DATA PROCESSING AND

ANALYTICS IN THE DIGITAL AGE

DATA PROCESSING AND ANALYTICS IN THE DIGITAL AGE: part 1*

Tijana Dragojević, Jovana Svičević, Sandro Radovanović

CLASSIFICATION OF LIFE INSURANCE USERS USING DATA MINING

Marija Kuzmanović, Dragana Makajić-Nikolić

PREFERENCES OF WINE CONSUMERS TOWARDS LOCAL WINE BRANDS: CASE OF SERBIA

Milica Maričić, Milica Bulajić, Marina Dobrota

EASE OF DOING BUSINESS AND GROSS DOMESTIC PRODUCT: IS THERE A RELATIONSHIP?

Ivana Ivković, Vesna Rajić

CONFIDENCE INTERVALS FOR THE POPULATION STANDARD DEVIATION: SIMPLE RANDOM SAMPLING VS. RANKED SET SAMPLING

Nikola Cvetković, Nebojša Dragović, Aleksandar Đoković

FIELD STRESS DETECTION ALGORITHM USING REMOTE SENSING

Nikola Zornić, Aleksandar Marković

CRYPTOCURRENCY PRICE FORECASTING USING TIME SERIES AND MONTE CARLO MODELINGAND SIMULATION

Nikola Vojtek, Ana Poledica, Bratislav Petrović

STATISTICAL AND SOFT COMPUTING TECHNIQUES IN AIRLINE INDUSTRY – A LITERATURE REVIEW

Jovana Kuljanin, Milica Kalić, Manuel Renold

THE IMPACT OF LOW COST CARRIER ON COMPETITION IN LONG HAUL MARKET: LONDON - NEW YORK ROUTE

Strahinja Radaković, Milan Radojičić, Milica Maričić

MULTIVARIATE APPROACH TO MAKING SPONSORSHIP DECISIONS: THE CASE OF EUROPEAN FOOTBALL LEAGUES

Marko Prodanović, Damjan Rovinac, Stefan Radibratović

MEASURES OF DIGITALIZATION IN EUROPEAN ENTERPRISES: LINEAR REGRESSION MODEL

Višnja Istrat, Dajana Matović, Milko Palibrk

RESEARCH OF ASSOCIATION RULES AS DECISION MAKING TOOL FOR MANAGERS

Boris Delibašić, Sandro Radovanović, Miloš Jovanović

SKI LIFT TRANSPORTATIONS AS PREDICTORS FOR INJURY OCCURRENCE

Anja Bjelotomić, Aleksandar Rakićević, Ivana Dragović

DECISION TREE-BASED ALGORITHM FOR THE CLASSIFICATION OF MUSICAL INSTRUMENTS

Slaviša Arsić, Dragan Pamučar, Milija Suknović

DETERMINING THE WEIGHTS OF CRITERIAIN MENU EVALUATION USING BEST-WORST METHOD

Stefan Vujović, Danijel Mišulić, Sofija Krneta

ANALYSIS AND PREDICTIONOF VIEWSIN YOUTUBE INTERVIEWS

Matija Milekić, Aleksandar Rakićević, Pavle Milošević

NEURAL NETWORKS IN MARKET SENTIMENT ANALYSIS FOR AUTOMATED TRADING: THE CASE OF BITCOIN

Dušica Stepić

EXPERIMENTALCOMPARISON OF MULTI-LABEL LEARNING METHODS

Ivan Rakić, Željana Milošević, Slađan Babarogić

DATA MINING USING ORACLE DATA MINER AND ANALYTIC FUNCTIONS WITH HADOOP

Jelena Ljubenović, Ognjen Pantelić, Ana Pajić Simović

BIG DATA ANALYSIS IN SOCIAL MEDIA

Sofija Prokić, Jelena Ljubenović

QUERY PROCESSING ASPECT IN HETEROGENEOUS DBMS

CLASSIFICATION OF LIFE INSURANCE USERS USING DATA MINING

Tijana Dragojević[1], Jovana Svičević[1], Sandro Radovanović[1]
[1]Faculty of Organizational Sciences, University of Belgrade
Corresponding author, e-mail:sandro.radovanovic@gmail.com

Abstract: *The significance of life insurance as the safest form of investment is reflected in the benefits to both the individual and society. The topic of this research includes creation and evaluation of a few different models for classification of potential and existing life insurance users into one of 8 possible risk categories. The dataset which was used to carry out the research has been downloaded from Kaggle.com competition, Prudential Life Insurance. The construction and evaluation of the aforementioned models have been conducted using CRISP-DM methodology. Due to the characteristics of the dataset, which included 128 attributes, and specific application area, attributes section will be applied, as well as algorithm parameters optimization. The whole research has been conducted using open source web application Jupyter Notebook, which uses Python code language. It comes to the conclusion that XGBoost algorithm outperforms other algorithms used in this paper.*

Keywords: *Life insurance, multiclass classification, attribute selection, parameter optimization, principal component analysis*

1. INTRODUCTION

The problem of predictions in the field of life insurance lies in the fact that life insurance is a type of insurance that must be maximally adjusted to each user. This requires detailed analysis of all aspects of human life. In addition, it must be taken into account that the prediction is carried out for a very long period of time, which is an additional aggravating circumstance.

The dataset which will be analyzed is Prudential Life Insurance Assessment, one of the most famous life insurance companies. In America, only 40% of the households own a life insurance, and since life insurance payouts are quite big, and the repayment period long, it is clear why the insurance companies give great importance to previously conducted detailed research which also includes medical examinations. People refuse to take the medical exam which takes 30 days on average, while the classical insuring of life insurance is considered outdated.

Therefore, as a principal business problem, in the specific case of the Prudential Life Insurance Company, the selection of corresponding indicators so that the classification of potential users would be carried out significantly quicker while respecting user's privacy and personal data, is pointed out. The goal is to predict the value of the Response variable for each user Id that is to assign one of 8 values of the nominal Response variable to each Id value.

The complexity of this problem lies in the fact that it is a large dataset, and what is specific is that most of the data contain medical and financial information which represents a sensitive area of human life. Therefore, the selection of attributes as one of the approaches can contribute to a great saving of time and money.

In the following chapters, more will be said about the preparation and analysis of data as well as the modeling of the solution using a variety of different algorithms

2. LITERATURE REVIEW

According to the literature on this topic, we have come to the conclusion that data mining in the field of life insurance is used in four situations:

- Acquiring new customers.
- Retaining existing customers.
- Performing sophisticated classification
- Correlation between Policy designing and policy selection

Significant conclusions were made in Devale and Kulkarni article Applications of data mining techniques in life insurance (Devale,A.B, Kulkarni, Dr.R.V., 2012). They recommend KNN algorithm for performing sophisticated classification. Also, they emphasize advantages of linear regression, which may be used to solve a problem because insurance industry regulators require easily interpretable models and model parameters.

Since the set of data contains a large number of attributes, the speed of execution of the algorithm is a very important component. Within his article on the potential for machine learning in the prediction of insurance policy sales, Adrian B.F. Ampt notes that the statistical analysis has two problems (Ampt, 2017). It is too slow and the analysis is prone to mistakes. Ultimately the Decision Tree algorithm and Logistic regression could generate model quickly enough and make accurate predictions. The results are not as accurate as statistical analysis in some cases, but the results were competitive and consistent. The models could also be generated much faster than the statistical analysis which takes weeks to set up for a new dataset, whereas machine learning could make predictions within the eight-hour window, although often within a minute and even seconds. Finally, machine learning was able to handle irrelevant features very well, which means that the data scientist does not necessarily have to comb through the data to pick relevant features.

2. METHODOLOGY

3.1. Data understanding and data preparation

The dataset was first imported into a Jupyter notebook (Kluyver, et al., 2016) and all the libraries which will be used in the project work were also imported.

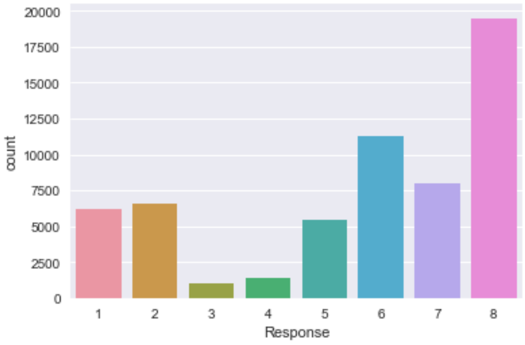


Figure 1: Distribution of classes of the output variable

It can be deducted from the image that there is an imbalance of classes (Satyasree, K.P.N.V., Murtthy J.V.R., 2013) and that classes 3 and 4 are significantly less represented than others, while class 8 counts the most.

A great number of missing values has been noted, which happens very often in real data sets and this presents a problem for algorithm application. In a Jupyter notebook, the presence of missing values can be checked by using the function *isnull()*.

The dataset contains following types of data:

- Categorical(nominal),
- Discrete and
- Continuous variables.

The variable which is predicted has an ordinal measure scale. This means that it is of a categorical type, with defined hierarchical relationship (ranked categories). Intervals between consecutive values are not necessarily the same.

So as to better understand the dataset, some of the variables will be presented and described:

- ID: key which uniquely identifies each line
- Product_Info_1-7: Normalized values pertaining to the type of product, that is, to the type of insurance
- Ins_Age: Normalized values of the age of potential clients
- Ht: Normalized height
- Wt: Normalized weight

- BMI: Normalized body mass index
- Employment_Info_1-6: A set of normalized values which refer to the employment history of the logged client
- Insured_Info_1-6
- Insurance_History_1-9
- Family_History_1-5
- Medical_History_1-41
- Medical_Keyword_1-48
- Response

Using the heat map graph, the correlation between all of the attributes is shown, so as to note which pairs of attributes are more correlated. It strives towards the elimination of highly related attributes since they can provide an unrealistic image about the efficiency of the applied algorithm.

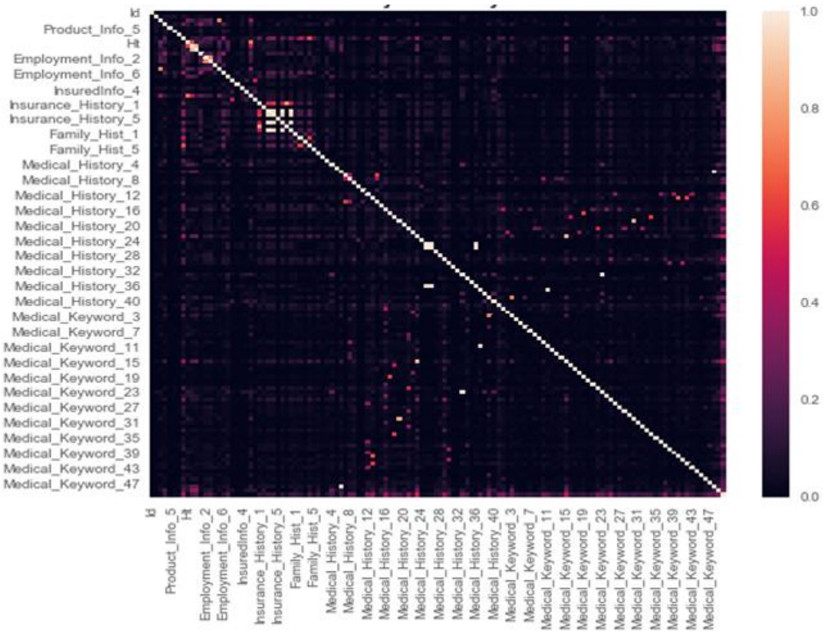
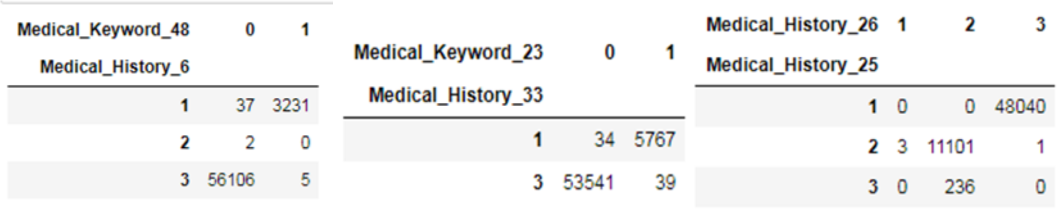


Figure 2: Display of the correlation between attributes

As expected, it is noted that there is a strong correlation between the pair of attributes Wt (weight) and BMI. Using analysis it has been determined that the three greatest correlations between pairs of attributes are:

- Medical_History_6, Medical_Keyword_48
- Medical_History_25, Medical_History_26
- Medical_History_33, Medical_Keyword_23

It can be concluded that the case of high correlation between attributes Medical History and Medical Keyword is probably about the medicinal state which is described using a specific term (Medical Keyword), therefore, about a kind of derived attributes. So as to better understand the above-mentioned correlation coefficients, it has been decided that the link between these attributes be shown in cross tables as well, which are shown in the images below.



Medical_Keyword_48	0	1
Medical_History_6		
1	37	3231
2	2	0
3	56106	5

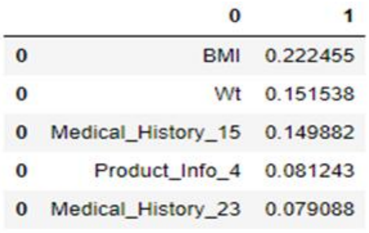
Medical_Keyword_23	0	1
Medical_History_33		
1	34	5767
3	53541	39

Medical_History_26	1	2	3
Medical_History_25			
1	0	0	48040
2	3	11101	1
3	0	236	0

Figure 3: Crosstables

Taking into consideration that the size of the dataset is 59,381, it is very significant that in most cases, as much as 56,106, the attribute Medical_History_6 assumes value 3 when Medical_Keyword_48 assumes value 0. The situation is similar to the rest of the attributes which have high correlation coefficients.

Mutual information was calculated. That is a measure which measures in what degree one attribute determines another attribute, specifically in this case, in what degree do the input variables determine the output variable (Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., 2016). The mutual information concept is connected with the entropy of an accidental variable, a basic term in the information theory, which determines the quantity of information that is kept in an attribute. The attribute which carries the most information, of all the output attributes, is BMI. It is interesting that the next ranking is Weight, which is, according to a previously calculated coefficient, highly correlated with the attribute BMI. Next, to these two attributes, three more are shown:



	0	1
0	BMI	0.222455
0	Wt	0.151538
0	Medical_History_15	0.149882
0	Product_Info_4	0.081243
0	Medical_History_23	0.079088

Figure 4: Mutual information

As part of the data preparation phase, the following steps were implemented:

1. Replacement of missing values

Since missing values (Kang, 2013) have been noticed inside of the dataset, the first step in the transformation phase is replacing missing values with some other values, other words, filling the lines. Since the occurrence of missing values can be a significant indicator too, here follow the attributes for which there is no data for all of the instances:

• Medical_History_10	99.061990
• Medical_History_32	98.135767
• Medical_History_24	93.598963
• Medical_History_15	75.101463
• Family_Hist_5	70.411411
• Family_Hist_3	57.663226
• Family_Hist_2	48.257860
• Insurance_History_5	42.767889
• Family_Hist_4	32.306630
• Employment_Info_6	18.278574
• Medical_History_1	14.969435
• Employment_Info_4	11.416110
• Employment_Info_1	0.031997

It is noticed that the missing values in a large number of cases occur with attributes which refer to data about the financial status of the subject, which could indicate wrong data collection methods or the sensitiveness of the subjects concerning this question group. Even though there are multiple methods to conduct the mentioned procedure, in this work, the method of filling the missing values with average values from the column in question was used. To that purpose, the *fillna()* method is used.

2. Defining the input and output variable

Set of the variable will be divided between a rising variable X(all variables, except the variable we are predicting and Id which does not carry information which helps predict the output variable) and Y(variable for which the predicting is done).

3. Adding new, derived variables

Also, it has been noticed that the attribute Product_Info_2 consists of a character part and a numerical part, therefore, two new columns will be created Product_Info_2_number and Product_Info_2_char, where categorical values will be placed in one column, and numerical in the other one, while the so-called dummy variables Medical_Keyword_1 –48 will be placed in a new variable Medical_keywords_sum. Besides that, for each of the attribute categories, new dummy variables have been derived, which expands the data set to a total of 942 attributes.

4. Principal component analysis

After that, for each of the attribute categories, principal component analysis has been conducted. For the attribute group InsuredInfo, 7 principal components define more than 90% variance in data, for MedicalKeyword 34, for MedicalHistory 2, for FamilyHistory 1, for EmploymentInfo 5, for InsuranceHistory 2 and for ProductInfo 3. All the principal components are then merged into a new dataset which will be the subject of algorithms.

3.2. Experimental Setup

When the data preparation phase is over, we move over to the next phase – modeling the solution and cross-validation. Since the output variable was categorical, the classification problem is to be solved. For that reasons, algorithms which are used to solve classification problems will be used. Since the Response variable has 8 values, that are 8 classes to which the corresponding instances need to be sorted, we are dealing with multiclass classification. Based on the dataset, its characteristics and the characteristics of the mentioned algorithms for multiclass classification, seven algorithms which will be used and which are expected to provide relatively good results have been selected, whether in the case of performance, execution speed that is or in the case of successfulness of the classification. The mentioned algorithms are:
Random Forest (Ho, 1995), Gradient Boosted Trees (Friedman, 2002), Extra Trees, Extreme Gradient Boosted Trees (Chen, T., He, T., Benesty, M., 2015), Logistic Regression, Ridge Logistic Regression and Lasso Logistic Regression (Tibshirani, 1996).

For the classification problem, the most commonly used measures of the classification of successfulness are:

- Confusion Matrix
- Accuracy
- Precision and Recall
- F measure

Since it has been determined that there is an imbalance between the classes, there measure F1 is selected (Powers, 2011), which combines precision and recall, with the same priority, is assigned to both. The usage of the accuracy measure is avoided since it can create a false image of the efficacy of the model and present it as better than it really is. The F1 score also enables for a simpler comparison of two or more algorithms.

$$F1 = 2\frac{precision * recall}{precision + recall} \tag{1}$$

Since we are dealing with multiclass classification, so we could use the F1 score, it is needed to add the parameter average and calibrate it. Possible values of this parameter are:

- micro (calculates metrics on the global scale, counting the total number of TP, FN, FP)
- macro (calculates metrics for each output attribute while ignoring their imbalance)
- weighted (calculates metrics for each label and restores their average)

Of the above mentioned, a parameter has been applied with all three possible values, but only for Random Forest, while, for measuring successfulness of classification for other algorithms, the focus is on the F1 score with a micro parameter.

4. RESULTS

In the first phase, the mentioned algorithms have been applied on a new dataset which comprises of principal components of all groups of attributes, while the parameter default values have been left as they are. In the table below, shown are the resulting values of the F1 score for all seven algorithms with default parameters, as well as values of the F1 score derived from applying cross-validation. It can be seen that the Gradient Boosted Trees algorithm proved best, thus, in the next phase which includes parameter optimization, the most attention has been dedicated to this algorithm.

Table 1: Algorithms before optimization

Algorithms:	F1 score(micro)	F1 score(macro)	F1 score(weighted)	Cross validation(F1 micro, cv =5)
Random Forest	0.4526	0.3401	0.3830	0.4312
Gradient Boosted Trees	0.4777	0.3970	0.4258	0.4611
Logistic Regression	0.4314	0.3041	0.3614	0.4282
Ridge Logistic Regression	0.4042	0.1847	0.3157	0.4030
Extra Trees	0.4767	0.3323	0.4145	0.4107
Extreme Gradient Boosted Trees	0.4605	0.3613	0.3965	0.4535
Lasso Logistic Regression	0.4315	0.3041	0.3614	0.4285

Parameter optimization and attribute selection have been carried out only on the few selected algorithms, above all, so as to improve algorithm performance and reduce the overfitting problem of the training set. Specifically, with algorithms such as Random Forest, Extra Trees, Extreme Gradient Boosted Trees, calibration of the number of trees for the algorithms has been carried out. As for the attribute selection, it has been carried out for Ridge Logistic Regression (mainly for practical reasons, since this algorithm proved as the quickest one). Also, the selection of attributes on the basis of the mutual information measure has been tried out, where only those attributes which carry the most information have been selected.

Parameter optimization implied setting the number of estimators and the maximum depth of the tree. In addition, there is a grid search (Bergstra, J., Bengio, Y., 2012) for algorithms with internal cross-validation. It is significant that the algorithm as the optimum value for the parameter number of trees has taken the highest value from the offered, 140 while the optimal depth of the tree is 6.

Results after optimization:

Table 2: Algorithms after optimization

Score for fold	1.RandomForest	3. Logistic Regression	4.Ridge Logistic Regression	5.ExtraTrees	6.Extreme Gradient Boosted Trees	7.Lasso Logistic Regression
1	0.454	0.430	0.405	0.439	0.473	0.427
2	0.457	0.434	0.409	0.442	0.476	0.435
3	0.451	0.427	0.401	0.432	0.464	0.426
4	0.445	0.427	0.398	0.429	0.463	0.424
5	0.445	0.425	0.402	0.430	0.463	0.425
average	0.450	0.429	0.403	0.434	0.468	0.427

5. CONCLUSION

The main problem when working on a dataset was a large number of attributes that significantly slowed down the execution of most algorithms. That's why the focus was on attribute selection. In addition, an important role in improving performance is taken in the detailed preparation of data. An analysis of the principal components was performed, but it is important to emphasize that this analysis was performed on a group of attributes that related to the same aspect of the life of the respondents. The resulting main components for each of the 7 groups are then merged into a new set of data. This step has significantly contributed to improving the results, reducing the problem of overfitting and execution time.

As previously mentioned, in the scope of this research different algorithms have been applied, including ensemble algorithm, and it proved that they, in fact, gave the best results. Based on the results, it was concluded that the selection of attributes has a very large impact on the efficacy of applied algorithms;

therefore, further workings should pay more attention precisely to this phase of research. Besides principal component analysis, which has proved as a very good solution for reducing the number of attributes to which the algorithm is applied, it would be interesting to explore the possibilities of attribute selection based on mutual information indicators as well. Additionally, even though, due to technical reasons, it was not possible to conduct a more detailed optimization of parameters and define more values, significant improvements is possible concerning this aspect.

Although performance metrics, namely F1, seems very low with values less than 0.5 it is worth to notice that there were 8 different classes. Having this in mind, a naïve approach to solving would yield approximately 0.125 values. Therefore, obtained performances are indeed good. Additionally, it is worth to notice that best-performing algorithms were state-of-the-art ensemble algorithms such as Gradient Boosted Trees, Extreme Gradient Boosted Trees and Random Forest which is an indicator that problem we tried to solve is highly challenging and non-linear since the F1 score was lower for optimized Lasso Logistic Regression and Ridge Logistic Regression.

REFERENCES

Ampt, A. (2017). On the potential for machine learning in prediction of insurance policy sales - Helping insurance intermediaries get insights in their clients' insurance needs. Eindhoven: Department of Mathematics and Computer Science, Eindhoven University of Technology.

Bergstra, J., Bengio, Y. (2012). Random Search for Hyper-Parameter Optimization. Journal of Machine Learning Research, 281-305.

Chen, T., He, T., Benesty, M. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2.

Devale,A.B, Kulkarni, Dr.R.V. (2012). Applications of data mining techniques in life insurance. International Journal of Data Mining & Knowledge Management Process, 35-37.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis - Nonlinear methods and data mining, 367-378.

Ho, T. (1995). Random decision forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 278-282.

Kang, H. (2013). The prevention and handling of the missing data. Korean Journal of Anesthesiology, 402-406.

Kluyver, T., Ragan-Kelley, B., Perez, F., Granger, B., Bussonnier, M., Frederic, J., i drugi. (2016). Jupyter Notebooks - a publishing for for reproducible computational workf. Positioning and Power in Academic Publishing: Players, Agents and Agendas (str. 87-90). Göttingen, Germany: IOS Press.

Powers, D. M. (2011). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. Journal of Machine Learning, 2(1), 37-63.

Satyasree, K.P.N.V., Murtthy J.V.R. (2013). An Exhaustive Literature Review on Class Imbalance Problem. International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), 110-115.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society,Series B (Methodological), 267-288.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

PREFERENCES OF WINE CONSUMERS TOWARDS LOCAL WINE BRANDS: CASE OF SERBIA

Marija Kuzmanović*[1], Dragana Makajić-Nikolić[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: marija.kuzmanovic@fon.bg.ac.rs

Abstract: *The wine market presents the consumer with a range of products and product attributes to consider when making a purchase decision. The study presented in this paper surveys the importance consumers in Serbia attach to both extrinsic and intrinsic attributes of the local wine brands. Five wine attributes were examined and tested through a discrete choice experiment. Two hundred forty wine consumers were asked to choose a wine they were most likely to purchase to have with friends and family. The results at the aggregate level indicate a high importance of the attributes brand and type of wine. Furthermore, preference-based segmentation based on individual respondents' utilities identified three distinct segments emerged from the overall sample.*

Keywords*: Wine, preferences, attributes, conjoint analysis, discrete choice experiment, segmentation.*

1. INTRODUCTION

In recent years, there is a growing interest in wine consumer behaviour. The dynamic nature of the wine sector increasingly attracts both practitioners and academics to undertake analysis of the different phases of the wine consumption process. This dynamics is reflected primarily in the growth and diversification of the offer, reduction of consumption in traditional countries, and emergence of new producer and consumer countries (Martínez, Mollá-Bauzá, Gomis, & Poveda, 2006). Moreover, in developed societies, consumer behaviour becomes diverse and creates continuous changes, both socio-economics and changes in modern lifestyles. These processes have led to a different patterns in the consumption of alcoholic beverages, especially wine (Caniglia, D'Amico, & Peri, 2006). Namely, while there has been a reduction in consumption of table wine, on the other hand there has been greater demand for quality wines and wines with regional specificity.

The consumer choice for wine is more complex than the choice for many other products, due to the large amount of different cues that may influence the purchase decision (Lockshin & Hall, 2003). The literature focusing on wine choice differentiates between two categories of choice attributes: extrinsic and intrinsic attributes (MacDonald, Saliba, & Bruwer, 2013). Intrinsic attributes are those associated with the vintage, type of grape, year and sensory characteristics of a wine (e.g. taste, flavour, sugar contentand colour), and extrinsic cues refer to non-sensory characteristics such as price, region of origin, brand and packaging (Lu, Rahman, & Chi, 2017).

Previous studies have identified numerous factors that have been found to have an impact on the wine selection process (Gustafson, Lybbert, & Sumner, 2016). Moreover, the specificity of wine as a product affects the willingness of consumers to even try out a particular type of wine (Everett, Jensen, Boyer, & Hughes, 2018). Most of the researches have focused on the importance different product attributes have on consumers when purchasing wine in the retail stores (Goodman, 2009), although there are also studies researching factors that influence the wine choice in the on-premise purchasing (restaurants, bar, café), depending on the price or information provided on the menu (Corsi, Mueller, & Lockshin, 2012). Surveys have suggested that price and grape variety are often the most influential variables when choosing a wine in retail stores, whereas, packaging and label design have been reported to be of lesser importance (Thomas & Pickering, 2003). The region of origin was found to be the most important driver for an on-premises wine choice by Australian, French and Italian consumers (Corsi, Mueller, & Lockshin, 2012). However, Lockshin and Corsi (2012) argued that consumers seem to be less confident when purchasing wine in a restaurant than in a store.

The vast majority of studies on wine consumer behaviour are focused on red (Mehta & Bhanja, 2018) or white (Saliba, Wragg, & Richardson, 2009) wine. However, recent market and industry trends show a growing popularity of rosé wine among consumers around the world, and thus, an increase in its production, but also researches (Kolyesnikova, Dodd, & Duhan, 2008).Price has always been considered one of the most important drivers of consumer choice, and wine is no exception. In particular, price allows customers to make inferences about the quality and value of a product (Gustafson, Lybbert, & Sumner, 2016; Chrea, et al., 2011). Lockshin, Rasmussen, and Cleary (Lockshin, Rasmussen, & Cleary, 2000) highlight the fact that the brand name acts a surrogate for a number of attributes including quality and acts as a short cut in dealing with risk and providing product cues. Bruwer, Li, and Reid(Bruwer, Li, & Reid, 2002) concluded that wine markets have been segmented based on nine major segmentation variables: quality, consumption, risk reduction, occasion based, cross-cultural, behavioural, involvement, geographic, wine-related lifestyles. Using Conjoint analysis, Mehta and Bhanja (2018) identified five factors (price, brand, taste, origin and type of the wine) as important in the choice of wine. Moreover, authors identify price as the most important factor, followed by the type of the wine whereby red was the most preferred type. Escobar, Kallas, and Gil(2018)used Generalised Multinomial Logit Model (GMNL)to determine the impact of the 2008 economic crisis on preferences of the citizens of Catalonia towards four wine attributes: wine origin, wine references, grape variety, and price. The research showed that the wine origin was the most important attribute before the crisis, while the price became the most important attribute during the crisis.

The purpose of this paper is to empirically determine the consumers' preferences towards both extrinsic and intrinsic attributes of local wines in Serbia, as well as to determine whether those preferences are heterogeneous. To our knowledge, several research related to wine consumers' preferences were conducted in Serbia. Different methods have been used to investigate the relative importance of cues on consumer wine choice and purchase behaviour. Vlahović, Potrebić, and Jeločnik(2012) implemented poll research survey on a sample of 150 respondents in Belgrade in 2011 with goal to perceive factors that influence on demand of wine; Radovanović, Đorđević, and Petrović(2017) used Descriptive statistics and Chi-test to analyse the data collected in 2015. In this paper, Conjoint analysis, the techniques commonly used to analyse consumer preferences will be used. Conjoint analysis implies that consumers have to show their preferences to a set of products or profiles created through a combination of product attributes and attribute levels. The growing wine market in Serbia presents a remarkable opportunity for marketers to formulate a strategy targeted at the Serbian consumers.

2. METHODOLOGY

This study focuses on Serbian wine consumers' habits and preferences. An online discrete choice experiment was conducted on individuals who consumed wine at least once in the past 12 months.

2.1. Conceptual framework

Conjoint analysis is one of the most widely used research techniques which helps uncover how people make choices and what they really value in products and services. It originated in mathematical psychology, and was first introduced in marketing research in the early 1970s to evaluate consumer preferences for hypothetical products and services. Nowadays, it is widely used to understand customers' preferences in various markets both service and manufactures (Kuzmanovic, Radosavljevic, & Vujoševic, 2013; Vukic, Kuzmanovic, & Kostic-Stankovic, 2015).The foundation of conjoint analysis is breaking a product or service down into its components (attributes) and then testing combinations of these components in order to find out what customers prefer. It is then possible to estimate the value of each component in terms of its effect on customer decisions.

The most common type of conjoint analysis is Discrete Choice Analysis (often called choice modelling or choice-based conjoint analysis). Rather than merely asking respondents what they like in a product, or what features they find most important, discrete choice experiment employs the more realistic task of asking respondents to choose between potential product concepts (i.e. combinations of attributes and levels) carefully assembled into choice sets. Each respondent is typically presented with 8 to 12 choice sets (Samoylov, 2017).

The output from discrete choice analysis is measurement of utility or value. The utility scores are numerical values that measure how much each attribute and level influenced the customer's decision to make that choice.

2.2. Attributes used in the experiment

The first step in Discrete Choice Analysis is identifying key product attributes and corresponding levels. As noted above, trying to model all the influences on wine purchase behaviour is complex and the total number of attributes could make the design of such an experiment much too large to be practical. The literature review highlighted that the most important wine attributes are price, region of origin, and brand name. In this paper, in accordance with the aim of the research, the following attributes were taken into account: Winery, Price, Type, Sweetness and Sparkling (see Table 1).

We developed the levels of the attribute Winery from available data on the market share of brands in the Serbian wine market. Namely, there is a considerable number of small family winery in Serbia, whose wine is considered to be of high quality and popular among the consumers. Therefore, in addition to an industrial winery with a tradition (Rubin), five more wineries has been taken into account. Prices on the Serbian wine market for these producers range from 400 RSD to 1600 RSD, thus the four price points were selected for the attribute Price. Depending on grape variety, three types of wine are distinguished. Thus, for an attribute Type we choose three levels: Red, White and Rosé. Wines can be made with a wide range of sweetness levels, from dry to sweet one. The subjective sweetness of a wine is determined by the interaction of several factors, including the amount of sugar in the wine, but also the relative levels of alcohol, acids, and tannins. According to EU regulation 753/2002 (eur-lex.europa, 2002), the following terms may be used on the labels both table and quality wines (depending of the sugar content): Dry, Medium dry, Medium and Sweet. Based on the amount of carbon dioxide in the wine, it can be sparkling or non-sparkling. Sparkling wine is a wine with significant levels of carbon dioxide in it, making it fizzy.

Table 1: Key attributes and corresponding levels

Attribute	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Winery	Rubin	Kovačević	Radovanović	Aleksandrović	Zvonko Bogdan	Matelj
Type	Red	Rose	White			
Sweetness of wine	Dry	Medium dry	Medium	Sweet		
Sparkling	Sparkling	Non-sparkling				
Price	400 RSD	800 RSD	1200 RSD	1600 RSD		

2.3. The choice experiment design and survey technique

Based on the attribute and their levels, total of 576 (=6x3x4x2x4) virtual wine concepts could be constructed. However, it is unrealistic for respondents to compare and select from such a great number of tasks. In general, respondents will fatigue after comparing more than 15 concepts. Experimental design methods and conjoint.ly platform (Samoylov, 2017) were used to structure the presentation of the levels of the five attributes in the choice sets. An implemented algorithm is used to insure each level of each attribute appears nearly an equal number of times across all surveys, but does not repeat in the other product concepts in each choice task. Because the number of choice sets is excessive for one respondent, the experiment is split into seven blocks, each consisted of 10 different choice tasks, i.e. 10 different purchase decisions.

The participants were provided with three product concepts plus the “None of the above” in each choice task and were asked to select the bottle of wine they would choose to buy to have with friends and family. Each product concept had one level of each of the five attributes. Respondents were given a short survey along with the choice tasks. Socio-demographic data (age, gender, household income, level of education) and wine consumption habits (frequency of drinking, place, quantity, wine type) were also collected.

2.4. Analytical and segmentation method

Discrete choice models (DCM) can be derived from utility theory. A DCM specifies the probability that an individual chooses a particular (wine) concept, with the probability expressed as a function of observed variables that relate to the concepts and the individual. The assumption is that the behaviour of the individual is utility-maximizing: individual *i* chooses the concept that provides the highest utility. The choice of the person is designated by variables y_{ij} for each alternative:

$$y_{ij} = \begin{cases} 1, & U_{ij} > U_{ik},\ \forall j \neq k \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

U_{ij} is the utility that person *i* obtains from choosing concept *j*.:

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{2}$$

$V_{ij}$is the part of utility associated with the observed factors influencing it, whereas ε_{ij}represents the unobserved sources of utility. These unobservables can be characteristics of the individuals and/or attributes of the item, and they can stand for both variation in preferences among members of a population and measurement error (Hanemann & Kanninen, 2001). Then the probability of individual *i* choosing alternative *j* from a set of *J* mutually exclusive alternatives choice is given by:

$$P_{ij} = P(U_{ij} \geq U_{ik} \mid k \in J)\ \ \forall k \neq j \tag{3}$$

To estimate the model parameters (part-worths), multinomial logit model or Hierarchical Bayes (HB) estimation can be used. Amongst other benefits of HB, this approach allows more parameters (attributes and levels) to be estimated with smaller amounts of data collected from each respondent (Samoylov, 2017).Estimated part-worths can be further used to assess the relative importance of each attribute for each respondent or group of respondents. These values are calculated by taking the utility range for each attribute separately, and then dividing it by the sum of the utility ranges for all of the attributes (Kuzmanovic, Savic, Popovic, & Martic, 2013).

Rather than use socio-demographic variables to define segments a priori and to test whether these groups differ in their preferences and behaviour, we thought the opposite approach, post hoc segmentation, would be more effective, because it derives segments from differences in their preferences. Segmenting on behavioural differences has been found to be more robust and stable over time. For choice experiments, segments that differ in their choice drivers can be found with K-means cluster analysis, which will be applied in this paper here to find consumer clusters that differ in their preferences.

3. RESULTS

3.1. Sample characteristics

Data were collected online using Conjoint.ly platform in June 2017. In total, 256 individuals answered the survey. After the elimination of incomplete surveys and ineligible participants, 240 eligible surveys were collected. The sample characteristics as well as respondents' habits concerning wine consumption are presented in Table 2.

Table 2: Socio-demographics data and respondents' habits

Demographic	Category	Number of respondents	Percent
Gender	Male	123	51.25%
	Female	117	48.75%
Age	18-20	17	7.08%
	21-40	194	80.83%
	41-60	21	8.75%
	>61	8	3.33%
Level of education	Primary school	9	3.75%
	High school	43	17.92%
	Undergraduate	140	58.33%
	Master degree	45	18.75%
	PhD degree	3	1.25%
Employment status	Student	103	42.92%
	Unemployed	16	6.67%
	Employed	116	48.34%
	Retired	5	2.09%
Averaged monthly income	do 25000	34	14.17%
	25000-50000	80	33.33%
	50000-75000	58	24.17%
	>75000	68	28.33%
Frequency of wine consumption	Every day	5	2.08%
	Several times a week	30	12.50%
	Once a week	62	25.83%
	Once a month	82	34.17%
	Rarely	61	25.42%
Quantity	1 glass	45	18.75%
	2 glasses	90	37.50%
	3 glasses	42	17.50%
	at least 4 glasses	63	26.25%
Place of consumption	At home	56	23.33%
	At friends'/family place	55	22.92%
	In clubs/pubs	47	19.58%
	In restaurant	82	34.17%
Type of wine	White	109	45.42%
	Red	66	27.50%
	Rose	65	27.08%

With regard to the frequency of wine consumption, ithas emerged that 34.17% respondents declared they drinkwine "once a month" and 25.83% of them declared they drink wine "once a week" (25.83%), while only 12.5% % of consumers declared that they drink wine "several times a week" or even every day (2.08%).Respondents declare that most often consume wine in restaurants (34.17%), and most of them drink white wine (45.42%). As much as 81.67% of respondents are willing to experiment with new brands and types of wine. Only one third of the respondents (32.5%) do not take into account the choice of the type of wine depending on the food they consume, while 31.65% of the respondents declare that the choice of the wine type affects the season. More than half of the respondents (60%) mostly buy domestic (Serbian) wine brands.

3.2. Aggregated preferences

Results from the analysis are shown in Figure 1 and Figure 2. Figure 1 is the graph description of the attributes importance. It can be seen that most important attribute to Serbian consumers is the Winery, and its average importance value at the aggregate level is even 59.43%. The attribute Type has shown to be second by importance (24.69%). The third ranked attribute is Sweetness, with a relative importance of 8.8%. The least important attributes are Sparkling and Price with importance values of just 4.92% and 2.16% respectively.

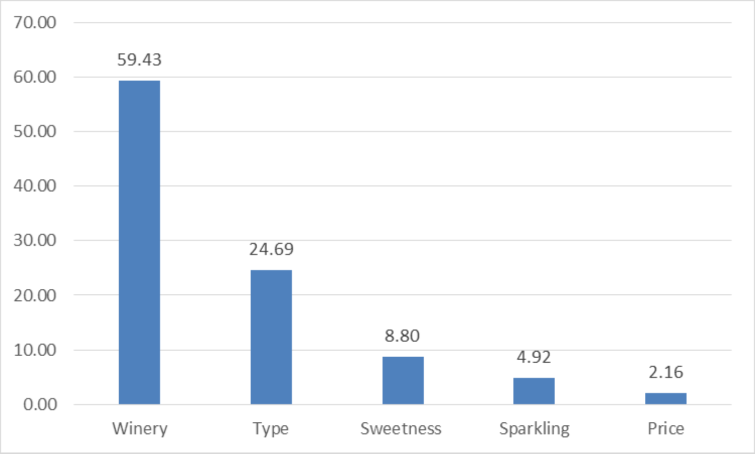


Figure 1: Relative importance of attributes (in %)

A more detailed insight into averaged preferences toward attribute levels (part-worths) is given in Figure 2.When it comes to the most significant attribute, Winery, respondents most prefer Kovačević, followed by Radovanović. The least desirable are Matalj and Rubin. On average, respondents almost equally prefer white and rose, and at least red wine, tilting with sweet wines, and most like medium and sparkling. When it comes to the attribute Price, at the aggregate level, there is no significant difference between the levels (price points), and at first glance the respondents are not price sensitive.

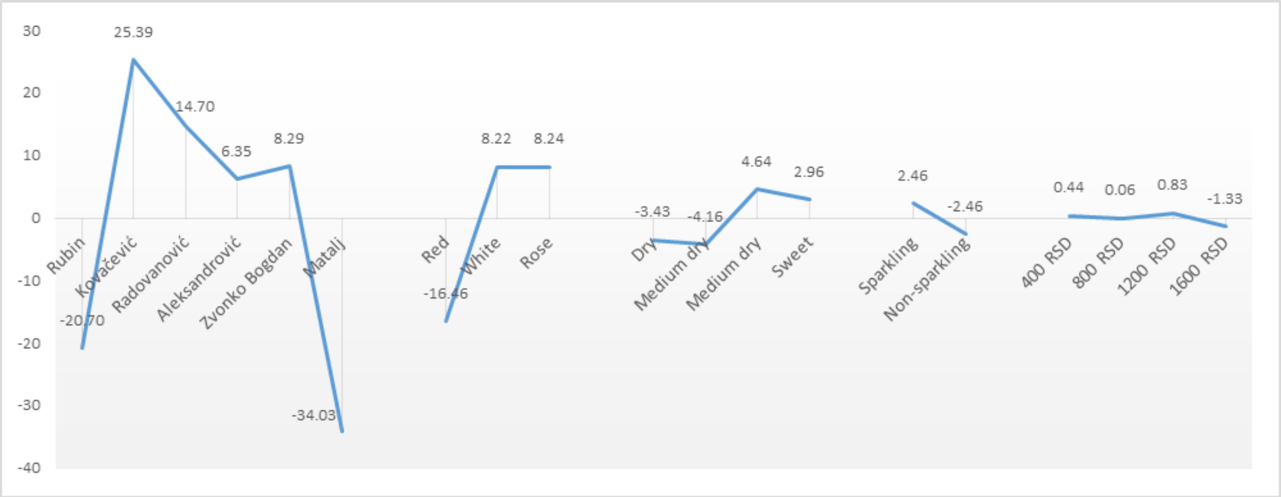


Figure 2: Averaged part-worth utilities

3.3. Post-hoc segmentation

A more detailed analysis of individual utilities revealed heterogeneity in consumer preferences, so three groups of consumers were isolated using *post hoc* segmentation (segmentation based on preferences). K-means cluster analysis was used for that purpose. Table 3 shows the importance of attributes for each of the isolated cluster.

Table 3: Relative importance of attributes (in %)

	Winery	Type	Sweetness	Sparkling	Price
Cluster 1	33.10	**35.16**	10.99	8.61	12.14
Cluster 2	22.17	**38.22**	17.78	4.40	17.43
Cluster 3	**39.57**	31.91	2.80	7.27	18.44

Although the price proved to be the least (negligible) important attribute on the whole sample, it can be noted that it is significantly more important at the segment level. This result indicates that averaging can cause loss of important information related to the actual preferences of the respondents, which may be reflected in the wrongly defined market strategy. Similar observations are made for other attributes as well. After the clusters were developed, socio-demographic data were used to further profile consumers.

The Cluster 1 covers 26.25% of the respondents and consists mostly of employed male respondents, who regularly drink wine, most often at home. This cluster includes respondents who especially prefer the red wine (which is in sharp contrast with the sample as a whole) and usually consume it (54%), so it is not surprising that they find the Type to be the most important attribute (35.16%). Somewhat less important is the attributes of Winery (33.10%), where respondentsprefer Kovačević, Radovanović and Aleksandrović wine brands. The other three attributes are significantly lessimportant to this cluster. However, they prefer non-sparkling dry wines, with the price of up to 800rsd (see Figure 3)

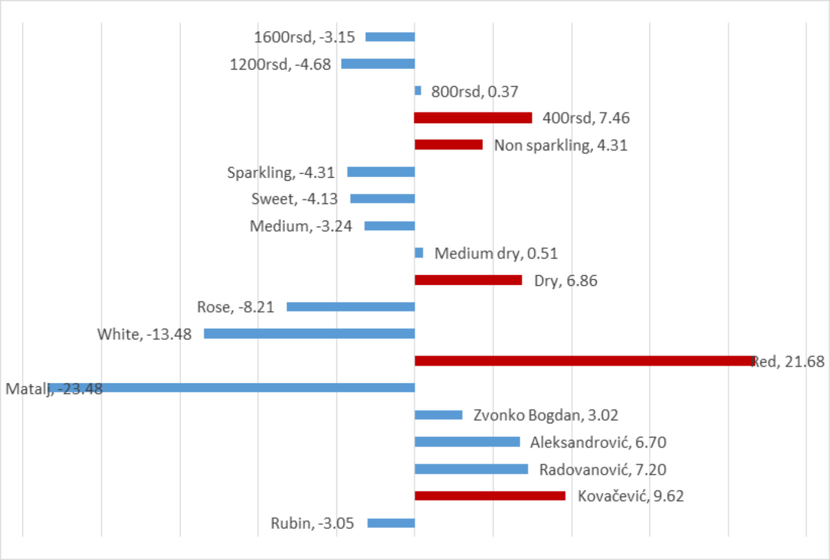


Figure 3: Preferences of Cluster 1

The second, somewhat larger cluster consists of 73 respondents (30.42%). The most important attribute for this segment is also Type (importance value = 38.22%), but respondents belonging to this cluster prefer rose wine, and to some extent white, while red wine is not preferred to them (see Figure 4). The Winery attribute is considerable less important for this cluster than for Cluster 1. In the same time, members of the Cluster 2 prefer the Kovačević brand wine by far more than the brands of other wineries. Sweetness and price are approximately equally important attributes (about 17%), whereby respondents prefer sweeter wines at a price of 400rsd. They are very price-sensitive, and expensive wines (from 800rsd) are negatively preferred. If they chose, they would rather choose sparkling wines.

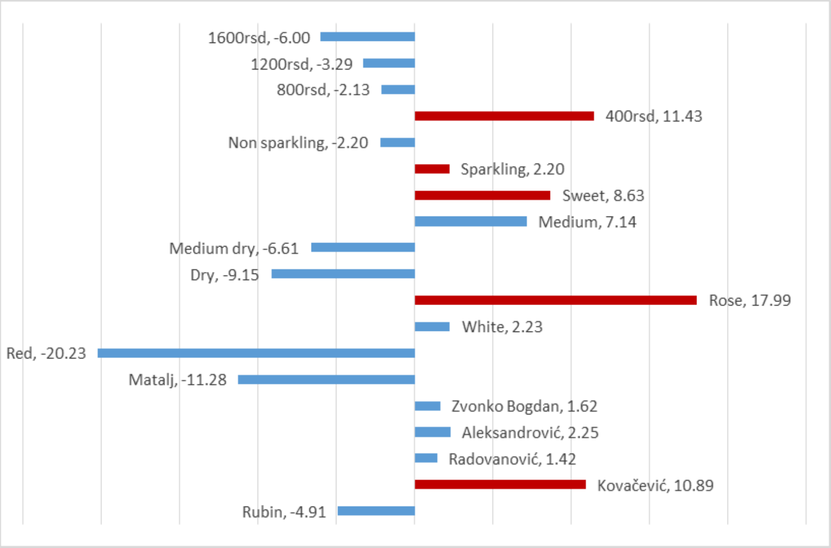


Figure 4: Preferences of Cluster 2

The demographic data shows that the majority of the respondents in this segment are younger women, with lower earnings and education than the remaining two segments. Members of this segment are less likely to drink wine than members of the Cluster 1, and when they drink it, they do it in a restaurant or in friends' home, and rarely at their own place. They usually drink rose or white wine, while only 15% drink red.

Cluster 3 is the largest one (43.33% of total sample) with Winery as the most important attribute (39.57%). Members of this cluster prefer Kovačević and Radovanović brands, followed by Zvonko Bogdan. Again, Matalj and Rubin are the least popular wine brands (see Figure 4). Type of wine is the second by importance attribute, where the respondents prefer white wine, but do not like red. This cluster is price-sensitive, but the respondents' preferences to price levels are unexpected. Namely, respondents prefer more expensive wines, which can be due to the fact that they use price to make inferences about the quality and value of a wine. Moreover, the quality wines of the favourite Serbian producers are exactly in the price range that the respondents prefer. Although sweetness is negligible important attribute for this cluster (only 2.8%), respondents prefer semi-sweet sparkling wines.

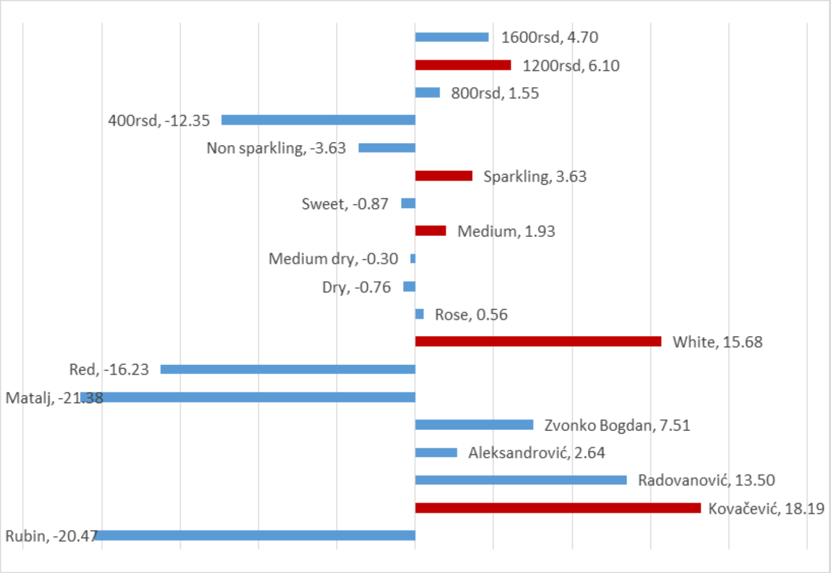


Figure 5: Preferences of Cluster 3

4. CONCLUSION

Wine is a difficult and confusing product for consumers to choose because of a large number of different cues that can influence purchasing decision. These cues are typically related to physical characteristics of wine as well as extrinsic attributes such as price and brand, labels and the like. Accordingly, understanding how consumers choose wine is a complex problem both for researchers and practitioners.

In this paper we have examined the importance consumers in Serbia attach to the key attributes of local wine brands. In addition to the brand, four attributes were also observed: price, type of wine, sweetness and whether it is sparking or not. To elicit consumer preferences, we used conjoint analyses, i.e. one of its forms known as discrete choice experiment.The average results on the whole sample indicated the high importance of the attributes brand and type of wine. However, heterogeneity in preferences were noticed and three unique segments were identified. These segments differ primarily in the type of wine their favour, but also whether they like sweet and sparkling wine or not. It turned out that the price is a moderately significant attribute in all three clusters, while Kovačević is the most preferred brand.

The findings in this study represent first empirical insights that examine the preferences of consumers towards wine characteristics using discrete choice analysis in Serbia.However, since 40% of the sample drinks wine at most once a week, the question arises as to whether the preferences of the true wine connoisseurs are the same. The future directions of the research should be towards eliciting the preferences of the connoisseurs of fine wines.

REFERENCES

Bruwer, J., Li, E., & Reid, M. (2002). Segmentation of the Australian wine market using a wine-related lifestyle approach. *Journal of Wine Research, 13*(3), 217-242.

Caniglia, E., D'Amico, M., & Peri, I. (2006). An analysis of consumers' perception of the quality of Doc Etna wine. *3rd International Wine Business & Marketing Research Conference.* Montpellier: France.

