criteria data envelopment analysis (MCDEA) for location–allocation problems in a fuzzy environment.Expert Systems with Applications. 38, 5687–5695.

Pendharkar P.2012.Fuzzy classification using the data envelopment analysis, Knowledge-Based Systems31(2012)183-192.

Samoilenko S, Osei-Bryson KM.2013. Using Data Envelopment Analysis (DEA) for monitoring efficiency-based performance of productivity-driven organizations: Design and implementation of a decision support system, Omega $41,131-142.$

Seol H, Choi J, Park G, Park Y.2007. A framework for benchmarking service process using data envelopment analysis and decision tree, Expert System with Application, 432-440.

Sohn SY, Moon TH.2004. Decision tree based on data envelopment analysis for effective technology commercialization, Expert Systems with Applications, 24,279-284.

Zadeh LA. 1965. Fuzzy sets, Inform. and Control 8 (3) 338–353.