Chrea, C., Melo, L., Evans, G., Forde, C., Delahunty, C., & Cox, D. N. (2011). An investigation using three approaches to understand the influence of extrinsic product cues on consumer behavior: An example of Australian wines. *Journal of Sensory Studies, 13*(24), 13-24.

Corsi, A. M., Mueller, S., & Lockshin, L. (2012). Let's see what they have... what consumers look for in a restaurant wine list. *Cornell Hospitality Quarterly, 53*(2), 110-121.

Escobar, C., Kallas, Z., & Gil, J. M. (2018). Consumers' wine preferences in a changing scenario. *British Food Journal, 120*(1), 18-32.

eur-lex.europa. (2002). Retrieved April 7, 2018, from COMMISSION REGULATION (EC) No 753/2002: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2002R0753:20071214:EN:PDF

Everett, C., Jensen, K., Boyer, C., & Hughes, D. (2018). Consumers' willingness to pay for local muscadine wine. *International Journal of Wine Business Research, 30*(1), 58-73.

Goodman, S. (2009). An International Comparison of Retail Consumer Wine Choice. *International Journal Of Wine Business Research, 21*, 41-49.

Gustafson, C., Lybbert, T., & Sumner, D. (2016). Consumer knowledge affects valuation of product attributes: Experimental results for wine. *Journal of Behavioral and Experimental Economics, 65*, 85-94.

Gustafson, C., Lybbert, T., & Sumner, D. (2016). Consumer sorting and hedonic valuation of wine attributes: exploiting data from a field experiment. *Agricultural economics, 47*(1), 91-103.

Hanemann, W., & Kanninen, B. (2001). The statistical analysis of discrete-response CV data. In I. Bateman, *Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation in the US, EC and Developing Countries* (pp. 302-441). Oxford University Press.

Kolyesnikova, N., Dodd, T., & Duhan, D. (2008). Consumer attitudes towards local wines in an emerging region: a segmentation approach. *International Journal of Wine Business Research, 20*(4), 321-334.

Kuzmanovic, M., Radosavljevic, M., & Vujoševic, M. (2013). Understanding Student Preferences for Postpaid Mobile Services using Conjoint Analysis. *Acta Polytechnica Hungarica, 10*(1), 159-176.

Kuzmanovic, M., Savic, G., Popovic, M., & Martic, M. (2013). A New Approach to Evaluation of University Teaching Considering Heterogeneity of Students' Preferences. *Higher Education, 66*(2), 153-171.

Lockshin, L. S., & Hall, J. (2003). Consumer purchasing behaviour for wine: what we know and where we are going. *International Colloquium in Wine Marketing* (pp. 1-21). Adelaide, South Australia: University of South Australia, Wine Marketing Research Group.

Lockshin, L., & Corsi, A. M. (2012). Consumer behaviour for wine 2.0: A review since 2003 and future directions. *Wine Economics and Policy, 1*(1), 2-23.

Lockshin, L., Corsi, A. M., Cohen, J., Lee, R., & Williamson, P. (2017). West versus East: Measuring the development of Chinese wine preferences. *Food Quality and Preference, 56*, 256-265.

Lockshin, L., Rasmussen, M., & Cleary, F. (2000). The nature and roles of a wine brand. *Australia and New Zealand Wine Industry Journal, 15*(4), 17-24.

Lu, L., Rahman, I., & Chi, C. (2017). Ready to Embrace Genetically Modified Wines? The Role of Knowledge Exposure and Intrinsic Wine Attributes. *Cornell Hospitality Quarterly, 58*(1), 23-38.

MacDonald, J., Saliba, A., & Bruwer, J. (2013). Wine choice and drivers of consumption explored in relation to generational cohorts and methodology. *Journal of Retailing and Consumer Services, 20*(3), 349-357.

Martínez, L. M., Mollá-Bauzá, M. B., Gomis, F. J., & Poveda, Á. M. (2006). Influence of purchase place and consumption frequency over quality wine preferences. *Food Quality and Preference, 17*(5), 315-327.

Mehta, R., & Bhanja, N. (2018). Consumer preferences for wine attributes in an emerging market. *International Journal of Retail & Distribution Management, 46*(1), 34-48.

Radovanović, V., Đorđević, D. Ž., & Petrović, J. (2017). Wine Marketing: Impact of Demographic Factors of Serbian Consumers On the Choice of Wine. *Economic Themes, 55*(2), 199-215.

Saliba, A. J., Wragg, K., & Richardson, P. (2009). Sweet taste preference and personality traits using a white wine. *Food Quality and Preference, 20*(8), 572-575.

Samoylov, N. (2017). Conjoint.ly, online discrete choice experimentation and conjoint analysis tool. Sydney: Conjoint.ly. Retrieved from http://conjoint.online/

Schäufele, I. &. (2017). Consumers' perceptions, preferences and willingness-to-pay for wine with sustainability characteristics: A review. *Journal of Cleaner production, 147*, 379-394.

Thomas, A., & Pickering, G. (2003). The importance of wine label information. *International Journal of Wine Marketing, 15*(2), 58-74.

Vlahović, B., Potrebić, V., & Jeločnik, M. (2012). Preferences of wine consumers on Serbian market. *Economics of Agriculture, 59*(1), 37-49.

Vukic, M., Kuzmanovic, M., & Kostic-Stankovic, M. (2015). Understanding the Heterogeneity of Generation Y's Preferences for Travelling: a Conjoint Analysis Approach. *Journal of Tourism Research, 17*(5), 482-491.

EASE OF DOING BUSINESS AND GROSS DOMESTIC PRODUCT: IS THERE A RELATIONSHIP?

Milica Maričić*[1], Milica Bulajić[1], Marina Dobrota[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: milica.maricic@fon.bg.ac.rs

Abstract: *Well-functioning legal and regulatory system is a prerequisite for creating an effective market economy. Therefore, it is in the interest of nations to improve their business regulation and stimulate business activity and growth. One of the means for measuring the ease of doing business is using the Doing Business Index (DBI) devised by the World Bank. In this paper, we examine the relationship amonggross domestic product (GDP) per capita and DBI topics which measure the ease of conducting business processes. Our approach is two-fold. First, we apply a machine learning algorithm, a clustering approach to cluster the countries ranked by the DBI, and second, we employ the Potthoff analysis to compare the regression models of GDP per capita made on retained clusters. The results of the conducted analyses indicate there are differences in the effects of factors of doing business on GDP per capita between clusters. We believe our research could provide additional insights on the topic of factors which influence GDP, application of composite indicator data, and comparison of regression models between groups.*

Keywords*: Doing Business Index, Gross domestic product, Potthoff analysis, Machine learning*

1. INTRODUCTION

Legal and regulatory system is a prerequisite for creating an effective market (Corcoran & Gillanders, 2015). Research has shown that more complicated and more costly procedures have a negative effect on the business environment and on the number of new companies. Namely, Klapper, Laeven, and Rajan (2006) showed that higher entry regulations decrease the number of new companies created. Next, Djankov et al. (2002) found that economies with higher entry costs have more corruption and larger percentage of unofficial economy. Ciccone and Papaioannou (2007) showed that more complex procedures reduce job creation. The same authors draw the conclusion that therefore the simplification of time-consuming government procedures related to doing and starting new businesses should be high on policy agendas.

In the last few decades, additional efforts have been placed to provide a ranking of countries, regions, or even cities based on their openness to business (Brunetti, Kisunko, & Weder, 1997; Davis, Kingsbury, & Merry, 2012). Besides simple performance indicators which are commonly used for ranking, recently composite indicators stood out as a reliable source of information and ranking, especially in the public sector (Jacobs & Goddard, 2007). One of the earliest such ranking of countries based on doing business and economy dates from 1990's. Namely, Cavusgil (1997) proposed a ranking of emerging markets using seven dimensions: Market size, Market intensity–economic intensity, Market growth–future market potential, Commercial infrastructure, Freedom (economic and political) and risk, Market receptivity –market accessibility, Market construction capacity. Mullen and Sheng(2006) aimed to extend and modernise the original study conducted by Cavusgil (1997) by employing novel and more specific indicators in their framework. Besides academics, the internationalorganisations also devise and publish composite indicators on the subject of ease of doing business. For this research, the most important one is the World Bank's Doing Business Index (DBI) published since 2003 (World Bank, 2017).

Gross domestic product (GDP) measures the monetary value of market production, of final goods and services produced for sale in a particular market and of nonmarket production, such as defence or education services provided by the government (International Monetary Fund, 2017). As such a metric, it provides information on the size of the economy and its performance. Therefore, it is used as an indicator of health of the economy whereas growth of the GDP implies the economy is doing good (International Monetary Fund, 2017). As the forecasting of future economic outcomes is a vital for decision-makers (Dritsaki, 2015), the question which arises is what influences the GDP and the growth of GDP. Different approaches have been taken to provide an answer. For example, time series analysis was used (Dritsaki, 2015). Schumacher & Breitung(2008) used expectation–maximization (EM) algorithm and principal component analysis (PCA) to forecast German GDP. Linear regression has also been used with success to model GDP. For example, Anghelache and coauthors(2015) aimed to regress final consumption and gross investment on GDP. Lakštutienė and Aušrinė(2015) modelled the GDP per capita using indicators of the capital market.

The academic community placed attention on the relationship between the GDP and the reform of business regulation and ease doing business. Namely, it is believed that the ease of doing business attracts foreign direct investments (FDI) which eventually have an impact on the GDP (World Bank, 2016). Many studies have been conducted to examine the relationship between FDI and gross domestic product (GDP) and showed positive relationship (Basu, Chakraborty, & Reagle, 2003; Hsiao & Shen, 2003; Nair-Reichert & Weinhold, 2001). Also, it has been shown that the ease of doing business and business regulation have an impact on the GDP. Namely, Djankov, McLiesh, & Ramalho(2006) in their model used as dependent variable the average GDP growth and as independent the normalized components of the World Bank's DBI. Also, Busse & Groizard(2008) used linear regression to model the GDP using FDI and the selected components of the World Bank's DBI. Their results indicate that the countries with lower level of regulation can stimulate growth by reforming. The two presented studies are important for the herein conducted research for two reasons. First, they show that there is relationship between GDP and ease of doing business which can be explored in more detail and second, that the values of World Bank's DBI

Our goal was to attempt to model the GDP per capita using the values of the DBI topics. However, our analysis aims to go a step further. Namely, we first propose the application of machine learning algorithm to divide countries ranked by the DBI. We collected the data on 187 countries, so our aim was to first group these countries into clusters and to compare the regression models built on each of the retained clusters. The proposed clustering approach is the Partition around medoid (PAM) algorithm (Kaufman & Rousseeuw, 1990) while the statistical analysis used to compare regression models is the Potthoff analysis (Potthoff, 1966). We believe that the obtained results can indicate whether there are differences between the regression models and which topics should policy makers from different clusters place more attention so as to increase their GDP.

The structure of the paper is as follows. Section 2 sees the introduction and a brief review of the DBI. The main concepts of the PAM clustering algorithm and the Potthoff analysis used to compare regression models are given in Section 3. Next, the research results are elaborated, while the discussion and the concluding remarks are given in the final section.

2. DOING BUSINESS INDEX (DBI)

Doing Business Index (DBI) is a complex composite indicator consisted of 10 topics made of 41 indicators in total. In our research, we focused solely on topic data. Herein, we will present the basics of each topic.

The first topic, *Starting a business*, is intended to measure the number of procedures, time, cost, and paid-in minimum capital to start a limited liability company. Next topic, *Dealing with construction permits* measures the number of procedures, time, and cost to complete all formalities to build a warehouse and the quality control and safety mechanisms in the construction permitting system. *Gettingelectricity*, quantifies procedures, time, and cost to get connected to the electrical grid, the reliability of the electricity supply and the transparency of tariffs. *Registering property* is related to procedures, time, and cost to transfer a property and the quality of the land administration system. The following topic, *Getting credit* measures movable collateral laws and credit information systems. *Protecting minority investors* deals with minority shareholders' rights in related-party transactions and in corporate governance. Payments, time and total tax and contribution rate for a firm to comply with all tax regulations as well as post-filing processes are quantified by the topic *Payingtaxes*. Time and cost to export the product of comparative advantage and import auto partsare measured in the topic *Trading a cross borders*. Succeeding topic, *Enforcing contracts*, measures how commercial disputes are resolved and the quality of judicial processes. The final topic, *Resolving insolvency*, deals with time, cost, outcome and recovery rate for commercial insolvency and the strength of the legal framework for insolvency(World Bank, 2017).So, to make the presentation of the results more clear (see Section 4), the topics have been coded from T1 to T10, in the order in which they have been briefly explained here.

The data collection process is based on a detailed reading of domestic laws and regulations as well as administrative requirements. The process itself is guided and overlooked by the World Bank experts. A signal that the data collected for the DBI is precise and of interest is the fact that there are 17 different data projects or indexes that use Doing Business Index data as one of their data sources (World Bank, 2017). For example, in a similar study, Morris and Aziz (2011) aimed to measure the relationship between the DBI topics and the FDI using the Pearson's correlation coefficient based on the results of Sub-Saharan Africa. Nevertheless, Arruñada (2007) argues that the procedure of indicator and topic data collection could be alteredso as to better represent the actual policies.

3. METHODOLOGY

3.1 PAM algorithm

Partitioning Around Medoids (PAM) is an implementation of the K-medoids algorithm. The algorithm partitions the entities into clusters and aims to minimise the distance between the entities assigned to a cluster and its centre, in this case, an entity (Kaufman & Rousseeuw, 1990). PAM has several favourable properties as it performs clustering with respect to any specified distance metric and it identifies clusters by the medoids. Thus, each element is considered as a potential medoid, while the other K-1 medoids are fixed (Van der Laan, Pollard, & Bryan, 2003). Nevertheless, the PAM algorithm has a slight drawback: it works inefficiently with large data set due to time complexity (Han, Kamber, & Tung, 2001). Namely, one of the advantages of the PAM method is the silhouette plot that shows how well cluster members are positioned within their respective clusters. Besides the calculation of the predefined number of clusters, it is possible to use the silhouette average widths for assessing the best number of clusters.

The silhouette plot shows how well cluster members are positioned within their clusters, and average silhouette widths enable us to assess the quality of the cluster structures. The silhouette width has a range [-1,1]. Values near +1 indicate that the entity is far away from any neighbouring clusters. A value of 0 indicates that the entity could be in the current cluster or in the neighbouring clusters and negative values indicate that the entity might be assigned to the wrong cluster. On the other hand, a value of the average silhouette below 0.5 shows a rather weak clustering structure, between 0.5 and 0.7 shows a reasonable cluster structure, and above 0.7 shows a strong structure(Kaufman & Rousseeuw, 1990).

3.2 Potthoff analysis

Multiple books and research articles aimed at introducing methods and tests to compare coefficients from ordinary least squares regression models (Howell, 2013; Potthoff, 1966). The Potthoff analysis stands out as a simple way of comparing linear regression models. Namely, so far it has been used with success in the fields of social psychology (Lawson & Lips, 2014), medicine (Akolekar, Syngelaki, Gallo, Poon, & Nicolaides, 2015), innovation management (Truong, Klink, Fort-Rioche, & Athaide, 2014), and others. Another evidence that the idea of comparing linear regression models is still developing, especially the Potthoff analysis, is the recently published SAS and SPSS code for related tests (Weaver & Wuensch, 2013).

In the simplest case, a Potthoff analysis is multiple regression analysis of the following model CGI: $Y = a + b_1C + b_2G + b_3C*G$, where Y is the criterion variable, C is the continuously distributed predictor variable, G is the dichotomous grouping variable, and C*G is the interaction between C and G. Grouping variables are commonly "dummy-coded" with *l*-1 dichotomous variables, where *l* is the number of groups (Wuensch, 2016). The same form would be made if there were more continuously distributed predictor variables. Namely, then there would be added *j* interaction variables, where *j* is the number of continuously distributed predictor variables.

Essentially, the Potthoff analysis consists of three tests: test of coincidence, test of intercepts, and test of parallelism. To conduct all three tests, four regression models should be made. Namely, the analysis is done as a series of multiple regressions with comparisons among the various models (Wuensch, 2016). The first model, named CGI, consists of the dependent variable explained through the continuous variable, grouping variable, and the interaction variable. The second model, C, models the dependent variable using only the continuous variable. The next model, CG, is formed using the continuous and grouping variable to explain the dependent variable. Finally, the fourth model is CI, where the explanatory variables are the continuous variable and the interaction variable. The model CGI is then compared to each of the three remained models using the partial F-test:

$$F = \frac{SS_{full} - SS_{reduced}}{(f - r)MSE_{full}} : F(m, n-k-1) \tag{1}$$

Where SS_{full} is the sum of squares of the model CGI, $SS_{reduced}$ is the sum of squares of the reduced model, which is C, CI, or CG depending on the test, *f* is the number of degrees of freedom of the full model, *r* is the number of degrees of freedom of the reduced model, while the MSE_{full} is the mean square error of the CGI model. The test has Fisher's distribution with *m* and *n-k-1* degrees of freedom, where *m* is the difference in the number of variables between the full and the reduced model, *n* is the number of observations, and *k* is the number of variables in the full model.

Test of coincidence compares CGI with the C model as it aims to explore whether there is difference between the intercept, slope or both between the observed groups. Namely, the null hypothesis that the regression line for predicting Y from C is the same at all levels of some grouping variable (Wuensch, 2016).

Further, the test of intercepts compares CGI with the CI model to test whether the intercepts are identical across groups. Finally, the test of parallelism compares CGI with the CG model. The null hypothesis states that the slope for predicting Y from G is the same for all observed groups of G.

The three presented tests can provide a detailed comparison of the two or more regression models created using the same variables on different groups. If there is a significant F result in the test of coincidence slopes and intercepts may then be assessed separately to determine whether they differ (Reynolds & Gutkin, 1980). So, the first test when conducting the Potthoff analysis is the test of coincidence, followed by the test of intercepts and test of parallelism if needed. However, before the Potthoff analysis, the grouping variable must be recoded meaningfully (West, Aiken, & Krull, 1996). In the conducted case study the grouping variable is the cluster the country belongs to.

4. RESULTS AND FINDINGS

The dataset on which the analysis was performed contained all ten topic values for 187 countries which are ranked by the DBI for the year 2016.The World Bank defines GDP per capita as the gross domestic product divided by midyear population (World Bank, 2018). The values of the GDP per capita have been collected from the World Bank database (World Bank, 2018) and normalisedusing *Min-Max* normalization. Afterwards, the first step in our analysis was to apply the clustering method, the PAM algorithm.

For our case study (187 entities, PAM algorithm, Euclidean distance), the “NbClust” package was able to obtain the results of 24 indices out of 30. The number of clusters suggested by the highest number of indexes was retained. In our case, seven indices proposed to retain two clusters. The most commonly used metric to evaluate PAM, the average silhouette width for the two retained clusters is 0.279, whereas the average silhouette width of each cluster is 0.190 and 0.451, respectively. Although the average silhouette of the first cluster is low, it is still positive. Countries which represent the cluster centres are Lebanon and Latvia.

Next, the descriptive statistics of the two retained clusters is presented in Table 2. The first column indicates the size of the clusters. We can observe that the clusters are not of similar size, 123 and 64 countries. Some of the countries which make cluster 1 are: Argentina, Brazil, China, India, Malta, Saudi Arabia, and South Africa. On the other side, some of the countries which make cluster 2 are: Australia, Belgium, Germany, Norway, United Kingdom, and the USA. As the clusters are large, the list of countries which make each cluster will not be listed completely, but it is available on demand. The next column indicates the mean value of topics in the specific cluster. According to the mean values in the first cluster, it can be observed that these countries underperform when it comes to certain topics and could reform to attract more FDI and investors. The topic with the lowest mean value in this cluster is T10, *Resolving insolvency*. On the other hand, the second cluster can be identified as a cluster of countries which perform well according to the DBI topics and whose laws aim to facilitate and stimulate doing business. It is also interesting to observe the minimum and maximum values of topics per cluster. In the first cluster, the lowest values of all topics expect T1 *Starting a business* and T9 *Enforcing contracts,* is 0, while in cluster 2, the situation is completely different. Namely, there is only one topic with the minimum value of 0, T2*Dealing with construction permits*. What can also be observed from Table 2 is that the standard deviation (StD) of topics per cluster is high.This could lead to the conclusion that although the countries have been grouped in two clusters and that the majority of indices suggested such a structure, the obtained clustering structure might not be that coherent.

Table 2 Basic descriptive statistics of clusters created using the PAM clustering method

Cluster 1	Size	Mean	Min	Max	StD		Cluster 2	Size	Mean	Min	Max	StD
T1	123	76.27	33.53	94.51	13.25		T1	64	89.94	63.60	99.96	5.81
T2	123	59.77	0.00	83.79	16.13		T2	64	70.77	0.00	86.30	12.20
T3	123	56.68	0.00	90.63	19.63		T3	64	78.83	43.42	99.88	12.13
T4	123	52.86	0.00	90.59	15.96		T4	64	75.64	49.62	94.46	11.00
T5	123	36.50	0.00	85.00	19.78		T5	64	66.72	35.00	100.00	15.02
T6	123	44.68	0.00	76.67	11.97		T6	64	63.46	38.33	81.67	8.79
T7	123	64.37	0.00	99.44	18.38		T7	64	79.89	48.60	99.44	10.13
T8	123	58.44	0.00	100.00	20.56		T8	64	88.77	59.61	100.00	11.21
T9	123	49.89	6.13	78.23	11.76		T9	64	65.20	32.43	84.15	9.51
T10	123	30.67	0.00	69.59	17.36		T10	64	66.32	20.30	93.81	16.79

To additionally inspect the clustering structure cluster means were compared using t-test as used by Russell and coauthors(2017)to compare means of clusters of countries. The results are presented in Table 3. As it can be observed, there is statistically significant difference in group means for all ten topics. This could lead to the conclusion that the clusters differ and that they are well separated. The absolute mean difference varies from 10.99 (T2) to 35.65 (T10). The high absolute mean difference can also be acknowledged for T5 (30.22) and T8 (30.33).

Table 3 Results of the cluster means comparison t-test

Topics	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
t^1	-9.78**	-5.22**	-9.51**	-11.45**	-11.67**	-12.19**	-7.44**	-13.05**	-9.61**	-13.47**
Abs mean difference^2	13.67	10.99	22.15	22.78	30.22	18.78	15.52	30.33	15.32	35.65

^{1} t – value of t-statistics
^{2}Abs mean difference – Absolute mean difference between cluster means
**p<0.01

Prior to conducting the Potthoff analysis we modelled the GDP per capita for all observed countries (Table 4). The obtained adjusted R^2 is 0.466and the model is statistically significant (F(10,176)=17.255, p <0.01).Taking a closer look on the values of the obtained coefficients and their significance, we conclude that the intercept, *Getting electricity, Getting credit, Paying taxes, Enforcing contracts,*and *Resolving insolvency*are statistically significant for modelling the GDP per capita. The model created on all of the observed countries is statistically significant and of a good quality, indicating that the GDP per capita could be modelled using 10 DBI topics.

Table 4 Regression model of GDP for all the observed countries

Topics	Intercept	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
$\hat{\beta}_i$	-33.022	-0.038	0.124	0.170	0.014	-0.151	-0.083	0.202	0.019	0.221	0.292
t	-5.053**	-0.398	1.65	2.479*	0.192	-2.894**	-0.861	2.888**	0.328	2.340*	5.166**

* p<0.05, ** p<0.01

The next step in the analysis was the application of the regression analysis on the retained clusters. The models for Cluster 1 and Cluster 2 are presented in Table 5. The obtained adjusted R^2 is 0.274 for Cluster 1 and 0.508 for Cluster 2.The model created for the Cluster 1 is of a bit lower quality as 27.4% variability of GDP *per* capita is explained, while the model created for Cluster 2 is of a better quality as it explains 50.8% variability of GDP per capita. Both models are statistically significant ($F_{cluster1}$(10,112)=5.595, p <0.01; $F_{cluster2}$(10,53)=7.506, p <0.01).

Table 5 Regression models of GDP for the two retained clusters

	Topics	Intercept	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Cluster 1	$\hat{\beta}_i$	-21.583	-0.039	0.114	0.126	0.008	-0.145	-0.096	0.165	0.031	0.211	0.100
	t	-3.232**	-0.466	1.631	1.994*	0.109	-2.747**	-1.06	2.708**	0.558	2.284*	1.655
Cluster 2	$\hat{\beta}_i$	-131.39	0.430	-0.039	0.564	0.078	0.039	-0.199	0.423	0.17	-0.089	0.543
	t	-3.846**	1.196	-0.182	2.749**	0.397	0.301	-0.788	1.539*	0.914	-0.386	4.233**

* p<0.05, ** p<0.01

When it comes to the model made for the countries in the first cluster, the intercept, *Getting electricity, Getting credit,Paying taxes,* and *Enforcing contracts* are statistically significant (Table 5). For*Getting electricity, Paying taxes,* and *Enforcing contracts*there is a positive relationship with the values of GDP per capita, while the ease of *Getting credit* decreases the value of GDP per capita. Topics which proved to be significant in Cluster 1 which attract attention are *Getting credit* and *Enforcing contracts.* The negative coefficient for *Getting credit* can be explained as that the credits given by banks in these countries might not generate enough positive return on investment (ROI).The negative impact of some indices seems to be logical (Messaoud & Teheni, 2014). When it comes to *Enforcing contracts,* the obtained coefficient can be explained as the more the country is prepared to enforce contracts, the foreign and domestic investors are more secure to invest knowing their investment can be timely solved in a quality and unbiased judicial process.

In the created model for Cluster 2, there are differences. Namely, the intercept is statistically significant, alongside *Getting electricity*, *Paying taxes* and *Resolving insolvency* (Table 5). All three topics have a positive relationship with the values of GDP per capita. Topic which proved to be positive and significant in

Cluster 2 which attracts attention is *Resolving insolvency.* The importance of this topics can be explained that if the investors know there is a legal framework for insolvency and that the commercial insolvency can be dealt in time and with a high recovery rate, they will invest more, which will eventually have an impact on growth.

Finally, we conduct the Potthoff analysis to inspect is there statistically significant difference between the two models. The analysis can provide valuable insights as if there is a difference in the regression models, it means that there is a difference in the importance of topics of doing business for the growth of GDP and GDP per capita. Therefore, the decision makers could be informed to use different approaches to increase the GDP per capita depending on the cluster they belong to.

The Potthoff analysis showed significant difference between the two models, (F(10,165)=4.000, $p< 0.01$), whereas that difference is in the slopes (F(10,165)=4.2264, $p< 0.01$), and in the intercept (F(1,377)=13.925, $p< 0.01$). This result indicates that the models significantly differ, and that both intercept and coefficients differ. Topics which are statistically significant for both models are *Getting electricity* and *Paying taxes*. Nevertheless, in the model for Cluster 1, *Getting credit* and *Enforcing contracts* are statistically significant, while in model for Cluster 2 *Resolving insolvency* has statistically significant impact on the GDP per capita.

These results indicate that government representatives in all analysed countries should place more effort to ease the procedures and cost of getting electricity and to create a reasonable total tax and contribution rate. The representatives in developing countries which are mostly in Cluster 1 should be stricter regarding the ease of allowing credit, while they should improve their laws on enforcing contracts. On the other hand, the more developed country, mostly in Cluster 2, should think of improving their laws regarding insolvency.

5. CONCLUSION

Extensive research was done on the subject of the effects of nations' political, legal, economic and social reforms on wealth and long-term growth(for example Acemoglu, Johnson, & Robinson, 2001; Knack & Keefer, 1995). Messaoud and Teheni (2014) went a step further and tried to model the relationship between business regulations and growth. Their results are also coherent with the results of Djankov, McLiesh, and Ramalho (2006) who stated that “good” business regulations lead to higher economic growth. Our paper is an attempt to extend the current literature on the topic.

The taken research approach was two-fold. We first clustered the countries ranked by the DBI. We retained two clusters, which are well separated. The first cluster is made of countries in which underperform when it comes to certain topics and could reform so as to attract more FDI and investors. On the other hand, the second cluster can be identified as a cluster of countries which perform well according to the DBI. In the next step, we performed the regression analysis of the GDP per capita and the Potthoff analysis. The chosen analysis has been used previously to compare regression models between groups (Lawson & Lips, 2014). Therefore, its results could indicate whether there is a difference in the models and should the decision makers use different approaches to increase the GDP per capita. The obtained results showed that the GDP per capita can be modelled using the DBI topics, that there is statistically significant difference between the two models. The observed difference indicates that policy makers in countries form different cluster should undertake different reforms to increase the GDP.

During our research we could identify two possible future directions of the study. One would be towards reducing the number of observed topics which make the DBI. For example, post-hoc I-distance could be implemented(Savic, Jeremic, & Petrovic, 2016). The second direction would be towards implementing hierarchical clustering methods (Miyamoto, 2012) or more advanced clustering methods, such as biclustering (Busygin, Prokopyev, & Pardalos, 2008) as the clustering results indicate that the currently suggested clustering scheme could be modified.

We believe our research could provide additional insights on the topic of modelling GDP, application of composite indicator data, and comparison of regression models between groups.

REFERENCES

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). The Colonial Origins of Comparative Development: An Empirical Investigation. *American Economic Review*, *91*(5), 1369–1401. https://doi.org/10.1257/aer.91.5.1369

Akolekar, R., Syngelaki, A., Gallo, D. M., Poon, L. C., & Nicolaides, K. H. (2015). Umbilical and fetal middle cerebral artery Doppler at 35-37 weeks' gestation in the prediction of adverse perinatal outcome. *Ultrasound in Obstetrics & Gynecology*, *46*(1), 82–92. https://doi.org/10.1002/uog.14842

Anghelache, C., Manole, A., Anghel, M. G., & Anghel, G. (2015). Analysis of final consumption and gross investment influence on GDP – multiple linear regression model. *Theoretical and Applied Economics*, *22*(3(604)), 137–142. Retrieved from http://store.ectap.ro/articole/1115.pdf

Arruñada, B. (2007). Pitfalls to avoid when measuring institutions: Is Doing Business damaging business? *Journal of Comparative Economics*, *35*(4), 729–747. https://doi.org/10.1016/j.jce.2007.08.003

Basu, P., Chakraborty, C., & Reagle, D. (2003). Liberalization, FDI, and Growth in Developing Countries: A Panel Cointegration Approach. *Economic Inquiry*, *41*(3), 510–516. https://doi.org/10.1093/ei/cbg024

Brunetti, A., Kisunko, G., & Weder, B. (1997). *nstitutional obstacles to doing business: region-by-region results from a worldwide survey of the private sector*.

Busse, M., & Groizard, J. L. (2008). Foreign Direct Investment, Regulations and Growth. *World Economy*, *31*(7), 861–886. https://doi.org/10.1111/j.1467-9701.2008.01106.x

Busygin, S., Prokopyev, O., & Pardalos, P. M. (2008). Biclustering in data mining. *Computers & Operations Research*, *35*(9), 2964–2987. https://doi.org/10.1016/j.cor.2007.01.005

Cavusgil, S. T. (1997). Measuring the potential of emerging markets: An indexing approach. *Business Horizons*, *40*(1), 87–91. https://doi.org/10.1016/S0007-6813(97)90030-6

Ciccone, A., & Papaioannou, E. (2007). Red Tape and Delayed Entry. *Journal of the European Economic Association*, *5*(2–3), 444–458. https://doi.org/10.1162/jeea.2007.5.2-3.444

Corcoran, A., & Gillanders, R. (2015). Foreign direct investment and the ease of doing business. *Review of World Economics*, *151*(1), 103–126. https://doi.org/10.1007/s10290-014-0194-5

Davis, K. E., Kingsbury, B., & Merry, S. E. (2012). Indicators as a Technology of Global Governance. *Law & Society Review*, *46*(1), 71–104. https://doi.org/10.1111/j.1540-5893.2012.00473.x

Djankov, S., La Porta, R., Lopez-de-Silanes, F., & Shleifer, A. (2002). The Regulation of Entry. *The Quarterly Journal of Economics*, *117*(1), 1–37. https://doi.org/10.1162/003355302753399436

Djankov, S., McLiesh, C., & Ramalho, R. M. (2006). Regulation and growth. *Economics Letters*, *92*(3), 395–401. https://doi.org/10.1016/J.ECONLET.2006.03.021

Dritsaki, D. C. (2015). Forecasting Real GDP Rate through Econometric Models: An Empirical Study from Greece. *Journal of International Business and Economics*, *3*(1). https://doi.org/10.15640/jibe.v3n1a2

Han, J., Kamber, M., & Tung, A. (2001). Spatial clustering methods in data mining: A survey. In H. Miller & J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery* (pp. 188–217). Taylor & Francis.

Howell, D. (2013). *Statistical methods for psychology*. Belmont, CA: Cengage Wadsworth.

Hsiao, C., & Shen, Y. (2003). Foreign Direct Investment and Economic Growth: The Importance of Institutions and Urbanization. *Economic Development and Cultural Change*, *51*(4), 883–896. https://doi.org/10.1086/375711

International Monetary Fund. (2017). Gross Domestic Product (GDP): An Economy's All - Back to Basics: GDP Definition. Retrieved May 15, 2018, from http://www.imf.org/external/pubs/ft/fandd/basics/gdp.htm

Jacobs, R., & Goddard, M. (2007). How Do Performance Indicators Add Up? An Examination of Composite Indicators in Public Services. *Public Money and Management*, *27*(2), 103–110. https://doi.org/10.1111/j.1467-9302.2007.00565.x

Kaufman, L., & Rousseeuw, P. J. (1990). Partitioning Around Medoids (Program PAM). In L. Kaufman & P. J. Rousseeuw (Eds.), *Finding groups in data: an introduction to cluster analysis* (pp. 68–125). https://doi.org/10.1002/9780470316801.ch2

Klapper, L., Laeven, L., & Rajan, R. (2006). Entry regulation as a barrier to entrepreneurship. *Journal of Financial Economics*, *82*(3), 591–629. https://doi.org/10.1016/j.jfineco.2005.09.006

Knack, S., & Keefer, P. (1995). Institutions and economic performance: cross-country tests using alternative institutional measures. *Economics & Politics*, *7*(3), 207–227. https://doi.org/10.1111/j.1468-0343.1995.tb00111.x

Lakštutienė, & Aušrinė. (2015). Correlation of the Indicators of the Financial system and Gross Domestic Product in European Union Countries. *Engineering Economics*, *58*(3). https://doi.org/10.5755/j01.ee.58.3.11532

Lawson, K. M., & Lips, H. M. (2014). The role of self-perceived agency and job attainability in women's impressions of successful women in masculine occupations. *Journal of Applied Social Psychology*, *44*(6), 433–441. https://doi.org/10.1111/jasp.12236

Messaoud, B., & Teheni, Z. E. G. (2014). Business regulations and economic growth: What can be explained? *International Strategic Management Review*, *2*(2), 69–78. https://doi.org/10.1016/j.ism.2014.03.001

Miyamoto, S. (2012). An Overview of Hierarchical and Non-hierarchical Algorithms of Clustering for Semi-supervised Classification. In V. M. Torra V., Narukawa Y., López B. (Ed.), *Modeling Decisions for Artificial Intelligence* (pp. 1–10). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-34620-0_1

Morris, R., & Aziz, A. (2011). Ease of doing business and FDI inflow to Sub-Saharan Africa and Asian countries. *Cross Cultural Management: An International Journal*, *18*(4), 400–411. https://doi.org/10.1108/13527601111179483

Mullen, M. R., & Sheng, S. Y. (2006). Extending and Comparing Cavusgil's Overall Market Opportunity Indexes. In S. Zou (Ed.), *International Marketing Research (Advances in International Marketing,* *Volume 17)* (pp. 219–249). https://doi.org/10.1016/S1474-7979(06)17008-8

Nair-Reichert, U., & Weinhold, D. (2001). Causality Tests for Cross-Country Panels: a New Look at FDI and Economic Growth in Developing Countries. *Oxford Bulletin of Economics and Statistics*, *63*(2), 153–171. https://doi.org/10.1111/1468-0084.00214

Potthoff, R. F. (1966). *Statistical aspects of the problem of biases in psychological tests*. Chapel Hill: University of North Carolina.

Reynolds, C. R., & Gutkin, T. B. (1980). A regression analysis of test bias on the WISC-R for Anglos and Chicanos referred for psychological services. *Journal of Abnormal Child Psychology*, *8*(2), 237–243. https://doi.org/10.1007/BF00919067

Russell, L. B., Bhanot, G., Kim, S.-Y., & Sinha, A. (2017). Using Cluster Analysis to Group Countries for Cost-Effectiveness Analysis: An Application to Sub-Saharan Africa. *Medical Decision Making*, 0272989X1772477. https://doi.org/10.1177/0272989X17724773

Savic, D., Jeremic, V., & Petrovic, N. (2016). Rebuilding the Pillars of Sustainable Society Index: A Multivariate Post Hoc I-Distance Approach. *Problemy Ekorozwoju – Problems of Sustainable Development*, *12*(1), 125–134.

Schumacher, C., & Breitung, J. (2008). Real-time forecasting of German GDP based on a large factor model with monthly and quarterly data. *International Journal of Forecasting*, *24*(3), 386–398. https://doi.org/10.1016/j.ijforecast.2008.03.008

Truong, Y., Klink, R. R., Fort-Rioche, L., & Athaide, G. A. (2014). Consumer Response to Product Form in Technology-Based Industries. *Journal of Product Innovation Management*, *31*(4), 867–876. https://doi.org/10.1111/jpim.12128

Van der Laan, M., Pollard, K., & Bryan, J. (2003). A new partitioning around medoids algorithm. *Journal of Statistical Computation and Simulation*, *73*(8), 575–584. https://doi.org/10.1080/0094965031000136012

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods*, *45*(3), 880–895. https://doi.org/10.3758/s13428-012-0289-7

West, S. G., Aiken, L. S., & Krull, J. L. (1996). Experimental Personality Designs: Analyzing Categorical by Continuous Variable Interactions. *Journal of Personality*, *64*(1), 1–48. https://doi.org/10.1111/j.1467-6494.1996.tb00813.x

World Bank. (2016). Doing Business. Retrieved January 1, 2017, from http://www.doingbusiness.org

World Bank. (2017). *Doing Business 2018: Reforming to Create Jobs*. Washington, D.C. Retrieved from http://www.doingbusiness.org/~/media/WBG/DoingBusiness/Documents/Annual-Reports/English/DB2018-Full-Report.pdf

World Bank. (2018). GDP per capita (current US$). Retrieved February 24, 2018, from https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Wuensch, K. L. (2016). *Comparing Regression Lines From Independent Samples*. Retrieved from http://core.ecu.edu/psyc/wuenschk/MV/multReg/Potthoff.pdf

CONFIDENCE INTERVALS FOR THE POPULATION STANDARD DEVIATION: SIMPLE RANDOM SAMPLING VS. RANKED SET SAMPLING

Ivana Ivković*[1], Vesna Rajić[1]
[1] University of Belgrade, Faculty of Economics, Serbia
*Corresponding author, email: ivanaivkovic@ekof.bg.ac.rs

Abstract. *Ranked set sampling (RSS) is the cost-efficient sampling procedure. This procedure gives more efficient estimators of the population parameters than the procedure based on simple random sampling (SRS), with the same sample size. In this paper, we compare the coverage probabilities of confidence intervals for the population standard deviation using the simple random sampling and the ranked set sampling. The following confidence intervals are considered: the exact, the Bonett, the Steve large sample normal approximations, the log asymptotic approximation and the adjusted degrees of freedom. The results for the Gamma, Log-normal and Exponential distributions and for the real data set are presented. The simulation study shows that the results obtained using the ranked set sampling are better than those using the simple random sampling.*

Keywords: *population standard deviation; confidence interval; coverage accuracy; ranked set sampling; simple random sampling*

1. INTRODUCTION

In this paper, we construct the confidence intervals for the population standard deviation. The population standard deviation is the most common scale parameter. The existing confidence interval for the estimation of the population standard deviation is the exact confidence interval, based on the statistic which has χ^2 distribution. This interval is appropriate if the distribution of the data is normal with no outliers. We are interested in confidence intervals which are appropriate if the data are not from the normal distribution, but from skewed distribution or have heavy tails. There are some alternatives to the exact confidence interval which can be used in such situations. There are no many authors who dealt with the confidence intervals that were less sensitive to the departure from normality and/or presence of outliers. Bonett (2006) proposed an approximate confidence interval for the population standard deviation which results were close to the exact confidence interval under the normality and had very good small-sample properties under the moderate non-normality. Cojbasic and Loncar (2011) and Cojbasic and Tomovic (2007) used the resampling methods for construction of confidence intervals for the population variance (taking the square root of the endpoints of that intervals gave the confidence intervals for the population standard deviation). Abu-Shawiesh et al. (2011) and Banik et al. (2014) conducted the large simulation studies in which compared the performances of the different confidence intervals for the standard deviation under the symmetric and skewed conditions. Hummel et al. (2005) proposed two alternative methods for finding the confidence interval for the standard deviation.

Ranked set sampling (or shortly RSS) is an alternative method of data collection and presents the cost-effective sampling procedure. The RSS is used for improving the estimators in the situations where the ranking of the units can be done easily compared to the effort required for the actual measurement of the variable of interest. This method was first proposed by McIntyre (1952). He estimated the mean of the population using the RSS instead of simple random sampling (or shortly SRS). Dell and Clutter (1972) showed that the ranked set sampling provided more precise estimates of the mean when the ranking of the units in the sample was easy. Using the RSS, Stokes (1980) concluded that an estimator of variance was asymptotically unbiased regardless of the errors that could occur during the ranking and that asymptotic efficiency of that estimator was better relative to the estimator based on the same number of the measured units from the random sample. Chen (2007) gave the review of the several variants of the ranked set sampling and presented some recent applications of that method. Samawi (1999) showed that the performance of the Monte Carlo methods, such as an importance or control variate sampling, were improved a lot using the ranked set sampling. Wolfe (2012) wrote the review article about the impact of the ranked set sampling on the statistical inference. Ganeslingam and Ganesh (2006) applied the ranked set sampling procedure on the estimation of the population mean and ratio using the real data set on the body measurements. Husby et al. (2005) used the crop production dataset from the United States Department of Agriculture to show the advantages of the RSS relative to the frequently used simple random sampling in the estimation of the mean and median of the population. Terpstra and Wang (2008) examined the several methods for construction of confidence intervals for the population proportion based on the RSS. Albatineh et al. (2014) performed the simulation study in which evaluated the performance of the several confidence intervals for the population coefficient of variation, using the coverage probabilities and the width of the intervals. Albatineh et al. (2017) constructed the confidence intervals for the Signal-to-Noise ratio using the RSS. More about the ranked set sampling methodology and its application can be found in Ozturk (2018), Zamanzade and Mahdizadeh (2017), Zamanzade and Vock (2015), Zhang et al. (2016), etc.

In this paper, we examine the confidence intervals for the population standard deviation which are more adequate to use when the data do not follow the normal distribution. We construct the confidence intervals using the simple random sampling and the ranked set sampling. The examined intervals are implemented using the R programming language. We generate random data from the Gamma, Log-normal and Exponential distributions, respectively and compare the coverage probabilities of the presented confidence intervals. Then, we apply the considered intervals to the measure of the systematic risk data.

The goal of this paper is to present the ranked set sampling procedure for construction of confidence intervals for the population standard deviation. The contribution of this paper is to emphasize the advantages of the ranked set sampling procedure over the simple random sampling procedure. The paper is organized as follows: in Section 2, we describe the ranked set sampling methodology; in Section 3, we present the confidence intervals for the population standard deviation; in Section 4, we conduct the simulation study for the data from the Gamma, Log-normal and Exponential distributions and for the real data set; in Section 5, we summarize results and draw the conclusions.

2. RANKED SET SAMPLING METHODOLOGY

The ranked set sampling procedures can be balanced or unbalanced. Each procedure can be with the perfect or imperfect ranking process. In this paper, we consider the balanced RSS with the perfect ranking process (see Ganeslingam and Ganesh (2006), Wolfe (2012), Albatineh et al. (2014)). The process of generating the balanced RSS involves drawing k^2 units at random from the population. After that, these units are randomly divided into *k* sets of *k* units each (we get *k* simple random samples of the size *k*). Within each set, the units are ranked according to the variable of interest. The perfect ranking process implies that actual measurements of the variable of interest on the selected units are done and that ranking is based on them. Opposite to the perfect ranking process, the imperfect ranking process includes visual comparisons of the units or the use of the auxiliary varables. After the ranking process, from the first set we select the unit with the smallest rank $X_{(1)}$ (if the ranking is perfect or $X_{[1]}$, if the ranking is imperfect). The remaining *k*-1 units are not considered further. Then, from the second set we select the unit with the second smallest rank $X_{(2)}$, and so on, until we select the unit with the largest rank from the *k*-th set, $X_{(k)}$. This procedure results in *k* observations $X_{(1)}, X_{(2)}, \dots, X_{(k)}$ and is called the cycle. The number of the units in each simple random sample, *k*, is called the set size. If we want to obtain the balanced ranked set sample of size $n = mk$, we repeat the cycle *m* times (see Table 1). The complete balanced RSS with set size *k* and *m* cycles is given by $\left\{X_{(j)i} : j = 1, 2, ..., k; i = 1, 2, ..., m\right\}$. The term $X_{(j)i}$ is called the *j*-th order statistic from the *i*-th cycle.

Table 1: The balanced RSS with *m* cycles and set size *k*

Cycle 1	$X_{(1)1}$	$X_{(2)1}$	…	$X_{(k)1}$
Cycle 2	$X_{(1)2}$	$X_{(2)2}$	…	$X_{(k)2}$
…	…	…	…	…
Cycle *m*	$X_{(1)m}$	$X_{(2)m}$	…	$X_{(k)m}$

Source: Wolfe (2012)

The estimators of the mean and the variance of the population, based on the RSS, are given with the following formulas (see Stokes, 1980):

$$\bar{X}_{RSS} = \frac{1}{km} \sum_{j=1}^{k} \sum_{i=1}^{m} X_{(j)i}, \tag{1}$$

$$S^2_{RSS} = \frac{1}{km-1} \sum_{j=1}^{k} \sum_{i=1}^{m} \left(X_{(j)i} - \bar{X}_{RSS} \right)^2 . \tag{2}$$

3. CONFIDENCE INTERVALS FOR THE POPULATION STANDARD DEVIATION

In this section, we report five confidence intervals for the population standard deviation.

- The exact confidence interval

Let $X_1,...,X_n$ be independent and identically distributed random variables from the normal distribution, i.e. $X_i \sim N\left(\mu,\sigma^2\right)$. Let $S^2 = \left(1/\left(n-1\right)\right)\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$ be a sample variance. The statistic $\left(n-1\right)S^2/\sigma^2$ has χ^2 distribution with $n-1$ degrees of freedom. The exact$\left(1-\alpha\right)\cdot 100\%$ confidence interval for the population standard deviation, based on the previous statistic, is of the form:

$$\sqrt{\frac{\left(n-1\right)S^2}{\chi^2_{\alpha/2,(n-1)}}} \le \sigma \le \sqrt{\frac{\left(n-1\right)S^2}{\chi^2_{1-\alpha/2,(n-1)}}}, \tag{3}$$

Where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ percentiles of the χ^2 distribution with $n-1$ degrees of freedom.

The exact confidence interval (3) is very sensitive to minor violations of the normality assumption. In the cases of violations of the normality assumption, there are the confidence intervals which present the alternatives to the exact confidence interval. In remain of the section, we consider that confidence intervals.

- The Bonett confidence interval

Let $X_1,...,X_n$ be continuous, independent and identically distributed random variables with$E\left(X_i\right) = \mu$, $\text{var}\left(X_i\right) = \sigma^2$ and the finite fourth moment. Bonett (2006) proposed the following estimator of the kurtosis, γ_4, which is asymptotically eqvivalent to the Pearson's estimator:

$$\bar{\gamma}_4 = n \cdot \sum_{i=1}^{n}\left(X_i - m\right)^4 / \left(\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right)^2,$$

Where *m* is a trimmed mean with trim-proportion equal to $1/\left\{2\cdot\left(n-4\right)^{1/2}\right\}$. This estimator tends to have less negative bias and smaller coefficient of variability than Pearson's estimator in the symmetric and skewed leptokurtic distributions. The $\left(1-\alpha\right)\cdot 100\%$ confidence interval for the population standard deviation can be written in the following form (see Bonett, 2006):

$$\sqrt{\exp\left[\ln\left(cS^2\right) - Z_{1-\alpha/2}se\right]} \le \sigma \le \sqrt{\exp\left[\ln\left(cS^2\right) + Z_{1-\alpha/2}se\right]}, \tag{4}$$

Where $c = n/\left(n - Z_{1-\alpha/2}\right)$, S^2 is the sample variance, $Z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile of the Z distribution and $se = c\cdot\left[\left\{\bar{\gamma}_4\cdot\left(n-3\right)/n\right\}/\left(n-1\right)\right]^{1/2}$.

- The Steve large sample normal approximations confidence interval

Steve proposed the following $\left(1-\alpha\right)\cdot 100\%$ confidence interval for the population standard deviation (see Banik et al., 2014):

$$\sqrt{\frac{S^2}{1-Z_{\alpha/2}\sqrt{\frac{\hat{\gamma}-1}{n}}}} \leq \sigma \leq \sqrt{\frac{S^2}{1+Z_{\alpha/2}\sqrt{\frac{\hat{\gamma}-1}{n}}}}, \qquad (5)$$

where S^2 is the sample variance, $Z_{\alpha/2}$ is the $\alpha/2$ percentile of the standardized normal distribution and $\hat{\gamma}=n\cdot\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^4/\left(\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right)^2$ is the kurtosis estimator.

- The log asymptotic approximation confidence interval (LOG CI)

The distribution of the sample variance, S^2, has the high skewness for small *n*. In order to reduce the skewness, Hummel et al. (2005) applied natural log to the sample variance in (5) and proposed the $\left(1-\alpha\right)\cdot 100\%$ confidence interval for the population standard deviation:

$$\sqrt{\left[S^2\exp\left(Z_{\alpha/2}\sqrt{\frac{\hat{\gamma}-1}{n}}\right)\right]} \leq \sigma \leq \sqrt{\left[S^2\exp\left(-Z_{\alpha/2}\sqrt{\frac{\hat{\gamma}-1}{n}}\right)\right]}, \qquad (6)$$

Where $Z_{\alpha/2}$ is the $\alpha/2$ percentile of the Z distribution and $\hat{\gamma}$ is the kurtosis estimator.

- The adjusted degrees of freedom confidence interval (ADF CI)

Hummel et al. (2005) adjusted the degrees of freedom of the exact confidence interval (3) and proposed the following $\left(1-\alpha\right)\cdot 100\%$ confidence interval for the population standard deviation:

$$\sqrt{\frac{\hat{r}S^2}{\chi^2_{\alpha/2,\hat{r}}}} \leq \sigma \leq \sqrt{\frac{\hat{r}S^2}{\chi^2_{1-\alpha/2,\hat{r}}}}, \qquad (7)$$

Where $\hat{r}=\dfrac{2n}{\hat{\gamma}_e+\left(\dfrac{2n}{n-1}\right)}$ and $\hat{\gamma}_e$ is the estimate of the kurtosis excess, which is defined as $\hat{\gamma}_e=\dfrac{n\left(n+1\right)}{\left(n-1\right)\left(n-2\right)\left(n-3\right)}\sum_{i=1}^{n}\dfrac{\left(X_i-\bar{X}\right)^4}{S^4}-\dfrac{3\left(n-1\right)^2}{\left(n-2\right)\left(n-3\right)}$. If the random sample is generated from the normal distribution, then $r=n-1$ and (7) reduces to (3).

The performance of all presented intervals will be considered using the SRS and the RSS. In order to get the estimators of the mean and variance of the population we will use the regular formulas for the SRS and Equations (1) and (2) for the RSS.

4. CASE STUDY

In this part of the paper, we present the results of applying the proposed methods on the simulated data and real-economic data.

4.1 Simulation study

In this part of our work we examine the coverage accuracy of two-sided confidence intervals for the standard deviation introduced in the Section 3. Our objective is to compare the performance of the confidence intervals for estimating the standard deviation using the SRS and the RSS. The nominal confidence level is set to 95% and we want to determine which of the proposed intervals will give the coverage probability that is the closest to 95%. For that purpose, we consider three scenarios. In the first scenario we deal with the Gamma distribution, while in the second scenario the subject of the consideration is the Log-normal distribution. In the third scenario we investigate the Exponential distribution. It is important to emphasize that we can deal with any other scenario with the other skewed distribution.

In the first scenario, we consider the Gamma distribution with the shape parameter 2 and with the scaling parameters 0.5, 1.6 and 3.2. For each combination of the parameter setting and the sample size (15, 20, 50, 80), we performed 1000 simulations. All examined intervals are implemented using the *R* programming language. In Table 2 we present the results of the coverage accuracy of 95% confidence intervals for the standard deviation of the Gamma distribution. It can be seen that for the small samples (size 15), the RSS Bonett interval gives the best results, i.e. the coverage probabilities that are the closest to 0.95 (the coverage greater than 0.932). For the moderate samples (size 20), the RSS Bonett interval gives the best coverage (below 0.961). In the case of big samples (size 50), depending on the scaling parameter of the Gamma distribution, the best coverage probabilities are obtained using the Bonett and the RSS Log intervals (the coverage greater than 0.939). For the samples of size 80, the Bonett interval and the RSS ADF interval give the best coverage accuracy (above 0.943).

Table 2: The coverage of 95% two-sided confidence intervals for the standard deviation of the Gamma distribution

a	***s***	***n***	χ^2	**RSS** χ^2	**Bonett**	**RSS Bonett**	**Steve**	**RSS Steve**	**Log**	**RSS Log**	**ADF**	**RSS ADF**
2	0.5	15	0.831	0.865	0.937	0.939	0.714	0.748	0.797	0.821	0.841	0.895
		20	0.838	0.881	0.909	0.960	0.769	0.773	0.788	0.864	0.857	0.907
		50	0.784	0.818	0.957	0.966	0.902	0.932	0.903	0.923	0.907	0.935
		80	0.803	0.861	0.945	0.958	0.913	0.936	0.923	0.943	0.926	0.952
2	1.6	15	0.824	0.887	0.911	0.936	0.750	0.757	0.767	0.784	0.835	0.883
		20	0.854	0.875	0.924	0.961	0.726	0.747	0.795	0.830	0.829	0.911
		50	0.805	0.898	0.933	0.972	0.900	0.923	0.885	0.939	0.908	0.934
		80	0.847	0.852	0.958	0.976	0.918	0.939	0.915	0.932	0.921	0.943
2	3.2	15	0.829	0.871	0.910	0.932	0.723	0.737	0.797	0.809	0.883	0.893
		20	0.807	0.868	0.923	0.952	0.750	0.832	0.840	0.845	0.853	0.905
		50	0.765	0.856	0.936	0.982	0.901	0.923	0.885	0.939	0.919	0.932
		80	0.758	0.840	0.952	0.965	0.907	0.941	0.905	0.938	0.922	0.936

In the second scenario, we deal with the Log-normal distribution with the shape parameter 2 and with the scaling parameters 0.25, 0.5 and 0.6. For each combination of the parameter setting and the sample size (15, 20, 50, 80), we generated 1000 samples. All considered intervals are implemented using the *R* programming language. In Table 3 we present the results of the coverage accuracy of 95% intervals for the standard deviation of the Log-normal distribution. For the small samples, the RSS χ^2 and the RSS Bonett intervals give the coverage probabilities that are the closest to 0.95 (above 0.924). In the case of the moderate samples, depending on the scaling parameter of the Log-normal distribution, the RSS χ^2 interval and the RSS Bonett interval give the best results (the coverage greater than 0.933). For the big samples (size 50) the best results are obtained with the Steve interval and the RSS Bonett interval (above 0.940). For the samples of size 80, the Steve, the RSS ADF, the Bonett and the RSS Bonett intervals give the coverage probabilities that are the closest to 0.95.

In the third scenario, we investigate the Exponential distribution with the rate parameters 0.5, 1.5 and 2.2. For each parameter setting and each sample size (15, 20, 50, 80), we performed 1000 simulations. All considered intervals are implemented using the *R* programming language. In Table 4 we present the results of the coverage accuracy of 95% confidence intervals for the standard deviation of the Exponential distribution. It can be seen that for the small samples, the RSS Bonett interval gives the best results (above 0.923). For the moderate samples, the RSS Bonett interval gives the best coverage (greater than 0.938). In the case of big samples (size 50) the best coverage probabilities are obtained using the RSS Bonett interval (above 0.945). For the samples of size 80, depending on the rate parameter of the Exponential distribution, the Bonett and the RSS Bonett intervals give the best coverage accuracy (above 0.942).

Table 3: The coverage of 95% two-sided confidence intervals for the standard deviation of the Log-normal distribution

μ	σ	***n***	χ^2	**RSS** χ^2	**Bonett**	**RSS Bonett**	**Steve**	**RSS Steve**	**Log**	**RSS Log**	**ADF**	**RSS ADF**
2	0.25	15	0.911	0.946	0.965	0.977	0.788	0.839	0.807	0.852	0.904	0.924
		20	0.920	0.937	0.964	0.980	0.833	0.883	0.864	0.874	0.915	0.935
		50	0.909	0.918	0.972	0.974	0.947	0.954	0.923	0.936	0.927	0.963
		80	0.912	0.938	0.969	0.980	0.945	0.964	0.929	0.943	0.942	0.955
2	0.5	15	0.817	0.846	0.922	0.942	0.658	0.690	0.747	0.759	0.769	0.838
		20	0.758	0.850	0.930	0.949	0.679	0.729	0.768	0.818	0.800	0.850
		50	0.669	0.828	0.937	0.953	0.911	0.927	0.872	0.910	0.904	0.914
		80	0.762	0.766	0.949	0.961	0.920	0.936	0.905	0.932	0.920	0.925
2	0.6	15	0.734	0.753	0.915	0.924	0.610	0.636	0.676	0.753	0.777	0.802
		20	0.709	0.737	0.900	0.933	0.579	0.612	0.749	0.760	0.785	0.840
		50	0.655	0.671	0.919	0.940	0.842	0.884	0.791	0.884	0.867	0.875
		80	0.556	0.682	0.939	0.941	0.866	0.888	0.862	0.886	0.884	0.897

Table 4: The coverage of 95% two-sided confidence intervals for the standard deviation of the Exponential distribution

λ	***n***	χ^2	**RSS** χ^2	**Bonett**	**RSS Bonett**	**Steve**	**RSS Steve**	**Log**	**RSS Log**	**ADF**	**RSS ADF**
0.5	15	0.757	0.799	0.904	0.930	0.614	0.637	0.763	0.769	0.781	0.797
	20	0.753	0.843	0.911	0.943	0.586	0.722	0.753	0.832	0.844	0.858
	50	0.720	0.762	0.938	0.958	0.813	0.907	0.848	0.891	0.904	0.910
	80	0.722	0.751	0.943	0.960	0.908	0.930	0.903	0.917	0.907	0.925
1.5	15	0.753	0.763	0.905	0.935	0.510	0.654	0.746	0.795	0.764	0.807
	20	0.769	0.786	0.931	0.953	0.631	0.763	0.770	0.836	0.797	0.836
	50	0.683	0.751	0.918	0.960	0.822	0.901	0.888	0.902	0.900	0.925
	80	0.658	0.744	0.942	0.963	0.916	0.934	0.914	0.920	0.918	0.933
2.2	15	0.746	0.755	0.904	0.923	0.625	0.652	0.720	0.726	0.780	0.839
	20	0.694	0.785	0.914	0.938	0.647	0.721	0.704	0.808	0.778	0.840
	50	0.687	0.734	0.931	0.945	0.883	0.897	0.875	0.908	0.893	0.906
	80	0.653	0.739	0.933	0.966	0.908	0.922	0.894	0.917	0.906	0.922

4.2. Application to the real data

In this part of the paper, we analyze the measure of the systematic risk data in 490 companies (it is about S&P500, but there are no data for some companies) on the 5th July 2017. The data are from the website http://finance.yahoo.com/.

In analysis of securities, the measure of the systematic risk (beta) takes a central place. The measure of the systematic risk represents the measure of the sensitivity of the yield of the securities to the changes in the yield on the market. Beta shows that if the yield on the market changes by one percent, by how many percentage points the yield of the securities will change.

Descriptive statistics for the analyzed data are given in Table 5. Figures 1 represents the histogram of the analyzed variable. We can see that the measure of the systematic risk is not normally distributed. Also, we used the Shapiro-Wilk normality test to examine whether the beta was normally distributed. The test showed the same result as the histogram (*p*-value is approximately 0).

Table 5: Descriptive statistics for the data set

Variable	***N***	**Mean**	**Std. deviation**	**Skewness coef.**
The measure of the systematic risk	490	1.02	0.52	0.43

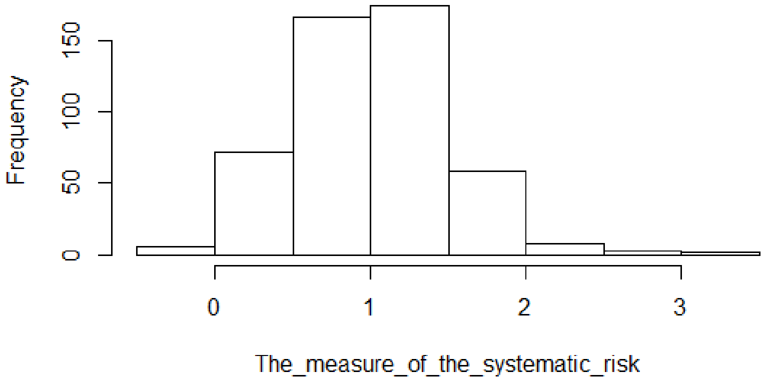


Figure 1: Histogram of the measure of the systematic risk

Results of the coverage accuracy of 95% confidence intervals for the standard deviation of the data set are given in Table 6. It can be seen that for the small samples, the RSS χ^2 interval gives the best coverage (0.945). For the moderate samples, the RSS ADF interval gives the coverage that is the closest to 0.95. For the samples of size 50, the RSS Steve interval is the best choice and in the case of the samples of size 80, the ADF interval gives the best coverage accuracy (0.946).

Table 6: The coverage of 95% two-sided confidence intervals for the standard deviation: the data set

n	χ^2	**RSS** χ^2	**Bonett**	**RSS Bonett**	**Steve**	**RSS Steve**	**Log**	**RSS Log**	**ADF**	**RSS ADF**
15	0.934	0.945	0.962	0.975	0.811	0.833	0.839	0.879	0.907	0.919
20	0.918	0.924	0.972	0.977	0.878	0.893	0.873	0.898	0.926	0.935
50	0.921	0.943	0.976	0.986	0.919	0.955	0.904	0.920	0.921	0.943
80	0.911	0.921	0.971	0.968	0.938	0.931	0.913	0.915	0.946	0.938

5. CONCLUSIONS

In this paper, we used the simple random sampling and the ranked set sampling to compare the coverage probabilities of confidence intervals for the population standard deviation. The exact confidence interval, the Bonett, the Steve large sample normal approximations, the log asymptotic approximation and the adjusted degrees of freedom confidence intervals were examined.

In the first scenario, we investigated the Gamma distribution. For the small and moderate samples, the RSS Bonett interval gave the coverage probabilities that were the closest to 0.95, whereas for the big samples the Bonett, the RSS Log and the RSS ADF intervals gave the best results. In the second scenario, we dealt with the Log-normal distribution. For the small and moderate samples, the best results were obtained using the RSS χ^2 and the RSS Bonett intervals. In most cases, for the big samples, the Steve interval and the RSS Bonett interval gave the best coverage. In the third scenario, we considered the Exponential distribution. For the small and moderate samples, the RSS Bonett interval was the best choice, whereas for the big samples, the Bonett and the RSS Bonett intervals gave the best results.

The analysis of the real data showed that for the small samples, the RSS χ^2 interval gave the best coverage accuracy, while for the moderate samples the best choice was the RSS ADF interval. For the big samples, the RSS Steve and the ADF interval gave the best results. We can see that using the RSS gives much better coverage probabilities, so we recommend using it when construct the confidence intervals for the population standard deviation.

REFERENCES

Abu-Shawiesh M. O. A., Banik S., & Golam Kibria, B. M. (2011). A simulation study on some confidence intervals for the population standard deviation. *SORT*, 35(2), 83-102. Retrieved from http://digitalcommons.fiu.edu/math_fac/9

Albatineh, A. N., Boubakari, I., & Golam Kibria, B. M. (2017). New confidence interval estimator of the signal-to-noise ratio based on asymptotic sampling distribution. *Communications in statistics – Theory and methods*, 46(2), 574-590. doi: 10.1080/03610926.2014.1000498

Albatineh, A. N., Golam Kibria, B. M., Wilcox, M. L., & Zogheib B. (2014). Confidence interval estimation for the population coefficient of variation using ranked set sampling: a simulation study. *Journal of applied statistics*, 41(4), 733-751. doi: 10.1080/02664763.2013.847405

Banik, S., Albatineh, A. N., Abu-Shawiesh, M. O. A., & Golam Kibria, B. M. (2014). Estimating the population standard deviation with confidence interval: a simulation study under skewed and symmetric conditions. *Journal of biometrics and biostatistics*, 5(2). doi: 10.472/2155-6180.1000190

Bonett, D. G. (2006). Approximate confidence interval for standard deviation of nonnormal distributions. *Computational statistics & data analysis*, 50, 775-782. doi: 10.1016/j.csda.2004.10.003

Chen, Z. (2007). Ranked set sampling: its essence and some new applications. *Environmental and ecological statistics*, 14, 355-363. doi: 10.1007/s10651-007-0025-0

Cojbasic, V., & Loncar, D. (2011). One-sided confidence intervals for population variances of skewed distributions. *Journal of statistical planning and inference*, 141, 1667-1672. doi: 10.1016/j.jspi.2010.11.007

Cojbasic, V., & Tomovic, A. (2007). Nonparametric confidence intervals for population variance of one sample and the difference of variances of two samples. *Computational statistics & data analysis*, 51, 5562-5578. doi: 10.1016/j.csda.2007.03.023

Dell, T. R., & Clutter, J. L. (1972). Ranked set sampling theory with order statistics background. *Biometrics*, 28(2), 545-555. Retrieved from http://www.jstor.org/stable/2556166

Ganeslingam, S., & Ganesh S. (2006). Ranked set sampling versus simple random sampling in the estimation of the mean and the ratio. *Journal of statistics & management systems*, 9(2), 459-472, doi: 10.1080/09720510.2006.10701217

Hummel, R., Banga, S., & Hettmansperger, T. P. (2005). Better confidence intervals for the variance in a random sample. Minitab Technical Report

Husby, C. E., Stasny, E. A., & Wolfe, D. A. (2005). An application of ranked set sampling for mean and median estimation using USDA crop production data. *Journal of agricultural, biological and environmental statistics*, 10(3), 354-373. doi: 10.1198/108571105X58234

McIntyre, G. A. (1952). A method for unbiased selective sampling, using ranked sets. *Australian journal of agricultural research*, 59(3), 385-390. doi: 10.1071/AR9520385

Ozturk, O. (2018). Ratio estimators based on a ranked set sample in a finite population setting. *Journal of the Korean statistical society*, 47, 226-238. doi: 10.1016/j.jkss.2018.02.001

Samawi, H. M. (1999). More efficient Monte Carlo methods obtained by using ranked set simulated samples. *Communications in statistics – Simulation and computation*, 28(3), 699-713. doi: 10.1080/03610919908813573

Stokes, S. L. (1980). Estimation of variance using judgement ordered ranked set samples. *Biometrics*, 36(1), 35-42. Retrieved from http://www.jstor.org/stable/2530493

Terpstra, J. T., & Wang, P. (2008). Confidence intervals for a population proportion based on a ranked set sample. *Journal of statistical computation and simulation*, 78(3-4), 351-366. doi: 10.1080/00949650601107994

Wolfe, D. (2012). Ranked set sampling: its relevance and impact on statistical inference. *ISRN Probability and statistics*, 2012. doi: 10.5402/2012/568385

Zamanzade, E., & Mahdizadeh, M. (2017). A more efficient proportion estimator in ranked set sampling. *Statistics and probability letters*, 129, 28-33. doi: 10.1016/j.spl.2017.05.001

Zamanzade, E., & Vock, M. (2015). Variance estimation in ranked set sampling using a concomitant variable. *Statistics and probability letters*, 105, 1-5. doi: 10.1016/j.spl.2015.04.034

Zhang, Z., Liu, T., & Zhang, B. (2016). Jackknife empirical likelihood inferences for the population mean with ranked set samples. *Statistical and probability letters*, 108, 16-22. doi: 10.1016/j.spl.2015.09.016

FIELD STRESS DETECTION ALGORITHM USING REMOTE SENSING

Cvetković Nikola*[1], Dragović Nebojša[2], Đoković Aleksandar[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
[2]The Ministry of Internal Affairs of the Republic of Serbia
*Corresponding author, e-mail: cvetkovic.nikola@fon.bg.ac.rs

Abstract: *As the human population is growing day by day, need for food is growing. However, possibilities for arable land are limited; it is necessary to improve the process of food production on the arable surfaces and use its full potential. With the development of remote sensing, it is possible to obtain information about agriculture surface without physical contact. Attention is devoted to monitoring the condition of crops and making different decisions based on the obtained data information in the electromagnetic spectrum. This paper presents an algorithm for detection of crop stress in the field using remote sensing. The algorithm is based on the extraction of vegetation indices from the image and comparing histogram of vegetation indices values of a healthy crop with all segments of the image. In this way, it is possible to notice the difference in vegetation indices between the healthy crop and stressed area.*

Keywords*: remote sensing, stress detection, agriculture, vegetation index, the algorithm*

1. INTRODUCTION

We live in the world where the current situation of global food security is one of the main issues. The balance between constantly growing food demand of the world population and global food production is alarming. From year to year, it is necessary to produce more food and raw materials which will be used in the further process of production, to feed growing population. As the climatic changes have a major impact on the production process, more efficient and acceptable production methods should be used. Given that the arable land has less, we must use their potential in the best possible way. Authors (Strange & Scott, 2005) explained that plant protection in general and the protection of crops against plant diseases in particular, have an obvious role to play in meeting the growing demand for food quality and quantity. Oerke, E.C., (2006) discovered that pathogens, animals and weeds, as direct yield losses are altogether responsible for losses between 20 and 40 % of global agricultural productivity. Zadoks, J. C., (1967) explained that crop losses due to pests and pathogens are direct, as well as indirect and they have some facets, some with short-term, and others with long-term consequences. Since the agricultural areas are very large, usually it is not possible to survey the whole field and detect all plant diseases. With the development and expansion of remote sensing in agriculture and using Unmanned Aerial Vehicles (UAVs) as part of it, this problem is overcome. Since every object on the earth, due to solar radiation, radiates the energy of a part of the electromagnetic spectrum, using remote sensing it is possible to obtain information of the observed part of the field that carries part of the electromagnetic spectrum. Analyzing changes in the spectrum are noticed. By measuring these changes and studying their relationships, information on plant health is obtained.

The literature about the use of remote sensing is extensive. It is used in various fields to obtain images without physical contact to object. Suresh, S. et al. (2018) explained how to improve remote sensing process by collecting better quality images from the satellite. Remote sensing found the application in different areas like geology, hazard assessment, oceanography, construction etc. In one paper, Rathje and Franke, (2016) found the use of remote sensing in geotechnical earthquake reconnaissance to document damage patterns and measure ground movements. Thus, the most important application was found in agriculture. The benefits of using remote sensing in agriculture over traditional were described on nitrogen stress detection in wheat by Wright Jr., D. L. et al. (2005). With the appearance of Unmanned Aerial Vehicles, use of remote sensing has experienced an expansion. Many authors (Konar & Iken, 2017; Omar et al., 2017; Xue et al., 2016) presented various possible areas of UAV’s application and their benefits, such as intertidal monitoring, concrete bridge decks surveying and development of automatic aerial spraying systems. Remote sensing application in agriculture was found thanks to multispectral images used for extraction of vegetation indices. Berni, J. et al. (2009) explained how to extract information about vegetation and plant stress using multispectral images. McDonald, A. J. et al. (1998) evaluated different vegetation indices obtained with multispectral images on coniferous forests. Combination of the multispectral image and variable rate technology in remote sensing was used to detect tree health problems and application of a pesticide in the citrus production, as described by Du., Q. et al. (2008).

In agriculture, remote sensing is used for a variety of purposes. Ballesteros, R. et al. (2014) noticed that remote sensing high-resolution images with proper treatment might be considered as a useful tool for precision in monitoring crop growth and development, water requirements, yield production, weed and insect infestations, among others. Vegetation indices, there are a lot of possibilities for monitoring condition of the field. Gitelson and Merzlyak (1998) used vegetation indices to determine the concentration of chlorophyll in higher plant leaves. Zhuang, S. et al. (2017) also used them to determine water stress in early phases. For stress detection in the field, in many papers, k-means is used for classification. Badnakhe and Deshmukh (2011) applied k-means clustering with AI to detect crop diseases. Also, Cheng, H. et al. (2013) presented k-means for image segmentation in their paper. Further research has found that there are many algorithms for stress detection, usually specialised for a particular type of stress.

Paper is organised as follows. After the introduction, Section 2 is devoted to the methodology and algorithm used for stress detection. Section 3 refers to the application of the stress detection algorithm on a wheat field. In Section 4, there is a conclusion and directions for further research.

2. METHODOLOGY

Remote sensing is the process of collecting information about an object or phenomenon without making physical contact with the observed phenomenon or object. It is a phenomenon that has numerous applications including photography, surveying, geology, forestry and other. Also, because of its ability to collect information about the huge area for less time and without physical contact, the remote sensing has found significant use in agriculture. There are very many applications of remote sensing in a field of agriculture like crop production forecasting, crop detection, identification all types of stress. In the field of agriculture, using unmanned aircraft sensor, the object is recorded, which is, in this case, the earth’s surface with the plant cultures. It is possible to determine plant stress using changes of the electromagnetic spectrum.

Plant stress is manifested as any state of the biological system that deviates from the optimum. Different intensity and length of negative effects affect the severity of stress, which always leads directly to a decrease in yield and quality of crops below the genetic potential. Stress causes a large number of natural and factors that are caused by man's actions. Some of the types of stress that are most commonly represented are stress from low temperature, stress from excess water, city and storms, stress from pesticides, whiskey mites, insects and plant diseases. It is thought that 90% of plant production worldwide is endangered by stress and that it has an increasing impact.

The algorithm for field stress detection using remote sensing shown in this paper consists of three main parts. First part includes the preparation of data, an image acquiring using unmanned aerial vehicles (UAV). The second part is pre-processing, which consists of selecting samples from the acquired image, selection of vegetation indices and extracting them, and input values for clusters. The last phase is processing of the image by comparing vegetation indices values of each image segment with the vegetation indices values obtained from samples and put it in the appropriate cluster defined in part two. Each part of the algorithm is explained below in more detail.

2.1. Preparation phase

The data preparation phase involves the process of collecting images from the air. One of the most advanced remote sensing technologies today is the use of UAVs equipped with a camera. This way, in a very quick and easy way, images of a certain surface are suitable for precision agriculture. Gathering images from the air is one of the essential steps of the entire image analysis process. When taking a certain surface, it is necessary to take care of more details to obtain a picture that is appropriate for the further analysis process. Image quality depends on several factors such as camera specifications, weather conditions and flying altitude.

In remote sensing in agriculture, several types of cameras are used to capture fields such as thermal camera, RGB and NIR cameras. Thermal technology is not so often used to getting images from the air because of low spatial and temporal resolution, low shooting and high price. Color RGB (Red, Green, Blue) present images that most closely represent how human eye would see a field from a plane. The advantage of this image type is that RGB is available from most aerial imaginary platforms. Thus, it has certain limitations. Generally, the crop needs to be significantly stressed to see a visual difference that can be identified in a colour image. Also, colour imaginary provides little opportunity to distinguish small differences in areas of high yield. RGB images are suitable for extraction of VARI and ExG vegetation indices. Near-infrared (NIR) imagery provides a greater assessment of plant health than traditional photos by visualising colour bands outside of what the human eye can see. It uses a false colour composite to display information that would normally be invisible to the human eye. The NIR map shows areas of highly vigorous crops in bright red and weak crops or bare soil in grey. The most used image type in agriculture is Normalized Difference Vegetative Index (NDVI). This image type is often obtained using BNDVI filter, and the present calculated index used to monitor crop health and photosynthetic activity.

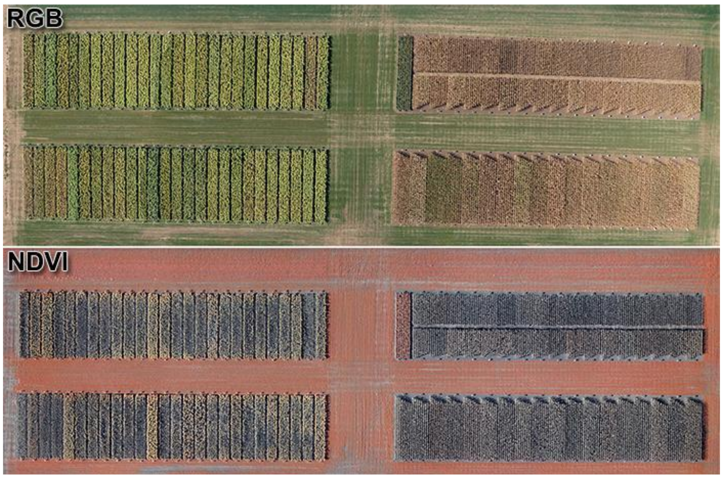


Figure 1: RGB and NDVI images of the field

2.2. Pre-processing phase

After image acquiring, it is necessary to select samples from the image of the field. The samples should be selected that represent a healthy plant crop. To achieve the efficiency of the algorithm, it needs to select at least ten samples.

Vegetation interacts with solar radiation in a different way than other natural materials. The vegetation spectrum typically absorbs in the red and blue wavelengths, reflects in the green wavelength, strongly reflects in the near-infrared (NIR) wavelength, and displays strong absorption features in wavelengths where atmospheric water is present. Different plant materials, water content, pigment, carbon content, nitrogen content, and other properties cause further variation across the spectrum. Measuring these variations and studying their relationship to one another can provide meaningful information about plant health, water content, environmental stress, and other important characteristics. These relationships are often described as vegetation indices.

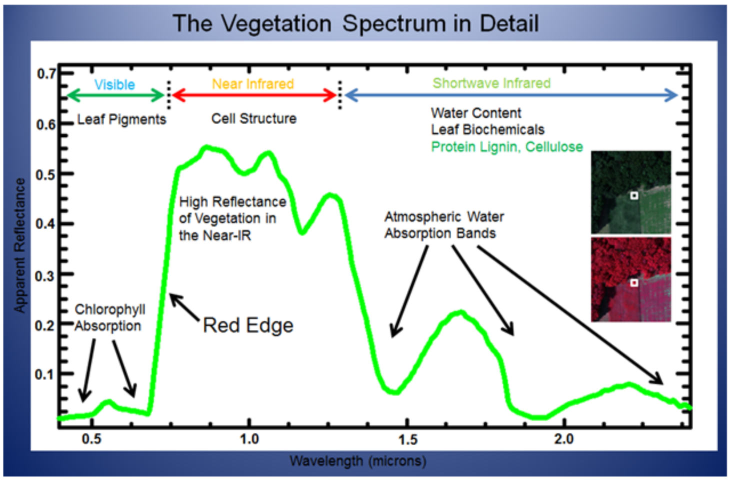


Figure 2: Vegetation Spectrum

The Normalized Difference Vegetation Index (NDVI) is perhaps the most well-known and often used vegetation index. It is a simple, but effective VI for quantifying green vegetation. The NDVI normalises green leaf scattering in the near-infrared wavelength and chlorophyll absorption in the red wavelength. The value range of an NDVI is -1 to 1 where healthy vegetation falls between values of 0.20 to 0.80.The vegetation atmospherically resistant index (VARIgreen) is based on the Atmospherically Resistant Vegetation Index (ARVI) and is used to estimate the fraction of vegetation in a scene with low sensitivity to atmospheric effects. The Excess Green Index (ExG) provides a near-binary intensity image outlying a plant region of interest, from which a segmentation can be accomplished with a suitable threshold.

Depending on the type of image (RGB or NDVI, plant species and type of possible stress, the appropriate vegetation index is selected to be extracted from each sample. The process of vegetation index extraction is following: every sample is divided into a certain number of segments (depending on the spatial resolution of the image), and vegetation index value is calculated. After that, for each sample, a histogram of segment vegetation indices is made. Passing through all the samples and calculating vegetation index histograms, one histogram which represents their average is calculated, and it will be used as a benchmark for the healthy plant for processing phase.

K-means is a widely used clustering algorithm in remote sensing for segmentation. The K-means clustering algorithm tries to classify objects based on a set of features into K number of classes. The classification is done by minimising the sum of squares of distances between the objects and the corresponding cluster or class centroid. Experimental results demonstrate that improved k-means clustering algorithm can reduce the computation amounts and enhance precision and accuracy of clustering (Ballesteros, R. et al., 2014). Thus, using k-means in many cases, it is not able to clearly distinguish a healthy plant from all types of stress in the field. The algorithm overcomes this problem explained in this paper. In the last step of pre-processing phase, the range of values of each cluster is entered. The algorithm is using three predefined clusters which represent: healthy plant, potential stress and stress in the field. Entered values should include a range of 0-100 which represents the percentage of overlapping histograms, so they do not match.

Example of cluster range:

- Healthy plant: 80-100% overlap
- Potential stress: 60-80% overlap
- Stress: 0-60% overlap

With vegetation index benchmark histogram calculated and entered cluster range values, the last phase of the algorithm is accessed.

2.3. Processing phase

Image processing phase is the finishing step of an algorithm for stress detection. Segments of the image are compared with it and depend on the percent of overlapping, clustered in one of the predefined clusters using benchmark vegetation index histogram.

In this phase, the whole image is divided into equal size segments (usually size 1m^2). Then, passing through each image segment, the histogram value of extracted vegetation index is calculated. By comparing the resulting histogram with the benchmark histogram, the percentage of the overlap is determined. Depending on this percentage, the segment will be classified to a particular cluster which includes an overlapping percentage. After classification of all image segments process of stress, detection is finished, and every segment of the image contains information to which cluster belongs. In this way, stress detection has been determined, and the data is ready for further processing and decision making.

3. CASE STUDY

Wheat is one of the most important plants used in the everyday nutrition of people. As the world is harvesting more than 218 million hectares, it is very important that crops are healthy to meet the people needs. Since it is a highly sensitive plant, various insect species can very often harm her. Also, black rust and ash are some of the diseases that can affect the crop due to the action of various insects and microorganisms. All of this points to the importance of monitoring these changes that affect the plant itself. Stress detection algorithm described in the previous section will be applied to the wheat field.

Image acquiring process was carried out by passing and 8.9 ha field of wheat located in the Sutton Bridge, United Kingdom. Aerial images are collected flying 70m from the ground, using DJI's drone, Phantom 4 Advanced. It is equipped with integrated still visible-light camera (RGB) with 20 megapixels objective and 3-axis gimbal stabilisation. After collection of images, stitching software is used to format picture of the whole field.

Figure 3: Healthy wheat sample images

After image stitching, next step is a selection of samples which represent a healthy plant. Ten samples size of 1m^2 is selected from the field image, and shown in Figure 3. For each sample, vegetation indices will be extracted. Since the image was collected using an RGB camera, the most appropriate vegetation index for stress detection of wheat is The Excess Green Index (ExG) which is calculated in the following manner:

$$ExG = 2g - r - b \tag{1}$$

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \qquad b = \frac{B}{R+G+B} \tag{2}$$

Each sample is then split into 10x10 segments, and ExG vegetation index is calculated for all of them. After creating histograms of vegetation index values for all samples, an average histogram was obtained which will be used as a benchmark in the processing phase.

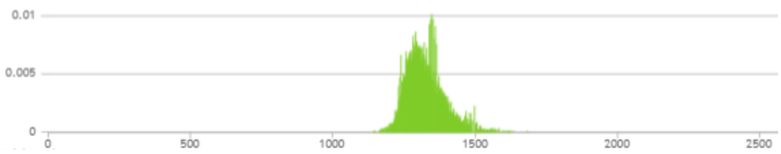


Figure 4: Benchmark vegetation index histogram

Before accessing image processing, cluster range values should be entered for percentages of overlapping. The empirical research found that the algorithm gives the best results for cluster values: Good (80-100% overlap of the histogram), Potential stress (50-80%) and Stress (0-50%).

When the cluster range values are defined, and benchmark histogram is calculated, the image processing step is accessed. Using software for image processing, by passing through each segment of the image, vegetation index histogram is calculated and then compared to benchmark histogram. Depend on the overlapping percent; the segment is classified to the appropriate cluster. In Figure 5, the original image, and image with classified segment are displayed by cluster colours. In the right picture, healthy plant is displayed with green colour, potential stress with a yellow and stressed area with red.

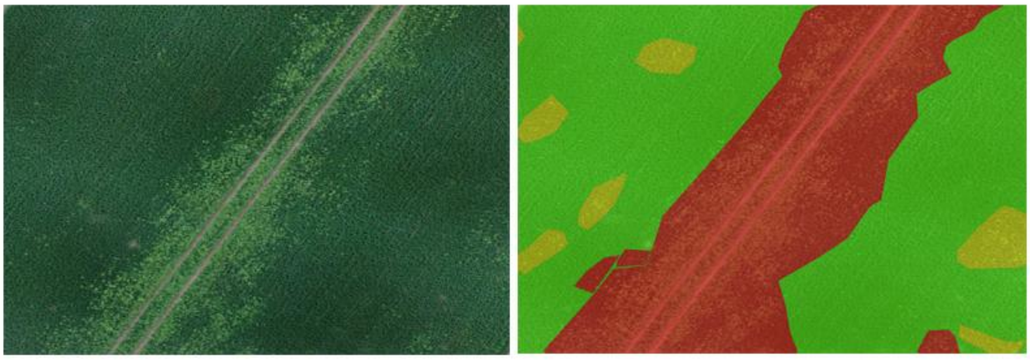


Figure 5: Original image and image with a clustered area

The stress detection algorithm has identified the areas that are affected by stress and need to be rehabilitated, as well as areas with potential stress to which attention should be paid.

4. CONCLUSION

It is possible to influence the health of the plants and yield significantly using and improving various algorithms for stress detection in the field with remote sensing. The algorithm shown in this paper is very convenient to display generalised stress in the field. With geographic information system (GIS), used to capture, store, manipulate, and present geographic data, it is possible to extract georeferenced data for each segment of the field. From these data, in a very easy way, exact coordinates of areas that are under stress should be extracted, and then healed appropriately. In the future, emphasis should be placed on comparing this algorithm with existing ones, with the aim of improving it. Also, the focus should be on possibilities of integration the data obtained with this algorithm with GPS based technologies.

REFERENCES

About stress and biostimulators. Retrieved April 29, 2018, from Ekoplant website, http://www.ekoplantserbia.com

Badnakhe, M. R., & Deshmukh, P. R. (2011). An Application of K-Means Clustering and Artificial Intelligence in Pattern Recognition for Crop Diseases. In *2011 International Conference on Advancements in Information Technology With a workshop of ICBMG*.

Ballesteros, R., Ortega, J. F., Hernández, D., & Moreno, M. A. (2014). Applications of georeferenced high-resolution images obtained with unmanned aerial vehicles. Part I: Description of image acquisition and processing. *Precision Agriculture*, *15*(6), 579-592.

Berni, J. A., Zarco-Tejada, P. J., Suárez, L., & Fereres, E. (2009). Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle. *IEEE Transactions on Geoscience and Remote Sensing*, *47*(3), 722-738.

Candiago, S., Remondino, F., De Giglio, M., Dubbini, M., & Gattelli, M. (2015). Evaluating multispectral images and vegetation indices for precision farming applications from UAV images. *Remote Sensing, 7(4),* 4026-4047.

Cheng, H., Peng, H., & Liu, S. (2013, March). An improved K-means clustering algorithm for agricultural image segmentation. In *PIAGENG 2013: Image Processing and Photonics for Agricultural Engineering (Vol. 8761, p. 87610G).* International Society for Optics and Photonics.

Corrigan, F. (2018, April 22). Multispectral Imaging Camera Drones In Farming Yield Big Benefits. [Web log post]. Retrieved from https://www.dronezon.com/learn-about-drones-quadcopters/multispectral-sensor-drones-in-farming-yield-big-benefits/

Du, Q., Chang, N. B., Yang, C., & Srilakshmi, K. R. (2008). Combination of multispectral remote sensing, variable rate technology and environmental modeling for citrus pest management. *Journal of Environmental Management*, *86*(1), 14-26.

Gitelson, A. A., & Merzlyak, M. N. (1998). Remote sensing of chlorophyll concentration in higher plant leaves. *Advances in Space Research*, *22*(5), 689-692.

Konar, B., & Iken, K. (2017). The use of unmanned aerial vehicle imagery in intertidal monitoring. *Deep Sea Research Part II: Topical Studies in Oceanography*.

McDonald, A. J., Gemmell, F. M., & Lewis, P. E. (1998). Investigation of the utility of spectral vegetation indices for determining information on coniferous forests. *Remote Sensing of Environment, 66(3)*, 250-272.

Oerke, E. C. (2006). Crop losses to pests. *Journal of Agricultural Science, 144*, 31–43.

Olson, K. (2017, August 17). RGB SENSORS VS. NIR SENSORS - WHICH IS BETTER FOR MEASURING CROP HEALTH? [Web log post]. Retrieved from https://www.botlink.com/blog/rgb-sensors-vs-nir-sensors-which-sensor-is-better-for-measuring-crop-health

Omar, T., & Nehdi, M. L. (2017). Remote sensing of concrete bridge decks using unmanned aerial vehicle infrared thermography. *Automation in Construction, 83*, 360-371.

Rathje, E. M., & Franke, K. (2016). Remote sensing for geotechnical earthquake reconnaissance. *Soil Dynamics and Earthquake Engineering, 91*, 304-316.

Samseemoung, G., Soni, P., Jayasuriya, H. P., & Salokhe, V. M. (2012). Application of low altitude remote sensing (LARS) platform for monitoring crop growth and weed infestation in a soybean plantation. *Precision Agriculture, 13(6),* 611-627.

Strange, R. N. & Scott, P. R. (2005). Plant disease: a threat to global food security. *Annual Review of Phytopathology, 43*, 83–116.

Suresh, S., Das, D., Lal, S., & Gupta, D. (2018). Image quality restoration framework for contrast enhancement of satellite remote sensing images. *Remote Sensing Applications: Society and Environment*, *10*, 104-119.

Vegetation Analysis: Using Vegetation Indices in ENVI (n.d). Retrieved April 29, 2018, from Harris geospatial solutions website http://www.harrisgeospatial.com

Wright Jr, D. L., Rasmussen Jr, V. P., & Ramsey, R. D. (2005). Comparing the use of remote sensing with traditional techniques to detect nitrogen stress in wheat. *Geocarto International, 20(1),* 63-68.

Xue, X., Lan, Y., Sun, Z., Chang, C., & Hoffmann, W. C. (2016). Develop an unmanned aerial vehicle based automatic aerial spraying system. *Computers and electronics in agriculture, 128*, 58-66.

Zadoks, J. C. (1967). Types of losses caused by plant diseases. In L. Chiarappa (Ed.), *FAO papers presented at the symposium on crop losses* (pp. 149–158).

Zhuang, S., Wang, P., Jiang, B., Li, M., & Gong, Z. (2017). Early detection of water stress in maize based on digital images. *Computers and Electronics in Agriculture, 140*, 461-468.

CRYPTOCURRENCY PRICE FORECASTING USING TIME SERIES AND MONTE CARLO MODELING AND SIMULATION

Nikola Zornić*[1], Aleksandar Marković[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: Nikola Zornić, nikola.zornic@fon.bg.ac.rs

Abstract: *Over the recent years cryptocurrencies have attracted a significant amount of attention. Everything started with Bitcoin and built up to the situation where we have over 1500 cryptocurrencies. Cryptocurrency market is the new stock market. Highly volatile, decentralised, open, widely accessible market. In this paper we will employ time series analyses together with Monte Carlo simulation to build financial simulation model for forecasting the price, analysing profitability and risk for some of the most popular cryptocurrencies. Although the cryptocurrency market has massive oscillations, any contribution to modeling its value increases the awareness of potential profits and losses.*

Keywords*: cryptocurrency, time series, simulation, model, Monte Carlo, profitability*

1. INTRODUCTION

Cryptocurrencies can be defined as digital, computer currencies whose implementation stands on the principles of cryptography, both to validate the realised transactions and to enlarge the currency in circulation (Cocco, Concas, & Marchesi, 2017).

The most well-known and widely used cryptocurrency is Bitcoin (BTC). At the same time, this cryptocurrency has the highest market valuation, usage, merchant acceptance and popularity (Hayes, 2015). Bitcoin was introduced in 2009, as the first decentralised digital currency platform, a currency which does not have a central authority to regulate it’s usage, validate, and settle transactions (Gandal & Halaburda, 2016).

Following Bitcoin’s footsteps, other cryptocurrencies were launched (Iwamura, Kitamura, & Matsumoto, 2014). Everyone can create its own cryptocurrency in minutes (Long, 2018). All cryptocurrencies formed after Bitcoin are called altcoins. Some of the most popular altcoins are Ethereum (ETH), Litecoin (LTC), Ripple (XRP), Zcash (ZEC), and Monero (XMR).

Bitcoin is one of the most studied cryptocurrencies, many authors analysed its pros and cons. Barber, Boyen, Shi, and Uzun (2012) pointed to several problems with Bitcoin, such as technical vulnerability, potential deflationary spiral, accidental loss of bitcoins, and malware attacks. Yermack (2013) analysed Bitcoin price against fiat currencies and showed that its volatility undermines its usefulness as a currency. Baek and Elbeck (2015) presented strong evidence to suggest that Bitcoin volatility is internally (buyer and seller) driven – leading to the conclusion that the Bitcoin market is highly speculative. Urquhart (2016) showed that Bitcoin market returns are significantly inefficient if observed at once, on the whole sample, but when sample is split into two subsample periods, test indicate that Bitcoin is efficient in the latter period. Bitcoin showed vast success and popularity since its creation (more in the recent years), thanks to its added value. Namely, some of the most important pros of Bitcoin are anonymity, decentralised nature, enhanced revenue, and use of proof-of-work mechanisms (Moore & Christin, 2013).

Cryptocurrencies have not been a topic of great interest for scientific papers until recently. Namely, only 193 articles have been published by the end of 2017 in journals indexed on Clarivate Analytics Web of Science Social Sciences Citation Index (SSCI) and Science Citation Index Expanded (SCIE) with the topic “cryptocurrency” OR “bitcoin” (Clarivate Analytics, 2018). Most of the authors tried to determine which factors influence the cryptocurrency price, trade volume, and volatility. Kristoufek (2013), for example, showed that there is a connection between the search queries and the Bitcoin price. Glaser, Zimmermann, Haferkorn, Weber, and Siering (2014) studied whether the reason for interest in cryptocurrencies on Wikipedia is looking for the new investment asset or the usage as currency itself. The result of the study showed that most of the interest is due to the asset aspect. It will be interesting to investigate their value if they ever become usable for their primary purpose – as medium of exchange for goods and services. Although concept of digital currency would highly increase efficiency of payments in digital world, it currently involves high risk for both consumers and businesses, due to its value instability.

This paper presents steps in modelling the cryptocurrency price and trying to forecast it. The first step was to apply time series analyses for fitting the cryptocurrency historical price data using specific time series model. Afterwards, the resulting function is used for forecasting the price with the help of Monte Carlo simulation.

After defining cryptocurrency and short literature review, in section two, the process of collecting data and building the model is presented. In the same section, the example of investment profitability analyses is demonstrated using the developed model as a framework and running simulation experiments. The results are analysed afterwards. Finally, in section three, conclusion and some future directions of the research are provided.

2. CRYPTOCURRENCY INVESTMENT SIMULATION MODEL

Three most popular cryptocurrencies have been chosen for building the model for forecasting cryptocurrency price: Bitcoin, Ethereum, and Litecoin. Data has been collected for the daily closing price in USD ($) for each of them. Bitcoin data covers the period from 17.07.2010 to 29.04.2018 (CryptoCompare, 2018a), Ethereum from 07.08.2015 to 29.04.2018 (CryptoCompare, 2018b) and Litecoin from 24.10.2013 to 29.04.2018 (CryptoCompare, 2018c). There are a lot of online cryptocurrency historical price databases, but CryptoCompare is selected as the one that has prices for all the mentioned cryptocurrencies. Simple descriptive overview of the data is presented in Table 1.

We can see that Bitcoin price has grown from an average of $0,14 in 2010 to $9.883,01 in 2018. The rate has been even higher for some periods of time, reaching the maximum value of $19.345,49 on 16.12.2017. Ethereum and Litecoin saw excessive growth in value, too.

Table 1: Cryptocurrency prices

Year		Bitcoin price [$]	Ethereum price [$]	Litecoin price [$]
2010	Days	168		
	Mean	**0,14**		
	Std. Dev.	0,09		
2011	Days	365		
	Mean	**5,64**		
	Std. Dev.	5,62		
2012	Days	366		
	Mean	**8,29**		
	Std. Dev.	3,21		
2013	Days	365		69
	Mean	**200,15**		**16,75**
	Std. Dev.	260,77		12,96
2014	Days	365		365
	Mean	**522,89**		**9,93**
	Std. Dev.	176,24		6,32
2015	Days	365	147	365
	Mean	**272,02**	**0,95**	**2,67**
	Std. Dev.	58,92	0,32	1,02
2016	Days	366	366	366
	Mean	**567,00**	**9,76**	**3,76**
	Std. Dev.	138,35	3,67	0,47
2017	Days	365	365	365
	Mean	**3981,07**	**221,68**	**49,85**
	Std. Dev.	3987,18	183,80	64,13
2018	Days	119	119	119
	Mean	**9883,01**	**776,79**	**175,53**
	Std. Dev.	2370,88	256,57	41,11

Table 2: Time series fit details

	Bitcoin price [$]			Ethereum price [$]			Litecoin price [$]		
Type	AR2	GARCH	ARMA	MA1	GARCH	MA2	GARCH	MA1	AR2
Function	Logarithmic	Logarithmic	Logarithmic	Logarithmic	Logarithmic	Logarithmic	Logarithmic	Logarithmic	Logarithmic
Detrend	First Order	First Order	First Order	First Order	First Order	First Order	First Order	First Order	First Order
Deseasonalize	None	None	None	None	None	None	None	None	None
Akaike (AIC) Rank	#1	#2	#3	#1	#2	#3	#1	#2	#3
Akaike (AIC) Fit	-7093.55	-7090.92	-7073.86	-2562.13	-2554.05	-2553.77	-3684.91	-3644.12	-3638.54
Bayesian (BIC) Rank	#2	#3	#4	#1	#3	#4	#1	#2	#4
Bayesian (BIC) Fit	-7069.75	-7067.12	-7050.06	-2562.13	-2534.47	-2534.20	-3663.30	-3644.12	-3616.94
Parameters	4	4	4	3	4	4	4	3	4
Parameter #1	Mu	Mu	Mu	Mu	Mu	Mu	Mu	Mu	Mu
Value	4.27E+11	4.27E+11	3.89E+11	5.27E+11	5.27E+11	5.27E+11	2.39E+11	2.39E+11	2.39E+11
Parameter #2	Sigma	Omega	Sigma	Sigma	Omega	Sigma	Omega	Sigma	Sigma
Value	6.94E-02	4.71E+11	6.90E+12	6.67E+12	4.39E+11	6.78E+12	6.18E+11	8.07E+12	8.01E+12
Parameter #3	A1	A	A1	B1	A	B1	A	B1	A1
Value	3.33E+12	2.78E+11	-1.90E-01	-7.34E+11	2.63E+11	-1.52E-02	3.78E+11	-1.07E-01	-1.15E-01
Parameter #4	A2	B	B1		B	B2	B		A2
Value	-1.77E-01	3.40E+11	2.74E-01		3.19E+10	1.23E+12	4.63E+10		-4.71E+12

Eleven time series algorithms have been employed on cryptocurrencies prices using Palisade @RISK 7.5.2 plug-in software for Microsoft Office Excel 2016. @RISK features automatic detection of required transformations to achieve stationarity. Each algorithm used logarithmic data transformation and first-order detrend. There have not been need to de-seasonalize the data. Akaike Fit (Akaike information criterion – AIC) and Bayesian Fit (Bayesian information criterion – BIC) are used as quality measures for time series fit. Results for the best ranked algorithms for each cryptocurrency are presented in **Error! Reference source not found.**.

AR2 – the autoregressive model showed the best result for representing Bitcoin price, and it will be used for creating the financial model for cryptocurrency profit analyses. This fit is presented in Figure 1, together with parameters for AR2 @RISK function.

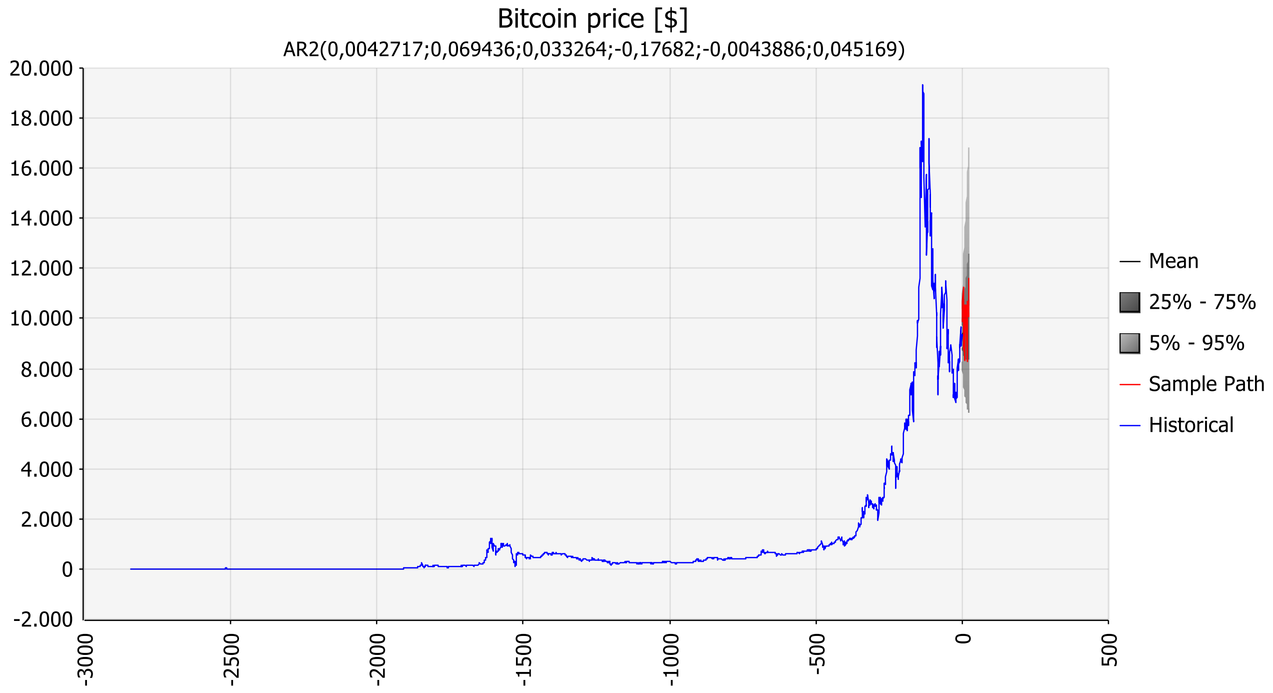


Figure 1: AR2 time series for Bitcoin price

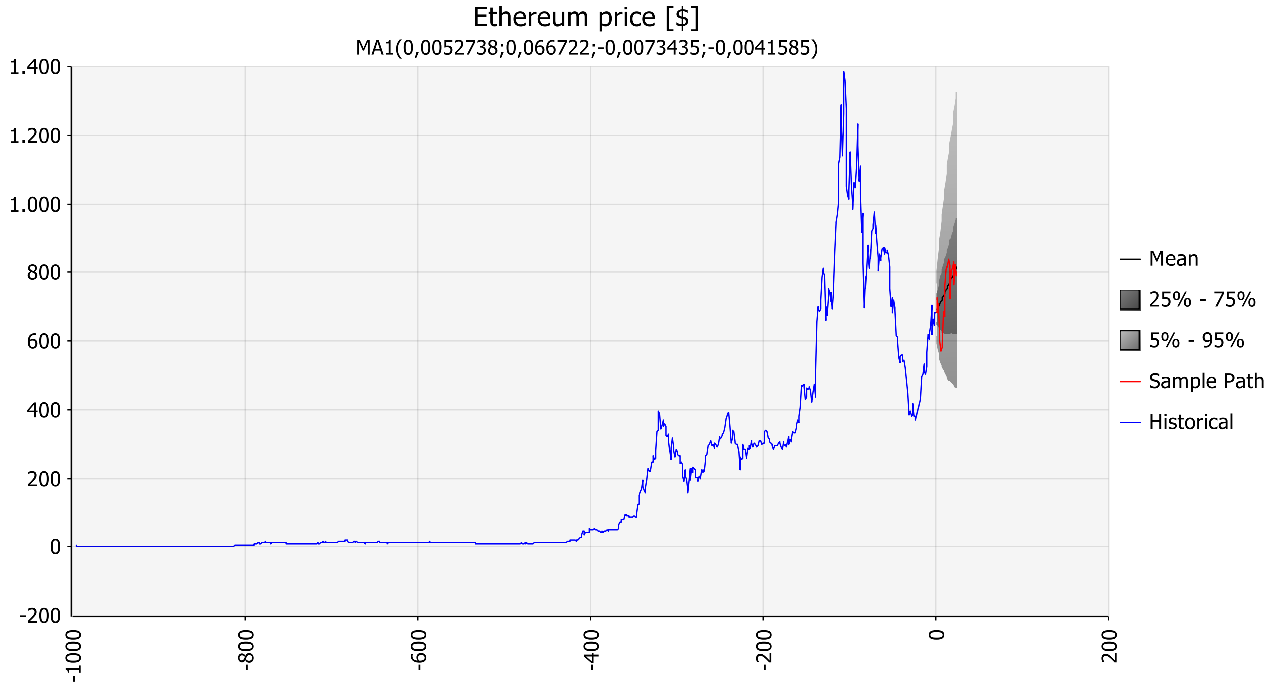


Figure 2: MA1 time series for Ethereum price

MA1 – moving average model showed the best results for representing Ethereum price according to both quality measures, and it will be used for simulating price in the financial model. This fit is presented in Figure 2, including parameters for MA1 @RISK function.

GARCH – generalised autoregressive conditional heteroskedasticity model is selected for representing Litecoin price trend. It has been ranked as the best one according to both quality measures, and it will be used for simulating price in the financial model. This fit is presented in Figure 3, as well as parameters for GARCH @RISK function.

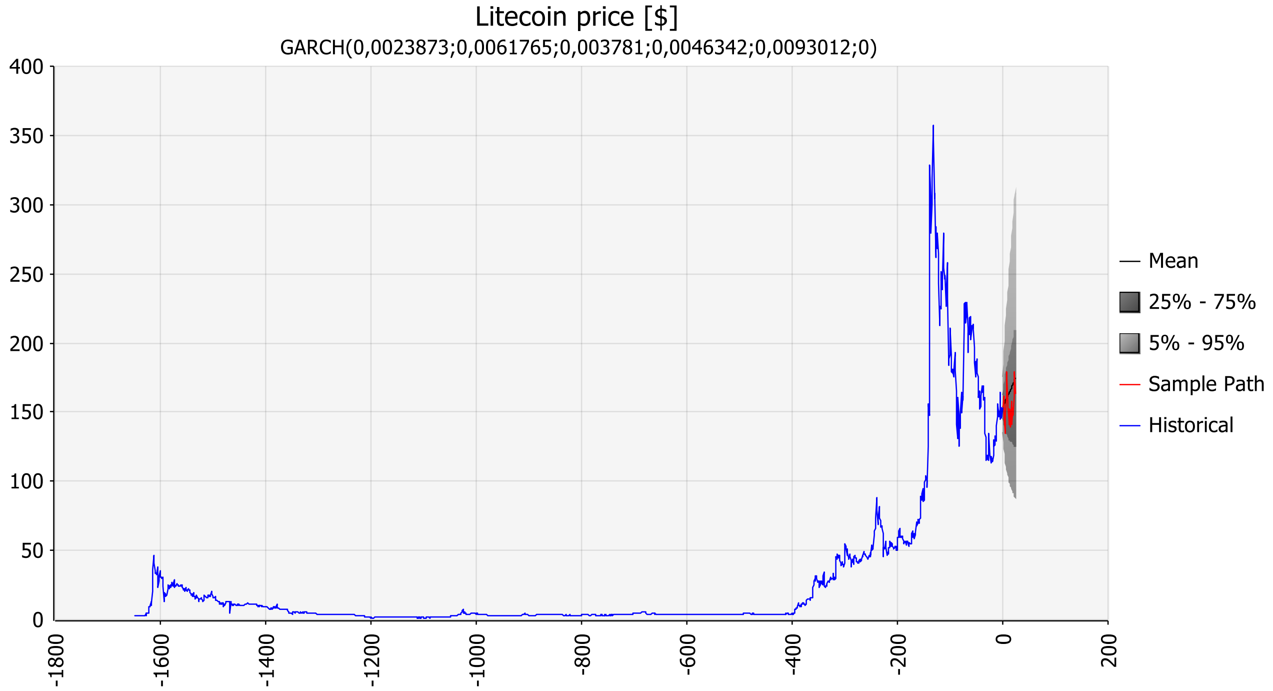
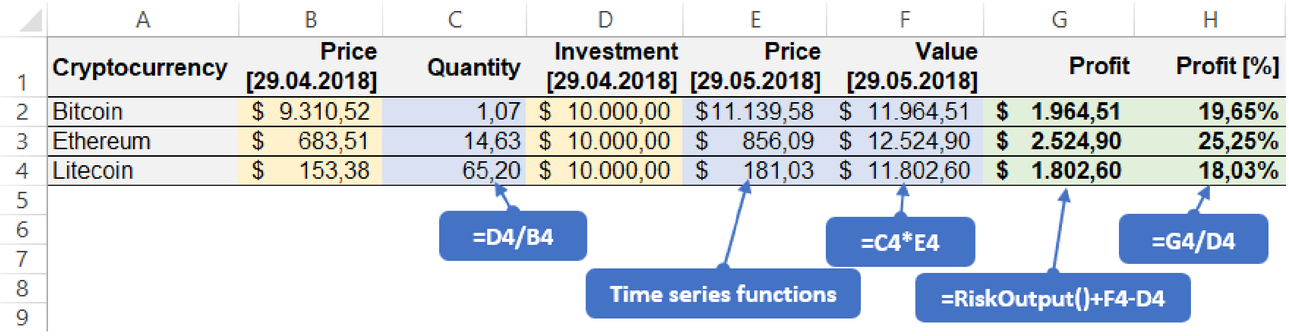


Figure 3: GARCH time series for Litecoin price

We will analyse a hypothetical situation where $10.000 is invested in each observed cryptocurrency on 29.04.2018 (Figure 4 – yellow fields are for input variables, blue for calculation steps, and green for output). Based on the investment, quantity of each cryptocurrency bought is calculated. In the next step, three time series fit functions are used for generating market price on 29.05.2018. Afterwards, the new cryptocurrency value and profit are calculated. Note that RiskOutput() is used for setting the profit as simulation output variable.

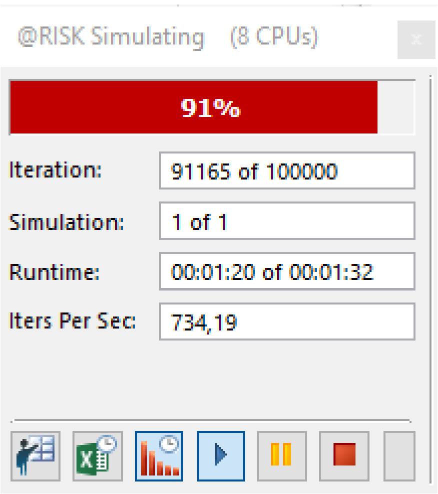


	A	B	C	D	E	F	G	H
1	**Cryptocurrency**	**Price [29.04.2018]**	**Quantity**	**Investment [29.04.2018]**	**Price [29.05.2018]**	**Value [29.05.2018]**	**Profit**	**Profit [%]**
2	Bitcoin	$ 9.310,52	1,07	$ 10.000,00	$11.139,58	$ 11.964,51	**$ 1.964,51**	**19,65%**
3	Ethereum	$ 683,51	14,63	$ 10.000,00	$ 856,09	$ 12.524,90	**$ 2.524,90**	**25,25%**
4	Litecoin	$ 153,38	65,20	$ 10.000,00	$ 181,03	$ 11.802,60	**$ 1.802,60**	**18,03%**
5								
6								
7								
8								
9								

Figure 4: Predicted cryptocurrency profit

After building the model, simulation is conducted using @RISK software with 100.000 iterations (Figure 5).

In each iteration, the cryptocurrency price on 29.05.2018 is generated using the selected time series model, other values are calculated and results are saved in database. When the simulation is finished, results are presented in form of probability distributions with extensive statistical indicators.


@RISK Simulating (8 CPUs)

91%

Iteration: 91165 of 100000

Simulation: 1 of 1

Runtime: 00:01:20 of 00:01:32

Iters Per Sec: 734,19

Figure 5: @RISK Monte Carlo simulation

We can say that investment into any cryptocurrency involves high risk, as the market is highly volatile (Yermack, 2013). Our results showed that the average profit on Bitcoin investment on 29.05.2018 should be $1.964,51, with a standard deviation of $4.135,37. Highest loses are set at $7.548,15 and the highest profit at $37.488,39. If we take a look at Figure 6, we can see that the profit will be between -$3.500 and $9.616 with 90% chance. Results for Ethereum and Litecoin are analysed in a similar manner.

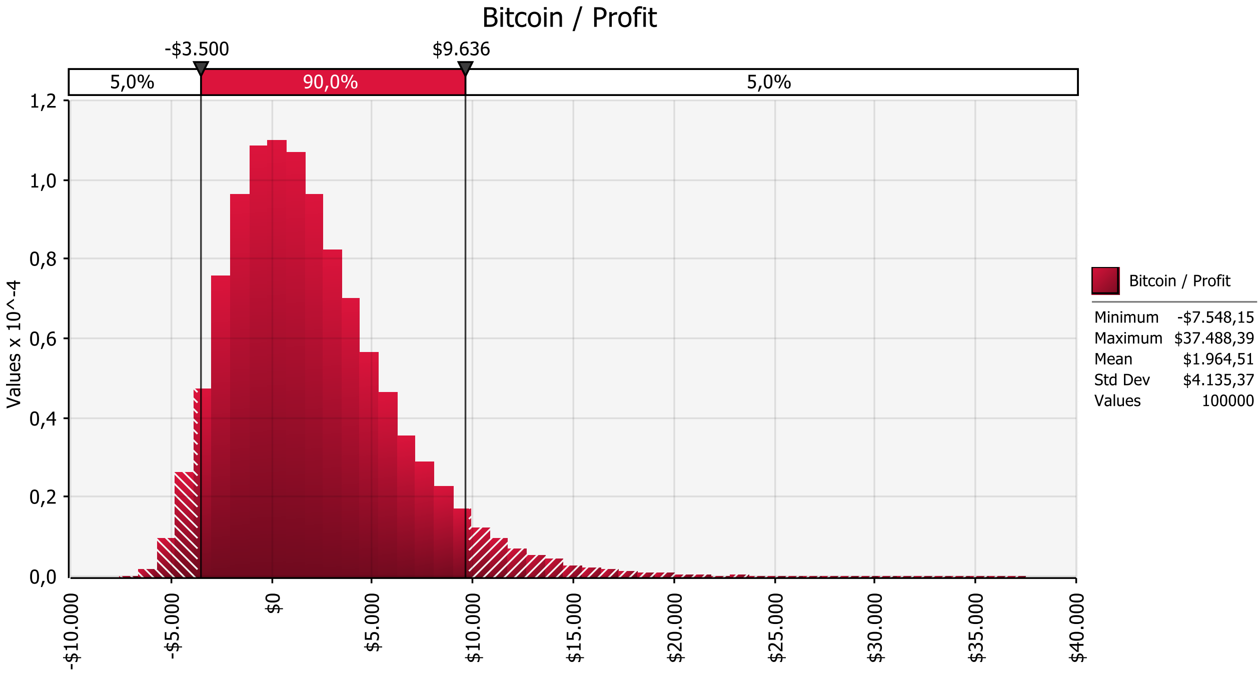


Figure 6: Bitcoin profit

Based on the current market situation and our predictions, Ethereum shows the highest potential for growth and profitability (Figure 7). Namely, mean profit on an investment of $10.000 into Ethereum is $2.524,90, and the maximum possible profit is $46.346,79.

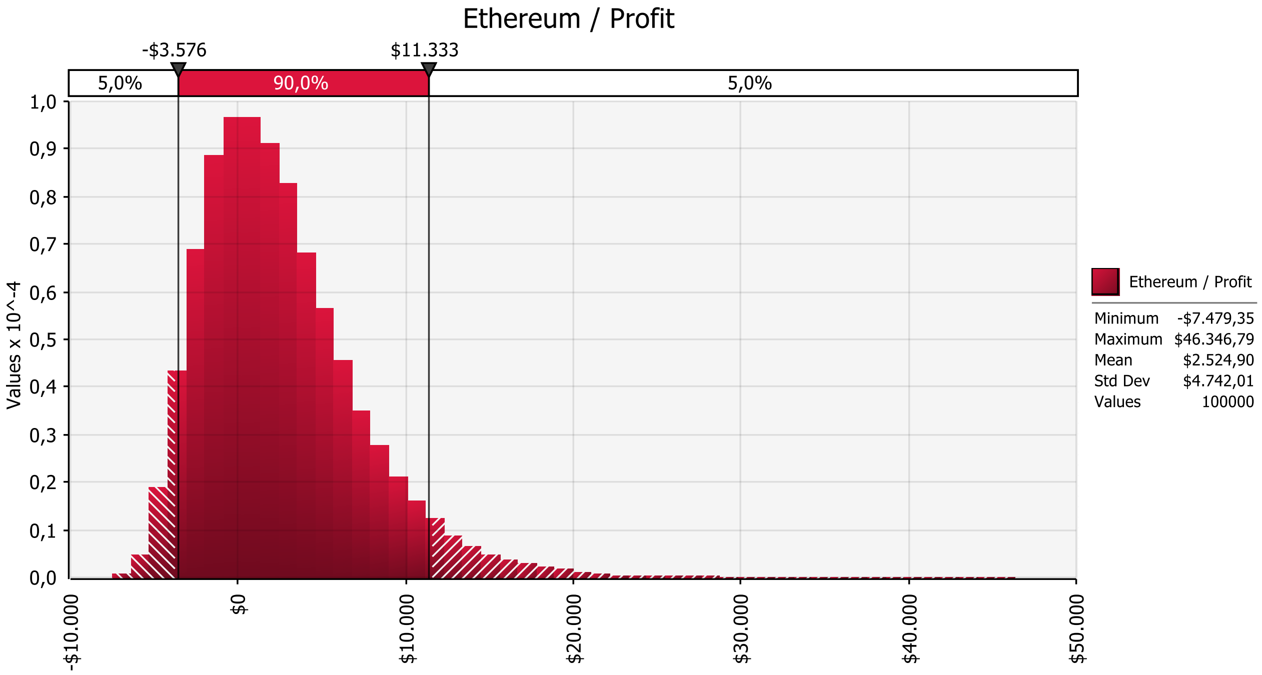


Figure 7: Ethereum profit

Litecoin carries the highest risk (Figure 8), with highest standard deviation and highest possible loses, but also it has the highest potential profit of $71.228,63.

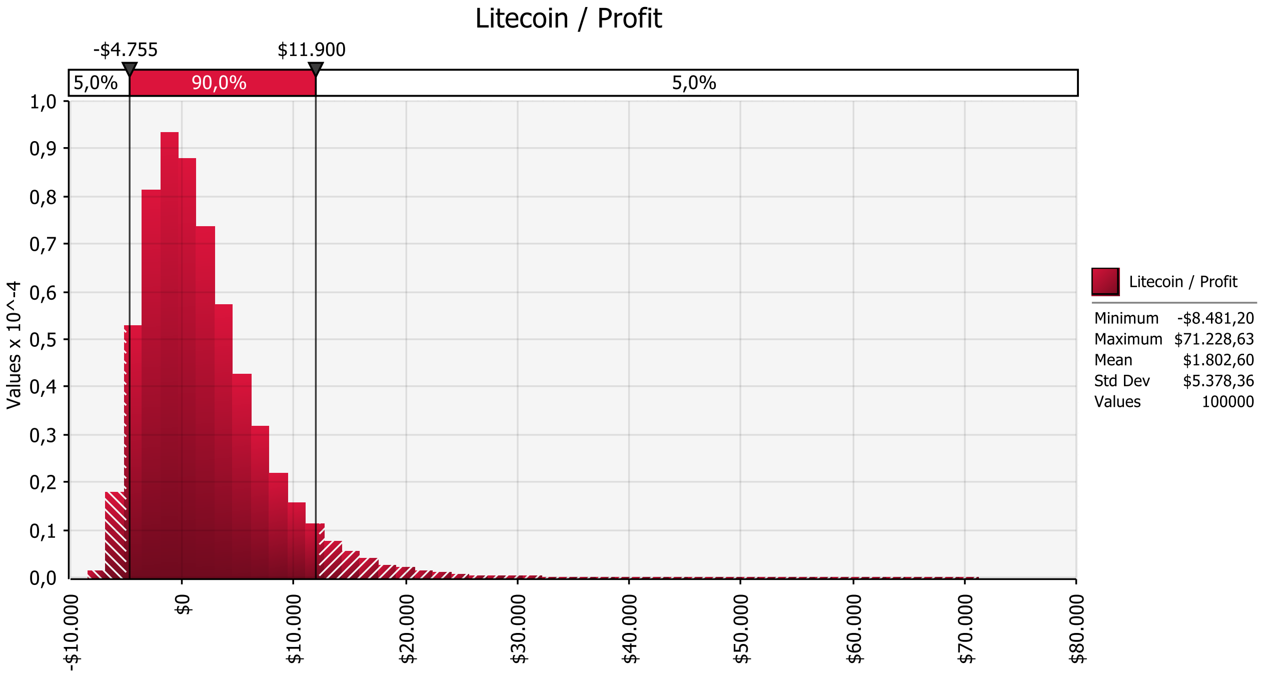


Figure 8: Litecoin profit

3. CONCLUSION

Cryptocurrency popularity is inevitably rising (Cheah & Fry, 2015). Our aim in this paper was to make a step towards equipping individual investors with a model for cryptocurrency price analyses using well-known time series analyses and Monte Carlo simulation.

To encourage other researchers on modelling and forecasting cryptocurrency price, the data sources and widely accessible tools for analyses are suggested. Afterwards, the model is created using the mentioned methodologies. Additionally, the model is employed on the example of investment and results are analysed.

Bitcoin price is prone to speculative bubbles and the bubble component contained within Bitcoin price is substantial (Cheah & Fry, 2015; Fry & Cheah, 2016). The market is highly volatile, but any kind of information gathering and analyses are giving higher insight into the possible price trends. Combination of time series with Monte Carlo simulation gives the investor possibility to analyse potential price in terms of probabilities and statistical indicators.

Monte Carlo simulation results for the three observed cryptocurrencies show high investment risks, but also high potential profits. Ethereum emerged as the cryptocurrency with highest mean and potential maximum profit, while Litecoin has the lowest mean profit and highest maximum losses. Bitcoin is in the middle, with lowest standard deviation, and solid mean profit.

During the research, we identified several future directions of the study. Firstly, the built model can be enabled with portfolio optimisation features. This would significantly reduce potential investors' risk and make investing more tempting. The other direction is regarding the number of cryptocurrencies analysed. The model can be simply expanded to involve a higher number of cryptocurrencies, especially the new and rising ones.

REFERENCES

Baek, C., & Elbeck, M. (2015). Bitcoins as an investment or speculative vehicle? A first look. *Applied Economics Letters*, *22*(1), 30–34. https://doi.org/10.1080/13504851.2014.916379

Barber, S., Boyen, X., Shi, E., & Uzun, E. (2012). Bitter to Better — How to Make Bitcoin a Better Currency. In *Financial Cryptography and Data Security* (pp. 399–414). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-32946-3_29

Cheah, E.-T., & Fry, J. (2015). Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters*, *130*, 32–36. https://doi.org/10.1016/J.ECONLET.2015.02.029

Clarivate Analytics. (2018). Web of Science. Retrieved March 25, 2018, from https://apps.webofknowledge.com

Cocco, L., Concas, G., & Marchesi, M. (2017). Using an artificial financial market for studying a cryptocurrency market. *Journal of Economic Interaction and Coordination*, *12*(2), 345–365. https://doi.org/10.1007/s11403-015-0168-2

CryptoCompare. (2018a). Bitcoin (BTC) - USD - Historical OHLC chart, social data chart and multiple chart indicators. Retrieved April 29, 2018, from https://www.cryptocompare.com/coins/btc/charts/USD?p=ALL&t=LC&e=CCCAGG

CryptoCompare. (2018b). Ethereum (ETH) - USD - Historical OHLC chart, social data chart and multiple chart indicators. Retrieved April 29, 2018, from https://www.cryptocompare.com/coins/eth/charts/USD?t=LC&p=ALL

CryptoCompare. (2018c). Litecoin (LTC) - USD - Historical OHLC chart, social data chart and multiple chart indicators. Retrieved April 29, 2018, from https://www.cryptocompare.com/coins/ltc/charts/USD?p=ALL&t=LC&fTs=1382565600&tTs=1525125600

Fry, J., & Cheah, E.-T. (2016). Negative bubbles and shocks in cryptocurrency markets. *International Review of Financial Analysis*, *47*, 343–352. https://doi.org/10.1016/j.irfa.2016.02.008

Gandal, N., & Halaburda, H. (2016). Can We Predict the Winner in a Market with Network Effects? Competition in Cryptocurrency Market. *Games*, *7*(3), 16. https://doi.org/10.3390/g7030016

Glaser, F., Zimmermann, K., Haferkorn, M., Weber, M. C., & Siering, M. (2014, April 15). Bitcoin - Asset or Currency? Revealing Users' Hidden Intentions.

Hayes, A. (2015, March 16). What Factors Give Cryptocurrencies Their Value: An Empirical Analysis. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2579445

Iwamura, M., Kitamura, Y., & Matsumoto, T. (2014). Is Bitcoin the Only Cryptocurrency in the Town? Economics of Cryptocurrency And Friedrich A. Hayek. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2405790

Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, *3*(1), 3415. https://doi.org/10.1038/srep03415

Long, E. (2018). How to Create Your Own Cryptocurrency. Retrieved April 29, 2018, from https://lifehacker.com/how-to-create-your-own-cryptocurrency-1825337462

Moore, T., & Christin, N. (2013). Beware the Middleman: Empirical Analysis of Bitcoin-Exchange Risk (pp. 25–33). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-39884-1_3

Urquhart, A. (2016). The inefficiency of Bitcoin. *Economics Letters*, *148*, 80–82. https://doi.org/10.1016/j.econlet.2016.09.019

Yermack, D. (2013). Is Bitcoin a Real Currency? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2361599

STATISTICAL AND SOFT COMPUTING TECHNIQUES IN AIRLINE INDUSTRY – A LITERATURE REVIEW

Nikola Vojtek*[1], Ana Poledica[2], Bratislav Petrovic[3]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: vojtekn@gmail.com

Abstract: *Statistical and soft computing techniques are widely used in airline industry for processing a large amount of data and assisting in the decision making process, since they are capable to capture both internal and external factors. Depending on the data, presence of vagueness, uncertainty and model existence, statistical and soft computing techniques could be applied in various processes. Aim of this paper is to provide the analysis of the literature covering the usage of these techniques in airline industry. We reflect on 44 papers published in a period 1961-2018. It consisted of 3 stages: initial analysis (publishing year), brief analysis of proposed methods and identification of operations/business processes. Results are indicating an extensive usage of statistical techniques. However, due to the ability to tolerate imprecision and uncertainty, there is a lot of opportunity for the application of soft computing techniques, especially for the passenger demand, overbooking and no-show forecasting.*

Keywords: *soft computing techniques, statistical techniques, airline, forecasting, no show*

1. INTRODUCTION

Nowadays, processing a large amount of data in airline industry is much easier than in previous few decades. Development of computational technologies allowed carriers to process data much faster and to make real time decision (Min & Joo, 2016; Pineda, Liou, Hsu & Chuang, 2018). Numerous techniques based on mathematics theory are applied in this complex process, and depending on the nature of the problem, there are - statistical or soft computing techniques. As pointed out by Lim and Balas (2013), the main difference is that statistical techniques (ST) are based on principles of precision and certainty, and soft computing techniques (SCT) tolerate imprecision and uncertainty. When it comes to the big data processing, ST need more time and a precisely formulated analytical model for computation. On the other hand, when observing vagueness and inconsistency in data, application of SCT could offer required and much needed solutions.

Two most commonly used SCT are fuzzy logic (FL) and neural networks (NN). As pointed out by Zadeh (1994), FL could be used to a highly extent in solving problems related to the vagueness and NN could be applied in situation in which uncertainty is highly present. When it comes to the ST, time series and statistical analyses are frequently used approaches. Both could be very efficient in capturing trends among data, seasonal components and could be a strong predictive tool with a precisely formulated analytical model. In airline industry, there are several fields in which both SCT and ST could be applied and contribute to the great extent. Fuzzy logic is mostly applied to measure quality of airline service (Li, Yu, Pei, Zhao & Tian, 2017) and neural networks are proposed for various predictions – aircraft maintenance operations (Luxhoj, Williams & Shyur, 1997), seat inventory management (Sun, Brauner & Hormby, 1998) and forecasting passenger demand (Chikr-El-Mezouar & Mohamed Hassan Gabr, 2011). Time Series analysis technique is used for building ARIMA models for forecasting passenger trends (Riddington, 1987) and statistical analysis is applied for building a model that is calculating the overbooking level (Amaruchkul & Sae-Lim, 2011). One of the overlapping area for both techniques is forecasting passenger demand. As highlighted by Cheng and Mengting (2018), various methods and their combinations could be found on this topic. Good example could be Faraway and Chatfield (1998) who combined neural networks and time series to create forecasting model. Aim of this paper is to provide the analysis of the literature covering the usage of both statistical and soft computing techniques for solving issues and challenges in airline industry. The analysis is covering 44 papers published in a 1961-2018 period. Next two sections contain the analysis of papers in which SCT and ST are used respectively. Focus of the analysis is to highlight proposed solutions and applied methods, and to link them to the carrier's business processes and operation areas in which they are used. Section four contains numerical analysis. Section five concludes the paper.

2. APPLICATION OF SOFT COMPUTING TECHNIQUES IN AIRLINE INDUSTRY

Due to the complexity and frequency of changes in the airline industry, carriers are often forced to take into account the uncertainty and ambiguity in the decision making process. Depending on the type and subject of the decision, SCT could be used to embrace uncertainty and deal with the incomplete and inaccurate information. Fuzzy logic could be applied to measure airline service quality, in new routes selection and supplier evaluation. Neural networks are finding their application in forecasting passenger demand, as a standalone technique or with the combination of other SCT and ST. Beside the FL and NN, other SCT were also considered as a part of this research - genetic algorithms and case based reasoning approach. In the following subsections, 17 papers dealing with the application of SCT in airline industry were analyzed.

2.1. Fuzzy logic

One of the key contribution of the fuzzy logic could be seen through the treatment of linguistic variables (Sugeno & Tanaka, 1991) and by applying fuzzy rules and fuzzy graphs (Zhu & Xu, 2014). With respect to the airline industry, an extensive usage of FL is recorded in measuring carrier's service quality. Chou, Liu, Huang, Yih and Han (2011) proposed weighted SERVQUAL method. Since human judgments are often vague, they stated that it is more adequate to use linguistic approach for describing the expectation value, perception value and important weight of evaluation criteria. Proposed method was applied on the Taiwanese airline and in the conclusion, recommendations were made for the carrier to improve current service quality. Li, Yu, Pei, Zhao and Tian (2017) developed a hybrid approach based on fuzzy analytic hierarchy process (AHP) and 2-tuple fuzzy linguistic method. The performance of the proposed three-stage model was measured on a three airlines' in-flight service quality in China. As stated by Perçin (2017), evaluation of the service quality in airline industry has become one of the most challenging tasks that could influence carriers' success on a long term basis. Using the combined fuzzy decision-making approach, Perçin proposed a method for evaluation of the service quality performance of airlines in Turkey. Combined approach, included fuzzy DEMATEL for dealing with the interactions among the evaluation criteria, fuzzy analytic network process for considering the interdependence and calculate the relative importance of each criterion and fuzzy VIKOR for evaluation and ranking.

Similar to addressed carriers' service quality, methods for evaluating service quality of an airport towards carriers were proposed. Pabedinskaitė and Akstinaitė (2014) also used SERVQUAL method, to assess the quality of airport services that has been provided to the carriers. Their method resulted in a set of criteria that are taking into account the changes in consumer needs. Evaluation of the service quality provided by the airports was also a subject of a research conducted by Pandey (2016). Author conducted investigation of service quality using the fuzzy multi criteria decision making method in order to identify the scope of improvements. Two airports in Thailand were assessed, and results indicated that the service quality of both airports is satisfactory, and some areas are marked for improvement.

In addition to the service quality evaluations, FL were also applied in other areas. Atli and Kahraman (2012) developed an aircraft maintenance planning system, in which for minimizing aircraft maintenance planning time, they proposed fuzzy critical path algorithm. Schedule planning was covered by Deveci, Demirel and Ahmetoğlu (2017). They proposed the interval type-2 fuzzy TOPSIS multi-criteria decision-making method for identifying the aspects and selecting the new routes. Rezaei, Fahim and Tavasszy (2014) analyzed suppliers within airline retail industry and emphasized the importance of selecting the most suitable supplier(s) that will meet a carrier's specific needs. Since this is a complex process characterized with the involvement of many, sometimes conflicting, qualitative and quantitative criteria, authors proposed a fuzzy AHP.

2.2. Neural Networks

A significant organizational and technological advantages are offered by neural networks within the decision making process (Aiken & Bsat, 1999). NN could be very useful with the imprecise and data with nonlinear relationship, and when exact model don't exist or it is poorly defined. When it comes to the airline industry, an extensive usage of NN is present in forecasting passenger flight demand. One of the earliest application in this field is proposed by Nam and Schaefer (1995). They highlighted forecasting passenger demand as crucial and necessary for carrier when making decision regarding seats allocation, hiring additional staff during the summer months, or even when ordering materials that have long delivery lead times. Authors also expressed their suspicion that the appropriate statistical technique will provide the best results. Thus, they proposed NN to be applied and included in the forecasting process. Chikr-El-Mezouar and Mohamed Hassan Gabr (2011) proposed iterated neural network models for time series analysis and performed validation on the two time series datasets - airline data and sunspot data. Comparing with the Box-Jenkins ARIMA model, NN gave slightly better results. Similar, Weatherford, Gentry and Wilamowski (2003) proposed NN for forecasting passenger demand and compared results with the traditional forecasting techniques - moving averages, exponential smoothing and regression. Results indicated that even basic NN structures provided better forecasts results.

The application of NN in the forecasting passenger flight demand is evident, but it is not limited to only that field. Luxhoj, Williams and Shyur (1997) developed NN model for prediction of the inspection profiles for aging aircrafts issue. The aim of the NN was to provide the number of abnormal and potentially unsafe conditions in aircraft and/or aircraft components/equipment that carrier could expect. Performance of the proposed NN model was tested against the performance of regression based model, and both models provided good prediction results. A more complex neural network model was applied by Sun, Brauner and Hormby (1998) to help in the decision making process within the carrier's revenue management system. Their findings implied that a significant improvement in accuracy can be achieved by applying NN.

2.3. Genetic algorithms and case based reasoning

A recent research conducted by Demirel and Deveci (2017) is proposing the application of genetic algorithm (GA) variants in the crew pairing process. Main objective of the GA is to generate a combination of crew pairings that will have minimal cost, will cover all flight legs and meet legal criteria. Similar, Kotecha, Sanghani and Gambhava (2004) proposed GA for solving the carriers' crew pairing problem. They added new cost-based uniform crossover (CUC) to the GA in order to solve set partitioning problem efficiently. Combined GA and CUC model was tested using the 28 real-world airline crew scheduling problems and promising results were obtained. Chiu, Chiu and Hsu (2004) proposed the usage of GA in the combination of case based reasoning (CBR) in the aircraft maintenance process. Objective of the proposed machine learning method is to retrieve similar cases for Boeing 747-400, and GA was applied for determining the dynamic weights and to introduce much needed non-similarity functions. In order to enhance the performance of the seat inventory management system, Chang, Hsieh, Yeh and Liu (2006) proposed a CBR seat allocation system in the combination with the dynamic probability method. Using the large carrier data, new combined system provided better results from the first-come first-served method that was currently applied.

3. APPLICATION OF STATISTICAL TECHNIQUES IN AIRLINE INDUSTRY

Processing a large amounts of data, using one of the ST require a significant amount of time and a precisely formulated analytical model. Considering a problem in which these conditions are met, ST could provide excellent results in various decision making processes. Among many techniques that could be categorized here, it can be said that time series and statistical analyses are frequently used approaches within the airline industry. In addition, other ST were also considered - operational research, stochastic programming and others. In the following subsections, 25 papers dealing with the application of ST in airline industry were analyzed.

3.1. Time series analysis

Time series analysis is widely applied ST in airline industry for forecasting passenger flight demand. One of the earliest research was conducted by Riddington (1987), who emphasized the financial benefits that carrier will achieve if forecasting accuracy is increased. Riddington proposed ARIMA model with the combination of cash management process in order to increase accuracy in predicting passenger flight demand. Research examined the effect of forecasting accuracy using the 'safety stock' approach and cash-management policy. Based on the obtained results, Riddington concluded that "*far greater returns can be achieved by adopting, in financial areas, approaches common in planning physical inventories*".

Forecasting method that combines time series analysis with additional approach was proposed by Benitez, Paredes, Lodewijks and Nabais (2013). With the modified Grey model authors introduced new aspect of time-varying coefficients which enabled recent data within database to be more influential. Solution has been tested on approximately 18000 routes from airport origin to airport destination, which is equal to 5857 routes, operated by different airlines, from city origin to city destination. Results indicated that modified Grey model worked properly for all airports and connections. To identify passengers demand trend, Tsui, Balli, Gilbey and Gow (2014) proposed a combination of two ARIMA models - the Box-Jenkins seasonal ARIMA (SARIMA) and the ARIMAX. Proposed solution was tested using the data of the Hong Kong airport. Different monthly time series were used for testing. SARIMA models were tested using the monthly time series between January 1993 and November 2010 and ARIMAX models were applied on the periods of January 2001 to November 2010. As concluded, even for different forecasting periods, both models provided accurate forecasting results with low MAPE and RMSE values.

3.2. Statistical analysis

Statistical analysis found its place within the revenue management area in airline industry. Various methods and solutions were created with the aim to optimize inventory, minimize the worst case regret, and forecast number of seats that carrier should allow to be overbooked. As explained by Somboon and Amaruchkul

(2017), maximization of profit could be achieved by seat inventory control and overbooking (accept number of reservations greater than plain capacity to compensate late cancellations and no-shows). Authors developed two-class overbooking model using the show-up probability, and conduct testing on a real life data. Based on the findings, they concluded that the profit is higher if overbooking limits are included.

To optimize inventory capacity, Chen, Günther and Johnson (2003) proposed optimal yield management policies that are considering statistical learning approach. As emphasized, a good yield management policy could be crucial to the carrier, since it allocates airplane seat capacity to various fare classes and maximizes revenue. Another capacity optimization method was proposed by Amaruchkul and Sae-Lim (2011). Their overbooking models comprised of three show-up rate distributions - normal, beta and deterministic. For testing and comparison purposes, binomial model was used as a benchmark. Based on the obtained results, model with beta show-up rate performed best. Lan, Ball, Karaesmen, Zhang and Liu (2015) developed a simultaneous overbooking model that minimizes the worst case regret. As important factor in airline network management area, Grammig, Hujer and Scheidler (2005) highlighted forecasting passenger demand. Thus, they introduced a multinomial probit specification model for forecasting number of passengers and tested it on a real airline booking data.

3.3. Other statistical techniques

One of the earliest researches of the booking problem with multiple fare classes was conducted by Thompson (1961). Later, Rothstein (1984) analyzed the same problem in order to emphasize the operational research techniques that carriers applied. After 60’s and early 80’s, this field gets more “deserved” attention from a both researchers and practitioners. Belobaba (1989) proposed a probabilistic decision model to be implemented as a part of carrier’s automated booking limit system. Smith, Leimkuhler and Darrow (1992) analyzed multiple models for revenue management in airline industry, and Robinson (1995) proposed probabilistic decision making model to set booking limits on future flights (seat inventory utilization). Van Ryzin and McGill (2000) recommended a simple approach based on adaptive algorithm and stochastic approximation theory to tackle the same problem - seat utilization. Similarly, other researchers proposed solutions with stochastic programming technique (Ahmed & Poojari, 2008; Gu & Zhu, 2016).All efforts towards the seat utilization should lead towards financial results. To measure airline financial performance, Weatherford and Belobaba (2002) proposed simulation analysis based on heuristic decision rule model. The same effort was made by Boyd and Bilegan (2003) who concluded that seat inventory utilization mechanisms are arguably the most important factors for carrier for achieving revenue. Fan and Wang (2013) even included risk and discount in the seat control model using the Markov decision process. To maximize revenue, another group of authors (Huang, Ge, Zhang & Xu, 2013) also used Markov decision process approach and applied on parallel flights with transference. As demand for flights increased progressively, focus slightly shifted from seat optimization to forecasting passenger flight demand. Under the limited demand information, Lan, Gao, Ball and Karaesmen (2008) proposed a new static and dynamic booking control policies. Filder, Wei and Ismail (2011) observed existing forecasting econometric models and evaluated their forecasting performance using multiple error measures. Zeng and Li (2015) analyzed passenger seat choice behavior and suggested overbooking policies to be applied. To determine the level of overbooking, the efficient way is to forecast no show, as suggested by several authors (Fildes, Nikolopoulos, Crone & Syntetos, 2008; Kunnumkal, Talluri & Topaloglu,2012).

4. NUMERICAL ANALYSIS

This section represents the numerical analysis of the literature review. Next figures represent analyzed papers by year and technique group.

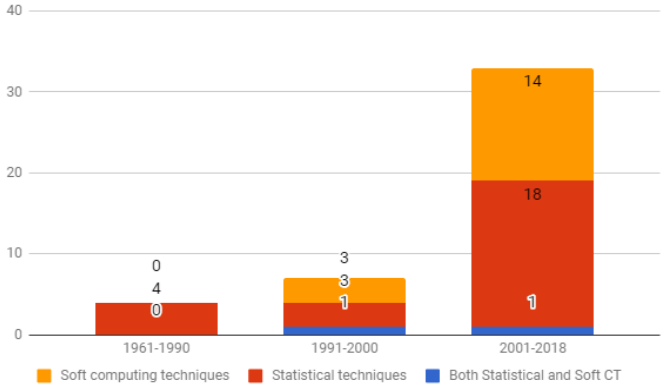


Figure 1: Analyzed papers by year and technique group

Among the total of 44 papers, 17 (38,6%) are covering SCT, 25 (56,8%) are covering ST and 2 (4,5%) are covering combination of both. Publishing years are grouped into three periods, based on the interests in the application of various techniques within the airline industry: (1) 1961-1990, early start; (2) 1991-2000, slow expansion; and (3) 2001-2018, progressive increase of interest. This can also be confirmed by observing the number of papers found and analyzed in those periods. This is not a finite number of papers, but it can definitely serve as a representative example how application of various techniques in solving airline industry issues moved through time. Next figure shows the methods proposed and analyzed in papers.

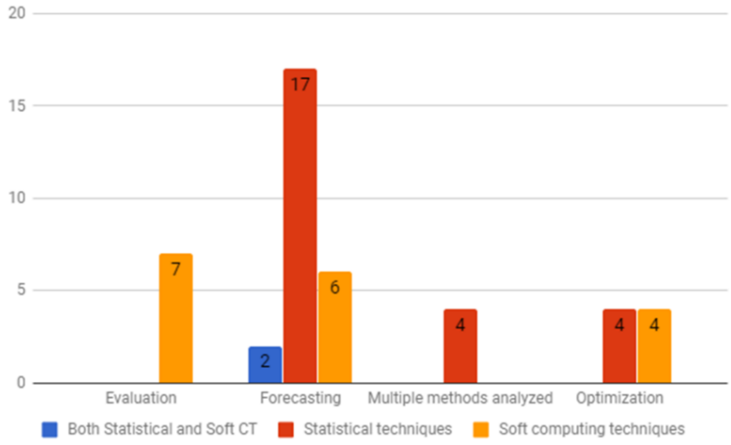


Figure 2: Analyzed papers by methods covered and technique group

Majority of papers are proposing a forecasting method (25) with an extensive usage of ST. In terms of the evaluation methods, a primary role of SCT is evident (7), which is opposite when it comes to papers that analysed multiple methods (4). For the optimization, equal number of statistical and soft computing techniques is captured. Next figure shows the business processes/operations within airline industry that are covered in papers.

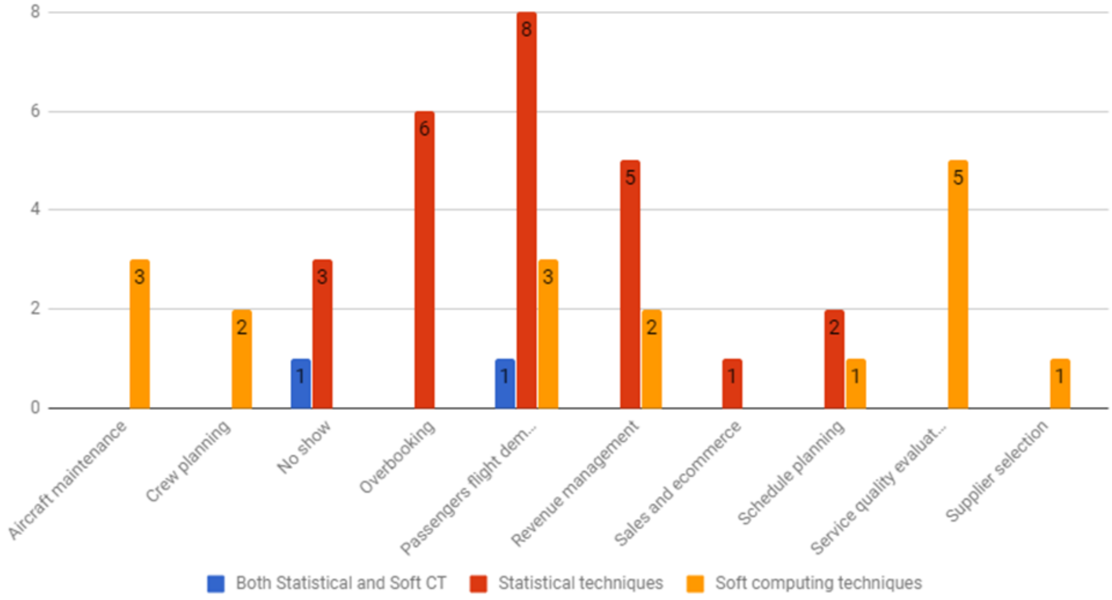


Figure 3: Analyzed papers by business processes/operations covered and technique group

The majority of analyzed papers are considering passenger flight demand forecasting (12), and among others, next business processes/operations are highlighted: seat utilization as a part of revenue management (7), overbooking (6), Service quality evaluation (5) and no show (4).

5. CONCLUSION

Complexity of business processes and operations, and the need for processing large amounts of data, made airline companies an interesting field for the application of computer techniques. Depending on the data, presence of vagueness, uncertainty and model existence, statistical and soft computing techniques could be applied in various decision making processes within airline industry.

Aim of this paper is to provide the analysis of the literature covering the usage of both statistical and soft computing techniques for solving issues and challenges in airline industry. For the purpose of this literature review, papers from various scientific journals were used. Analysis covered three stages: initial analysis (focus on publishing year and techniques), brief analysis of the proposed methods and identification of operations/business processes. Based on the initial analysis, it can be said that the airline industry is receiving a lot of attention nowadays when it comes to the application of SCT. A majority of founded and analyzed papers are covering a period of 2001-2018, and just few of them could be found in earlier periods. More than half of the papers are proposing a forecasting method (25) with an extensive usage of statistical techniques. Passenger flight demand forecasting is still one of the most interesting topics (12), as well as inventory seat utilization as a part of revenue management (7).

Based on the analysis, it can be concluded that there is still a lot of opportunity for the application of SCT, especially when it comes to the passenger demand, overbooking and no show forecasting. Although this research is not final and certainly not covering all papers that are dealing with the application of SCT and ST in airline industry, it could be used as a guide for additional researches and reviews.

REFERENCES

Ahmed, A., H. &Poojari, C., A. (2008). An Overview of the Issues in the Airline Industry and the Role of Optimization Models and Algorithms. *The Journal of the Opr Research Society*, 59(3), 267-277.

Aiken, M., & Bsat, M., (1999). Forecasting market trends with Neural Networks. *Information Systems Management*, 42-48.

Amaruchkul, K., & Sae-Lim, P. (2011). Airline overbooking models with misspecification. *Journal of Air Transport Management*,17, 143-147.

Atli, O., & Kahraman, C. (2012). Aircraft Maintenance Planning Using Fuzzy Critical Path Analysis. *International Journal of Computational Intelligence Systems*, 5(3), 553-567.

Belobaba, P., P. (1989). Application of a Probabilistic Decision Model to Airline Seat Inventory Control. *Operations Research*, 37(2), 183-197.

Benítez, R.B.C., Paredes, R.B.C., Lodewijks, G., & Nabais, L.J. (2013). Damp trend Grey Model forecasting method for airline industry. Expert Systems with Applications, 40, 4915–4921.

Boyd, E., A., &Bilegan, I., C. (2003). Revenue Management and ECommerce. Management Science, *Special Issue on E-Business and Management Science*, 49(10), 1363-1386.

Chang PC., Hsieh JC., Yeh CH., & Liu CH. (2006) A Case-Based Seat Allocation System for Airline Revenue Management. In: Huang DS., Li K., Irwin G.W. (eds) *Intelligent Computing. ICIC 2006. Lecture Notes in Computer Science*, vol 4113. Springer, Berlin, Heidelberg.

Chen, C., P., V., Günther, D., & Johnson, L., E. (2003). Solving for an Optimal Airline Yield Management Policy via Statistical Learning. *J of the Royal Statistical Society. Series C (Applied Statistics)*, 52(1), 19-30

Cheng, L. & Mengting, X. (2018). A Review of Research on Airline Passenger Volume Forecasting. *4th International Conference on Machinery, Materials and Computer (MACMC 2017). Published in Advances in Engineering Research*, 150, 3112-319.

Chikr-El-Mezouar, Z. & Mohamed Hassan Gabr, M. (2011). Iterated neural network models for time series analysis. *International Journal of Statistics*, 69(2), 129-149. doi: https://doi.org/10.1007/BF03263553

Chiu, C., Chiu, N.-H., & Hsu, C.-I. (2004). Intelligent aircraft maintenance support system using genetic algorithms and case-based reasoning. *Int. J. Adv. Manuf. Technol.*, 24, 440–446. doi: 10.1007/s00170-003-1707-x

Chou, C-C., Liu, L-J., Huang, S-F., Yih, J-M., &Han, T-C. (2011). An evaluation of airline service quality using the fuzzy weighted SERVQUAL method. *Applied Soft Computing*, 11(2), 2117-2128. doi: https://doi.org/10.1016/j.asoc.2010.07.010

Demirel, Ç. N., & Deveci, M. (2017). Novel search space updating heuristics-based genetic algorithm for optimizing medium-scale airline crew pairing problems. *Int.J. of Comp Intel. Sys.*, 10, 1082–1101.

Deveci, M., Demirel, N. C., & Ahmetoğlu, E. (2017). Airline new route selection based on interval type-2 fuzzy MCDM: A case study of new route between Turkey - North American region destinations. *Journal of Air Transport Management*, 59, 83-99. doi: https://doi.org/10.1016/j.jairtraman.2016.11.013

Fan, W., & Wang, J. (2013). A model for airline seat control considering risk and discount. *Proceedings of the 2nd Int. Conf. on Computer Science and Electronics Engineering (ICCSEE 2013)*, 0114-0117.

Faraway, J. &Chatfield, C. (1998). Time Series Forecasting with Neural Networks: A Comparative Study Using the Airline Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 47(2), 231-250.

Filder, R., Wei, Y., &Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *Int.J.of Forecasting*, 27, 902–922.

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and Operational Research: A Review. *The Journal of the Operational Research Society*, 59(9), 1150-1172.

Grammig, J., Hujer, R., &Scheidler, M. (2005). Discrete Choice Modelling in Airline Network Management. *Journal of Applied Econometrics*, 20(4), 467-486.

Gu, Y., & Zhu, J. (2016). Study on Seat Capacity Allocation in Airline Alliance. *Advances in Social Science, Education and Humanities Research*, 63, 358-364.

Huang, Y., Ge, Y., Zhang, X., and Xu, Y. (2013). Overbooking for parallel flights with transference. *Int. J. Production Economics*, 144, 582–589.

Kotecha K., Sanghani G., & Gambhava N. (2004) Genetic Algorithm for Airline Crew Scheduling Problem Using Cost-Based Uniform Crossover. In: Manandhar S., Austin J., Desai U., Oyanagi Y., Talukder A.K. (eds) Applied Computing. AACC 2004. *Lecture Notes in Computer Science*, vol 3285. Springer, Berlin, Heidelberg. doi: https://doi.org/10.1007/978-3-540-30176-9_11

Kunnumkal, S., Talluri, K., & Topaloglu, H. (2012). A randomized linear programming method for network revenue management with product-specific no-shows. *Transportation Science*, 46(1), 90–108.

Lan, Y., Ball, O. M., Karaesmen, Z. I., Zhang, X. J., & Liu, X. G. (2015). Analysis of seat allocation and overbooking decisions with hybrid information. *European J. of Operational Research*, 240, 493–504.

Lan, Y., Gao, H., Ball, O. M., &Karaesmen, I. (2008). Revenue Management with Limited Demand Information. *Management Science*, 54(9), 1594-1609.

Li, W., Yu, S., Pei, H., Zhao, C., and Tian, B. (2017). A hybrid approach based on fuzzy AHP and 2-tuple fuzzy linguistic method for evaluation in-flight service quality. *Journal of Air Transport Management*, 60, 49-64. doi: https://doi.org/10.1016/j.jairtraman.2017.01.006

Lim, C.P., Balas, V.E., & Do, Q., (2013). Special issue recent advances in soft computing: Theories and applications. *Journal of Intelligent & Fuzzy Systems*, 24, 415-416. doi:10.3233/IFS-2012-0562

Luxhoj, J.T., Williams, T.P. & Shyur, Hj. (1997). Comparison of regression and neural network models for prediction of inspection profiles for aging aircraft. *IIE Transactions*, 29(2), 91-101. doi: https://doi.org/10.1023/A:1018585211110

Min, H. & Joo, S-J. (2016). A comparative performance analysis of airline strategic alliances using data envelopment analysis. *Journal of Air Transport Management*, 52, 99-110. doi: https://doi.org/10.1016/j.jairtraman.2015.12.003

Nam, K., & Schaefer, T. (1995). Forecasting international airline passenger traffic using neural networks. *Logistics and Transportation Review*, 31(3), :239.

Pandey, M. M. (2016). Evaluating the service quality of airports in Thailand using fuzzy multi-criteria decision making method. *Journal of Air Transport Management*, 57, 241-249. doi: https://doi.org/10.1016/j.jairtraman.2016.08.014

Pabedinskaitė, A., & Akstinaitė, V. (2014). Evaluation of the Airport Service Quality. *Procedia - Social and Behavioral Sciences*, 110, 398-409. doi: https://doi.org/10.1016/j.sbspro.2013.12.884

Perçin, S. (2017). Evaluating airline service quality using a combined fuzzy decision-making approach. *Journal of Air Transport Management*. In press. doi: https://doi.org/10.1016/j.jairtraman.2017.07.004

Pineda, P.J.G., Liou, J.J.H., Hsu, C-C. & Chuang, Y-C. (2018). An integrated MCDM model for improving airline operational and financial performance. *Journal of Air Transport Management*, 68, 103-117. doi:https://doi.org/10.1016/j.jairtraman.2017.06.003

Rezaei, J., Fahim, P. B. M., &Tavasszy, L. (2014). Supplier selection in the airline retail industry using a funnel methodology: Conjunctive screening method and fuzzy AHP. *Expert Systems with Applications*, 41(18), 8165-8179. doi: https://doi.org/10.1016/j.eswa.2014.07.005

Riddington, G., L. (1987). Forecast Accuracy: The Financial Benefits to a Small Airline. *The Journal of the Operational Research Society*, 38(6), 479-485.

Robinson, L., W. (1995). Optimal and Approximate Control Policies for Airline Booking with Sequential Nonmonotonic Fare Classes. *Operations Research*, 43(2), 252-263.

Rothstein, M. (1984). OR and the Airline Overbooking Problem. *Operations Research*, 33(2), 237-248.

Smith, C., Leimkuhler, J.F., &Darrow, R.M., (1992). Yield management at American Airlines. *Interfaces*, 22, 8–31.

Somboon, M., & Amaruchkul, K. (2017). Applied Two-Class Overbooking Model in Thailand's Passenger Airline Data. *The Asian Journal of Shipping and Logistics*, 33(4), 189-198.

Sugeno, M., & Tanaka, K., (1991). Successive identification of a fuzzy model and its applications to pre-diction of a complex system. *Fuzzy Sets and Systems*, 42, 315-334.

Sun X.S., Brauner E., & Hormby S. (1998) A Large-Scale Neural Network for Airline Forecasting in Revenue Management. In: Yu G. (eds) Operations Research in the Airline Industry. *International Series in Operations Research & Management Science*, 9. Springer, Boston, MA. doi: https://doi.org/10.1007/978-1-4615-5501-8_2

Thompson, H.R., (1961). Statistical problems in airline reservation control. Operations *Research Quarterly*, 12, 167–185.

Tsui, K.H.W., Balli, O.H., Gilbey, A., & Gow, H. (2014). Forecasting of Hong Kong airport's passenger throughput. *Tourism Management*, 42, 62-76.

Van Ryzin, G., & McGill, J. (2000). Revenue Management Without Forecasting of Optimization: An Adaptive Algorithm for Determining Airline Seat Protection Levels. *Management Science*, 46(6), 760-775.

Weatherford, L., R. & Baobaba, P., P. (2002). Revenue Impacts of Fare Input and Demand Forecast Accuracy in Airline Yield Management. *The Journal of the Operational Research Society*, 53(8), 811-821.

Weatherford, L., Gentry, T. & Wilamowski, B. (2003). Neural network forecasting for airlines: A comparative analysis. *Journal of Revenue and Pricing Management*, 1(4), 319-331. doi: https://doi.org/10.1057/palgrave.rpm.5170036

Zeng, X., & Li, Y. (2015). Research on Passenger Seat Choice Behavior in Airline Revenue Management. *3rd International Conference on Management, Education, Information and Control (MEICI 2015)*. Atlantis Press. 1407-1413.

Zhu, B., & Xu, Z., (2014). A fuzzy linear programming method for group decision making with additive reciprocal fuzzy preference relations. *Fuzzy Sets and Systems*, 246, 19-33. doi: http://dx.doi.org/10.1016/j.fss.2014.01.001

THE IMPACT OF LOW COST CARRIER ON COMPETITION IN LONG HAUL MARKET: LONDON - NEW YORK ROUTE

Jovana Kuljanin*[1], Milica Kalic[1], Manuel Renold[2]
[1]University of Belgrade – Faculty of Transport and Traffic Engineering, Vojvode Stepe 305, Belgrade
[2]ZHAW Zurich University of Applied Sciences, Gertrudstrasse 15, 8400 Winterthur, Switzerland
*Jovana Kuljanin, e-mail: j.kuljanin@sf.bg.ac.rs

Abstract: *The introduction of low cost model into long-haul service has been a challenging task which is proved to be successful for several carriers across the globe. Norwegian Air Shuttle has been the pioneering carrier in Europe that adopted the low-cost model in long-haul market mainly focusing its network on high density transatlantic routes from Scandinavian cities as well as from United Kingdom. This paper performs analysis on some aspects of competition in the route that connects two metropolitan cities London and New York. By employing descriptive statistics, the comprehensive analysis on certain competition indices such as market share, market concentration index (HHI) and price comparison of fares has been performed in this paper. The findings approve that Norwegian expansion at London Gatwick has became a serious threat to well established carrier on this route, particularly British Airways that put substantial effort to create counter strategy to efficiently combat the rival.*

Keywords: *Low cost long haul model, Norwegian Air Shuttle, transatlantic route, British Airways, secondary airport*

1. INTRODUCTION

The Deregulation Act stipulated in 1978 in United States has become a milestone for revolutionary changes bringing the number of innovations. The most important among them are regulatory changes that certainly shape the market structure allowing the entrance of new players on the market and providing the level playing field for all participants. The wave of changes was riding at phenomenal speed serving as an impetus to other regions, specifically Europe, to liberalize its market across national borders. The emergence of low-cost business model can be perceived as a “pure fruit” of market liberalization (Francis et al., 2006) in Europe that permanently alter the landscape of competition. In such new circumstances, the full-service network carriers (FSNCs) initially protected by the state, coped the severe competitive pressure induced by low-cost carriers (LCC) and their business model that aims at keeping the costs as lower as possible. Although the LCCs organize their network as point-to-point system opposed to FSNCs that highly rely on hub-and-spoke network concept, they still have a large portion of overlapping market in which a fierce battle is evident among them.

Thanked to its simple operating model with no-frill product (Cento, 2009) and their network focused on connecting intra-European destinations , the LCCs successfully survived the crisis occurred as an outcome of terrorist attack in 2001. Moreover, the emergence of LCCs with their flights operated from secondary, less congested airport (in the vicinity of primary airports) become a serious threat to FSNCs that operate their flights from primary (hub) airports. In that sense, the competition must be light at a broader context taking into account not only competition among competitors on specific airport pairs, but also between city pairs or even region levels.

According to Dobruszkes (2006) who used exhaustive data from 2004, the LCCs contained 18% of regular intra-European seats at the time, and that these seats were limited almost exclusively to Western Europe (98%). In the following years, the LCCs put substantial effort to diversify its networks towards markets of Central and Eastern Europe (Dobruszkes, 2009). Finally, in recent years some European carriers have attempted to incorporate the low-cost business concept into long-haul markets mainly by connecting large European cities with large metropolis in North America. The aim of this paper is to investigate some aspects of competition on long-haul route where low-cost carrier offer its service in addition to several full-service carriers that are traditionally dominant in these markets.

The paper is divided as follows. After a brief Introduction, the Section 2 provides literature review on long-haul low cost business model. Section 3 outlines some essential information on long-haul low cost model that emerge as a new business concept by exploring Norwegian Air Shuttle as a representative of the airline group that successfully applies this business model. The Section 4 highlights some aspects of competition between on the route between London and New York characterized by the presence of both LCC and FSNCs. In this section, several measures that represent some aspect of competition will be calculated such as market share and HHI index. Additionally, the price competitiveness of airlines on this route will be given through comparison of air fares obtained by Internet search. Finally, Section 5 concludes the paper.

2. LITERATURE REVIEW

The long-haul low cost business model has been relatively novel in airline industry. In recent few years, abundant of literature investigate the viability of such concept that require the radical enhancement in order to be profitable. Bearing in mind that large number of airlines adhered to this model in recent past, the investigation of cost and revenue aspect of their business model become essential factor for their sustainability on the dynamic competitive market. Low cost business model has been reserved for short and medium-haul routes for years with 50-60% cost savings compared to FSNCs, the advantage that is impossible to achieve on long-haul routes.

The viability of low cost business concept into long-haul service has raised the fierce debate among scholars and airline experts who broadly investigated the financial aspects of such model. Morrell (2008) was among the first who examined the applicability of well-established LCC business model advantages into long-haul service. The author took a rather pessimistic approach and questioned the problem of generating demand (due to the lack of connecting passengers) to support the existence of hub by-pass service. The author also claimed that lowering long-haul fares significantly from current fares is not feasible for LCCs. On the other hand, Daft and Albers (2012) held optimistic side emphasizing the importance of revenue consideration as a key factor of feasible existence of LCC long-haul service. The authors found that ancillary revenues can significantly contribute to airline's profitability. The recent study conducted by De Poret et al. (2015) who performed a detailed financial assessment of low-cost operation on the transatlantic market leads to similar conclusions. Namely, the authors revealed that higher seating densities, higher cargo revenues and additional ancillary revenues can ensure the economic viability of long-haul LCC operation

Despite previous work that focused on revenue side of long-haul low cost model, Soyk et al. (2017) focused solely on evaluation of cost differences between 37 airlines that operates transatlantic routes, among which there are those who adopt low-cost business model. The authors found that emerging long-haul LCC carriers within derived cluster achieved 33% lower unit costs compared to legacy hub carriers, of which 24 percentage points were evaluated as sustainable. Finally, in the contrast to the previous researches, Soyk et al. (2018) found that the emerging North Atlantic long-haul LCCs do not have a revenue disadvantage compared to FSNCs particularly on dense routes approved by the application of new metric proposed.

In addition to already well-established European long-haul carriers, one can anticipate that prominent LCCs such as Ryanair could pave the way for successful acquiring of long-haul operation in the future, the idea which is thoroughly investigated in van den Hoek (2017). Having in mind that there is a potential of between 20 and 113 out of 442 routes of up to 12,000 km that do not have non-stop flight to and from Europe (Wilken et al., 2016) combined to the tendency of LCCs to enter the charter airline long-haul territory (Rodríguez and O'Connell, 2017), the development of future network of LCC long-haul carriers will be challenging task.

3. THE EVOLUTION OF NORWEGIAN LONG HAUL LOW COST BUSINESS MODEL

As stated in Wensveen and Leick (2009), the concept of low-cost long haul flying dated back to 1977 when Skytrain, the company founded by Freddie Laker, operated between New York and London offering airfares substantially lower than its legacy competitors. However, it took several decades that this concept becomes well established in airline industry. The current iteration of the long haul low cost model has been around for a decade, with Jetstar (2006) and AirAsia X (2007) the initial pioneers (CAPA, 2017). Concerning European floor, Norwegian Airlines was the pioneering company that launched its first long haul flight in 2013 between Oslo and New York and shortly after between Stockholm and New York. In addition to these transatlantic flights from Scandinavia (including Copenhagen), the carrier introduced long-haul links from three large European cities: London (2014), Paris (2016) and Barcelona (2017). By the end of October 2017, the airline long haul network encompassed 26 destinations and 28 routes that place it as a largest long haul low cost operator in terms of network size and in the second place in terms of weekly seats (Table 1).

However, this network expansion is supported by the strategy of exploit new Boeing 787-9 with 294 seats onboard. The long-haul strategy for the 787 fleet currently highly relies on dense routes into popular leisure markets that can easily be stimulated by low-fares. These routes include New York (via JFK and now Newark), Los Angeles, San Francisco (via Oakland) and Miami (via Fort Lauderdale).

As abovementioned, Norwegian put substantial effort to position itself on U.K. market since 2014 by operating its transatlantic flights from the less congested airport in London airports system, London Gatwick (LGW), which provides more flexibility in terms of airport charges as well as slots allocations. Table 2 outlines the characteristics of six competing routes between Norwegian and legacy carriers British Airways which traditionally been dominant in these markets. On the other hand, British Airways flights are mainly concentrated at London Heathrow (LHR), although it directly competes with Norwegian on San Francisco and Orlando routes from London Gatwick.

Table 1: Long haul low cost operations ranked by weekly seat capacity (2nd -8th October 2017)

Rank	Airline	Weekly seats	Number of routes
1.	AirAsiaX	133458	21
2.	Norwegian	87337	48
3.	Scoot	69144	18
4.	Jetstar Airways	46900	14
5.	Air Canada rouge	37923	20
6.	Thai AirAsia X	31668	4
7.	NekScoot	24070	6
8.	Cebu Pacific	13080	5
9.	Azul	12466	4
10.	Eurowings	11780	12

Source: CAPA (2017)

However, not all routes presented are characterized by the same level of competition between those two airlines. It is evident that competition will be intensified when there is overlapping characteristics of service offered. However, the market share on specific routes will be highly affected by passenger's personal preferential towards. The competition will be certainly intensified when both carriers offer their flights with similar schedule (the same day and approximately the similar departure times) such as the case with New York and Los Angeles routes. On the other hand, Norwegian operates Miami route only twice a week, while the same route with one stopover is served by British Airways in the code share with its partner American Airlines every day with high frequency accounting up to ten daily flights. Since direct flight is not offered by British Airways, it is reasonable to expect that certain amount of passengers will be attracted by Norwegian direct flight if the timetable fits specific passenger's requirement.

Table 2: The characteristics of the competing transatlantic routes between Norwegian Air Shuttle (DY) and British Airways (BA) in 2017/2018 winter timetable

Destination airport	Origin airport	Days of operation	Frequency	Schedule departure
New York (JFK Airport)				
DY	London Gatwick	1234567	2 daily flights	06:00; 17:10
BA	London Heathrow	1234567	13 daily flights	From 08:25 to 19:50
BA	London Gatwick	1234567	1 daily flight	16:45
Los Angeles (LAX Airport)				
DY	London Gatwick	1234567	1 daily flight	12:50
BA	London Heathrow	1234567	5 daily flights	From 10:35 to 15:30
Boston (BOS Airport)				
DY	London Gatwick	1234567	1 daily flight	16:00 (4, 7); 16:20 (1,3); 16:50 (5)
BA	London Heathrow	1234567	4 daily flights	From 11:15 to 19:10
Fort Lauderdale-Miami (FLL Airport)				
DY	London Gatwick	1234567	1 daily flight	16:20 (1); 14:55 (3,5); 14:50 (7)
BA	London Gatwick	1234567	1 daily flight	09:05 (1, 4); 09:10 (6)
Florida Orlando (MCO Airport)				
DY	London Gatwick	1234567	1 daily flight	14:05
BA	London Gatwick	1234567	1 daily flight	11:35 (1,2,3); 11:00 (7)
		1234567	2 daily flights	11:35; 13:20
San Francisco – Oakland (OAK Airport)				
DY	London Gatwick	1234567	1 daily flight	12:55 (2, 6); 12:45 (4); 14:20 (7)
BA	London Gatwick	1234567	1 daily flight	08:45 (1); 10:10 (3); 11:00 (6)

It is evident that Norwegian's less-price based strategy aims at diverting the portion of those price sensitive passengers away from legacy carriers, and to secure some portion of business traffic in order to increase its yield. For example, on New York routes, traditionally perceived as the most popular touristic market, the Norwegian's average fare is approximately 30% less than its rival British Airways price. However, British Airways offers fourteen daily flights spread across entire day at LHR compared to one daily flight offered by

Norwegian. With its well-established flights to New York, British Airways can capture the portion of passengers that highly regards early departure times. With its late afternoon flight, Norwegian could count on the price sensitive segment of passengers who are willing to arrive late afternoon in New York and it is likely that certain portion of these passengers previously used the legacy carrier's flight. Afternoon departure times are generally characteristics for all Norwegian transatlantic routes from London Gatwick, as it is probably seen as a good strategy that provides the balance between different passenger segments' requirements.

The next section provides the comprehensive assessment of different aspects of competition that exists or has existed in the high-density market such as those that connects two metropolitan cities London and New York. The influence of Norwegian on legacy carriers, specifically British Airways as a dominant player in London-New York market, will be thoroughly discussed through traffic and capacity statistics.

4. THE OUTLOOK OF LONDON – NEW YORK ROUTE

The London New York route has always been perceived as one of the high density route in the world. Two metropolises have historically had important links that catalyze the mobility of people, trade and service. Moreover, London stands out as an important gateway that consolidates a large portion of traffic from different part of Europe, Africa and Asia that terminate in New York area. Both London and New York are served by multiple airport system, which allow the large number of possible connections between two cities Table 3 provides the list of airports that serves these two metropolitan cities along with their respective number of passengers.

Table 3: Airports serving the metropolitan areas of London and New York (Port Authority, 2015)

City	Airports	IATA code	Number of passengers in 2015 (mil.)
London Airport System	London Heathrow	LHR	74.9
	London Gatwick	LGW	40.3
	London Stansted	STN	22.5
	London Luton	LTN	12.3
	London City	LCY	4.3
New York metropolitan area	John F. Kennedy International	JFK	56.8
	Newark Liberty International	EWR	37.5
	LaGuardia Airport	LGA	28.4

For example, London is served by five airports among which London Heathrow is the largest one and serves as a hub of full-service carrier British Airways. On the other hand, New York metropolitan area is served by three airports, among which JFK is the largest and most important one from which large number of intercontinental flights have been performed. In other words, the same city pair (i.e. London-New York) can be realized throughout different airport pairs offering the potential passengers possibility to make selection between different airports as well as between different airlines.

4.1. Number of passengers and airlines' market share

In order to have better insight into the different degree of competition, Table 4 outlines the characteristics of three routes (airport pairs) that serve these two metropolitan areas in 2016. The data on the number of passengers have been retrieved from U.S. DoT's T-100 database provided by Bureau of Transportation Statistics that allows the free access.

Table 4: Some characteristics of London –New York route in 2016

Airport pairs	Airlines operated	Number of passengers	Total number of passengers (% of all pax)	Market share per route per airline
LHR-JFK	British Airways (BA)	651 397	1 434 051 (69.9%)	45.4%
	Virgin Atlantic (VS)	406 229		28.3%
	American Airlines (AA)	238 756		16.6%
	Delta Air Lines (DL)	135 282		9.4%
	Kuwait Airways (KU)	2 387		0.2%
LGW-JFK	Norwegian Air Shuttle (DY)	112 469	168 480 (8.2%)	66.8%
	British Airways (BA)	56 011		33.2%
LHR-EWR	United Airlines (UA)	232 118	449 963 (21.9%)	51.6%
	British Airways (BA)	134 599		29.9%
	Virgin Atlantic (VS)	74 050		16.5%
	Air India (AI)	9 176		2.0%

Source: Bureau of Transportation Statistics (2018)

As seen from Table 4, London – New York route was operated by two European carriers (BA and DY), four American carriers (VS, AA, UA, DL), one Middle Eastern carrier (KU) and one Asian carrier (AI). The high density route among these three options is certainly one that connects two major airports (LHR and JFK) within the cities. This route encompassed the market share of 69.9% in terms of passengers that commenced their trip at London. BA is dominant carrier on this route accounting for 45.4% of all passengers transported and followed by three American carriers – VS, AA and DL which market share accounts for 28.3%, 16.6% and 9.4% respectively. The second dense route is one that connects LHR with EWR, the second largest airport in New York area, with market share of almost 22%. Similar to LHR-JFK route, this route is operated by BA as a dominant player (51.6%), two American carriers that combined have a market share of almost 50% and Air India with only 2% of share. Finally, the route that connects second largest airport (LGW) and major airport in New York (JFK) is characterized by the presence of low-cost carrier DY that started its operation in the third quarter of 2014. In order to reduce the competitive pressure induced by low fares offered by DY, BA has recently announced its flights to JFK from LGW. However, the market share of BA is significantly lower compared to DY that held almost 70% of share and transported more than hundred thousand passengers in 2016.

Fig. 1 depicts the historical trends in terms of number of passengers on the route London New York including the traffic on all three possible airport cities pairs. It was felt that 2010 was an appropriate start year because this period coincides with recovery of airline industry after the severe world economic crisis occurred in 2008.

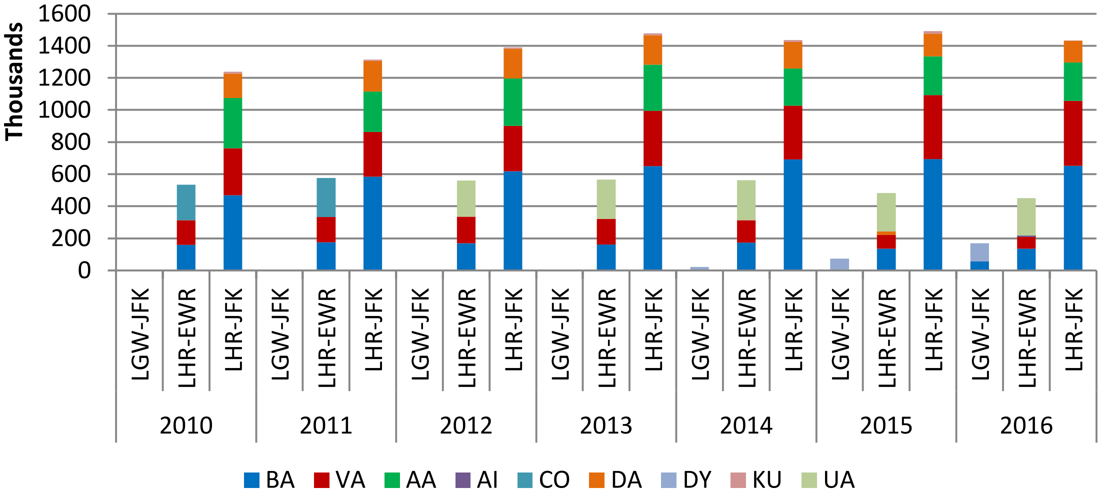


Figure 1: Number of passengers at three routes connecting London and New York

As observed from Fig. 1, the number of passengers from London Heathrow to New York JFK had seen the steady growth in the period from 2010 to 2013. Afterwards, the number of passengers fluctuated around 1.45 million reaching its peak in 2015 when almost 1.5 million people were transported between these two points. Among five airlines operated these routes, it is evident that BA has the highest market share. BA is currently the most dominant carrier on this route with market share encompassing approximately 45% of total passengers carried, followed by three American carriers: VA, AA and DA. However, not all of them have been the real competitors. For example, AA offers some portion of its flight in the cooperation with BA through code-share agreements, while other portion operates independently. Over the time horizon, VA has radically increased its market from 23.7% in 2010 to 28.3%, whereas AA reduced its share from 25.2% in 2010 to 16.6% in 2016. Finally, Kuwait Airways has been present on this route during the observed period, but its market share is not significant (less than 1%). On the other hand, the number of passengers at LHR-EWR was stable during the period from 2010 to 2014 accounting for around half million passengers with slight decrease in 2015 and 2016. Continental Airlines (CO) was present at this route with significant market share in the past, but it withdrew the market in 2012. Finally, the flights to New York have been introduced from LGW for the first time in 2014 when DY offered its service by affordable prices. Since then, this carrier records the rapid expansion with number of passengers exceeding one hundred thousand in 2016.

4.2. The market concentration of London – New York route

The Herfindahl–Hirschman index (HHI) is defined as the sum of squared market shares of airlines in a market and thereby provides an easily interpretable measure of concentration (Lijesen, 2004). The HHI for specific route can slightly vary depending on the measure used to express the market share and thus it can be calculated by either capacity offered or number of passengers transported. The HHI ranges from the value close to zero, indicating nearly perfect competition to ten thousand, indicating a monopoly. For the purpose of this paper we calculated the value of HHI based on market share expressed through number of seats (i.e. capacity of aircraft) offered by specific airline. It is worth mentioning that depending on the configuration of the aircraft operated, the values of HHI can slightly differ. As a carrier with the most carried passengers, BA's configuration of two most dominant aircraft type B777 (two version – V1 and V2) and B747 (three versions – V1, V2 and V3) (Table 5) can have impact on the value of HHI. Fig. 2 shows the different value of HHI based on maximum, medium and minimum configuration of aircraft operated this route.

Table 5: British Airways B777 and B747 configuration

	B777			**B747**		
	B777-200 (V1)	**B777-200 (V2)**	**B777-300**	**B747-400 (V1)**	**B747-400 (V2)**	**B747-400 (V3)**
Standard seats	203	122	185	243	185	145
Recliner seats	24	40	44	36	30	30
Flat bed seats	48	48	56	52	70	86
Open suites	-	14	14	14	14	14
Total	**275**	**224**	**299**	**345**	**299**	**275**

The U.S. Department of Justice considers a market with an HHI of less than 1,500 to be a competitive marketplace, an HHI of 1,500 to 2,500 to be a moderately concentrated marketplace, and an HHI of 2,500 or greater to be a highly concentrated marketplace (Investopedia, 2018). As observed from Fig. 2 the market concentration index ranges from less than 1,600 in 2010 to more than 1,800 in the case of moderate capacity. Thus, the London New York route can be characterized as a route with medium level competition, which is certainly benefit for potential passengers perceived through lower fares and higher level of service.

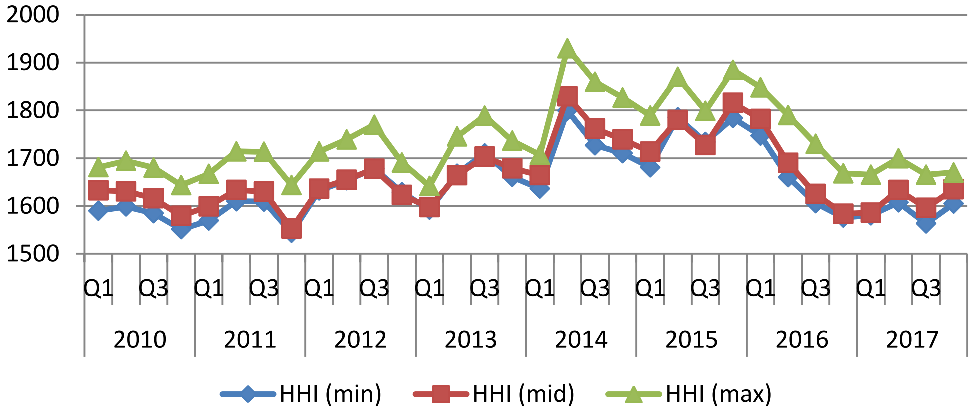


Figure 2: HHI index for different aircraft configuration for London New York route

4.3. Price competitiveness between BA and DY

In order to investigate the price competitiveness of two major rival BA and NY, the fares from both carriers are collected throughout time-horizon starting from fixed points in the time (2nd October) encompassing twelve points in time ranging from 7 days before departures to 6 months (7 observations in total). For each of time point, the average value in economy travel class (the lowest tariff) are calculated and presented in Fig 3.

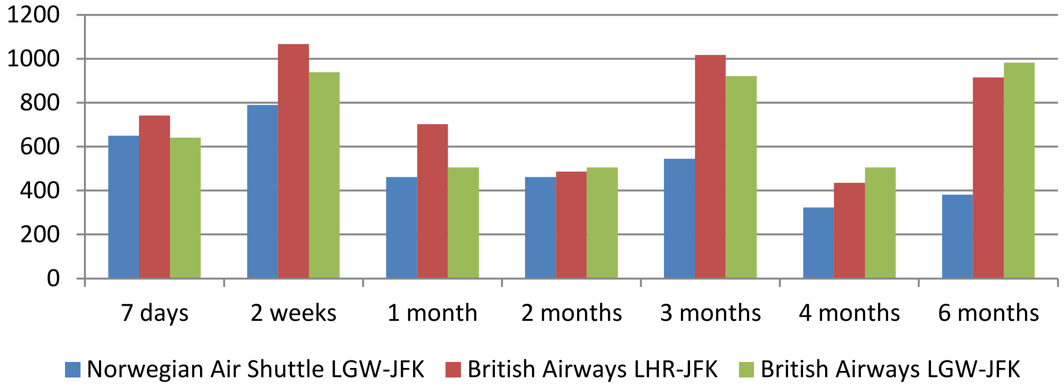


Figure 3: Average fares of BA and DY on routes connecting London and New York (in EUR)

DY average ticket price is even two times lower for this route than that offered by its major competitor BA. This makes the LGW-JFK route appealing to price sensitive passengers. Moreover, the differences in price is significant enough to seriously harm BA market position, although it still can count on its well established hub-and-spoke network that serves as a “feeder” to these long-haul flights. The evidence of rising competition can be found in British Airways strategy that has been recently announced through “squeezing” 52 seats on its Boeing 777 flights from Gatwick by shrinking seating space (Telegraph, 2016). The configuration of British B777 counts 275 seats in 3 classes, and with additional seats BA services will be more in accordance with Norwegian’s B787 with 294 seats onboard. In this way, the airline attempts to reduce its unit cost and to offer even more low fare in order to retain its market position on competitive routes. According to comparative analysis of fares provided by BA from LGW provided in Fig. 3, it seems that BA successfully applied the mentioned strategy.

5. CONLUSION

The long haul routes have been traditionally reserved for full-service carriers that successfully manage to collect a large number of passengers over their hub-and-spoke network system. However, in recent few years, the situation has been radically changed with several low-cost carriers that started to efficiently operate long-haul routes. One of the most successful one among them in Europe was certainly Scandinavian carrier, Norwegian Air Shuttle with its aggressive strategies that aims at diverting a large portion of passengers, particularly those price sensitive, from well-established airlines. The Norwegian routes are still focused on touristic destinations that characterized the high demand during the entire year.

The paper investigates some aspects of competition on London New York route as one of the highest density route in the world. Norwegian initially used the opportunity to solely operate from London Gatwick, less congested airport in London Airport System offering one daily flight and offering the fares which are almost double lower than its major competitor British Airways from London Heathrow. In order to efficiently combat the fierce competition imposed by low-cost rival, British Airways started to operate the flights from the London Gatwick implying the strategy of “squeezing extra seats”. In such way, British Airway succeeded to offer lower fares which are highly in line with those offered by Norwegian. However, the introduction of low-cost service in this market stimulating the demand generating the new growth in market that is well matured.

Acknowledgement

This research has been supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, as a part of the project TR36033 (2011–2018).

REFERENCES

Bureau of Transportation Statistics. (2018). Retrieved from https://www.bts.gov/topics/airlines-and-airports (accessed on March, 2018)

Daft, J., & Albers, S. (2012). A profitability analysis of low-cost long-haul flight operations. *Journal of Air Transport Management, 19,* 49-54. doi: 10.1016/j.jairtraman.2012.01.010

De Poret, M., O’Connell, J. F., & Warnock-Smith, D. (2015). The economic viability of long-haul low cost operations: evidence from the transatlantic market. *Journal of Air Transport Management, 42*, 272-281. doi: 10.1016/j.jairtraman.2014.11.007

Dobruszkes, F. (2006). An analysis of European low-cost airlines and their networks. *Journal of Transport Geography, 14 (4)*, 249–264. doi: 10.1016/j.jtrangeo.2005.08.005.

Dobruszkes, F. (2009). New Europe, new low-cost air services. *Journal of Transport Geography, 17 (6)*, 423–432. doi: 10.1016/j.jtrangeo.2009.05.005

CAPA - Center for Aviation. (2017). Long haul low cost becomes mainstream as full service airlines gradually embrace new business models. Retrieved from https://centreforaviation.com/insights/analysis/long-haul-low-cost-becomes-mainstream-as-full-service-airlines-gradually-embrace-new-business-models-348105 (accessed on March, 2018)

Cento, A. (2009). *The Airline Industry: Challenges in the 21st Century*. Heidelberg: Physica-Verlag Springer.

Francis, G., Humphreys, I., Ison, S., & Aicken, M. (2006). Where next for low cost airlines? A spatial and temporal comparative study. *Journal of Transport Geography, 14* (2), 83–94. doi: 10.1016/j.jtrangeo.2005.05.005

Investopedia. (2017). Herfindahl-Hirschman Index – HHI. Retrieved from https://www.investopedia.com/terms/h/hhi.asp (accessed on March, 2018).

Lijesen, G. M. (2004). Adjusting the Herfindahl index for close substitutes: an application to pricing in civil aviation. *Transportation Research Part E*, *40* (2). 123-134, doi: 10.1016/S1366-5545(03)00045-0

Morrell, P. (2008). Can long-haul low-cost airlines be successful?, *Research in Transportation Economics, 24 (1)*, 61-67. doi: 10.1016/j.retrec.2009.01.003

Port Authority of NY & NJ. (2016). 2015 Airport Traffic Report. Retrieved from https://www.panynj.gov/airports/pdf-traffic/ATR_2015.pdf (accessed on March, 2018)

Rodríguez, A. M., & O'Connell, j. F. (2017). Can low-cost long-haul carriers replace Charter airlines in the long-haul market? A European perspective. *Tourism Economics, 24 (1)*, 64-78. doi: 10.1177/1354816617724017

Soyk, C., Ringbeck, J., & Spinler, S. (2017). Long-haul low cost airlines: characteristics of the business model and sustainability of its cost advantages. *Transportation Research Part A: Policy and Practice, 106,* 215-234. doi: 10.1016/j.tra.2017.09.023

Soyk, C., Ringbeck, J., & Spinler, S. (2018). Revenue characteristics of long-haul low cost carriers (LCCs) and differences to full-service network carriers (FSNCs). *Transportation Research Part E: Logistics and Transportation Review, 112,* 47-65. doi: 10.1016/j.tre.2018.02.002

Telegraph. (2016). British Airways 'to squeeze extra 52 seats onto Boeing 777 economy flights'. Retrieved from https://www.telegraph.co.uk/news/2016/11/07/british-airways-to-squeeze-extra-52-seats-onto-boeing-777-econom/ (accessed on March, 2018).

Van den Hoek, M. (2017). A Competitive Analysis of Ryanair in the Long-Haul Airline Industry. *Master thesis at Seoul National University*. Retrieved from http://s-space.snu.ac.kr/bitstream/10371/129130/1/000000141295.pdf (accessed on May, 2018).

Wensveen, G. J., & Leick, R. (2009). The long-haul low-cost carrier: A unique business model. *Journal of Air Transport Management, 15* (3), 127-133. doi: 10.1016/j.jairtraman.2008.11.012

Wilken, D., Berster, P., & Gelhausen, M. C. (2016). Analysis of demand structures on intercontinental routes to and from Europe with a view to identifying potential for new low-cost services. *Journal of Air Transport Management, 56*, 79-90. doi: 10.1016/j.jairtraman.2016.04.018

MULTIVARIATE APPROACH TO MAKING SPONSORSHIP DECISIONS: THE CASE OF EUROPEAN FOOTBALL LEAGUES

Strahinja Radaković*[1], Milan Radojičić[1], Milica Maričić[1],
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: radakovic.strahinjaa@gmail.com

Abstract: *Sponsorship, as a communication activity, has significantly developed in the last several decades. Although the budgets that the companies are willing to allocate for sponsorship are growing, the sponsor's attitude towards sponsorship and sponsoring events is changing. The decision on which athlete, sports club or competition to sponsor is nowadays a highly important decision which can have a significant impact on the brand image, brand awareness, brand preference, and attitude towards the brand. Therefore, analytical approach should be taken. Herein we suggest a multivariate approach to making sponsorship decisions. The proposed approach was employed on the case of European football leagues: Premier League (United Kingdom), La Liga (Spain), and Ligue 1 (France). We hope the presented research will provide additional insights on differences between the three European football leagues and on the application of multivariate analysis in decision making in sport sponsorship.*

Keywords*: Sport sponsorship, Football, Data analysis, Clustering, Decision making*

1. INTRODUCTION

Sponsorship can be defined as an investment, in cash or products or services, in return for exploitable commercial rights (Smith, Graetz, & Westerbeek, 2008). The possibility of displaying advertisement messages to wider public has attracted the attention of marketing managers to turn to sponsorship as the mean of marketing communication. Sponsorship of sport, arts, and entertainment-related events and competitions has slowly, but surely, become an important part of the marketing mix (Jensen & Cornwell, 2017). According to the International Events Group (2017), companies spent $60.1 billion on sponsorship in 2016, whereas the most common type of sponsorship is sport sponsorship.

With the development of information systems and technology it became possible to use distinct methods of tracking and collecting data from the sport matches. The collected data could later be used by various interest groups: team managers, journalists, data analysts, and of course, the fans (Radojičić, Djokovic, & Jeremić, 2016). The application of statistical and mathematical methods, data visualization, and data mining to the collected data from the pitch to gain information on previous play and predict future results is considered sports analytics (Alamar, 2013). Experts in sport marketing are taking the advantage of the acquired data and the results provided by sports analytics to assess their current marketing activities, but also to predict new ones (Fried & Mumcu, 2017). Namely, the obtained results can help the experts to make the decision which event or talented individual to sponsor.

The big five of the European football are the English Premier League, Spanish La Liga, German Bundesliga, Italian Serie A, and French Ligue 1. In the last several years the Serie A has witness a decline in both results and attendance (Kate Langshaw, 2017; UEFA, 2016), leaving Premier League and La Liga to dominate. In the last four years of the last 15 major European and International major club titles, Spanish clubs have won 14 and English clubs just one. In our analysis, we chose the Premier League, La Liga and Ligue 1. Bundesliga was listed out of the analysis as their league consists of 18 clubs compared to 20 in the chosen three and Serie A was ruled out due to the current decline of the League.

The aim of this paper is to try to assist marketing and brand managers in the process of deciding which football league and football club to sponsor. Namely, using statistical multivariate analysis, we strive to inspect the differences between three leading European football leagues based on seasonal post-match data. The research hypothesis is that there are differences in the football style that these clubs play within their leagues, and that the clubs although from three different leagues can be grouped. The analysis could assist the decision makers to choose the football clubs which play similar style when making their sponsorship mix of clubs and leagues in which they will be present.

The structure of the paper is as follows. Section 2 sees a brief literature review of sport sponsorship and decision making in sponsorship. The research results are elaborated in the next section while we finish the paper with the discussion and the concluding remarks.

2. LITERATURE REVIEW

In the following two subsections we will place our attention on the role of sport sponsorship in the integrated marketing communications of a company and the possible benefits the sponsoring company might gain from such activities. Accordingly, we give an overview of the process of decision making in sponsorship.

2.1. Sport sponsorship

Sport events attract the attention of a large number of spectators and thus are increasingly being broadcast worldwide. As a result of this high media coverage, sport sponsorship takes on an international character, and the budgets allocated for becoming an official sponsor of a global sports event are rising and have no tendency to decline (O'Reilly, Lyberger, McCarthy, Séguin, & Nadeau, 2008). For many world acknowledged brands, sponsorship of sports event plays a central role in their integrated marketing communication (Santomier, 2008).

Sport sponsorship has become the most common form of sponsorship (Quester, 1997). Sports events and sporting competitions have traditionally been given significant support by the sponsors (McCarville & Copeland, 1994). However, the popularity of sport sponsorship has grown significantly in the past century thanks to drastic social changes, i.e. the development of information technologies, the Internet, social media, media, and transport. Namely, the marketing experts saw the possibility to use the attention of spectators during sports events to show them their sponsorship messages. Therefore, sport sponsorship has become a long-term marketing strategy which is used to communicate with a large external and internal audience to gain competitive advantage (Papadimitriou & Apostolopoulou, 2009).

Another major benefit of sponsorship sport events is the presence of heightened emotions. Namely, visitors and viewers of sport competitions are emotionally involved with the game, their club, favourite athlete, or national team (Biscaia, Correia, Rosado, Ross, & Maroco, 2013). Global brands are aware of the emotional impact of sport and utilize it to connect with the consumers (Santomier, 2008). Namely, the best world-class athletes and teams are sponsored by the largest companies in the world under the motto "the best with the best" (Šurbatović, 2014). In this way, companies want their brand to be identified with athletes and teams and want to transfer the positive sports results to their products and services. Therefore, sport sponsorship has become an effective marketing strategy for companies, but at the same time it is an important source of revenue for sports organizations.

2.2. Decision making in sponsorship

Decision making is sponsorship is a topic which is slowly, but surely gaining importance. As the decision on sponsorship activities can impact the brand, the question "whom to sponsor" still remains a challenging practical and academic question (Lee & Ross, 2012).

There are several studies which have employed statistical, data mining, and decision-making techniques in the field of sponsorship. Lee & Ross (2012) employed Analytic Hierarchy Process (AHP) to identify the decision-making factors of sport sponsorship. Their analysis covers sponsors of clubs competing in different sports, therefore providing insights on the importance of factors from different perspectives. Research also shows that AHP could be easily combined with other multiple attribute decision making (MADM) methods such as TOPSIS (Isik, Ozaydin, Burnaz, & Topcu, 2016).

Another interesting approach, although not yet applied in the sphere of sponsorship, is the application of Data envelopment analysis (DEA) and efficiency analysis the efficiency of sports teams. According to the received results of efficiency analysis, managers can choose whether a player should be given more play time or not, the board can decide to decrease or increase player's salary, or sponsors can make a decision which player or team to sponsor. For example, an interesting distance-based approach to efficiency initially suggested by Jeremic et al. (2012) was applied to evaluating efficiency of basketball players (Radovanović, Radojičić, Jeremić, & Savić, 2013).

Multivariate statistical analyses were also employed to assess the effectiveness of sponsorship, which can be used in decision making in sponsorship. For example, linear regression was employed to model the accuracy and certainty of brand recognition (Olson & Mathias Thjømøe, 2009). On the other side, multivariate statistical analyses were used in sports analytics. Multiple linear regression was used to calculate the efficiency of football players in the UEFA Champions League (Radojičić, Djokovic & Jemremić, 2016), while cluster analysis and Principal component analysis (PCA) were used to assess players of Real Madrid and Barcelona (Radojičić, Milenković, Totić, Bijelić, Djoković, 2013).

The presented literature review shows that the field of application of statistical and mathematical methods on data related to sport is developing and that the obtained results can be of use for decision making in sponsorship.

3. RESEARCH RESULTS

In the proposed research we collected the data on the Premier League, La Liga, and Ligue 1 for the season 2016/17. The data was collected from the official web sites of the three leagues which provide aggregate statistics (Soccer Stats, 2018). The chosen indicators which we observed in the analysis can be grouped as follows:

- Attractiveness of the game – Measured through indicators *Goals for, Number of assists, Total shots, Shots on goal, Goals scored from direct play*
- Roughness of the game – Measured through indicators *Number of yellow cards, Number of Red cards, Number of fouls against*.

We chose the indicators of attractiveness and roughness of the game as we believe these features of the game can attract or turn away football fans and spectators. Accordingly, we consider that these features might be of interest for sponsors, so as they could choose the league and/or club which will attract the most viewers.

3.1. Comparison of the three football leagues

In this part of our analysis we aimed to inspect whether there is statistically significant difference between the values of the chosen eight indicators within the three leagues. The first step was to determine whether the observed indicators are normally distributed or not. The Kolmogorov-Smirnov test showed that none of the indicators is Normally distributed, therefore, in the next steps of the analysis we used nonparametric tests.

When it comes to the comparison of attractiveness indicators between the three leagues the Kruskal-Wallis test indicates that there is no statistically significant difference. This means that the spectator or viewer of the French League can expect to enjoy the same attractiveness of the game as the spectator or viewer of the Spanish or English league. There is no statistically significant difference between the total number of goals for (KW=2.483, p>0.05) and for the number of assists (KW=0.346, p>0.05). Also, goals are scored after a similar number of shots on goal (KW=4.159, p>0.05) and total shots (KW=2.946, p>0.05). In the three observed leagues, there is no difference in the number of goals scored from direct play (KW=4.601, p>0.05). All this could lead to an important conclusion that the French league, which might not be financially and marketing supported as English or the Spanish League, is not missing out on-field excitement. Namely, Premier League clubs' revenue in 2016 was €110m per club, compared to €44m in La Liga, and €21m in Ligue 1 (UEFA, 2016).

Analysis of the roughness indicators shows there is difference between the leagues. It has been established that there is statistically significant difference: *Number of yellow cards* (KW=32.256, p<0.01)*, Number of Red cards* (KW=16.559, p<0.01)*,* and *Number of fouls against* (KW=26.051, p<0.01). All indicators are related to the subjective factor reflected through referees, because they are the ones who assess whether a particular start or duel will be defined as a foul and whether it will be sanctioned by a card. As there has been noted that there is difference between the three leagues in terms of roughness, we examined more closely the Medians of the three indicators per league. The results are given in Table 1.

Table 1: Median of the chosen three indicators of roughness per three observed leagues

League	**Number of yellow cards**	**Number of red cards**	**Number of fouls**
Premier League	71.5	2.0	430.5
La Liga	97.5	4.0	537.5
Ligue 1	65.0	4.0	470.0

According to the median, the most yellow cards are awarded in the Spanish league. However, the situation is different when the number of red cards is observed. In La Liga the red cards are awarded more often, and the same accounts for France. The results of French League have the lowest median of *Number of yellow cards* and the highest median of *Number of red cards.* This could indicate that style of the League 1 is rougher with more tackles which are sanctioned. On the other hand, this could indicate that the referees in the league do not allow rough play. As far as the number of fouls is concerned, they are most often awarded in the Spanish league, which has the median of 537.5. This result could be interpreted as that the referee organization in Spain suggests their referees to protect the players from possible injury and stop any signs of dangerous play. It is presumed, that such a decision has been made because La Liga, especially Real Madrid and Barcelona, has the best and the most expensive players of today's. Namely, their injury and absence from the field would mean a drawback for the team. On the other hand, it would also be a financial loss for the entire league reflected through marketing activities and sponsorship contracts, and primarily by selling broadcast rights. The least fouls are awarded in Premier League, which means that their referees allow a rougher game. It can also be concluded that the Premier League is therefore the most dynamic of the three leagues as it has the least stoppage time during matches.

To additionally inspect the relationship between the chosen three indicators, we conducted the correlation analysis per league (Table 2). It can be observed that there is high positive correlation between fouls and yellow cards over all leagues, which ranges from 0.702 (Premier League) to 0.619 (La Liga). Such a result indicates that a large number of fouls is accompanied by a larger number of yellow cards. Interestingly, the correlation between the number of yellow and red cards is only observed in the Spanish league. The relationship is such that a greater number of yellow cards leads to a greater possibility of obtaining red card. This indicates that in the French and the English league there is a greater possibility of getting a direct red card. The Pearson's correlation coefficients show that there is no statistically significant relationship between number of fouls and red cards, which may indicate that most red cards are not received after two fouls, whereas that players after the objection or inappropriate tackles often earn the most severe warning.

Table 2: Correlation analysis between the chosen three indicators of roughness per three leagues

League		**Number of yellow cards**	**Number of red cards**	**Number of fouls**
Premier League	**Number of yellow cards**	1	0.226	0.702**
	Number of red cards	0.226	1	-0.026
La Liga	**Number of yellow cards**	1	0.506*	0.619**
	Number of red cards	0.506*	1	0.168
Ligue 1	**Number of yellow cards**	1	0.382	0.646**
	Number of red cards	0.382	1	-0.045

Note: *p<0.05, **p<0.01

3.2. Clustering clubs from the three football leagues

The second direction of our research was to cluster clubs of the three observed leagues according to "roughness" and "attractiveness". To conduct the analysis, we performed the K-means clustering (Hartigan & Wong, 1979). K-means clustering algorithm partitions n observations into k predetermined number of clusters in which each observation belongs to the cluster with the nearest mean, serving as the centre of the cluster. Therefore, on our analysis, the cluster centre is a fictive club. To cluster clubs according to "roughness" we used *Number of yellow cards, Number of red cards,* and *Number of fouls.* We retained three clusters which we named Rough, Medium rough, and Soft clubs.

The distance between the cluster Rough clubs and Medium rough clubs is 74.682, between Rough clubs and Soft clubs 151.281, and between Medium rough and Soft clubs 77.336. The cluster centres are presented in Table 3. It can be observed that the clubs in the cluster Rough clubs tend to receive more yellow cards, red cards, and commit more fouls. Interestingly, the Medium rough clubs almost receive the same number of red cards as the Rough clubs. According to the cluster centre of the third cluster, the clubs in cluster tend to play "clean" and avoid making fouls.

Table 3: Cluster centres of the three retained clusters of clubs based on "roughness"

Cluster	**Number of yellow cards**	**Number of red cards**	**Number of fouls**
Rough clubs	102.3	4.4	564.7
Medium rough clubs	77.2	4.2	494.3
Soft clubs	66.0	3.2	417.8

The analysis which would be of interest for the decision-makers is to observe how many clubs from each league is in each of the retained clusters and to provide additional information on the retained clusters. Table

4 provides valuable insights. The cluster Rough clubs consists of 12 clubs of which 11 are from Spanish league, and only one from French league (Toulouse). Interestingly, there is no club which plays in the Premier league that is characterized as rough. For the European football it is a good result that this is the smallest cluster, meaning that the clubs from the first leagues of these three countries primarily practice “fair play”. When it comes to the cluster Medium rough clubs, the clubs in this cluster have a similar average of red cards as Rough clubs, but they have considerably fewer fouls made and yellow cards obtained. This cluster is mostly made of French clubs (10), followed by Spanish (7) and English clubs (5). The last cluster includes 15 clubs from Premier League, nine from Ligue 1 and only two clubs from La Liga. Interestingly, the two clubs from La Liga are the two best Spanish clubs, Real Madrid and Barcelona. Therefore, according to the number of fouls that their players make, the number of red and yellow cards received, Real Madrid and Barcelona are more similar to English than Spanish clubs.

Table 4: Cluster centres of the three retained clusters of clubs based on “roughness”

Cluster	**League**			**Total**
	Premier League	La Liga	Ligue 1	
Rough clubs	0	11	1	**12**
Medium rough clubs	5	7	10	**22**
Soft clubs	15	2	9	**26**

However, since the previously conducted Kruskal-Wallis test indicated that there might be difference in the referees’ criteria between the observed leagues, these assigned cluster names should be taken with caution. Namely, the "soft" clubs do not have to play without contact yet that referees hesitate to give their players red or yellow cards. One thing is certain, Real Madrid and Barcelona in the Spanish league have a similar treatment from the referees as clubs from the English League.

For potential further, in-depth analysis, we provide the list of clubs within each cluster (Table 5).

Table 5: Clubs which make each of the three retained clusters on based on roughness indicators

Rough teams	**Medium rough teams**		**Soft teams**	
Sevilla (ESP)	Middlesbrough (UK)	Watford (UK)	Chelsea (UK)	Hull City (UK)
Espanyol (ESP)	Everton (UK)	Monaco (FRA)	Tottenham (UK)	Sunderland (UK)
Alaves (ESP)	Real Sociedad (ESP)	Olympique Lyonnais (FRA)	Stade Malherbe Caen (FRA)	Montpellier (FRA)
Malaga (ESP)	Las Palmas (ESP)	Olympique de Marseille (FRA)	West Ham United (UK)	Barcelona (ESP)
Valencia (ESP)	Manchester United (UK)	Nantes (FRA)	Arsenal (UK)	Paris Saint Germain (FRA)
Real Betis (ESP)	Villarreal (ESP)	Saint-Etienne (FRA)	Southampton (UK)	Nice (FRA)
Deportivo La Coruna (ESP)	Atletico Madrid (ESP)	Avant de Guingamp (FRA)	West Bromwich Albion (UK)	Girondins Bordeaux (FRA)
Leganes (ESP)	Crystal Palace (UK)	Lille (FRA)	Bournemouth (UK)	Rennes (FRA)
Sporting Gijón (ESP)	Athletic Bilbao (ESP)	Metz (FRA)	Liverpool (UK)	Angers (FRA)
Osasuna (ESP)	Eibar (ESP)	Dijon (FRA)	Leicester City (UK)	Real Madrid (ESP)
Granada (ESP)	Celta Vigo (ESP)	Nancy (FRA)	Stoke City (UK)	Manchester City (UK)
Toulouse (FRA)			Swansea City (UK)	Lorient (FRA)
			Burnley (UK)	Bastia (FRA)

To cluster clubs according to “attractiveness” we used *Goals for, Number of assists, Total shots, Shots on goal,* and *Goals scored from direct play.* We again retained three clusters which we named Attractive, Watchable, and Boring clubs.

The distance between the clusters Attractive and Watchable clubs is 144.595, between Attractive and Boring clubs 237.800, and between Watchable and Boring clubs 93.869. Comparing the distances between clusters retained in case of “roughness” and “attractiveness”, it can be observed that in the case of “attractiveness” the clusters are more separated. The cluster centres are presented in Table 6. According to the cluster centres it can easily be concluded that there is difference in the values of attractiveness indicators which achieve the clubs in the three clusters. Attractive clubs dominate within all five attractiveness indicators.

Table 6: Cluster centres of the three retained clusters of clubs based on "attractiveness"

Cluster	Goals for	Number of assists	Total shots	Shots on goal	Goals scored from direct play
Attractive clubs	86.3	57.1	601.3	238.4	67.2
Watchable clubs	53.9	37.1	479.3	176.9	38.8
Boring clubs	41.8	27.5	399.9	130.3	28.7

Additional analysis of the three retained clusters is presented in Table 7. Among the 60 observed football clubs, more than half of them can be classified as Boring clubs. Namely, the play of 34 clubs is mostly oriented to defence and actions to save their goal. The league with most Boring clubs is the French league, where 43.33% of them practices such tactic, followed by Spanish league (36.67%), and English league (33.33%). Interestingly, the least boring and the most attractive clubs come from the English league. The six attractive clubs from the Premier league are the clubs which finished the season 2016/17 on the top six places (Arsenal, Liverpool, Manchester United, Manchester City, Tottenham, Chelsea). Expectedly, the two attractive clubs from the Spanish league are Barcelona and Real Madrid. However, the analysis of the attractive clubs from the French league provides interesting insights. Olympique Lyonnais has been grouped as an attractive club, while Nice was not. However, Nice completed the championship as third, and Olympique Lyonnais as fourth. Therefore, although Olympique Lyonnais played more attractive football, Nice had better defence which eventually led to its better rank. The provided clustering structure indicates that the Premier League is the most attractive league, where the spectators could expect more goals and more shots on target.

Table 7: Cluster centres of the three retained clusters of clubs based on "attractiveness"

Cluster	**League**			**Total**
	Premier League	La Liga	Ligue 1	
Attractive clubs	6	2	3	**11**
Watchable clubs	4	7	4	**15**
Boring clubs	10	11	13	**34**

For potential further, in-depth analysis, we provide the list of clubs within each cluster (Table 8).

Table 8: Teams which make each of the three retained clusters based on attractiveness indicators

Attractive teams	**Watchable teams**	**Boring teams**		
Arsenal (UK)	Southampton (UK)	Leicester City (UK)	Stoke City (UK)	Hull City (UK)
Chelsea (UK)	Everton (UK)	Swansea City (UK)	Burnley (UK)	Sunderland (UK)
Tottenham (UK)	Bournemouth (UK)	Middlesbrough (UK)	Crystal Palace (UK)	Watford (UK)
Paris Saint Germain (FRA)	West Ham United (UK)	Las Palmas (ESP)	Deportivo La Coruna (ESP)	Leganes (ESP)
Real Madrid (ESP)	Sevilla (ESP)	Alaves (ESP)	Eibar (ESP)	Celta Vigo (ESP)
Manchester City (UK)	Villarreal (ESP)	Real Betis (ESP)	Sporting Gijón (ESP)	Nice (FRA)
Olympique Lyonnais (FRA)	Atletico Madrid (ESP)	Stade Malherbe Caen (FRA)	West Bromwich Albion (UK)	Girondins Bordeaux (FRA)
Manchester United (UK)	Real Sociedad (ESP)	Espanyol (ESP)	Saint-Etienne (FRA)	Rennes (FRA)
Monaco (FRA)	Athletic Bilbao (ESP)	Granada (ESP)	Metz (FRA)	Lorient (FRA)
Barcelona (ESP)	Malaga (ESP)	Bastia (FRA)	Toulouse (FRA)	Nancy (FRA)
Liverpool (UK)	Valencia (ESP)	Lille (FRA)	Osasuna (ESP)	Nantes (FRA)
	Olympique de Marseille (FRA)			Dijon (FRA)
	Angers (FRA)			
	Montpellier (FRA)			
	Avant de Guingamp (FRA)			

4. CONCLUSION

The field of study of application of statistical, mathematical, and data-mining techniques in sponsorship and in sports analytics is developing (Radojičić, Djoković & Jeremić, 2016). In this paper, we aimed to enlarge the current literature on the topic of decision making in marketing, particularly in sport sponsorship. Herein, we applied statistical multivariate analyses to ease the sponsors' decision making when choosing which league or club from the Premier League, La Liga or Ligue 1 to sponsor.

Our results show that there is no statistically significant difference between the three observed leagues when it comes to attractiveness, meaning the matches in all three leagues are of same quality. However, difference was spotted when it comes to roughness. Maybe there really is the difference in the style of play, but maybe there is difference in the referees' approach to what *a foul is* and what is not. This result might indicate that the game in La Liga might be a little bit slower and with more stoppage time due to referees' decisions. The results of clustering showed that Real Madrid and Barcelona are more similar to English than Spanish clubs and that most of the clubs from the three leagues are oriented to defence and actions to save their goal.

Several future directions of the study could be determined. One direction could be towards the introduction indicators of popularity of football clubs on social media. Namely, football clubs have large, committed communities of fans who can easily be communicated with via sponsorship. The second direction of the analysis could be the application of Structural Equation Measurement (SEM) analysis. Namely, SEM has been used with success in the analysis of data related to football. For example, it was used to examine the linkages between financial performance, sporting performance and stock market performance of English football clubs (Samagaio, Couto, & Caiado, 2009). Also, other statistical multivariate analysis such as Principal Component Analysis could be employed (Moura, Martins, & Cunha, 2014).

We hope the presented research will provide additional insights on differences between the three European football leagues and on application of multivariate analysis in decision making in sport sponsorship.

REFERENCES

Alamar, B. (2013). *Sports analytics : a guide for coaches, managers, and other decision makers*.

Biscaia, R., Correia, A., Rosado, A. F., Ross, S. D., & Maroco, J. (2013). Sport Sponsorship: The Relationship between Team Loyalty, Sponsorship Awareness, Attitude Toward the Sponsor, and Purchase Intentions. *Journal of Sport Management*, *27*(4), 288–302. https://doi.org/10.1123/jsm.27.4.288

Fried, G., & Mumcu, C. (2017). *Sport Analytics: A Data-driven Approach to Sport Business and Management*. Routledge.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

International Events Group. (2017). Sponsorship Spending Forecast: Continued Growth Around The World. Retrieved April 1, 2018, from http://www.sponsorship.com/IEGSR/2017/01/04/Sponsorship-Spending-Forecast--Continued-Growth-Ar.aspx

Isik, M., Ozaydin, O., Burnaz, S., & Topcu, Y. I. (2016). A Multi Criteria Decision Analysis Approach to Measure the Effectiveness of Sports Sponsorship. In *Looking Forward, Looking Back: Drawing on the Past to Shape the Future of Marketing. Developments in Marketing Science: Proceedings of the Academy of Marketing Science* (pp. 564–573). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-319-24184-5_143

Jensen, J. A., & Cornwell, T. B. (2017). Why Do Marketing Relationships End? Findings From an Integrated Model of Sport Sponsorship Decision-Making. *Journal of Sport Management*, *31*(4), 401–418. https://doi.org/10.1123/jsm.2016-0232

Jeremic, V., Bulajic, M., Martic, M., Markovic, A., Savic, G., Jeremic, D., & Radojicic, Z. (2012). An Evaluation of European Countries' Health Systems through Distance Based Analysis. *Hippokratia*, *16*(2), 170–174.

Kate Langshaw. (2017). La Liga vs Premier League – Which Is Better? | Olive Press News Spain. Retrieved April 1, 2018, from http://www.theolivepress.es/spain-news/2017/10/20/la-liga-vs-premier-league-which-is-better/

Lee, S., & Ross, S. D. (2012). Sport sponsorship decision making in a global market. *Sport, Business and Management: An International Journal*, *2*(2), 156–168. https://doi.org/10.1108/20426781211243999

McCarville, R. E., & Copeland, R. P. (1994). Understanding Sport Sponsorship through Exchange Theory. *Journal of Sport Management*, *8*(2), 102–114. https://doi.org/10.1123/jsm.8.2.102

Moura, F. A., Martins, L. E. B., & Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of Sports Sciences*, *32*(20), 1881–1887. https://doi.org/10.1080/02640414.2013.853130

O’Reilly, N., Lyberger, M., McCarthy, L., Séguin, B., & Nadeau, J. (2008). Mega-Special-Event Promotions and Intent to Purchase: A Longitudinal Analysis of the Super Bowl. *Journal of Sport Management*, *22*(4), 392–409. https://doi.org/10.1123/jsm.22.4.392

Olson, E. L., & Mathias Thjømøe, H. (2009). Sponsorship effect metric: assessing the financial value of sponsoring by comparisons to television advertising. *Journal of the Academy of Marketing Science*, *37*(4), 504–515. https://doi.org/10.1007/s11747-009-0147-z

Papadimitriou, D., & Apostolopoulou, A. (2009). Olympic Sponsorship Activation and the Creation of Competitive Advantage. *Journal of Promotion Management*, *15*(1–2), 90–117. https://doi.org/10.1080/10496490902892754

Quester, P. G. (1997). Awareness as a measure of sponsorship effectiveness: the Adelaide Formula One Grand Prix and evidence of incidental ambush effects. *Journal of Marketing Communications*, *3*(1), 1–20. https://doi.org/10.1080/135272697346014

Radojičić, M., Djokovic, A., & Jeremić, V. (2016). Evaluating football players efficiency using different multivariate analysis approaches. In *XLIII Simpozijum o operaiconim istraživanjima – SYM-OP-IS 2016* (pp. 607–610).

Radojičić, M., Milenković, N., Totić, S., Bijelić, A., & Đoković, A. (2013). Statistička analiza performansi fudbalskih timova. In *XL Simpozijum o operaiconim istraživanjima – SYM-OP-IS 2013* (pp. 845–850).

Radovanović, S., Radojičić, M., Jeremić, V., & Savić, G. (2013). A Novel Approach in Evaluating Efficiency of Basketball Players. *Management Journal for Theory and Practice Managemen*, *67*, 37–45. https://doi.org/10.7595/management.fon.2013.0012

Samagaio, A., Couto, E., & Caiado, J. (2009). *Sporting, financial and stock market performance in English football: an empirical analysis of structural relationships*.

Santomier, J. (2008). New media, branding and global sports sponsorship. *International Journal of Sports Marketing and Sponsorship*, *10*(1), 9–22. https://doi.org/10.1108/IJSMS-10-01-2008-B005

Smith, A., Graetz, B., & Westerbeek, H. (2008). Sport sponsorship, team support and purchase intentions. *Journal of Marketing Communications*, *14*(5), 387–404. https://doi.org/10.1080/13527260701852557

Soccer Stats. (2018). Soccer Statistics. Retrieved February 20, 2018, from www.soccerstats.com

Šurbatović, J. (2014). *Menadžment u sportu*. Zavod za udžbenike.

UEFA. (2016). *The European Club Footballing Landscape*. Retrieved from http://www.uefa.com/MultimediaFiles/Download/OfficialDocument/uefaorg/Clublicensing/02/53/00/22/2530022_DOWNLOAD.pdf

MEASURES OF DIGITALIZATION IN EUROPEAN ENTERPRISES: LINEAR REGRESSION MODEL

Marko Prodanović*[1], Damjan Rovinac[1], Stefan Radibratović[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: marcusdee012@gmail.com

Abstract: *The main idea of this research is to show how digital world expansion that is ubiquitous nowadays affects the economies. We observed the financial measures of digitalization, as total turnover from e-commerce in enterprises, and nonfinancial indicators such as percentage of enterprises that are purchasing or selling online, that own a website and advertise through the internet, that employ ICT specialists, as well as the percentage of employees that use a computer or mobile internet connections. In this paper, we mainly focused on countries in EU. A backward linear regression model was employed in the analysis. Our findings show that the total turnover from e-commerce is mostly affected by Internet advertising, number of ICT specialists employed and the percentage of enterprises that are selling online. The obtained model accounts for 73% of the variability. The presented research might provide insights on the factors which influence digitalization of enterprises.*

Keywords*: European Union, digitalization, enterprises, e-commerce, the linear regression model*

1. INTRODUCTION

Economic growth, as a very important feature for every country, depends on many factors such as natural resources, implemented laws, available technology, human capital. The digitalization is one of them, especially important nowadays. Because of this, the whole world, including European Union (EU), started to measure digitalization and to collect data for analysis in order to increase its positive effects. The term digitalization could be defined as a process of creating new value through the aspects of digital technology (Schallmo & Williams, 2018a). It focuses on capabilities which support the whole business idea, and with the fast development in the field of ICT, countries, industries, and companies compete and create value in completely new ways (Schallmo & Williams, 2018b).

World institutions are continually reporting on how to stimulate the digital transformation of European economies and enterprises, and proposing short and long-term strategies on digital entrepreneurship to maximize its impact (Ardolino et al., 2017). Before any new processes implemented, it is important that we have a clear picture of how the new process affect other parts of the industry. More particularly, EU has noticed that digitalization can affect massively on traditional businesses and industries, by creating and destroying jobs (European Economic and Social Committee, 2017). Logically, by implementing new methods of work and new technologies, not all the people in the industry, especially older generations, can follow up new trends and be ready to change their work routine. Accordingly, it is clear that implementing digitalization can have a positive effect, but it is also really important to be aware of the downsides and to find a reasonable solution to minimize the possible negative effect.

It is also important to mention that not all EU countries are developed, and according to that, digitalization is not the main focus for all. Research showed that in developing countries population age and urban population are positively associated with the ICT adoption (Billon, Marco, & Lera-Lopez, 2009). Additional research showed that in developing countries Internet costs are negatively associated with ICT adoption (Chen, Jaw, & Wu, 2016). However, a number of researches argue that ICT and the Internet decrease production costs, enhance the creation and spread of new ideas, support knowledge, sharing, and improve R&D processes (Kamalipour & Friedrichsen, 2017; Matt, Hess, & Benlian, 2015; Meijers, 2014). The main belief of these authors is that ICT is tightly linked to higher economic growth (Czernich, Falck, Kretschmer, & Woessmann, 2011; Degryse, 2016; Heeks, 2010). Therefore, countries should create strategies in order to use the potential that the digitalization has (O’Donnell, 2016).

The Digital Economy is a global phenomenon, and in order to see the whole potential of the digital economy, it is important to review EU on a global level. In order to do that, European Commission has introduced Digital Economy and Society Index - DESI (European Commission, 2018), which evaluates the performance of both the individual EU countries and EU as a whole in comparison to 15 other countries: Australia, Brazil,

Canada, China, Iceland, Israel, Japan, Korea (Rep.), Mexico, New Zealand, Norway, Russia, Switzerland, Turkey, and the United States.
When we discuss about digitalization, NRI index is something worth to mention. NRI represents a key tool in assessing countries' preparedness to reap the benefits of emerging technologies and capitalize on the opportunities presented by the digital transformation and beyond. More particularly, the World Economic Forum experts assess the factors, policies, and institutions that enable a country to fully leverage information and communication technologies (ICTs) for increased prosperity and crystallizes them into a global ranking of networked readiness at the country level in the form of the NRI (The Global Information Technology Report , 2016).

The research conducted in this paper aims to find the connections between different digitalization indicators. Connection between indicators could be very useful for society in global, especially for companies and counties in developing stage. We observed the financial measures of digitalization, as total turnover from e-commerce in enterprises, and nonfinancial indicators such as percentage of enterprises that are purchasing or selling online, that own a website and advertise through the internet, that employ ICT specialists, as well as the percentage of employees that use a computer or mobile internet connections. In this paper, we mainly focused on countries in Europe, specifically the 28 countries that are the members of the EU, EU-28.

The paper is organized as follows. The next section gives the list of indicators that are used in this study, as well as the employed methodology. The third section presents the results and discusses the findings of the research. The final section gives some concluding remarks on the presented topic.

2. METHODOLOGY

This paper focuses on investigating and modelling the relationship among different measures of digitalization in the enterprises in Europe. The data for this analysis are collected from Eurostat and are publicly available (Eurostat, 2017). The data are collected for 28 European countries (EU-28). We observed a set of eight variables that measure the digitalization of the enterprises in the economies of European Union countries:

- *Value of E-Commerce Sales of Enterprises* – This indicator measures enterprises' total turnover from e-commerce. It is given as a percentage of total turnover and includes all the enterprises, without financial sector (ten persons employed or more).
- *E-commerce Purchases of Enterprises* – The percentage of enterprises that have ever made any purchase through computer-mediated networks.
- *E-commerce Sales of Enterprises* – The percentage of enterprises that are selling their products online (which covers at least 1% of their turnover).
- *Internet Advertising of Enterprises* – The percentage of enterprises that use any social media for advertising over the internet.
- *Computer Internet Connections used by the Employees in Enterprises* – Persons employed that are using computers with access to World Wide Web; a percentage of total employment.
- *Mobile Internet Connections used by the Employees in Enterprises* – This indicator counts persons employed in a company (a percentage of total employment), which were provided with a portable device that allows a mobile connection to the internet for business use.
- *Enterprises that have a Website* – Percentage of enterprises that own a website, enterprises without financial sector (ten persons employed or more).
- *Enterprises that Employ ICT Specialists* – A percentage of enterprises that employ ICT specialists, without financial sector.

In order to investigate and model the relationship among digitalization indicators in enterprises in EU, we observed the variable *Value of E-Commerce Sales of Enterprises* as a dependent variable. We have chosen this specific variable because it can be defined as a financial indicator that describes the current turnover from digital commercialization. We aimed to investigate whether the set of seven other variables, which could be defined as the explanatory variables, significantly influence the e-commerce sales in Europe.

We used backwards multiple least squares linear regression (Narula & Wellington, 1982; Rencher & Christensen, 2012; Seal, 1967) to create a model, which would automatically exclude non-significant variables from the observation (Gujarati, 2002). The original multivariate regression model is given by the following formula:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki} + \varepsilon_i \qquad (1)$$

where Y_i is an *i*-th observation of the dependent variable *Y*, β_j is a *j*-th regression coefficient (j=1,…k), X_{ji} is an *i*-th observation for the *j*-th independent variable, and ε is a residual.

3. RESULTS AND DISCUSSION

Table 1 gives the descriptive statistic of the variables used in the research. As described in Section 2, variables are mainly given as the percentage of turnover/enterprises/employment/etc. Thus, all the variables are normalized. Among the observed variables, the variables *E-commerce Purchases of Enterprises* and *Enterprises that have a Website* have the widest range. The largest mean value is for the *Enterprises that have a Website*, 76.11%, which tells us that this is the mostly spread digitalization indicator in the countries of EU. Other two digitalization indicators that are spread across Europe are *Internet Advertising of Enterprises*, with the mean value 49.96%, and *Computer Internet Connections used by the Employees in Enterprises*, with the mean value 48.86%. The largest standard deviation is noticed with the variable *E-commerce Purchases of Enterprises*, SD=16.41%, followed by *Internet Advertising of Enterprises*, SD=13.13%.

Table 1: The descriptive properties of variables defined in Section 2

Employment rates of recent graduates	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
Value of E-Commerce Sales of Enterprises	4	33	16.360	7.445
E-commerce Purchases of Enterprises	11	75	38.571	16.419
E-commerce Sales of Enterprises	7	30	17.250	6.598
Internet Advertising of Enterprises	27	74	49.960	13.128
Computer Internet Connections used by the Employees in Enterprises	27	75	48.860	11.891
Mobile Internet Connections used by the Employees in Enterprises	10	51	25.071	10.846
Enterprises that have a Website	45	96	76.110	11.955
Enterprises that Employ ICT Specialists	10	33	20.960	5.175

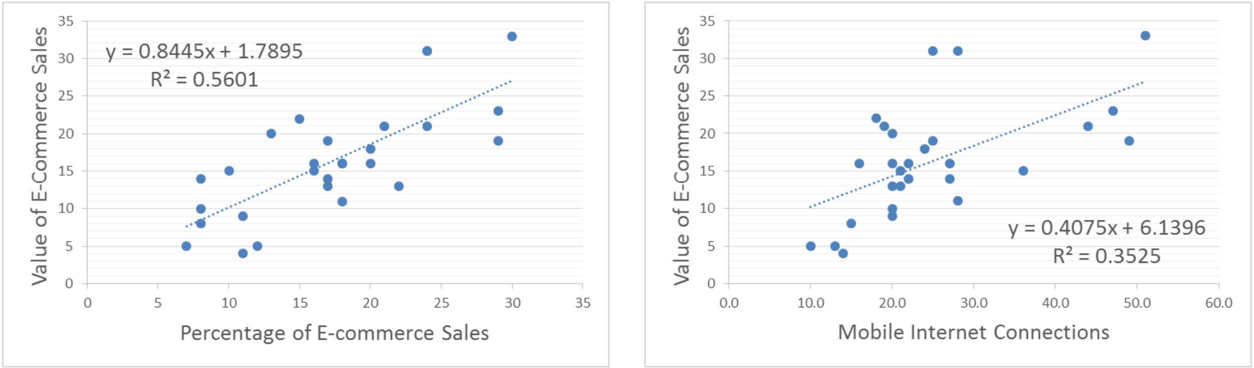
The basic objective of this study was to determine the relationships between *Value of E-Commerce Sales of Enterprises* and the overall digitalization performance. For this purpose, we calculated Pearson correlation coefficients that are given in Table 2.

Table 2: Pearson's correlation coefficients between variables

Employment rates of recent graduates	Value of E-Commerce Sales	E-commerce Purchases	E-commerce Sales	Internet Adv.	Computer Internet Conn.	Mobile Internet Conn.	Enterprises with a Website
E-commerce Purchases	0.556**						
E-commerce Sales	0.748**	0.603**					
Internet Advertising	0.205	0.454*	0.516**				
Computer Internet Connections by Employees	0.444*	0.785**	0.666**	0.648**			
Mobile Internet Connection by Employees	0.594**	0.669**	0.747**	0.575**	0.822**		
Enterprises that have a Website	0.525**	0.720**	0.661**	0.618**	0.821**	0.618**	
Enterprises that Employ ICT Specialists	0.495**	0.420*	0.469*	0.733**	0.401*	0.495**	0.503**

*p<0.05, **p<0.01

From Table 2, we can see that the variable *Value of E-Commerce Sales (percentage of turnover)* is strongly correlated with *E-commerce Sales (percentage of enterprises that are selling online)* and *Mobile Internet Connections used by the Employees in Enterprises*. These relationships are graphically presented in Figure 1, *a* and *b*. Pearson's correlation coefficient ranges from 0.205 to 0.822 and we have only one coefficient which is not statistically significant, the correlation coefficient between *Internet advertising* and *Value of E-Commerce sale* (r=0.205, p>0.05).



(a) *(b)*

Figure 1: Relationship between *Value of E-Commerce Sales* and (a) Percentage of E-commerce sales and (b) Mobile Internet Connection used by the Employees in Enterprises

From Figure 1*a* and Table 2, it can be seen that the Pearson's correlation coefficient between *Value of E-Commerce Sales (percentage of turnover)* and *E-commerce Sales (percentage of enterprises that are selling online)* is r=0.748 (p<0.001). The coefficient of determination is R^2=0.5601, indicating that the percentage of enterprises that are selling online alone explains 56.01% of the variability in the percentage of total turnover for e-commerce. In Figure 1*b* and Table 2, we see that the Pearson's correlation coefficient between *Value of E-Commerce Sales (percentage of turnover)* and *Mobile Internet Connections used by the Employees in Enterprises* is r=0.594 (p=0.001). The coefficient of determination R^2=0.3525, which means that 35.25% of total turnover for e-commerce is explained solely by mobile internet connections.

In order to model the presented relationships, we have built the linear regression model. As mentioned previously, we used backwards multiple least squares linear regression model. The original regression model is presented in Table 3.

Table 3: The original multiple linear regression model for the dependent variable *Value of E-Commerce Sales of Enterprises*

Explanatory variables	**B**	**SE**	**Beta**	**t**	**95%CI B**	
Intercept	-2.174	5.785		-0.376	-14.241	9.894
E-commerce Purchases of Enterprises	0.055	0.082	0.121	0.670	-0.116	0.225
E-commerce Sales of Enterprises	0.687	0.191	0.609	3.598**	0.289	1.086
Internet Advertising of Enterprises	-0.396	0.111	-0.698	-3.555**	-0.628	-0.164
Computer Internet Connections used by the Employees in Enterprises	-0.067	0.204	-0.107	-0.327	-0.493	0.359
Mobile Internet Connections used by the Employees in Enterprises	0.117	0.161	0.171	0.730	-0.218	0.452
Enterprises that have a Website	0.110	0.134	0.177	0.822	-0.170	0.391
Enterprises that Employ ICT Specialists	0.775	0.260	0.539	2.978**	0.232	1.318
F	10.974**					
R^2	0.793					
Adjusted R^2	0.721					

*p<0.05, **p<0.01

In the original linear regression model, all of the variables defined for this research were included in the analysis. From Table 3 it can be seen that, among all of the listed variables included in the model, only *E-commerce Sales of Enterprises, Internet Advertising of Enterprises,* and *Enterprises that Employ ICT Specialists* influence the dependent variable *Value of E-Commerce Sales of Enterprises*. All other variables' influences are not significant (p>0.05). With the adjusted coefficient of determination R^2=0.7210, the model reveals that 72.1% of the variability of total turnover for e-commerce is explained by the given combination of explanatory variables. The whole model is significant at 0.01 level of significance (F=10.974, p<0.001). The model also exhibits a large level of multicollinearity among explanatory variables.

Table 4 presents the reduced backwards multiple linear regression model. This model was used to exclude the non-significant variables from the model, step by step automatically, and thus to create the model that would better fit the given data.

Table 4: The reduced multiple linear regression model for the dependent variable *Value of E-Commerce Sales of Enterprises*

Explanatory variables	**B**	**SE**	**Beta**	**t**	**95%CI B**	
Intercept	1.144	3.255		0.352	-5.573	7.862
E-commerce Sales of Enterprises	0.906	0.133	0.803	6.796**	0.631	1.181
Internet Advertising of Enterprises	-0.363	0.087	-0.639	-4.165**	-0.542	-0.183
Enterprises that Employ ICT Specialists	0.844	0.214	0.587	3.941**	0.402	1.286
F	25.335**					
R^2	0.760					
Adjusted R^2	0.730					

*p<0.05, **p<0.01

From Table 4, it can be seen that the reduced model includes only three, of initially seven explanatory variables. As indicated by the original model, only E-c*ommerce Sales of Enterprises, Internet Advertising of Enterprises,* and *Enterprises that Employ ICT Specialists* significantly influence *Value of E-Commerce Sales of Enterprises*. The most influential indicator is *E-commerce Sales of Enterprises* (percentage of enterprises that are selling online). What is interesting to see from result is that the values of *Internet Advertising of Enterprises* have a negative impact on the dependent variable. This result could be interpreted as that the use social networks for advertising does not automatically mean the sales are going to rise.

The estimated model is presented as follows in the given formula:

$$\hat{Y} = 1.144 + 0.906X_1 - 0.363X_2 + 0.844X_3 \qquad (2)$$

where Y is the dependent variable *Value of E-Commerce Sales of Enterprises*, X_1 is *E-commerce Sales of Enterprises*, X_2 is *Internet Advertising of Enterprises*, and X_3 is *Enterprises that Employ ICT Specialists*. With the adjusted coefficient of determination R^2=0.730, the model reveals that 73% of the variability of the total turnover of e-commerce in European enterprises is explained by the given combination of explanatory variables. The whole model is significant at 0.01 level of significance (F=25.335).

4. CONCLUSION

This research focused on creating a model that would analyze the dependence of a total turnover of e-commerce in European enterprises, as a financial indicator, from a set of indicators that measure the state of digitalization transformation.

After four iterations in a backwards linear regression model analysis, the final model shows that the dependent variable *Value of E-Commerce Sales of Enterprises* depends on *E-commerce Sales of Enterprises, Internet Advertising of Enterprises,* and *Enterprises that Employ ICT Specialists.* Other four variables (*E-commerce Purchases of Enterprises, Computer Internet Connections used by the Employees in Enterprises, Mobile Internet Connections used by the Employees in Enterprises,* and *Enterprises that have a Website*) are not statistically significant in creating the model. The final model, presented in table 4, has the coefficient of determination R^2=0.73, which means that 73% of the variability of total turnover for e-commerce is explained by the given combination of explanatory variables.

This analysis could be useful for researchers and companies which have implemented digitalization or have a plan to do so in a future. According to this data, all the enterprises can focus only on three variables and consequently, reduce costs in the process of digital transformation. Besides companies, research can be helpful for policy makers and governments as it provides insights that ICT specialists are needed in companies so as to increase the value of e-commerce sale and thus increase the level of economic activity. For example, policy makers could promote higher education in the ICT field.

As a direction of future research, it would be interesting to analyze the same regression model, using the same variables, but on different groups of countries (divided for example according to different life standards). It is possible to get certain changes in scores because of the market diversity and different people priorities according to their living standards (developed countries, developing countries, or undeveloped countries).

REFERENCES

Ardolino, M., Rapaccini, M., Saccani, N., Gaiardelli, P., Crespi, G., & Ruggeri, C. (2017). The role of digital technologies for the service transformation of industrial companies. *International Journal of Production Research*, 1–17. https://doi.org/10.1080/00207543.2017.1324224

Billon, M., Marco, R., & Lera-Lopez, F. (2009). Disparities in ICT adoption: A multidimensional approach to study the cross-country digital divide. *Telecommunications Policy*, *33*(10–11), 596–610. https://doi.org/10.1016/j.telpol.2009.08.006

Chen, Y.-Y. K., Jaw, Y.-L., & Wu, B.-L. (2016). Effect of digital transformation on organisational performance of SMEs. *Internet Research*, *26*(1), 186–212. https://doi.org/10.1108/IntR-12-2013-0265

Czernich, N., Falck, O., Kretschmer, T., & Woessmann, L. (2011). Broadband Infrastructure and Economic Growth*. *The Economic Journal*, *121*(552), 505–532. https://doi.org/10.1111/j.1468-0297.2011.02420.x

Degryse, C. (2016). Digitalisation of the Economy and its Impact on Labour Markets. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2730550

European Commission. (2018). The Digital Economy and Society Index (DESI). Retrieved from https://ec.europa.eu/digital-single-market/en/desi

European Economic and Social Committee. (2017). Impact of digitalisation and the on-demand economy on labour markets and the consequences for employment and industrial relations, 1–76. Retrieved from https://www.ceps.eu/system/files/EESC_Digitalisation.pdf

Gujarati, D. (2002). *Basic Econometrics*. McGraw-Hill/Irwin.

Heeks, R. (2010). Do information and communication technologies (ICTs) contribute to development? *Journal of International Development*, *22*(5), 625–640. https://doi.org/10.1002/jid.1716

Kamalipour, Y. R., & Friedrichsen, M. (2017). Introduction: Digital Transformation in a Global World. In *Digital Transformation in Journalism and News Media* (pp. 1–4). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-27786-8_1

Matt, C., Hess, T., & Benlian, A. (2015). Digital Transformation Strategies. *Business & Information Systems Engineering*, *57*(5), 339–343. https://doi.org/10.1007/s12599-015-0401-5

Meijers, H. (2014). Does the internet generate economic growth, international trade, or both? *International Economics and Economic Policy*, *11*(1–2), 137–163. https://doi.org/10.1007/s10368-013-0251-x

Narula, S. C., & Wellington, J. F. (1982). The Minimum Sum of Absolute Errors Regression: A State of the Art Survey. *International Statistical Review / Revue Internationale de Statistique*, *50*(3), 317. https://doi.org/10.2307/1402501

O'Donnell, S. (2016). Digital skills: unlocking the information society. *Information, Communication & Society*, *19*(12), 1770–1772. https://doi.org/10.1080/1369118X.2016.1235719

Rencher, A., & Christensen, W. (2012). Chapter 10, Multivariate regression – Section 10.1, Introduction. In *Methods of Multivariate Analysis* (p. 19). John Wiley & Sons.

Schallmo, D. R. A., & Williams, C. A. (2018a). Digital Transformation of Business Models (pp. 9–13). https://doi.org/10.1007/978-3-319-72844-5_3

Schallmo, D. R. A., & Williams, C. A. (2018b). History of Digital Transformation (pp. 3–8). https://doi.org/10.1007/978-3-319-72844-5_2

Seal, H. (1967). Studies in the History of Probability and Statistics. XV The historical development of the Gauss linear model. *Biometrika*, *54*(1–2), 1–24. https://doi.org/10.1093/biomet/54.1-2.1

Silja Baller, Attilio Di Battista, Soumitra Dutta, Bruno Lavin(2016). The Global Information Technology Report 2016. *The Networked Readiness Index 2016* (pp. 3). http://www3.weforum.org/docs/GITR2016/WEF_GITR_Full_Report.pdf

RESEARCH OF ASSOCIATION RULES AS DECISION MAKING TOOL FOR MANAGERS

Višnja Istrat*[1], Dajana Matović[1], Milko Palibrk[2]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
[2]Administration for joint services of the Republic Bodies, Serbian Government, Belgrade, Serbia
[*]Corresponding author, e-mail: visnja.istrat@gmail.com

Abstract: *In modern business it is a challenge to find possibilities to improve business decision-making of managers. Managers' decisions directly affect the profit and success of companies at the market. As very complex process that should result with right managers' decisions, there is need to improve methods and techniques of modern decision-making. In the paper the significance and application of association rules will be analysed on the example of car sales. Research will be dealing with data in large databases of car sales by use of data mining software Orange. Main results provide recommendation to create suitable promotional activities according to existing and potential buyers, with car offers and additional equipment that are bought together. One of advantages of this model, when compared to others, is that it could also be applied in other industries. Associations have been used in order to make convenient and interesting product offer for* the market.

Key words: *decision-making, data mining, association rules.*

1. INTRODUCTION

Business decision-making is choosing the best solution of all available alternatives, according to (Suknović and Delibašić 2010). According to (Agrawal et al, 1993), Agrawal first established the association rules with the purpose of analysis the market basket.

Although the market basket is mostly being used in sales, it is important to highlight that there are also other segments where its application is significant: analysis of finances, medical analysis, then analysis of insurance companies or telecommunications services.

The example of company Catalogue sales is one of the examples of successful application of method of association rules in marketing, especially on improvement of sales (Verhoef et al, 2010). The most important result of the application of association rules is oriented towards the need to create thematic catalogues that will contain certain items of products that will adapt to market segments.

In order to solve the problem of company Colorful world of colors the method of association rules had been used (Verhoef et al, 2010). Company recognized the fall of sales of certain products. The goal was to find out if these are products that are sold together with some other products or on their own. Analysis was designed so that it was analysed the preferences of buying non-popular set of products in combination with the main set of products. In main set of products are the ones that bring the biggest sales profit. Research showed that the buyers continued to buy products from the main set, and the sales of non-popular products stayed at the same level.

Another example of application of method of association rules has been described in the frame of searching the web (Verhoef et al, 2007). Searching the association rules were used to reveale the web pages that are jointly used. For instance, information that users who approach to pages A and B also do it for page C, can be used to create certain links (i.e. from A to C), which would result for further analysis of e-business.

Extensive research on association rules mining has been conducted in (Martin et al, 2014). This research extends evolutionary algorithm based on decomposition to perform learning of the intervals of the attributes and a condition selection for each rule. Application of association rules on market basket analysis had been described in (Omondi, A. O., and Mbugua, A. W., 2017). The methodology used contain minimum spanning trees. According to (Jain, S. et al, 2018), the importance of data mining and business strategy have been described with patterns that emerged as results. In (Griva, A. et al, 2018) there is business analytics approach that describes customer visit segments from basket sales data.

Researchers were dealing with the extensive application of method of association rules in business decision-making. This method has been proved to be the most successful in the field of market basket. This paper presents the application of association rules with the purpose of discovering patterns of the customers' buying habits in the car industry.

2. RESEARCH

Data mining is mostly oriented towards the model creation. Under right circumstances, model can provide the explanations how output elements of certain interest, such as precise order or unsuccessful payment of bills, are connected and can be predicted by available facts.

Association rules determine the items that are bought together. According to (Grabmeier, J., and Lambe, L., 2007), the term of association rules was firstly used by Agrawal. The task of association is to find out rules that exist between the cases in the dataset. According to (Grabich, M., 1997), knowledge is accumulation of the facts and information, whilst wisdom is the synthesis of knowledge and experience that deepen our understanding of the connections between different entities and eventual some hidden message in their existing. We can say that knowlege is tool, while wisdom is set of skills where we use knowledge as the tool. Cases for association rules are grouping the items that are bought together in shopping in supermarket – the field of market basket is dedicated to this issue (Fernando, B., and Susanto, B., 2011).

The importance of rules, that is the quality of rules discovered by the association, is determined based on the markers of support and confidence. Support is the procentual part from the sample that applies to certain rule, that is the number of instances that satisfy a rule. Confidence tells what percentage of cases posess attribute A, also posess attribute B. The biggest fault of the association rules is that in most cases of analysis on large databases many rules are not interlinked, so their simplicity can go to unconsistency (Doko A., 2010). There are numerous approaches that can be used for revealing certain rules, and the most basic one is the Apriori algorhitm. This algorhitm works on the preassumption that if certain number of items is present in the dataset, than each item is also present (Grabmeier, J., and Lambe, L., 2007).

This paper shows the structure of data, the evaluation process and the application of association rules. Based on the structure of research data, as well as the use of tools of business intelligence, the end-result will be the model of business decision-making and knowledge discovery.

According to (AL-Zawaidah and 2011), there are three dimensions that determine the complete development of modern decision-making. Those are: qualitative aspect, quantitative aspect and information-communication aspect. In this paper the multidisciplinary approach will be described. Data will be shown in quantitative way and will be processed by numerical values. Methods and techniques of business intelligence will be applied. Attributes will be described in order to ensure the qualitative aspect of modern decision-making.

Description of the result will ensure qualitative aspect. Use of software architecture and visual overview of data will ensure information-communication aspect of modern decision-making. Data that was used in research is connected to car sales. Total of 1728 transactions are being processed. Attributes are connected to buying (very high rate of transactions, high, medium and low), maintenance of cars (very high rate, high, medium, low), number of car doors (two, three, four, five), number of persons (two, three, four), seating space (small, medium, large) and safety (low, medium, high).

All data are divided into four classes: accurate, unaccurate, good and very good. Orange has options for each phase of CRISP-DM methodology. From large database of car sales transactions, data will be processed in software Orange Canvas (Figure 1), in order to find the association rules. Associations could help buyers with provided knowledge about the best car offer. Buyers help the process of association rules in the selection of cars, based on relavant sales data.

This paper describes the model of application of business intelligence on business problem, with the use of modern software architecture. Orange, as very suitable and user-friendly software, contains numerous options for data mining, such as model creation, testing, data vizuelization, application of model, etc. One of the popular technologies for data mining is CRISP-DM methodology.

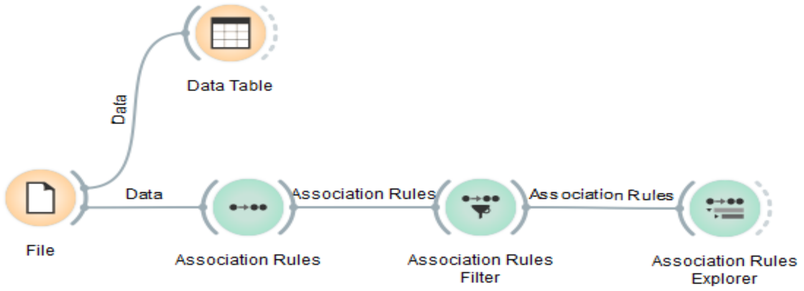


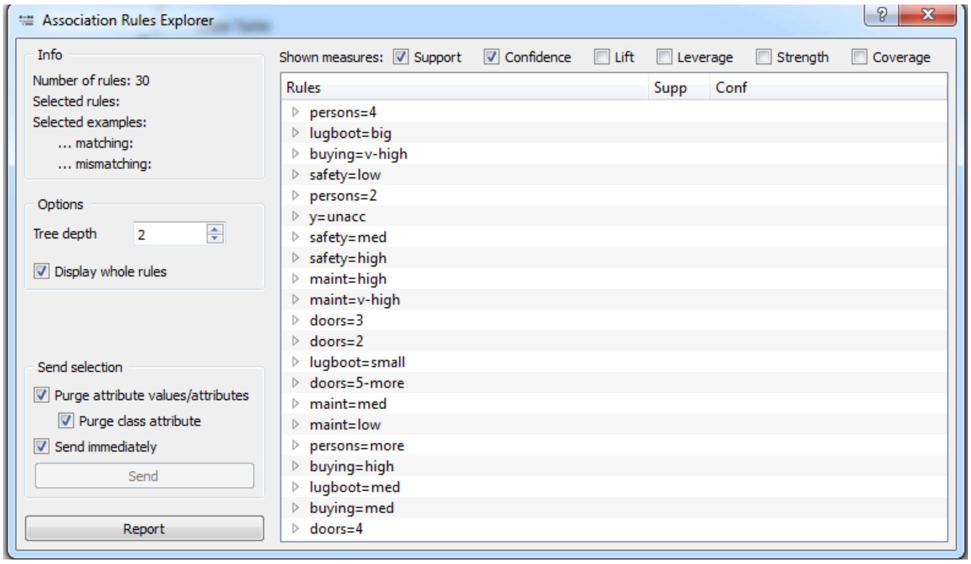
Figure 1: Model of association rules created in software Orange Canvas

3. DISCUSSION

The parametars for support and confidence are being defined. In this case minimal support is defined at 30%, and minimal confidence is 50%. In further research this parametars were changed in order to show how their change affect the end-result of finding the association rules.

Considering that the minimal support is set at 30%, there are only two association rules derived. First association with provided parametars is that car for two persons belongs to class that is called unaccurate. Moreover, 30% of all car sale transactions for two persons were from the class unaccurate. Support of 50% means that when customers buy from the class unaccurate, there is 50% probability that they will buy cars for two persons. It is predicted that in all future transactions, cars from this class would be offered to customers. Second association rule is that when safety is defined as low, then it follows that the most often car is from class unaccurate. These changes of parametars show the strength of created association rules. The higher the parametars are, the stronger association rules are (with highest probability to show).

Orange is convenient for goods visual overview of gained results. Report shows the number of gained association rules, as well as their description (Figure 2).


Association Rules Explorer

Info
Number of rules: 30
Selected rules:
Selected examples:
... matching:
... mismatching:

Options
Tree depth 2
☑ Display whole rules

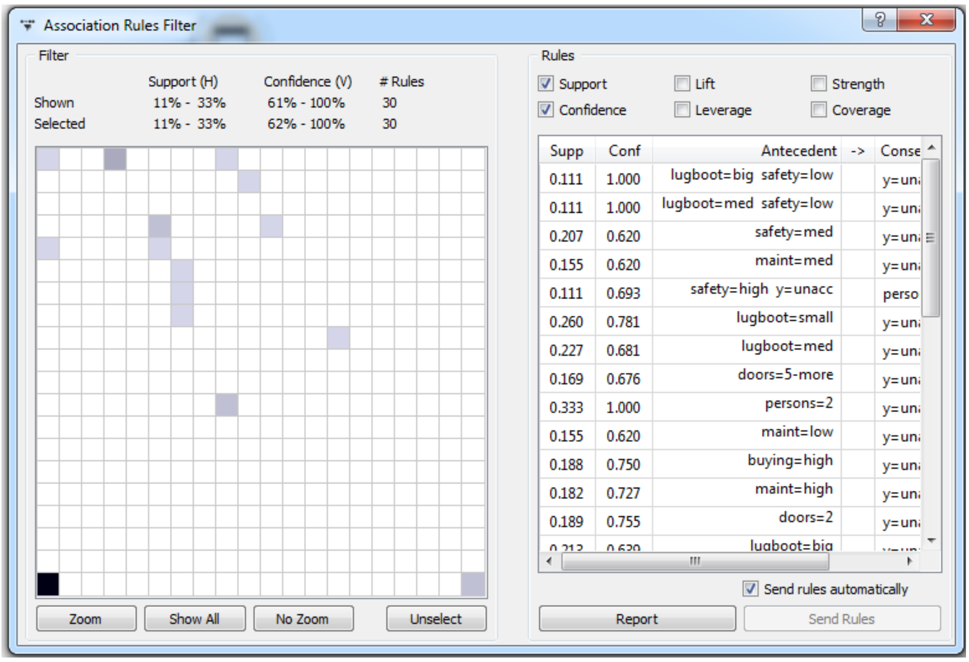
Send selection
☑ Purge attribute values/attributes
☑ Purge class attribute
☑ Send immediately

Shown measures: ☑ Support ☑ Confidence □ Lift □ Leverage □ Strength □ Coverage

Rules	Supp	Conf
persons=4		
lugboot=big		
buying=v-high		
safety=low		
persons=2		
y=unacc		
safety=med		
safety=high		
maint=high		
maint=v-high		
doors=3		
doors=2		
lugboot=small		
doors=5-more		
maint=med		
maint=low		
persons=more		
buying=high		
lugboot=med		
buying=med		
doors=4		

Figure 2: Association rules

By reducing minimal support from 30% to 10 %, and increasing minumal confidence from 50% to 60% different results are achieved. The number of revealed association rules has increased. There are thirty new association rules with support rate from 11% until 33% and confidence rate from 62% until 100%. Further explanations show patterns that are used to create best set of products for sales which results with highest degree of sales and customer satisfaction. On the left-hand side of the graphicon (Figure 3) there is overview of gained association rules. More details in (Istrat, 2017).



	Support (H)	Confidence (V)	# Rules
Shown	11% - 33%	61% - 100%	30
Selected	11% - 33%	62% - 100%	30

☑ Support ☐ Lift ☐ Strength
☑ Confidence ☐ Leverage ☐ Coverage

Supp	Conf	Antecedent	->	Conse
0.111	1.000	lugboot=big safety=low		y=un
0.111	1.000	lugboot=med safety=low		y=un
0.207	0.620	safety=med		y=un
0.155	0.620	maint=med		y=un
0.111	0.693	safety=high y=unacc		perso
0.260	0.781	lugboot=small		y=un
0.227	0.681	lugboot=med		y=un
0.169	0.676	doors=5-more		y=un
0.333	1.000	persons=2		y=un
0.155	0.620	maint=low		y=un
0.188	0.750	buying=high		y=un
0.182	0.727	maint=high		y=un
0.189	0.755	doors=2		y=un
0.213	0.620	lugboot=big		y=un

☑ Send rules automatically

Figure 3: Overview of obtained association rules

From the obtained association rules one was highlighted with biggest support of 0.333 and biggest confidence of 1. It is the rule when customers buy car for two persons they most often decide to buy from the class that is defined as unaccurate. Analysis of each association rule show precise patterns of buyers' behaviour. Another association rule that is distinguished by importance is the one with support of 0.260 and confidence of 0.781 which shows that when smaller space inside the car is needed, then buyers most often buy from class unaccurate. Nonstandard data (outliers) can cause problems in creating the model. Outliers represent rare events that are extemptions from the ruled in data. They can affect that the model applied on all data won't be with high quality, because it is difficult for algorhitms to find the regularities in the presence of extemption. There are no outliers in this research. More details in (Istrat and Lalić, 2017).

Putting data through Apriori algorhitm provided association rules that are then checked and only the most important are chosen to show and create meta rules, in order to get the most acceptable solution. This solution is applicable not only in finding the original universal rule for improvement of car sales, but for other items as well. Recommendation is creating the marketing campaigns based on knowledge that was revealed by use of tools of business intelligence. It is recommended that customers are provided by questionnaire that would search for what car characteristics are the most important.

In the case there are car characteristics (attributes): price, maintanance costs, number of doors, number of seats, size of the truck and safety. Gained association rule showed that buyers that buy car for two persons and which find safety at low level, they most often do it from the class unaccurate.. It is recommended to create promotional activities according to existing and potential buyers, with car offers and additional equipment that are bought together. Advantage of application of this model when comparing to others is that it could be applied also in other different sectors (banking sector, education, etc.). The topic of research is very interesting and provides extensive space for further elaboration of association rules application. Real-term settings and practical implication of research provide advantage when compared to other similar research. Experienced business intelligence analytics were used to define the results of the research.

4. CONCLUSION

It is challenge for researchers to provide sigificant scientific and expert contribution. Innovation of research of application of business intelligence is shaped with knowledge and creativity, as well as the use of modern software architectures from the field of data mining. Gained associations were used with purpose to make the best products' offer at the market. The process of improvement of model was shown with the help of knowledge that managers get through association rules about habits and buyers' behaviour in shopping. Better choice of the product offer for launching at the market was improved by efficiency of management of sales with the help of business intelligence. Developed model points out at the significance of methods and techniques of business intelligence for support of management of business systems in creating maximization of profit. The goal of further research would be creating the model of business decision making that is applicable to market in business systems of different areas and with commercial purpose.

REFERENCES

Agrawal, R., Imielinski, T., and Swami, A., (1993). Mining association rules between sets of items in large databases. In Proc. 1993 ACM-SIGMOD. Int. Conf. Management of Data. pp.207-216. Washington. D.C.

AL-Zawaidah, F., H., Jbara, Y.H., AL-Abed Abu-Zanona, M. (2011). An improved algorithm for mining association rules in large databases. World of Computer Science and Information Technology Journal WCSIT. ISSN: 2221-0741 Vol. 1. No. 7, 311-316.

Doko A. (2010). Automatic genesis of onthology and browsing the Web. Split. Croatia.

Fernando, B., Susanto, B. (2011).The implementation of association rules in analyzing the sales of Amigo group. Journal Informatika. Vol 7. No 1.

Grabich, M. (1997). Alternative representations of discrete fuzzy measures for decision making. International Journal of Uncertainty. Fuzziness and Knowledge-Based Systems. No5.

Grabmeier, J., & Lambe, L. (2007). Decision trees for binary classification variables grow equally with the Gini impurity measure and Pearson's chi-square test. International Journal of Business Intelligence and Data Mining. Volume 2. pp.213-226.

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. Expert Systems with Applications, 100, 1-16.

Istrat, V. P. (2017). Unapređenje modela poslovnog odlučivanja sistemom asocijativnih pravila (Doctoral dissertation, Univerzitet u Beogradu-Fakultet organizacionih nauka).

Istrat, V., & Lalić, N. (2017). Creating a Decision-Making Model Using Association Rules. Applied Artificial Intelligence, 31(5-6), 538-553.

Jain, S., Sharma, N. K., Gupta, S., & Doohan, N. (2018). Business Strategy Prediction System for Market Basket Analysis. In Quality, IT and Business Operations (pp. 93-106). Springer, Singapore.

Martin, D., Rosete, A., Alcalá-Fdez, J., & Herrera, F. (2014). A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules. IEEE Transactions on Evolutionary Computation, 18(1), 54-69.

Omondi, A. O., & Mbugua, A. W. (2017). An Application of association rule learning in recommender systems for e-Commerce and its effect on marketing.

Suknović, M., & Delibašić, B. (2010). Poslovna inteligencija i sistemi za podršku odlučivanju. FON, Beograd.

Verhoef, P.C., Venkatesan, R., McAlister, L., Malthouse, E.C., Krafft, M., Ganesan, S. (2010). CRM in Data-Rich Multichannel Retailing Environments: A Review and Future Research Directions. Journal of interactive marketing 24. p. 121-137.

Verhoef, P., C., van Doorn, J., Dorotic, M. (2007). Customer Value Management: An Overview and Research Agenda. Marketing, Journal of Research in Management. 2. 51-69.

SKI LIFT TRANSPORTATIONS AS PREDICTORS FOR INJURY OCCURRENCE

Boris Delibašić*[1], Sandro Radovanović[1], Miloš Jovanović[1]
[1]University of Belgrade – Faculty of Organizational Sciences
*boris.delibasic@fon.bg.ac.rs:

Abstract: *Ski lift transportation data is readily available in ski resorts, still heavily underused. There are many ways this data can be used, but in this paper, we study the applicability of this data for injury occurrence prediction on the data from ski resort Kopaonik. We propose a decision tree model for studying injury occurrence. We find that the most relevant predictor for injury occurrence is the number of ski lift transportations. While the number of ski lift transportations is below a certain threshold we can predict the number of injuries on a day almost linearly with the number of ski lift transportations. When the number of ski lift transportation increases, i.e. when the ski resort gets congested, the occurrence of injuries depends on the level of satisfaction skiers achieve. If skiers have an above average or even a below average level of satisfaction, the injury occurrence risk is high. If skiers experience average level of experience then the risk of injury is average or low.*

Keywords*: Ski injury occurrence, data mining, ski lift transportation data, Ski resort Kopaonik*

1. INTRODUCTION

The problem studied in this paper is the daily occurrence of ski injuries based on ski lift transportation data, a readily available resource in ski resorts. Studying this problem is important as ski injuries pose a major public healthcare problem with significant medical costs worldwide. Ski resorts are interested to accurately predict the daily level of injury occurrence. This is important for allocation of medical resources in ski resorts. Insurance companies are also interested to know more accurate statistics for ski injury occurrence for defining ski insurance prices. The problem of predicting injuries has usually been assessed by assessing the number of skiers in the resort and knowing the injury rate of ski injury occurrence which is usually 0.2% (Ruedl et al. 2013). Bohanec and Delibašić (2015) have proposed several models for daily ski injury occurrencebased on ski lift transportation data. Their models were concerned whether injuries will occur on a certain day or not, and whether an above average number of injuries will occur. This paper extends the problem defined in the paper from Bohanec and Delibašić (2015) by predicting whether injuries will occur at an expected level (within one standard deviation of the normalized residual between expected injury occurrence and real injury occurrence) or injuries will occur at a higher or lower level. Our results suggest that the number of ski lift transportations is a great predictor for situations when there is a specific amount of skiers in the ski resort. Above this amount of skiers, i.e. when assumably congestion in ski resort occurs, ski lift transportation number are not a reliable predictor of ski injury occurrence. To understand the behaviour of skiers better we introduce a skiers satisfaction measure which gives better insight whether injuries will occur at lower, higher, or expected values.

The remainder of the paper is structured as follows. In Section 2 we provide background on ski injury research based on ski lift transportation data. In Section 3 we explain the data and the methods used in this research. In Section 4 we present the obtained results. We make the conclusion in Section 5.

2. BACKGROUND

The study of ski injury based on ski lift transportation data is recently getting more attention (e.g. Dallagiacoma, M. 2017). Ski lift transportation data is a resource kept within ski resorts and usually only used for simple reporting for ski resort management. Delibašić, et al. (2017a) have proposed spatial and temporal clustering models based on ski lift transportation data. The identified clusters provide motivation for the marketing team of ski resorts to enable more intelligent ski ticket formats to their users. Delibašić et al. (2017c) have extended the previous work of skiing clusters by investigating hidden factors that motivate skiers to choose a specific path.

Ski injury is a well-studied area (e.g. Dalipi, F., & Yayilgan, S. Y. 2015), however, the data from ski lift transportation has only been recently started to be used for studying ski injury.

Delibasic, B., & Obradovic, (2012) proposed a decision support system prototype for early warning for ski injuries. The same authors also suggested (Delibašić, B., & Obradović, Z. 2015) that skiing in groups has a significant influence on ski injury occurrence. Bohanec, M., & Delibašić, B. (2015) proposed severaldata-mining and expert models for predicting daily injury risk for a ski resort. Dobrota, et al. (2016) clustered skiing trajectories and found clusters with varying levels of injury risk. Delibašić, et al. (2017b) have proposed a ski injury assessment model for individual skiers. This study suggested ski injuries to be early failure events, i.e. it was shown that the first couple of hours of skiing are more critical to get injured than the later hours.

3. DATA AND METHODS

The data for this research stems from the ski resort Kopaonik, Serbia, which is Serbia's largest and most significant ski resort bordering its southern province Kosovo & Metohija. The data is from the period 2006 to 2011 and belongs to the company Ski resorts of Serbia. The injury data is provided with courtesy of the Mountaineer rescue service of Serbia which covers the ski resort of Kopaonik with its ski patrol. In this research also data from the Serbian hydrometeorological service was used.

The basic format of the ski lift transportation data is:

- Date and time of ski lift gate entrance
- Location of ski lift gate
- Skier ID

The complete dataset for this research had the following attributes:

- Average daily temperature
- Average wind speed
- Average cloudiness
- Daily number of ski lift transportations in the ski resort
- The total daily sum of vertical meters of all skiers in the ski resort
- Total number of skiers

The output attribute was the risk level which could have been:

- Low injury occurrence
- Expected injury occurrence
- High injury occurrence

The output attribute was calculated as a normalized residual. For each skier day, the expected level of injury was calculated based on the number of ski lift transportations, as ski lift transportations showed a better correlation to the number of injuries than the vastly used measure which is the number of skiers.

Then the residual was calculated as the difference between the expected injury occurrence and the observed injury occurrence. As it was not the same if there was a difference when there is a small number of ski lift transportations in the ski resort or a huge number of ski lift transportations, the residual was normalized taking into account the largest number of ski lift transportations.

For the analysis in this paper, we used the data mining software Orange from the University of Ljubljana (Demšar et al 2013). We used Sieve diagrams which find a correlation between attributes based on the strength of the chi-square statistics. We also applied decision trees and other popular data mining algorithms to test the strength of our results.

4. RESULTS

We found a strong correlation between the number of ski lift transportations and the risk level. In Figure 1 this relationship can be seen. This relationship indicates that while the number of ski lift transportations is below 38,193 then mostly the expected level of injury risk can be noticed, i.e. we can safely use the number of ski lift transportations to predict ski injury. However, if the number of ski lift transportation is higher than this number then either a low or above average number of injuries can be expected.

On the decision tree on Figure 2, this point is made clearer. The most important indicator of risk injury prediction is the number of ski lift transportations (the rootof the tree). If the number of ski lift transportations is 36,523 or below than an expected level of injuries will occur. We are confident of this finding with 97.9% percent at the training set. However, when there is a larger number of ski lift transportations it is much more difficult to get an overall good prediction. The decision tree on Figure 2 has an overall classification accuracy at the test set 88.9% while the majority class prediction (which is the expected value) has an occurrence of 89.6% which means that the decision tree makes a worse prediction than if using only the expected value as the forecast. Although accuracy is low, it is more important to identify injury behavior. In other words, false positive (predicting injury that did not occurred) prediction is of less cost compared to false negative predictions (predicting that injury will not occur, but it occurred). Therefore, recall of prediction model is more important.

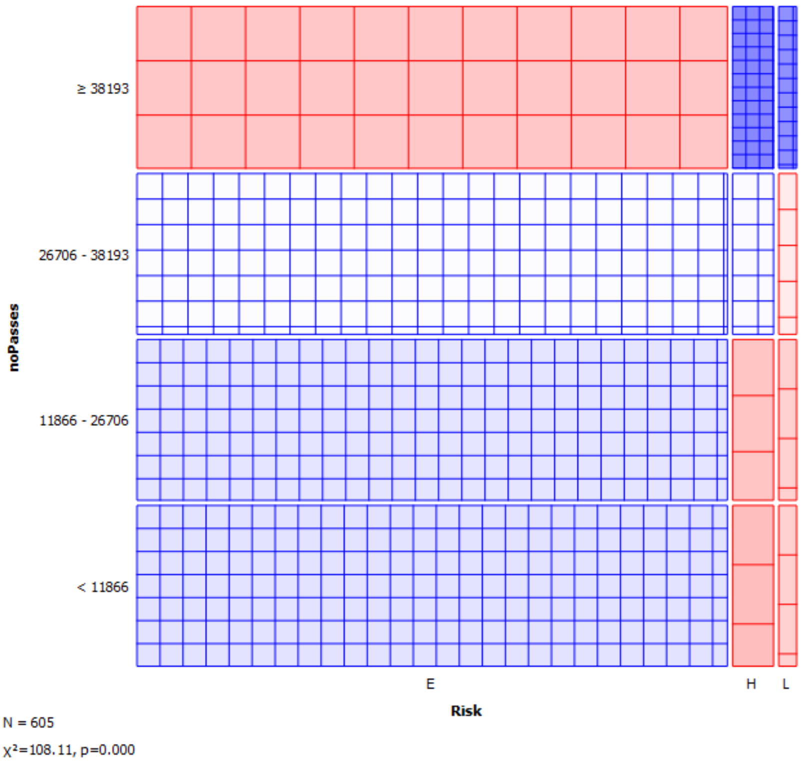


Figure 1: Sieve diagram showing a correlation between the number of ski lift transportations (noPasses) and the risk level (E-expected, H-high, L-low)

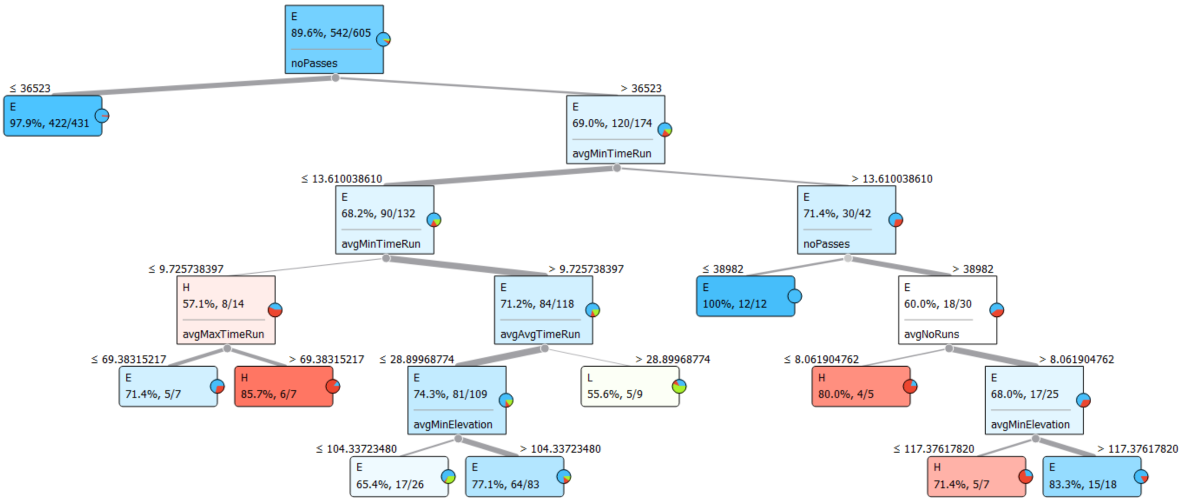


Figure 2: Basic decision tree model

In order to understand what is happening when there are a lot of transportations in the ski resort (the ski resort has a lot of skiers in the system) we propose a measure for quantifying success in the ski resort. We assume that each skier’s goal is to achieve as much skiing in the resort as possible. We measured this success with achieved vertical meters (the sum of height above sea level between ski lift exit points and ski lift entrances) within a certain time period.

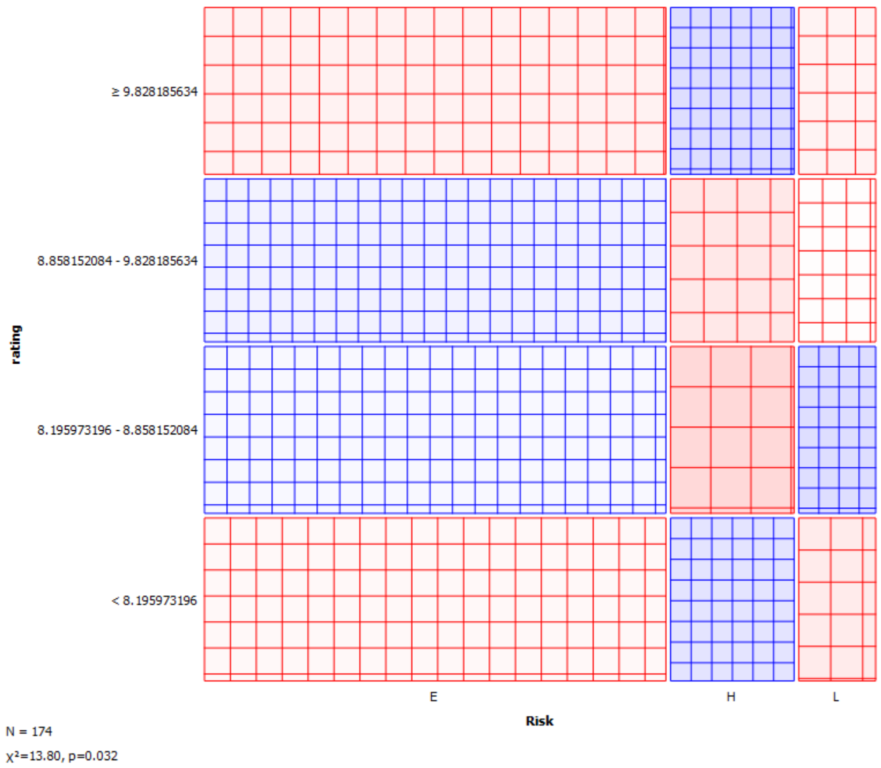


Figure 3: The relationship between the rating of the success of skiers and the risk level

The rate of success is shown in (1). It is calculated for each day, and is the ratio of the total amount of vertical meters (vm) skiers have achieved and the total time (t) skiers have spent in the ski resort:

$$SR(d) = \frac{\sum vm}{\sum t} \tag{1}$$

From Figure 3 it can be noticed that if the rating of success is below 8.19 or above 9.83 (vertical meters per minute) an above average level of risk can be expected. If this level is between 8.86 and 9.83 an average level of injury can be expected. If this level is between 8.19 and 8.86 either an average or low level of ski injuries can be expected.

This finding is interesting as it suggests that a smaller, but also greater, a rating of success influences high risk of injury. This can be explained as follows. If skiers' population have a high rate of success, they ski at faster speeds. Faster speeds induce greater levels of risk, which is a well-known result known from road traffic injuries. On the other hand, if the rate of success is below a certain threshold either the conditions of skiing are not optimal (low visibility, bad conditions of slopes) or the ski resort is overcrowded. This way injury due to collisions is more likely to occur. In any case, it worth noticing that safer skiing can be expected when this ratio is between 8.2 and 9.8 vertical meters per minutes.

We have also analyzed the performance of several data mining algorithms on the dataset that only included days with ski lift transportations above 36,523. The results are shown in Table 1.

Table 1: Data mining algorithm performance is shown in percentages

Algorithms	**Accuracy**	**Recall (Expected Risk)**	**Recall (High Risk)**	**Recall (Low Risk)**
AdaBoost	56.9	70.8	30.3	4.8
Logistic Regression	69.5	100	3	0
Naïve Bayes	54	71.7	15.2	1.43
Random Forest	59.8	85.8	3	0
Tree	61.5	79.2	33.3	1.43

It can be noticed that algorithms have it overall difficult to predict the occurrence of the low-risk level. On the other hand, algorithms don't achieve a better accuracy of prediction over 69.5% which is the majority class predictor (Logistic regression models uses this method for prediction). However, if we inspect recall of the models we can observe that lower accuracy models had better recall for high risk of injury. For predicting high-risk occurrence two algorithms showed to be the best, AdaBoost and the Decision tree where the decision tree had a slight advantage of 3%.

5. CONCLUSION

This paper has studied the occurrence of ski injury based on ski lift transportation data. It was shown that this data can only be used a good predictor until a certain amount of ski lift transportation happens in the ski resort. After this level, the injury risk can vary significantly. We, therefore, introduced a success measure of skiers which showed to be significantly correlated (p=0.032 from Figure 3) to the risk level occurrence. From data, it can be noticed that the safest skiing occurs when skiers have a balanced success score. If this score is above or below the average values high risk of injuries could occur. Ski resorts could influence the safety of the ski resorts by controlling that skiers achieve the optimal success scores. This could be either done by intelligently slowing down the ski lifts if skiers speed too much. Ski lift speeds could be also better adjusted to achieve faster transportations when there is congestion in the ski resorts. Of course, better warnings of skiers to ski safely could also be of great help.

This paper suggested some first insights into studying the daily level of injury occurrence in ski resorts. Further efforts should be made in understanding what is happening in ski resorts which are congested and how this influences the level of injury. We propose to introduce new features which could probably be better predictors, such as quality of snow or other quality of success measures, for data mining models, but also to test these findings on other ski resorts to prove the validity of these findings.

REFERENCES

Bohanec, M., & Delibašić, B. (2015, May). Data-mining and expert models for predicting injury risk in ski resorts. In International conference on decision support system technology (pp. 46-60). Springer, Cham.

Dalipi, F., & Yayilgan, S. Y. (2015, July). The impact of environmental factors to skiing injuries: Bayesian regularization neural network model for predicting skiing injuries. In Computing, Communication and Networking Technologies (ICCCNT), 2015 6th International Conference on (pp. 1-6). IEEE.

Dallagiacoma, M. (2017). Predicting the risk of accidents for downhill skiers. School of Information and Communication Technology, KTH Royal Institute of Technology. Stockholm, Sweden.

Delibasic, B., & Obradovic, Z. (2012, April). Towards a DGSS prototype for early warning for ski injuries. In Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on (pp. 68-72). IEEE.

Delibašić, B., & Obradović, Z. (2015). Identifying High-Number-Cluster Structures in RFID Ski Lift Gates Entrance Data. Annals of Data Science, 2(2), 145-155.

Delibašić, B., Marković, P., Delias, P., & Obradović, Z. (2017a). Mining skier transportation patterns from ski resort lift usage data. IEEE Transactions on Human-Machine Systems, 47(3), 417-422.

Delibašić, B., Radovanović, S., Jovanović, M., Obradović, Z., & Suknović, M. (2017b). Ski injury predictive analytics from massive ski lift transportation data. Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology, 1754337117728600.

Delibašić, B., Radovanović, S., Jovanović, M., Vukićević, M., & Suknović, M. (2017c, September). An Investigation of Human Trajectories in Ski Resorts. In International Conference on ICT Innovations (pp. 130-139). Springer, Cham.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., ... & Štajdohar, M. (2013). Orange: data mining toolbox in Python. The Journal of Machine Learning Research, 14(1), 2349-2353.

Dobrota, M., Delibašić, B., & Delias, P. (2016). A Skiing Trace Clustering Model for Injury Risk Assessment. International Journal of Decision Support System Technology (IJDSST), 8(1), 56-68.

Ruedl, G., Kopp, M., Sommersacher, R., Woldrich, T., & Burtscher, M. (2013). Factors associated with injuries occurred on slope intersections and in snow parks compared to on-slope injuries. Accident Analysis & Prevention, 50, 1221-1225.

DECISION TREE-BASED ALGORITHM FOR THE CLASSIFICATION OF MUSICAL INSTRUMENTS

Anja Bjelotomić*[1], Aleksandar Rakićević[1], Ivana Dragović[1]
[1]University of Belgrade, Faculty of Organizational Sciences
*Corresponding author, e-mail: anjinimejl@gmail.com

Abstract:*This paper examines the use of binary decision tree for classification of musical instruments, using mel-frequency cepstral coefficients as timbre features. We analyze 4094 samples of bass clarinet, contrabassoon, flute, oboe, trumpet and violin to get their timbre coefficients. The dimensions of these coefficients are reduced using principal component analysis. A binary tree is created and optimized to predict the instrument classes based on these timbre coefficients.*

Keywords:*Binary decision trees, musical instrument classification, mel-frequency cepstral coefficients*

1. INTRODUCTION

The human capability to perceive differences between musical instruments is a subject of research for a number of years. Even with minimal musical knowledge, most people can easily make distinction between familiar musical instruments, even played at the same loudness and pitch. By definition of American National Standards Institute (1951) the quality of auditory sensation by which a listener can distinguish between two sounds of equal loudness, duration and pitch are known as timbre. Hence it could be said that musical instrument recognition is strongly dependent on timbre. Unfortunately, unlike pitch and loudness, timbre has proven to be somewhat difficult to measure or quantify (Loughran, Walker, O'Neil & O' Farrell, 2000).

Over the last decade, numerous studies have been investigating the nature of speech and speaker recognition research. Progress has been made on the analysis of speech waveforms, in its perception by humans, and in the use of different statistical methods for classification. On the other hand, the topic of instrument classification and recognition has been studied less. In this paper, we will be relying on knowledge gained in speech research, such as the *mel-frequency cepstral coefficients.*

This paper focuses on the automatic recognition of musical instruments, where the idea is to build a binary decision tree that can "listen" to the musical sounds and recognize which instrument is playing. The experimental material consists of 4094 single notes, the timbre of which has been studied comprehensively. Combined with the state-of-the-art automatic sound source recognition systems, these form the foundation for the most important part of this work: the extraction of perceptually relevant features of acoustic musical signals and their classification into instrument groups.

Binary trees are one of most common and powerful data structures in the computer science The computational cost of making a tree is fairly low, but the cost of using it is even lower $-O(\log N)$ (Castan, Ortega & Lleida, 2010). Binary trees are used as a cornerstone of classification analysis, which is the main reason it is discussed in this work.

The literature review and motivation part of this paper introduces the studies of the sound timbre and classification methods used for their recognition. Following, Section 3 reflects the method used in the experiment and explains the core idea of the work, while details on implementation of binary tree are given in Section 4. Finally, results are concluded in Section 5.

2. LITERATURE REVIEW

As mentioned, the field of speech/music classification was studied by many researchers. Firstly, there are numerous studies that analyze the behavior of different classification algorithms used for music and speech other than binary trees. In the Tables 1 and 2, a brief overview is given on general studies already done in the field of music recognition.

Large amount of data is in the audio format, from several resources such as broadcasting channels, databases, Internet streams and commercial CDs. This has made space for a new field of research, audio content analysis (ACA), or machine listening, whose purpose is to analyze the audio data and extract the content information directly from the acoustic signal (Burred & Lerch, 2004). Speech and music classification is also useful in other forms of applications, for example, content-based audio coding or indexing other data, such as classification of video content through the accompanying audio. Spurred by the growth of annotated datasets and the democratization of high-performance computing, feature learning has enjoyed a renewed interest in recent years within the MIR community, both in supervised and unsupervised settings (MATLAB, 2015).

Table 1: Summary of studies in the field of audio classification

Authors	Classification method	Main application	Audio materials	Results
Saunders (1996)	Multivariate Gaussian classifier	Automatic real-time FM radio monitoring	Different types of music, commercials, talk	95-96% of accuracy
Footer (1997)	MFCCs, short-time energy	Using acoustic similarity to retrieve audio documents.	409 sounds and 255 (7sec long) clips of music	High rate of success, although no specific accuracy rate provided
Burred and Leech (2004)	KNN classifier, 3 component GMM classifier, MFCC	Classification of music into genres and audio classification	Speech, 13 genres of music, background noise	94,6% hierarchical approach and 96,3% direct approach
Marques and Moreno (1999)	Gaussian mixture model, Support Vector Machines, mel-frequency feature set	Automatic annotation system	Bagpipes, clarinet, flute, harpsichord, organ, piano, trombone and violin	GMM had an 75% accuracy rate, SVM 70% accuracy
Eronen (2001)	Distance-based algorithms, probabilistic classifiers	Listen and recognize the instrument playing	29 instruments	35% accuracy between 29 instruments, 77% accuracy between six instrument families
Toghiani-Rizi and Windmark (2017)	Multilayer perceptron	Musical instrument recognition	Feature vector of length 50, containing a normalized frequency spectrum of the audio signal	Average accuracy of 93,5%

Numerous studies have been done in the field using binary decision trees (Table 2).

Table 2: Summary of studies in the field of audio classification using binary decision trees

Authors	Main application	Audio materials	Results
Wold, Blum and Wheaton (1996)	Building a system to distinguish between sound classes	laughter, bells, synthesizer, different instruments	98,6% accuracy
Han, Pen, Jeon, Lee and Ha (1998)	Genre classification	Songs from three genres: jazz, classical and popular music.	55% accuracy
Lavner and Ruinsky (2009)	Segmentation of audio signals into speech and music	12 hours of speech, 22 hours of music	99,4% accuracy for speech, 97,8% for music, quick adjustment to altering speech/music sections
Castan, Ortega and Lleda (2010)	Classification of audio frames into speech or music	20 tracks, that alternate segments of music, speech or both	99,56-99,94% accuracy

Nowadays, many Internet search sites, such as AltaVista and Lycos, evolved from purely textual indexing to multimedia indexing. It is estimated that there are approximately thirty million multimedia files on the Internet with no effective method available for searching their audio content (Swain, 1998). If every sound file had a corresponding text file that accurately described human perceptions of the given audio content, audio files could be easily searched. For example, in an audio file containing only speech, the text file could include spoken text, names, most frequent words etc. In a music file, the annotations could include instruments recognized in the file or parts of the songs used. There is an uncountable amount of audio data stored off and online with no adequate classification technique based on its content, so the interest of creating computer systems and algorithms to classify instruments is evident. In that manner, automatic methods able to effectively index multimedia files are key.

With a goal of gaining broader knowledge about audio information retrieval, we chose binary decision trees as the cornerstone of our analysis, as it is easy to use, but also provides a lot of useful information. Binary trees are closely related to information theory. The idea behind trees is to choose the most informative feature at each step. The principle is to quantify how much information is provided by knowing certain facts. Additionally, binary trees usually give highly accurate classification rates, which is the reason they will be used in this work. The goal is to provide basic classification insight in order to carry out our analysis further.

3. THE PROPOSED MODEL

In this study, we investigate a problem of classifying samples of six instruments: bass clarinet, contrabassoon, flute, oboe, trumpet and violin. The first part of the experiment is focused on extraction of relevant timbre features from original input dataset. The second one regards the construction of a binary decision tree, which will be used to classify input data into instrument groups.

3.1. Feature extraction

The original data is represented as a set of tones played on different instruments. Instruments used in the experiment are bass clarinet, contrabassoon, flute, oboe, trumpet and violin. Each instrument covers a range of different dynamics and scales, ranging from pianissimo to fortissimo, from second to eighth octave, respectively.

Since the goal of this paper is to determine the timbre similarity between tones of the same instruments, first we must extract and explicitly represent timber features of the sound. This is done using **Mel-Frequency Cepstral Coefficients (MFCCs)**. It is well known that the MFCCs are a compact and efficient representation of speech (Castan *et al.*, 2010).Each coefficient has a value for each frame of the sound. We examine the changes of these coefficients across the range of the sound. In order to obtain MFCCs we must:

1. Divide signals into frames,
2. Get the amplitude spectrum of each frame,
3. Take the log of these spectrums,
4. Convert to mel-scale,
5. Apply the discrete cosine transform (DCT).

First, we divide signals into frames. This is done using a windowing technique. In most applications, the audio signal is analyzed by means of *short term* or *short-time* processing technique, according to which the signal is broken into, in our case, overlapping windows (frames) and the analysis is carried out on a frame bases (Giannakopoulos & Pikrakis, 2014). The reason we use windowing is because sound signals aren't stationary and their properties vary usually very dynamically over time. So in order to catch the switch in time between the frames, we use overlapping by some percent (25%, 50%, 75% etc.). In this experiment, the overlapping step is set to 50%.

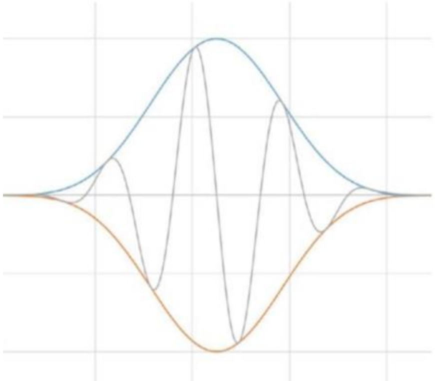


Figure 1. Windowed sample (Marsland, 2009)

Mel-scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another (Stevens, Volkmann, & Newman, 1937). We project our signals onto mel-scale in order to represent the computed sounds in a way that's similar to human perception of the tones.

Sampled, these tones were split into windows, or frames. Windowing reduces the amplitude of the discontinuities at the boundaries of each point of the sampled signal. This makes the endpoints of the waveform meet and, therefore, results in a continuous waveform without sharp transitions. Applying a window minimizes the effect of spectral leakage (National Instruments, 2015).

This way, we are able to catch changes between the signal frames and apply the discrete Fourier transform (DFT), which is kind of a cornerstone of digital signal processing. Discrete Fourier transform gives us a frequency-domain (spectral) representation of the signal:

$$X[k] = \sum_{n=0}^{N} x[n] e^{\frac{-2j\pi kn}{N}}. \tag{1}$$

Or, if we rewrite it:

$$x(n) = \frac{1}{N} \sum_{n=0}^{N-1} X(k) y_k(n), \tag{2}$$

where $y_k(n) = \frac{-2j\pi kn}{N}, \quad n = 0, ..., N-1.$

It can be seen that the original signal can be represented as a weighted average of a family of fundamental signals (basis functions), where each signal, $y_k(n)$ is a complex exponential and its weight is equal to the *k*th DFT coefficient (Giannakopoulos & Pikrakis, 2014).

After these transformations our data is represented in frequency or spectral domain, where we get the most of our timbre information from. Most commonly used and very efficient algorithm for the computation of DFT coefficients is fast Fourier transform. We are particularly interested in the magnitude of the *k*th DFT coefficient, since it represents a measure of intensity with which the respective frequency participates in the signal $x(n)$ (Giannakopoulos & Pikrakis, 2014).

Discrete Cosine Transform is used to reduce the data. DCT represents our input as a sum of cosine waves oscillating at each frequency of the signal. Discrete Cosine Transform can be expressed as:

$$y(k) = a(k) \sum_{n=0}^{N-1} x(n) \cos(\frac{\pi(2n+1)k}{2n}), \quad k = 0, 1, ..., N-1. \tag{3}$$

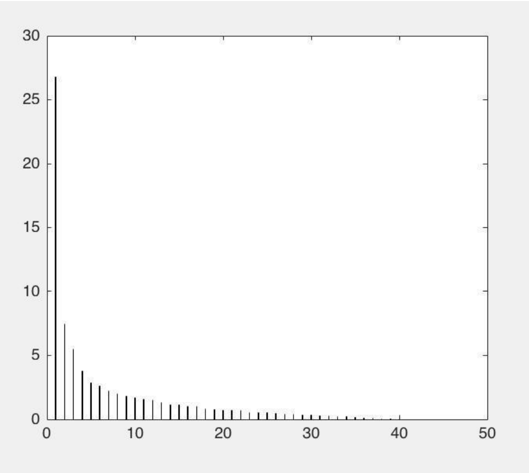


Figure 2. Eigen spectrum of the violin tone A3 *(forte)*

Afterthisalgorithmhas been applied, we have got a matrix of values for each sample sound. That is, every sound is represented with the *number of coefficients observed*, in this case **40**, by the *number of frames* in size. Since we must represent our data it in a more compact way, we are using **Principal Component**

Analysis (PCA). We apply PCA to the calculated coefficient data. PCA provides a roadmap for how to reduce a complex data set to a lower dimension to reveal the sometimes hidden, simplified dynamics that often underlie it (Wold, Blum, Keislar & Wheaton, 1996). This way, we get our final matrix of input data. It shows that one principal component is enough to describe the data needed for this model.

In Figure 2 werepresenttheeigenspectrumof a tone A from thethirdoctave, playedwith forte dynamics on the violin. Wecansee, thatthefirst principal componentoutof 40 describesthe most ofthe signal. Three, orfour at most, principal components are enough to describethe data in most efficientway. In thisexperiment, we are using one principal component. Finally, weapplybinarydecisiontreeontoour set ofsignals.

3.2. Decision tree

Tree-based algorithm used in this study is **Classification and Regression Trees (CART)** algorithm. As the name indicates, it can be used both for classification and regression (Saunders, 1996). CART uses **Gini impurity** as information measure. The ‘impurity’ in the name suggests that the aim of the decision tree is to have each leaf node represent a set of data points that are in the same class, so that there are no mismatches. This is known as purity (Marsland, 2009).

The CART classification method is implemented using MATLAB *fitctree* function. Once fitctree function returned classified tree, there are a number of parameters which can be used in order to optimize the tree in the best manner. In this experiment, we focused on getting the best values for name-value arguments *Maximum number of categories (MaxNumCategories)* and *Minimum Leaf Size (MinLeafSize)*. After setting up these arguments, another method for optimizing the decision tree is used, called pruning. Pruning optimizes tree depth (leafiness) is by merging leaves on the same tree branch (Marsland, 2009). Now that the tree is improved it can be used to classify and predict outputs of our test dataset.

3.3. Classification details

Since the feature extraction is a big and complex process and isn't quite the focus of the work, we are not going to overview that part of the experiment. We are using the output matrix of the feature extraction process as our input matrix. It is represented as a set of 4094 sound signals and 40 cepstral coefficients describing the timbre for each of them.

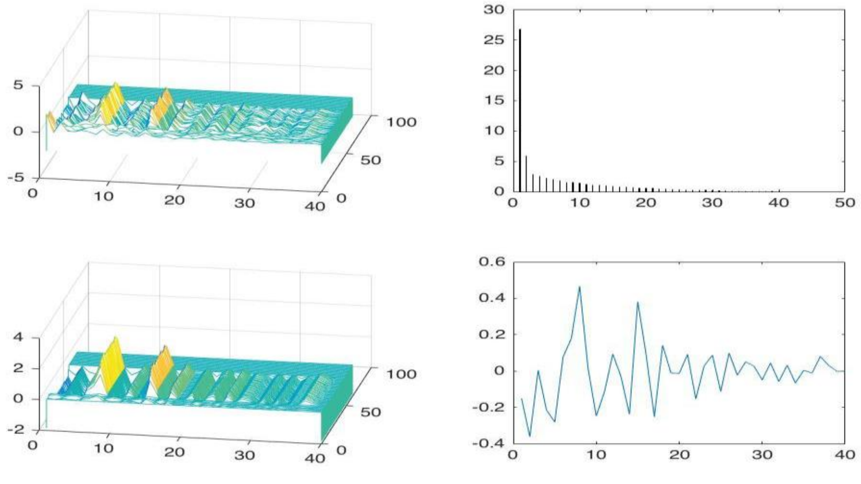


Figure 3: Representation of tone A4, *mezzo forte,* played on the oboe

Figure 3 is a representation of the oboe tone A, from the fourth octave, played in mezzo forte dynamics. In the upper left corner, we can see the spectral envelope of the sound, in its original form. Figure from the lower left corner shows the same tone, but with dimensions reduced using Singular Value Decomposition. We can see from the similarity of two left figures, that even with reduced dimensionality we still get a good approximation of the analog music signal. On the right part of the picture, in the upper corner is eigen spectrum of the tone. It allows us to see that two or three principal components are enough to describe the whole signal, while even with one we get all the information needed. Lower left graph is a 2-D representation of the signal with reduced dimensions.

In order to construct the tree, we first separated the data into training and testing sets. We are going to optimize our binary tree to fit our problem adjusting parameters *minimum leaf size* and *maximum number of* *categories.* In order to grow an efficient tree, we must consider its simplicity and predictive power. Usually, a deep tree with many leaves is highly accurate on the training data. On the other hand, the tree is not guaranteed to show a comparable accuracy on an independent test set. This leafy tree tends to overfit, or overtrain, and test accuracy is often far less than its training (resubstitution) accuracy. A shallow tree does not attain high training accuracy. But it can turn out more robust – meaning, its training accuracy could be close to that of a representative test set and shallow tree is easy to interpret.

The argument that we used to control the depth of decision trees in MATLAB is *MinLeafSize*. Function *fitctree* splits a categorical predictor using the exact search algorithm, if the predictor has at most *MaxNumCategories* levels in the split node. Passing a small value can lead to loss of accuracy and passing a large value can increase computation time and memory overload (Marsland, 2009). Commonly used method for optimization is called pruning which is commonly used in construction of binary decision trees. Pruning optimizes tree depth or leafiness by merging leaves on the same tree branch. Pruning is implemented with MATLAB function *prune*. Now our tree is final, optimized, depth-adjusted binary decision tree that we can now use to classify our test, using MATLAB function *predict*.

4. RESULT ANALYSIS

There are several possible uses of a learning tree. Creating tree to classify unknown sounds is the first one. Others include the analysis of the tree: which questions, or we could say attribute, splits the data? In our case: which coefficient has the biggest informative value? Furthermore, the tree can be used to improve the analysis, by finding out what's wrong with the misplaced data (Marques & Moreno, 1999).The main focus of this paper is to classify the sounds using binary decision tree. In addition, we will analyze which timbre attribute split the data into instrument classes.

The argument that we used to control the depth of decision trees in MATLAB is *MinLeafSize*. Function *fitctree* splits a categorical predictor using the exact search algorithm in case the predictor has at most *MaxNumCategories* levels in the split node. Otherwise, *fitctree* finds the best categorical split using one of the inexact algorithms. Passing a small value can lead to loss of accuracy and passing a large value can increase computation.

In order to test our tree, we will feed it with test data. As we mentioned, this is done using predict function in MATLAB. Function returns predicted class labels for our classification tree and predictor, in our case testing dataset. This vector is crucial in understanding how well our tree classified data. Output argument LABEL is a vector of the same type as the response data used in training tree, ctree. Each entry of LABEL corresponds to the class with minimal expected cost for the corresponding row of testing input features. We used two measures to calculate the error rate of the tree: resubstitution loss and cross-validation error. Both values are fairly low, where resubstitution loss is 0,065% and cross-validation error 0,13%.

Table 3: Confusion matrix for binary tree classifier

	Bass clarinet	**Contra-bassoon**	**Flute**	**Oboe**	**Trumpet**	**Mandolin**
Bass clarinet	163	0	0	0	0	0
Contra-bassoon	0	166	0	0	0	0
Flute	0	0	147	0	0	0
Oboe	0	0	0	130	0	0
Trumpet	0	0	0	0	99	0
Mandolin	0	0	0	0	0	316

With respect to our low classification error, we can see that tree predicted only 2 samples out of class. As confusion matrix shows us, tree misclassified two flute samples as the violin tones. This makes sense, because the flute and violin are two instruments that can reach to the highest tones. Also, this confirms the rules our binary tree defined for this problem. As we can see in the figure 4 below, the first question our tree asks is: Does our input signal's 40th cepstral coefficient hold a value that's larger than -0,0344922? If it does have a value larger than -0,0344922 it is directed to the class six, which is enumeration for the violin, and further classified. On the other hand, if it doesn't, then it is classified as class three, in our case the flute.

This is also an example of another use of binary trees – analyzingthe factors influencing classification. Unlike neural networks, binary trees offer the possibility of insight in our system, which can give us a lot of useful information for method improvement.

Spectral component is any of the waves that range outside the interval of frequencies assigned to a signal. By extracting these spectral features, meaning each spectral component of the tone, and calculating their amplitude, we are presented with the occurring frequency of each spectral component in a single tone. In other words, we are able to see how much of each non-fundamental frequencies of the tone are represented in the signal. Differences in these frequencies define what we call timbre of a tone.

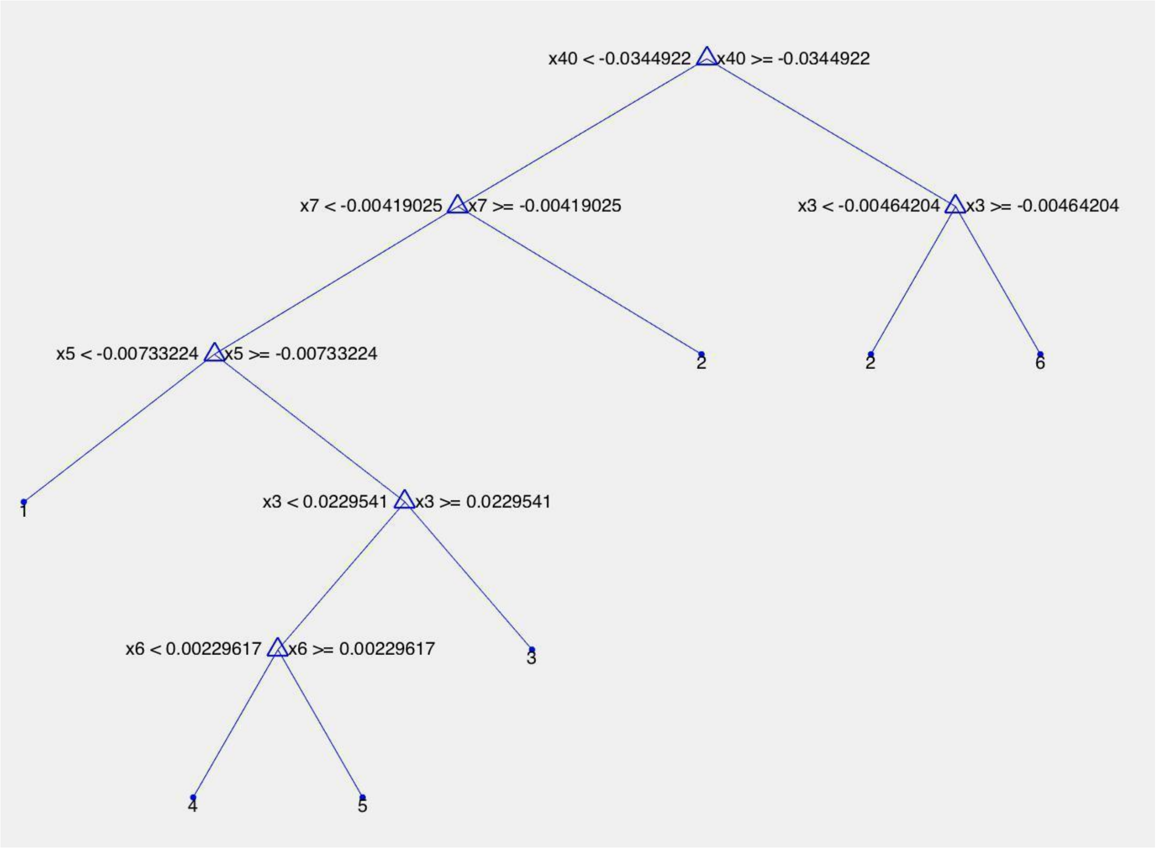


Figure 4: Representation of classification binary decision tree for our problem

We log these frequencies, actually mel-scale it, and by doing it we adjust it to human perception of the sound, since we perceive it logarithmically. Human ear perceives frequencies lower than 1 kHz on a linear scale, after which it switches to log scale. Meaning, we perceive less the difference between higher tones, ones over 1 kHz, then we do with the lower ones.

Analyzing the tree to see which attributes it used to split the data, we can see in which range of the frequencies bears the most information. We can see that the lower indexed filters are more present than the higher ones, but the *highest indexed* one is the most informative one. This makes sense, since the lower filters present lower non-fundamental frequencies, and together they represent a lot of spectral information, but the amount of highest frequencies is what makes the big difference. The bass clarinet could never get the high notes of the violin. It is important to notice, that the first 13 filters stand for lower frequencies, while the rest 27 stand for higher frequencies, in log space.

5. CONCLUSION

This paper has presented the classification of musical sounds in instrument families using binary decision trees with the accuracy of 99,87%. The automatic tree construction, using *fitctree* function, has created well-balanced, efficient trees. Binary decision trees are easy to use and usually give accurate results for classification and regression, so we often use them as a cornerstone of classification analysis.

As we can see, binary decision trees give a high accuracy in musical signal classification, not just for our problem type. We notice that in all of the mentioned studies in the field accuracy is rarely lower than 90%. Since the data is represented in a compact manner, with most important features represented directly and our data vectors are decorrelated, it makes it much easier for our tree classifier to segment the data. This is a good sign for our information retrieval method. In addition, optimization of the tree parameters benefits the high accuracy. Optimal parameters have been set for this specific problem.

We used cross-validation error and resubstitution error to measure the accuracy of classification tree. These measures were also used for evaluating the efficiency of the tree supplied with different-valued parameters

MinLeafSize and MaxNumCategories. We have created a high accuracy tree, with our features extracted from unified audio data, with dimensions reduced. Analyzing the tree, we can see that a major part of timbral information lies in the lower frequency range, but the filter bearing most valuable information is the highest indexed one, meaning the first question and the crucial information is the amount of the highest frequencies presented in a single tone.

There are a lot of questions and space for future work. Analyses should be done on how cepstral coefficients behave in the same instrument family, as well as changes in cepstrals just for one instrument. The effect of playing techniques could also be taken into account. These analyses would bring us closer to understanding and using timbre of a sound with more clarity.

Further work should also check accuracy of binary tree with different parameterization, more instrument families, and different sound representation and see the effect on the result and accuracy. Using different filter shapes within the first layer seems crucial for an efficient learning with spectrogram-based CNNs (Shlens, 2003), which is another direction for research.

REFERENCES

American National Standards Institute (1951). American Standards Association Incorporated, New York, USA.Retrieved from: https://www.ansi.org

Burred, J., & Lerch, A. (2004).Hierarchical automatic audio signal classification, *Journal of the Audio Engineering Society,* 52(7-8):724-739.

Castan, D., Ortega, A., & Lleida., E. (2010). Speech/music classification by using the C4.5 decision tree algorithm, FALA 10, *VI JornadasenTecnologíadelHabla and II Iberian SLTech Workshop*, (pp. 197-200).

Eronen, A. (2001). Automatic musical instrument recognition, Tampere University of Technology, Department of Information Technology. Doi:10.1.1.79.6635.

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to Audio Analysis A MATLAB Approach* (1st ed.), Academic Press.

Han, K.P., Park, Y.S., Jeon, S.G., Lee, G.C., & Ha, Y.H. (1998, Feb 1). Genre classification system of TV sound signals based on a spectrogram analysis. *IEEE Transactions on Consumer Electronics*, 44(1): 33-42. Doi: 10.1109/30.663728.

Lavner, Y. and Ruinskiy, D. (2009, Jan). *A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation*, EURASIP Journal on Audio, Speech, and Music Processing. Doi:110.1155/2009/239892.

Loughran, R., Walker, J., O'Neil, M., & O' Farrell, M. (2000).The use of Mel-frequency Cepstral Coefficients in Musical Instrument Identification, *International Computer Music Association.* Doi: 10.1.1.331.2898.

MATLAB (2015). *Statistics and Machine Learning Toolbox* (1st ed.), User's guide.

Marsland, S. (2009). *Machine Learning: An Algorithmic Perspective*, Chapman & Hall.

Marques, J., & Moreno, P. J. (1999, Jun). A study of musical instrument recognition using Gaussian Mixture Models and Support Vector Machines, Cambridge Research Laboratory, Technical Report Series.

National Instruments (2015). Understanding FFTs and Windowing. Retireved from: http://download.ni.com/evaluation/pxi/Understanding%20FFTs%20and%20Windowing.pdf

Pons, J., Slizovskaia, O., Gong, R., Gomez & E., Serra, X. (2017, Jun). Timbre Analysis of Music Audio Signals with Convolutional Neural Networks. 25th European Signal Processing Conference (EUSIPCO). Kos island, Greece.

Saunders, J. (1996, May). Real-time discrimination of broadcast speech/music. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP '96), (Vol. 2, pp. 993-996). Atlanta, Ga, USA.

Shlens, J. (2003). A tutorial on Principal Component Analysis: Derivation, Discussion and Singular Value Decomposition, Version 1. Retrieved from: http://www.snl.salk.edu/~shlens/notes.html

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch, *Journal of the Acoustical Society of America*, 8(3): 185-190. Doi:10.1121/1.1915893.

Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-Based Classification, Search and Retrieval of Audio, *IEEE Computer Society*, 3(3): 27-36. Doi: 10.1109/93.556537.

DETERMINING THE WEIGHTS OF CRITERIA IN MENU EVALUATION USING BEST-WORST METHOD

Abstract: *The evaluation of dishes represents basic activity in the structuring of the menu, which allows the optimal use of resources in order to fully satisfy the expectations of users and restaurant management. Models for menu analysis allow systematic evaluation, by comparing individual dishes according to previously selected criteria. This paper presents a new approach for determining the weights of criteria in menu evaluation using Best-Worst Method (BWM) which objectivize the inconsistencies in expert judgment. The model was successfully tested on the menu of a collective nutrition restaurant, where nine relevant experts have influenced on forming the weight coefficients of three groups of criteria.*

Keywords: *menu evaluation, collective nutrition, restaurant management, BWM, MCDM, expert judgment*

1. INTRODUCTION

A high-quality menu allows optimal utilization of capacity and resources of a restaurant in order to fully satisfy the expectations of users and restaurant management. The menu represents a worthwhile synthesis of needs of the target group of users, on one hand, and the ability of a restaurant to prepare a dish in a cost-effective way according to the defined standards, on the other. The menu is structured based on the production capacities of a restaurant (technical, technological and organizational), available resources and preferences of restaurant service users.

In the process of menu optimization, the evaluation of dishes constitutes a basic activity by which the dishes with lesser performance and a smaller contribution to the set goals, are innovated or substituted with better ones. All business decisions of restaurants related to production and placement are derived from the menu. (Taylor, J., Brown, D., 2007). The menu interprets to the guest restaurant's offer, kindness, and a type of service and influences the creation of a unique experience in dining (McCall, M., Lyn., A., 2008). For this reason, researchers in many studies are working to create a model for optimizing menus to increase efficiency, customer satisfaction and profit (Taylor, J., Reynolds, D., Brown, D., 2009).

An early attempt to menu analysis, as first introduced by Miller (1980), employed a four quadrant matrix with vectors associated with sales and popularity, measured as sales velocities. Kasawana and Smith (1982), later, using the Boston Consulting Group Portfolio Analysis as the basis for the Menu Engineering Matrix approach, incorporated contribution margin defined as the difference between the sales price of an item and the cost of food product to produce that item. They considered high gross profit and low food cost as mutually exclusive. Pavesic (1983), used wighted gross profit/contribution margin to replace the individual menu item gross profit and included “popularity” as an indirect third variable. Hayes and Huffman (1985) developed an individual profit and loss statement for all menu components in an attempt to alloacate all costs including labor and fixed costs to individualy menu items. Bayou and Bennet (1992) developed a profitability analysis model to evaluate the financial strength of menu items in an attempt to allocate variable costs such as labor. Horton (2001) proposed different approach and modified Kasawana and Smith’s Menu Engineering Model by including estimated labor into the contribution margin (gross profit).
Tom and Annaraud (2017) applied fuzzy multi-criteria decision making techniques to Kasavana and Smith model in order to reduce inaccuracies in the evaluation of alternatives presented by linguistic expressions and providing more relevant information to the decision makers.
Previous researchs did not adventage multi criteria decision making methods based on non-matrix access, wich disable aggregation of wide range relevant criteria for evaluation menu items, both quantitative and qualitative, in a peculiar comparable figure. A new approach, presented in this paper exceeds specified limits and allows for more accurate menu evaluation and an optimization of menu assortiment by elimination or substitution of weak ranked menu items with a new better ones.

In this study, the authors have chosen to apply BWM for determining the weight coefficients of criteria due to following advantages (Rezaei, 2015): (1) A smaller number of comparisons in pairs, e.g. The Analytical Hierarchy Process (AHP) method (Satty, 1980) requires n (n-1) / 2 comparisons, while BWM requires 2n-3 comparisons; (2) The weight coefficients obtained by using BWM are more reliable - comparisons are made with a higher degree of consistency; (3) In most models for multi-criteria decision-making (MCDM), e.g. AHP method, the degree of consistency represents a check whether the comparison of criteria is consistent or not; in BWM degree of consistency is used to determine the level of reliability since BWM outputs are always consistent; (4) When comparing in criterion pairs BWM uses only integer values, unlike AHP which also requires the use of fractional values.

2. BEST-WORST METHOD

This section shows BWM algorithm that includes the following steps:

Step 1. Identification of the evaluation criteria set $C=\{c_1,c_2,...c_n\}$, where n represents the total number of criteria.

Step 2. Identification of a single criterion with the most dominant and most inferior impact provided that if there are two or more criteria of the same importance only one is arbitrarily chosen.

Step 3. Determining the dominance of the most important criteria from the set C in relation to other criteria of the same set, this is measured on the scale of numbers 1-9. The measurement result is represented by vector "best in relation to others" (BO):

$$A_B=(a_{B1},a_{B2},...,a_{Bn}) \tag{1}$$

where a_{Bj} represents the advantage of the most dominant criterion B in relation to criterion j, where $a_{BB}=1$.

Step 4. Determining the dominance of all the criteria from the set C in relation to the most inferior criterion of the set, expressed by a number on scale 1-9. The result of measurement is represented by vector "others compared to the worst" (OW):

$$A_W=(a_{1W},a_{2W},...,a_{nW}) \tag{2}$$

where a_{jW} represents the dominance of criterion j in relation to the worst criterion W, where $a_{WW}=1$.

Step 5. Calculation of optimal values of weight coefficients of the criteria from set C, $(w_1^*,w_2^*,...,w_n^*)$, whereby the condition should be satisfied that maximum absolute values of differences (3)

$$\left|\frac{w_B}{w_j}-a_{Bj}\right| and \left|\frac{w_j}{w_w}-a_{jW}\right| \tag{3}$$

for all values of j be minimized. This condition can be represented by the following minimax model:

$$\begin{aligned}&\min\max_j\left\{\left|\frac{w_B}{w_j}-a_{Bj}\right|,\left|\frac{w_j}{w_w}-a_{jW}\right|\right\}\\&s.t.\\&\sum_{j=1}^{n}w_j=1\\&w_j\geq 0\ \ \forall j\end{aligned} \tag{4}$$

The previous model (4) can be represented by an equivalent model in the following way:

$$\begin{aligned}&\min\xi\\&s.t.\\&\left|\frac{w_B}{w_j}-a_{Bj}\right|\leq\xi,\forall j\\&\left|\frac{w_j}{w_w}-a_{jW}\right|\leq\xi,\forall j\\&\sum_{j=1}^{n}w_j=1\\&w_j\geq 0\ \ \forall j\end{aligned} \tag{5}$$

By solving the system of equations and inequations of the model (5), optimal values of the evaluation weight coefficients $(w_1^*,w_2^*,...,w_n^*)$ and ξ^* are obtained.

For each value $a_{BW}\in\{1,2,...,9\}$ the values of consistency index are calculated $CI(\max\xi)$, (Rezaei, 2015), Table 1

Table 1: Consistency index values

a_{BW}	1	2	3	4	5	6	7	8	9
$CI(\max\xi)$	0.00	0.44	1.00	1.63	2.30	3.00	3.73	4.47	5.23

By solving the system of equations and inequations of the model (13), we get the optimal values of evaluation weight coefficients $(w_1^*, w_2^*, ..., w_n^*)$ and ξ^*.

It is considered that the values of evaluation weight coefficients are reliable if the condition represented by the expression (6), (Rezaei, 2015) is satisfied,

$$CR = \frac{\xi^*}{CI} \le 0,25 \tag{6}$$

$$CR \in [0,1]$$

where CR is the degree of consistency. From the expression (6) it can be noticed that as the value ξ^* increases, the value of CR increases, that is, the reliability of comparison results of the defined criteria by experts is decreased. If the condition represented by expression (6) is not satisfied, the optimal weight coefficients of criteria are calculated in the form of interval numbers by solving the model (7)

$$\begin{array}{ll}
\min w_j & \max w_j \\
s.t. & s.t. \\
\left|\frac{w_B}{w_j} - a_{Bj}\right| \le \xi^*, \forall j & \left|\frac{w_B}{w_j} - a_{Bj}\right| \le \xi^*, \forall j \\
\left|\frac{w_j}{w_w} - a_{jW}\right| \le \xi^*, \forall j & \left|\frac{w_j}{w_w} - a_{jW}\right| \le \xi^*, \forall j \\
\sum_{j=1}^{n} w_j = 1 & \sum_{j=1}^{n} w_j = 1 \\
w_j \ge 0 \ \ \forall j & w_j \ge 0 \ \ \forall j
\end{array} \tag{7}$$

For each interval value, the center of the interval is determined, which is used to rank the criteria of alternatives (Reazei, 2016).

3. DETERMINING THE WEIGHTS OF CRITERIA USING BWM

In the first step, based on the expert experience, nine relevant criteria for menu evaluation have been identified: Time required for preparation, Technical-technological and organizational requirements (TTOR) for storage of meal components, TTOR for meal preparation, Price, Energy value, Digestibility, Sensory properties, Elan for work and Possibility of preparation in unforeseen circumstances, for which the labels and explanations are given in Table 2

Table 2: Evaluation Criteria

Criterion label	Criterion name	Explanation
C1	Time needed for preparation	The duration of technological process
C2	TTOR storage of meal components	Space, appliances and storage facilities
C3	TTOR for meal preparation	Space, machines, appliances, accessories, recipes, qualifications of people
C4	Price	Cost of fresh foods
C5	Energy value	Caloric value of a meal
C6	Digestibility	A subjective feeling in the body after consuming a meal
C7	Sensory properties	Appearance, smell, taste, texture
C8	Elan for work	Mental and physical work after the meal
C9	The possibility of preparation in unforeseen circumstances	In case of TTOR dysfunctionality

In the second step, BWM (Rezaei, 2015) was applied and a comparison of the best criterion with the others was made using the nine-degree scale [1.9], where 1 is the same significance, and 9 is a distinct dominance. The survey included nine reference experts with a minimum of ten years experience who carried out a comparison and the obtained results were presented with nine BO vectors, (Table 3)

Table 3: BO vector of compared criteria

	Criterion	E1	E2	E3	E4	E5	E6	E7	E8	E9	GM	Rank
BO	C1	4	6	7	4	8	9	4	5	5	5,53	8
	C2	2	2	8	8	7	8	1	9	9	4,72	6
	C3	1	1	5	7	2	2	2	6	8	2,88	2
	C4	3	7	3	9	3	6	3	8	7	4,93	7
	C5	7	3	1	2	1	5	7	1	1	2,25	1 (B)
	C6	8	4	4	6	4	3	9	2	3	4,30	4
	C7	5	5	2	3	9	7	6	3	6	4,65	5
	C8	9	8	9	1	5	1	8	4	2	3,90	3
	C9	6	9	6	5	6	4	5	7	4	5,60	9 (W)

In the next step, the experts compared the worst criteria with the others and the results were presented with nine OW vectors, (Table 4)

Table 4: OW vector of compared criteria

	Criterion	E1	E2	E3	E4	E5	E6	E7	E8	E9	GM	Rank
OW	C1	6	4	3	6	2	1	6	5	5	3,70	2
	C2	8	8	2	2	3	2	9	1	1	2,88	4
	C3	9	9	5	3	8	8	8	4	2	5,55	8
	C4	7	3	7	1	7	4	7	2	3	3,82	3
	C5	3	7	9	8	9	5	3	9	9	6,34	9 (B)
	C6	2	6	6	4	6	7	1	8	7	4,44	6
	C7	5	5	8	7	1	3	4	7	4	4,27	5
	C8	1	2	1	9	5	9	2	6	8	3,49	7
	C9	4	1	4	5	4	6	5	3	6	3,82	1 (W)

The values of BO and OW vectors were aggregated using the expression for geometric mean (GM) calculation, and then they were assigned the ranks which were used to form the model (5). Thus, for BO vector in table 3 for criterion C1, averaging was carried out as follows

$GM_1 = \sqrt[9]{E_1 \cdot E_2 \cdot ,..., \cdot E_9} = \sqrt[9]{6 \cdot 4 \cdot 3 \cdot 6 \cdot 2 \cdot 1 \cdot 6 \cdot 5 \cdot 5} = 3.70$

In the same way, the ranks of remaining criteria in Tables 3 and 4 were obtained. Based on the acquired values of criteria ranks, the model was set (5)

$$
\begin{gathered}
\min w_j \\
s.t. \\
\left\{
\begin{array}{l}
\left| \frac{w_5}{w_1} - 8 \right| \le \xi \;; \\
\left| \frac{w_5}{w_2} - 6 \right| \le \xi \;; \\
\left| \frac{w_5}{w_3} - 2 \right| \le \xi \;; \quad \left| \frac{w_5}{w_4} - 7 \right| \le \xi \;; \\
\left| \frac{w_5}{w_6} - 4 \right| \le \xi \;; \\
\left| \frac{w_5}{w_7} - 5 \right| \le \xi \;; \quad \left| \frac{w_5}{w_8} - 3 \right| \le \xi \;; \\
\left| \frac{w_5}{w_9} - 9 \right| \le \xi \;; \\
\sum_{j=1}^{9} w_j = 1 \\
w_j \ge 0, \quad \forall j = 1,2,...,9
\end{array}
\right.
\qquad
\left\{
\begin{array}{l}
\left| \frac{w_1}{w_9} - 2 \right| \le \xi \;; \\
\left| \frac{w_2}{w_9} - 4 \right| \le \xi \;; \\
\left| \frac{w_3}{w_9} - 8 \right| \le \xi \;; \quad \left| \frac{w_4}{w_9} - 3 \right| \le \xi \;; \\
\left| \frac{w_6}{w_9} - 6 \right| \le \xi \;; \quad \left| \frac{w_7}{w_9} - 5 \right| \le \xi \;; \\
\left| \frac{w_8}{w_9} - 7 \right| \le \xi \;; \\
\sum_{j=1}^{9} w_j = 1 \\
w_j \ge 0, \quad \forall j = 1,2,...,9
\end{array}
\right.
\end{gathered}
$$

The model shown is solved using the Lingo 17.0 software. By solving this model, the final values of weight coefficients were acquired

$$w_1 = 0.03687;$$
$$w_2 = 0.06155;$$
$$w_3 = 0.19627;$$
$$w_4 = 0.04988;$$
$$w_5 = 0.26313;$$
$$w_6 = 0.11567;$$
$$w_7 = 0.08035;$$
$$w_8 = 0.17174;$$
$$w_9 = 0.02453.$$

Using the expression (6), the value of consistency degree (CR) is calculated,

$$CR = \frac{\xi^*}{CI} = \frac{1,725083}{5,23} = 0.3298$$

Since the minimum consistency condition is not satisfied ($CR > 0.25$), the optimal weight coefficients of criteria are calculated in the form of interval values by solving the model (7). In the following section a model for obtaining the interval values of weight coefficient of the first criterion is shown.

$$\min w_1$$
$$s.t.$$

$$\begin{cases} \left|\frac{w_5}{w_1}-8\right| \le \xi^*; \\ \left|\frac{w_5}{w_2}-6\right| \le \xi^*; \\ \left|\frac{w_5}{w_3}-2\right| \le \xi^*; \quad \left|\frac{w_5}{w_4}-7\right| \le \xi^*; \\ \left|\frac{w_5}{w_6}-4\right| \le \xi^*; \\ \left|\frac{w_5}{w_7}-5\right| \le \xi^*; \quad \left|\frac{w_5}{w_8}-3\right| \le \xi^*; \\ \left|\frac{w_5}{w_9}-9\right| \le \xi^*; \\ \sum_{j=1}^{9} w_j = 1 \\ w_j \ge 0, \quad \forall j = 1,2,...,9 \end{cases}$$

$$\begin{cases} \left|\frac{w_1}{w_9}-2\right| \le \xi^*; \\ \left|\frac{w_2}{w_9}-4\right| \le \xi^*; \\ \left|\frac{w_3}{w_9}-8\right| \le \xi^*; \quad \left|\frac{w_4}{w_9}-3\right| \le \xi^*; \\ \left|\frac{w_6}{w_9}-6\right| \le \xi^*; \quad \left|\frac{w_7}{w_9}-5\right| \le \xi^*; \\ \left|\frac{w_8}{w_9}-7\right| \le \xi^*; \\ \sum_{j=1}^{9} w_j = 1 \\ w_j \ge 0, \quad \forall j = 1,2,...,9 \end{cases}$$

$$\max w_1$$
$$s.t.$$

$$\begin{cases} \left|\frac{w_5}{w_1}-8\right| \le \xi^*; \\ \left|\frac{w_5}{w_2}-6\right| \le \xi^*; \\ \left|\frac{w_5}{w_3}-2\right| \le \xi^*; \quad \left|\frac{w_5}{w_4}-7\right| \le \xi^*; \\ \left|\frac{w_5}{w_6}-4\right| \le \xi^*; \\ \left|\frac{w_5}{w_7}-5\right| \le \xi^*; \quad \left|\frac{w_5}{w_8}-3\right| \le \xi^*; \\ \left|\frac{w_5}{w_9}-9\right| \le \xi^*; \\ \sum_{j=1}^{9} w_j = 1 \\ w_j \ge 0, \quad \forall j = 1,2,...,9 \end{cases}$$

$$\begin{cases} \left|\frac{w_1}{w_9}-2\right| \le \xi^*; \\ \left|\frac{w_2}{w_9}-4\right| \le \xi^*; \\ \left|\frac{w_3}{w_9}-8\right| \le \xi^*; \quad \left|\frac{w_4}{w_9}-3\right| \le \xi^*; \\ \left|\frac{w_6}{w_9}-6\right| \le \xi^*; \quad \left|\frac{w_7}{w_9}-5\right| \le \xi^*; \\ \left|\frac{w_8}{w_9}-7\right| \le \xi^*; \\ \sum_{j=1}^{9} w_j = 1 \\ w_j \ge 0, \quad \forall j = 1,2,...,9 \end{cases}$$

In the same way, non-linear mathematical models for the remaining criteria (C2-C9) are constructed, with limitations. By solving these limitations interval values of weight coefficients of the remaining criteria are acquired. The obtained intervals, the lower boundary (LB) and the higher boundary (HB), are shown in Table 5

Table 5: Boundary values of weight coefficient intervals

ww_j	LB	HB
w_1	0.02535418	0.04736751
w_2	0.05185670	0.07025333
w_3	0.15435290	0.24982500
w_4	0.02941330	0.05611102
w_5	0.24318140	0.30230740
w_6	0.09790565	0.13126080
w_7	0.07425574	0.09230993
w_8	0.12876410	0.21785330
w_9	0.02267408	0.02818695

Since evaluation of menu dishes (alternative) could be done in the next stage using the selected MCDM model, the obtained interval values of weight coefficients will be considered as interval or rough numbers.

The concept of rough numbers introduced by Zhai (Zhai et al., 2008) was derived from the theory of rough (Pawlak, 1982).

4. CONCLUSION

The evaluation of menu is based on an objective and precise understanding of relevant factors that affect the satisfaction of users and restaurant employees.

User satisfaction is reflected in sensory perception and subjective feeling during and after the meal, while employee satisfaction reflects the synthesis of various impressions such as personal income, working conditions, working atmosphere and other feedback effects that imply user satisfaction. In a stimulating working environment, users and employees in the restaurant stimulate each other by raising the aspiration level over a longer period of time (Arsic, S., 2014).

BWM based approach allows the elimination of uncertainty of subjective assessment about importance of criteria of experts (Pamucar, D., et al., 2018) in various areas related to nutrition (nutrition, food technology, nutrition organization, quality of life) and precise positioning of dishes according to value rank which was acquired by evaluation. During the testing of the introduced approach 9 relevant experts have influenced on forming the weight coefficients of three groups of criteria: 1.group of criteria related to the subscribers (C6-C8) was evaluated by 35 users; 2.group of criteria related to food preparation (C1-C3 and C9) was evaluated by 15 experts involved in the food preparation process; 3.group consists of criteria related to the price of fresh foods required for food preparation (C4) acquired from market analysis and the caloric value of a ready meal (C5) measured in a renowned state institution.

The model is adequate for analysis and predictions in the following time period for the purpose of quality business decision-making by top management (Suknović, M., et al., 2012). The approach can be very effectively applied when it is necessary to form a menu based on the defined effects that nutrition should manifest on the target group of subscribers (favoring the quantity of nutrients and their digestibility, elan for physical work after the meal, preparation in unforeseen circumstances, etc.) especially in restaurants for collective nutrition of students, athletes, security forces and rekonvalescents, so future research should be aimed in that direction. In educational terms, the approach helps decision makers to better understand the complexity of process of identifying relevant criteria, dish evaluation and creating an optimal menu in given time, space, social, economic and other circumstances.

REFERENCES

Arsic, S. (2014). Possibilities for improving the food system at the Military Academy - the economic aspect. Military Technical Courier, 4, 168-186.

Bayou, M.E., Bennett, L.B. (1992). Profitability analysis for table-service restaurants. Cornell Hotel and Restaurant Administration Quarterly, 33(2), 49-45.

Hayes, D.K., Huffman, L. (1985). Menu analysis: a better way. Cornell Hospitality Quarterly, 25(4), 64-70.

Horton, B.W. (2001). The effect of labor and menu category on menu classifications. Hospitality Review, 19(2), 35-46.

Kasawana, M.L., Smith, D.J. (1982). Menu enguneering. Lansing, MI: Hospitality Publications Inc.

McCall, M., Lyn., A. (2008). The effects of restaurant menu item descriptions on perceptions of quality, price, and purchase intention. Journal of food service Business Research, 11(4), 439-445.

Miller, J.E. (1980). Menu Pricing and Strategy. Boston: CBI Publishing.

Pamucar, D., Petrovic I., Cirovic G. (2018). Modification of the Best-Worst and MABAC methods: A novel approach based on interval-valued fuzzy-rough numbers. Expert systems with applications, 91, 89-106.

Pavesic, D. (1983). Cost-margin analysis: a third approach to menu pricing and design. International Journal of Hospitality Management, 2(3), 127-134.

Pawlak, Z. (1982). Rough Sets. International Journal of Computer and Information Sciences, 11(5), 341-356.

Reazei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. Omega, 64, 126-130.

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. Omega, 53, 49-57.

Satty, T. (1980). The Analytical Hierarchy Process. New York: McGraw-Hill.

Suknović, M., Delibasic, B., Jovanovic, M., Vulicevic, M. (2012). Reusable components in decision tree induction algorithms. Comput Stat, 127–148.

Taylor, J., Brown, D. (2007). Menu analysis: a review of techniques and approaches. Hospitality Review, 25(2), 74-82.

Taylor, J., Reynolds, D., Brown,D. (2009). Multi-factor menu analysis using data envelopment analysis. International Journal of Contemporary Hospitality Management, 20(4), in press.

Tom, M., Annaraud, K. (2017). A fuzzy multi-criteria decision making model for menu engineering. 2017 IEEE International Conference. Naples, Italy.

Zhai, L.Y., Khoo, L.P., Zhong, Z.W. (2008). A rough set enhanced fuzzy approach to quality function deployment. International Journal of Advanced Manufacturing Technology, 37(5-6), 613-624.

ANALYSIS AND PREDICTION OF VIEWS IN YOUTUBE INTERVIEWS

Stefan Vujović[1*], Danijel Mišulić[1], Sofija Krneta[1]
[1]University of Belgrade, Faculty of Organizational Sciences
*Corresponding author, e-mail: stefanvujovic93@gmail.com

Abstract: *YouTube videos carry great number of data about the viewers habits, topics of interest and attention catchers. Several YouTube channels in Serbia stream specialized content in form of interviews. By analyzing channel content, it can be seen that there are some impacts that lead to popularity of a certain video. This paper is written in order to show correlation between video guests popularity on Wikipedia and view count on YouTube, as well as a connection between the words used in the title of the video and view count.*

Keywords*: YouTube, Wikipedia, interviews, correlation, popularity, analysis*

1. INTRODUCTION

In the last few years, YouTube as a video streaming service gained more television channel formed content. According to YouTube fact sheet, there are over 300 hours of content being uploaded every minute, many of them being interviews, weekly shows that are only available on YouTube. There are many channels in Serbia with specialized content, such as one to one interviews. These videos carry a large number of data within which gives the opportunity to understand content popularity, trends in online marketing and change of habits of the targeted audience.

Analyzing the content of Serbia's most popular YouTube interview focused channel, the purpose of this paper is to answer following questions: Is the view count influenced by the guest interview and/or the title of the video? Are there any "trigger" words that influence the popularity of a video and are there any trends in selection of guests? The questions will be answered through detailed analysis of statistics collected from YouTube API, information about guests obtained from Wikipedia and by exploring information about guests in DBpedia.

First part of the analysis will cover pure statistics and try to mine the correlations between popularity of the video on the characteristics of the guest, such as guest's gender and profession. Second part of the analysis will focus on the topic of the interview, keywords that are mentioned in the title and to correlate it with the video's popularity. Finally, information about the guests, the topic and the popularity of the video will be put into clustering algorithm in order to generate clusters of guests that might implicate the popularity of future videos.

The main goal of the analysis should lead to concrete advice to YouTube channel owners on how to improve their statistics and popularity.

2. RELATED WORK

Several papers in the last 5 years have written about impact on popularity of YouTube videos. Mekouar, Zrura and Bouyakhf (2017) had used two simple regression models for applying machine learning in predicting the popularity of videos based on videos parameters and they proposed a popularity function with good performance over the tested set of videos. Shuxina, Chenyu and Xueming (2017) applied more in depth techniques for analysis of similar video service provider, Youku, where they defined popularity patterns based on initial popularity of the video. Similar approach was taken in Pinto, Almeida and Goncalves (2013) where authors created models to advise users on actions needed for reaching a greater number of views. Initially, idea of defining popularity patterns of videos over time was used in Figueriedo, Benevenuto and Almeida (2011) for the first time. Slightly different approach was taken in Brodersen Scellato and Wattenhofer (2012) by analyzing content of the video, its geographical origin and geographical popularity.

Slightly different goal, but the same means was used recently in Woo, BKP, & Chung, JOP (2018). The authors used YouTube statistics to create a model for evaluation of an educational video. In Tackett et al. (2018) YouTube statistics was used for predicting viewing pattern for popular medical YouTube videos.

Over the years, targeted audiences had grew, but also had changed. The need to have personal relationship with a viewer, to target specific needs resulted in appearance of a great number of vlogs - video blogging channels, interview specialized channels and similar content. It is now more than the view count in preceding days that impacts the popularity of the video. In Cheng, Dale and Liu (2008) similar statements from the overall YouTube statistics were taken out, but even though the approach to analyzing content to find the connections between videos and popular content is great idea, over the 10 years period the statistics had changed, and so did the statements concluded. 10 years later, in Bärtl, M (2018), decent analysis of YouTube videos was done, but the paper was solely focused on analysis, without creating any conclusions about the connection between popularity and the content.

3. DATA EXPLORATION

3.1. Data structure

The initial data was obtained by using YouTube API from Balkan Info channel. Titles of videos on this channel are well structured and it was easy to collect guest names from the interviews. This way, it was possible to collect the main characteristics of a published video clip such as video title, number of views, number of comments, number of likes and other.

Some columns of the dataset obtained in this way are the following: title, published, tags, comments, likes, dislikes, favourites, views, guest_name, quote and page_name. The gender of interviewed guests was extracted using an automated approach combining name-based and image-based gender inference methods, as done in Karimi, Wagner, Lemmerich, Jadidi, & Strohmaier (2016).

The existing data set was then expanded with data from the Wikipedia api. For each guest, a number of reviews of their wiki pages have been collected, if their page exists, for 2018, 2017. and 2016. from Serbian wikipedia and English as well. It was necessary to collect this data to find out more about the popularity of the guest and thus compare it with the number of views of his clip. In order to get more information about guests, their professions were collected for a later analysis of the impact of what they do on number of views of their videos. Eleven different profession groupings were identified, and data were obtained from dbpedia, writing Sparql queries. The collection of the entire data set is done using the Python programming language due to easy manipulation of the data and corresponding libraries that it offers, and all data values are saved in the CSV (comma separated file) after collection. Data was collected for every published video clip in the period from July 15, 2015. to March 6, 2018.

3.2. Guest statistics

The number of unique guests is 284 of the total number of interviews collected - 474, and the total number of different data columns is 38. Further analysis of data shows that women appear much less frequently as speakers in the show, only 11 of them were guests, and they appeared in total 24 times in the observed period. Furthermore, it can be argued that their interviews are significantly less frequently reviewed, on average 60 449 times per video clip, and that same average for men is 80 716. On average, male guests average 91 likes more than female per video clip. Clips in which men appear on average have more dislikes and comments, but those differences are much less compared to the number of views. Men have 152 dislikes per clip, and women 139, on average. The number of comments where the guest is a male is an average of 235, and for women 225. It is interesting to visualize some of these data that is available after the initial collection phase. Comparison of male and female guests is shown in the Figure 1.

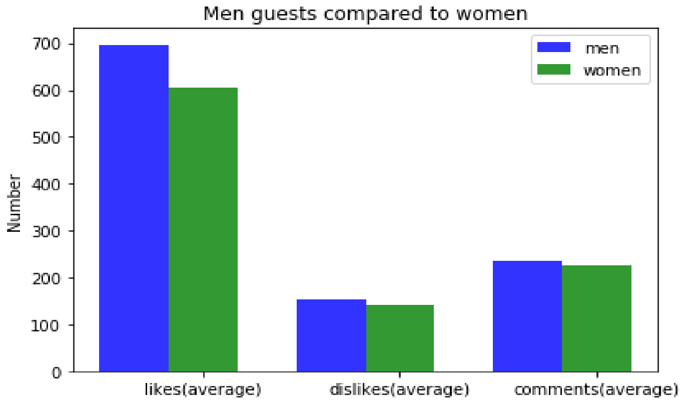


Figure 1: Female and male guests compared

Comparison of average views by gender is shown in the Figure 2.

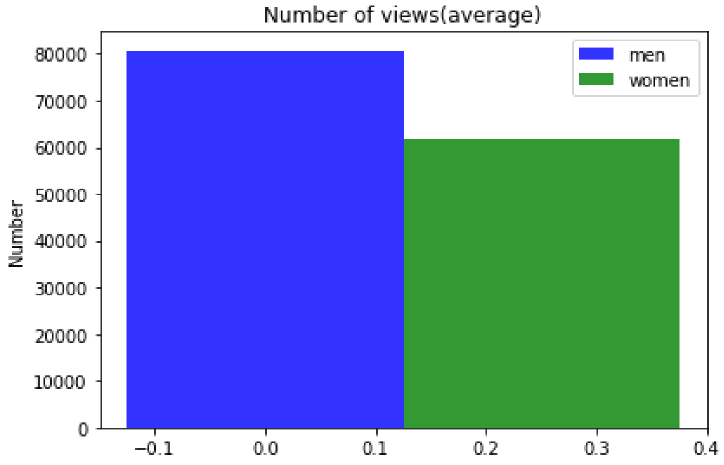


Figure 2: Views compared by gender

As for the professions differentiation of guests, The largest part of guests comes from scientists and politics. Profession with most views per video clip on average is Artist, and Music and Performer professions are positioned high also. Politics is the profession with least number of views on average. It is interesting to note that guests from the sphere of politics are the second most often welcomed to the show despite of the lack of people viewing their clips. On the other side, there is a small number of guests that are artists by profession despite to the fact that their clips are viewed the most.

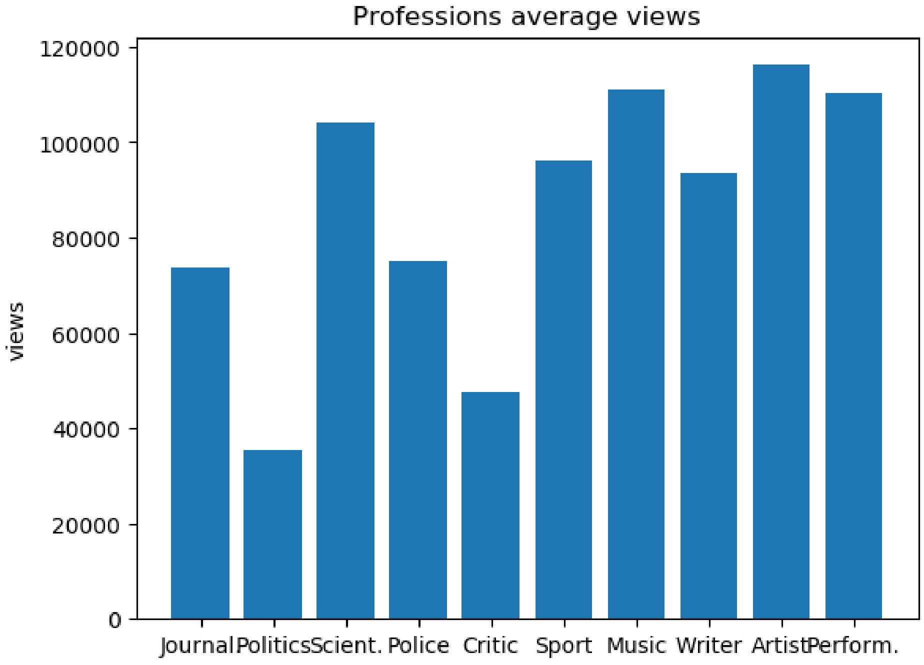


Figure 3: Average views by professions

The guest who is most often invited to the show in the observed period is Jugoslav Petrušić and this number is 14 times. The average number of views of his interviews is 115 054 per clip. In Figure 4 you can see how many views on average have five guests who were most often invited to give interview.

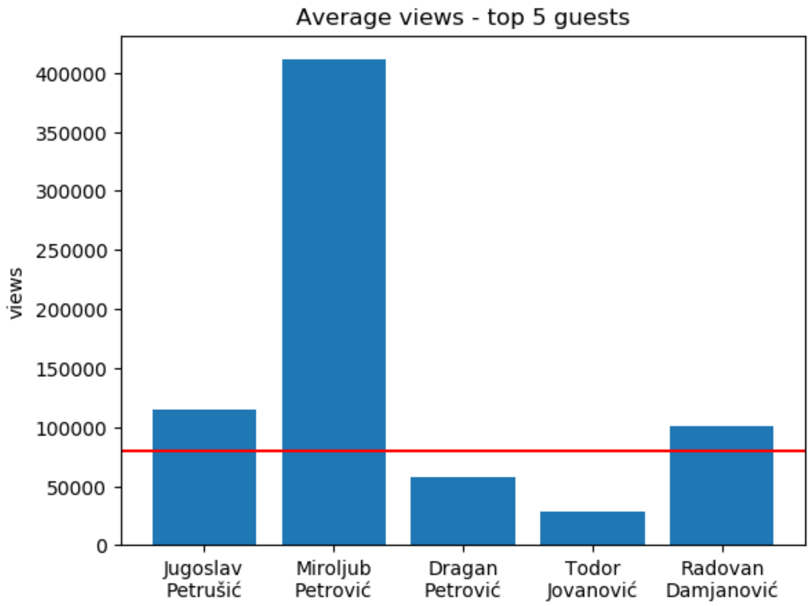


Figure 4: Average views of five most often guests

Red line in Figure 4 represents average views of all video clips collected in the observed period - 79 744. Three of five most frequent guests are above the average views line, and other two are significantly below. A person who certainly justifies a large number of invites to the show is Miroljub Petrovic with an average number of views with over 300 000 above average. This guest also has the most cumulated views - 5 007 144, more than next four guests combined together.

Of the total number of guests, only 89 of them have a page on en.wikipedia in all three observed years 2016, 2017, and 2018. The number of views of these guests interviews is above average - 73 902 which is in contrast to the assumption that guests who have wiki pages are more popular and will have more views on their show appearances. The number of guest pages is considerably smaller when looking at sr.wikipedia where only 19 guests have their own wiki pages in the observed years. There is only 10 different guests that have wiki pages both on en.wikipedia as well as on sr.wikipedia through these 3 years. The average number of views that these guests get is 45 790 per video. Also, guests that are not present in wikipedia pages, do not have any SparQL knots, which made profession based analysis unreliable from this source. That is why our profession based classification of guest was made solely on YouTube data.

In the Figure 5 is shown correlation between guest pages yearly searches on Wikipedia and views of their videos. Guests that do not have wiki page are excluded from consideration. There is a strong correlation between the number of views on Youtube and sr.wikipedia searches and very little between en.wikipedia and video clip views. This is a good indicator that en.wikipedia can be excluded in some future work for our particular case.

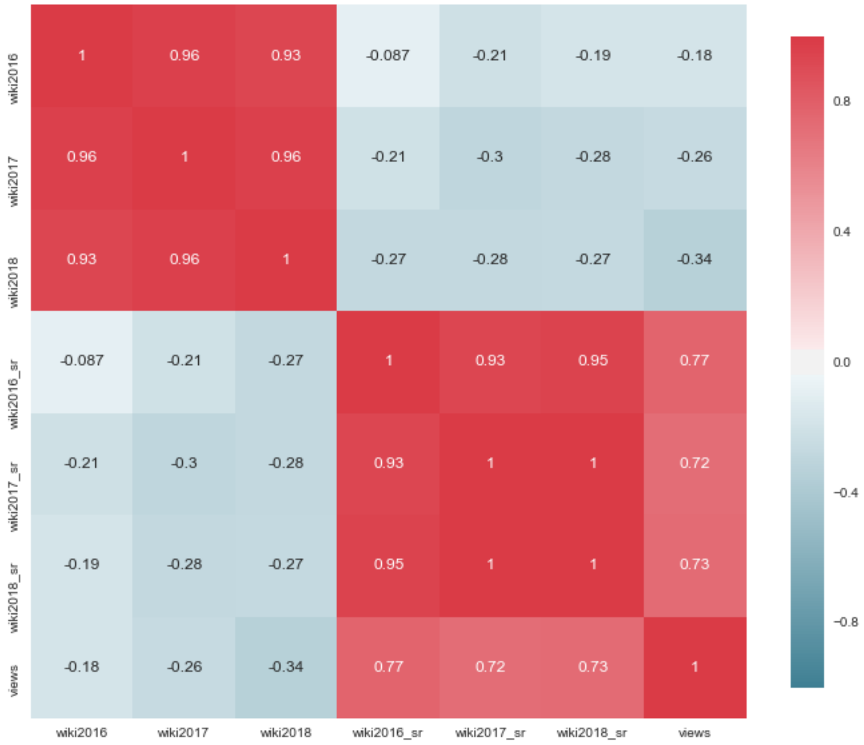


Figure 5: Correlation between views and wiki searches

3.3. Title statistics

Text processing focused on title of videos and tags, related to each of the video. Using text processing libraries in Python and extensions in RapidMiner, a set of most used words and phrases was created. Analysing most popular words, it can be concluded which are the most popular words and phrases in videos (Figure 6 and 7).

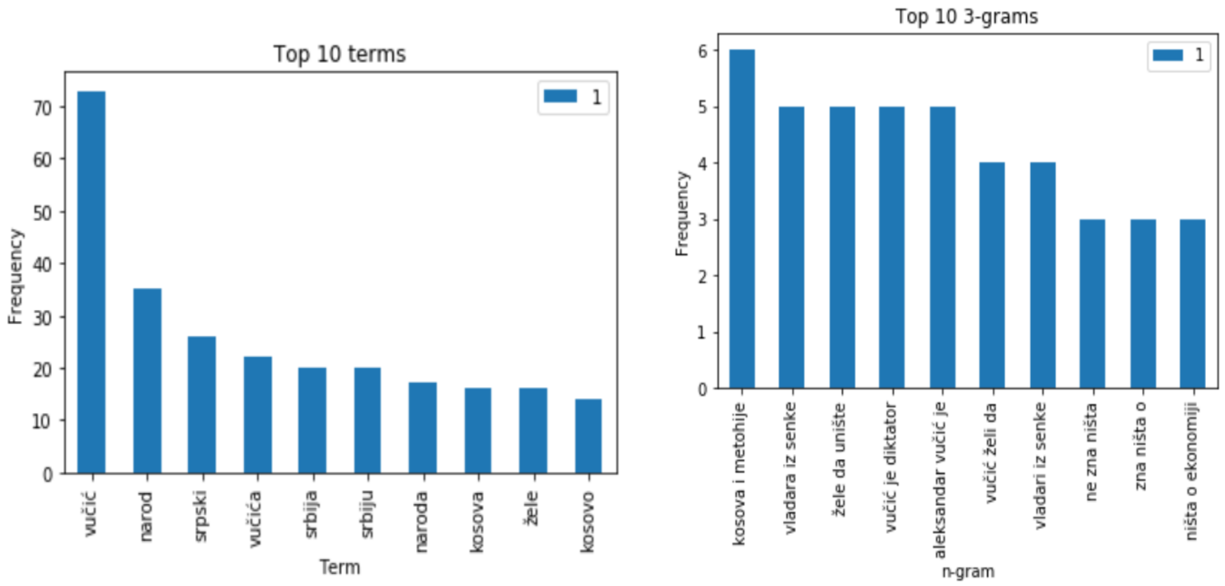


Figure 6: Top 10 terms (left) and Figure 7: Top 10 phrases (right)

Above mentioned words are connected to the popularity of videos, which is shown on the following figure:

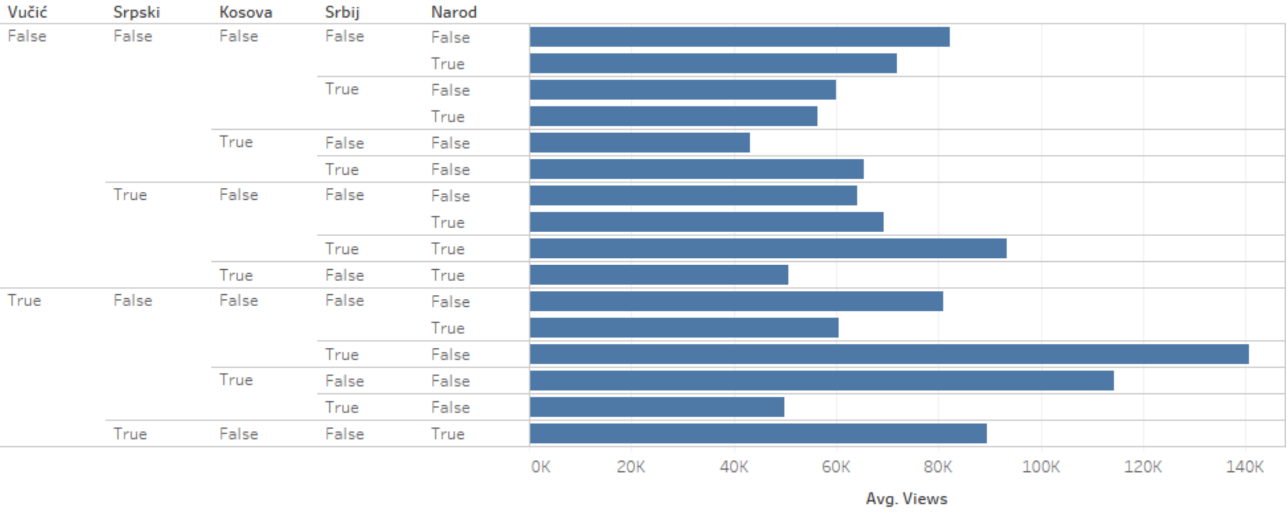


Figure 8: Correlation between views and use of keywords

It can be concluded that use of words "Vučić" and "Srbija" in the video title attract more attention than use of other words. Also, "Vučić" and "Kosovo" have the same influence on the reach of the video.

3.3. Clustering Guests

As claimed in the channel description, "Balkan Info strives to provide unbiased reporting and has opened its doors to all guests and all political parties". As presented above (Figure 3), the interviewed guests come from different professional backgrounds and they discuss about different topics. However, some guests could be considered more 'serious' than others, as some guests tend to share information that could be described as conspiracy theories. An interest application of clustering would be to divide guests into the 'serious' ones and the 'conspiracy theorists'. Having assumed that this could save time for the viewers when making the decision to watch the video or not (as they tend to be between 1 and 3 hours long), an attempt for clustering was made.

As features for clustering, we used the following predictors, out of the earlier mentioned set attributes: comments, likes, dislikes, favourites, views, wiki2016, wiki2017, wiki2018 and the name of the guest. We used k-means algorithm to train the clustering model. However, just around 8 percent of them were assigned to the 'conspiracy' cluster, which we do not consider as a proper distribution. It is worth mentioning that a higher number of dislikes could be considered as a signal of a guest being a 'conspiracy theorist', but there are exceptions where the guest is perceived negatively by the audience even though being 'serious'.

In the following figure, it can be seen that the characteristics of the first cluster is high number of dislikes, while the second cluster has a high number of popular words with a lower number of dislikes.

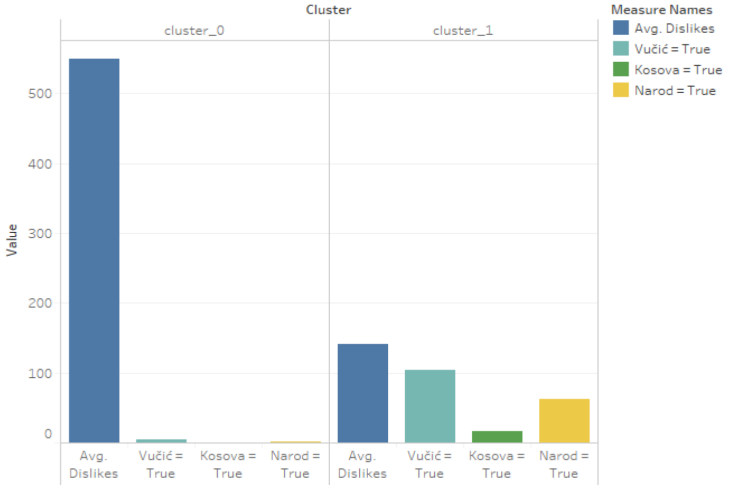


Figure 9: Main characteristics of the clusters

4. CONCLUSIONS AND FUTURE WORK

From the explorative analysis done, several conclusions can be outlined. First, guest from the sphere of politics are not interesting content for viewers. On the other hand, performers and artists are interesting and attractive content. Taking into consideration that people without any information on wikipedia pages are more view than others, lead to conclusion that locally familiar people that do not have the chance to appear on other media do have greater popularity online, on YouTube. Certain guests in Serbia have great popularity and can bring a lot of popularity to the channel if invited.

As for the timing, it can be recommended to publish video clips on Sundays, since that is the day with the largest number of views cumulative.

Topics which need to be covered in the interview need to contain several of the popular words such as: vučić, kosovo, srbija, žele, narod etc. in order to attract viewers and it needs to be mentioned in the title of the video.

We have seen that en.wikipedia gives very small correlation with views of published video and can be excluded from consideration. Sr.wikipedia gives strong correlation and in some future work data searches will be collected not only for the past 3 years, but also for years before 2016.

For the purpose of this research it was only taken into consideration title of the video in the part of text processing. Next step would be to dig deeper into semantics of the whole conversation. Another idea for future work is to include other sources for data collecting, besides dbpedia and wikipedia.

REFERENCES

“BALKAN INFO - Zvanični kanal - YouTube,” 2018. [Online]. Available: https://www.youtube.com/channel/UCLG5Qu54Q7gywaCeDs5etuQ/about. [Accessed: 20-May-2018].

YouTube channels, uploads and views: A statistical analysis of the past 10 years. Convergence, journals.sagepub.com

Brodersen, A., Scellato, S., Wattenhofer, M. (2012) YouTube around the world: geographic popularity of videos. In Proceedings of the 21st international conference on World Wide Web

Cheng, X., Dale, C., Liu, J. (2008) Statistics and Social Network of YouTube videos. 2008 16th International Workshop on Quality of Service

Figueiredo, F., Benevenuto, F., Almeida, JM. (2011) The tube over time: characterizing growth of youtube videos. WSDM ‘11 Proceedings of the fourth ACM international conference on the Web search and data mining, p. 745 - 754

Karimi, F., Wagner, C., Lemmerich, F., Jadidi, M., & Strohmaier, M. (2016). Inferring Gender from Names on the Web: A Comparative Evaluation of Gender Detection Methods, 8–9. *https://doi.org/10.1145/2872518.2889385*

Mekouar, S., Zrira, N., Bouyakhf EH. (2017). Popularity Prediction of Videos in YouTube as Case Study: A Regression Analysis Study. *BDCA'17 Proceedings of the 2nd international Conference on Big Data, Cloud and Applications*

Pinto, H., Almeida, JM., Goncalves, AM. (2013). Using early view patterns to predict the popularity of youtube videos. *WSDM '13 Proceedings of the sixth ACM international conference on Web search and data mining, pages 365-374*

Shuxina, O., Chenyu, L., Xueming, L. (2017) Analyzing the dynamics of online video popularity. *The Journal of China Universities of Posts and Telecommunications Volume 24, Issue 3, June 2017, Pages 58-69*

Tackett, S, Slinn, K, Marshall, T, Gaglani, S(2018). Medical Education Videos for the World: An Analysis of Viewing Patterns for a YouTube Channel, europepmc.org

YouTube fact sheet, http://youtube.com/t/fact-sheet

Woo, BKP, & Chung, JOP (2018). Using YouTube analytics to evaluate a Chinese video-based lecture regarding Parkinson's disease. Journal of Clinical Neuroscience, jocn-journal.com

NEURAL NETWORKS IN MARKET SENTIMENT ANALYSIS FOR AUTOMATED TRADING: THE CASE OF BITCOIN

Matija Milekić*[1], Aleksandar Rakićević[1], Pavle Milošević[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: matija.milekic@outlook.com

Abstract: *Cryptocurrency mining and trading have attracted much attention lately. There are numerous studies and experiments performed to predict the future prices of cryptocurrencies and identify factors that influence the cryptocurrencies' price movement. The aim of this research is to investigate the potential relationship between market sentiment and Bitcoin price movement. The market sentiment analysis is performed on Reddit social network using neural networks. Furthermore, we propose an intelligent system for automated cryptocurrency trading based on neural network outputs. The proposed intelligent trading system is tested over a 4-month period (December 2017 – April 2018), with daily records. The trading simulations are performed multiple times in order to obtain credible results. It is shown that the proposed system achieves a positive average profit that significantly outperforms Bitcoin performance. The results may be seen as a confirmation of the assumption that market psychology is an important factor in the still emerging cryptocurrency market.*

Keywords: *neural networks, automated trading, market sentiment, cryptocurrency, Bitcoin, Reddit.*

1. INTRODUCTION

The price movements of cryptocurrencies have attracted much attention in the year 2017 and 2018 from both practitioners and academia. The problem of prompt estimation whether the prices will go up or down, i.e. should investors buy or sell assets from their portfolio, has been in the spotlight. These kinds of researches have been conducted for a very long time in different markets. Numerous studies have been done in order to predict movements of the stock or foreign exchange markets. These markets are characterized with relatively low volatility compared to the cryptocurrency market where the price can either suddenly go down or rapidly rise by more than 100 percent on a weekly or even daily basis. Being this unpredictable, the factor that has been added to the formation of the prices and has become one of the most dominant is the factor of faith of investors. Bearing this in mind, the aim of this paper was to show how investors' approval and interaction within the social network can affect the future price movements of cryptocurrencies.

Cryptocurrency trading is currently a hot topic. Numerous researches have been conducting experiments to predict the future prices of cryptocurrencies. Brauneis and Mestel (2018) showed that cryptocurrency prices become less predictable as liquidity increases. Gandal, Hamrick, Moore and Oberman (2018) dealt with the problem ofsuspicious trading activities identification, and the estimation of their impact on the price. Gangwal and Longin (2018) examine doscillations of Bitcoin prices based on extreme value theory. Dynamic investment strategies based on the rational expectations bubble model of prices were proposed by Kreuser and Sornette (2018). Detzel, Liu, Strauss, Zhou, and Zhu (2018) documented that Bitcoin returns are predictable by moving the average technical indicator. On the other hand, the influence of the social media on cryptocurrencies' price movements was thoroughly analyzed by Bollen, Mao and Zeng (2011), Garcia, Tessone, Mavrodiev, and Perony (2014) and Garcia and Schweitzer (2015). Their results imposed a new perspective on researchers and practitioners. Further, Kim, Kim, Kim, Im, Kim, Kang and Kim (2016) estimated the fluctuations in the prices and the number of transactions of cryptocurrencies based on user comments in online cryptocurrency communities. Phillips and Gorse (2017) proposed a class of novel online social media indicators to deal with the price prediction problem.

In this paper, we used social media data generated on the Reddit network. The Reddit network has a large number of subscribers who generate an even larger number of posts and comments each day about many subjects. This network has become especially popular among investors as a place for discussing the cryptocurrency market. We saw a potential in analyzing posts and comments from Reddit network to extract trading signals for Bitcoin. Similarly, this was done by Garcia and Schweitzer (2015) who presented a framework to derive general knowledge from social network data. To do this, they combined economic signals related to market growth, trading volume, and the use of Bitcoin as a means of exchange, with social signals (including search volumes), word-of-mouth levels, emotional valence and opinion polarization.

In order to predict trends in the cryptocurrency price movement, we used an artificial neural network. For the purpose of this paper, we used astandard one-layered feed-forward neural network with back-propagation learning algorithm and the inputs that describe the sentiment of the Reddit network and technical indicators of price movement. In fact, the sentiment of the Reddit network is depicted using the daily difference of ups and downs and the number of posts on Reddit. On the other hand, daily Relative Strength Index and Volume-to-Market-Capitalization are technical indicators used in the experiment. The trading signal obtained from neural network is further extracted using another technical indicator, i.e. ZigZag. The experiment is performed on the dataset of daily records collected during the period of 4 months.

The paper is arranged as follows. In Section 2 we provide a detailed overview of the literature. In Section 3 we present the research methodology. Results are presented in Section 4. Finally, Section 5 concludes the paper and gives a brief discussion of future research direction.

2. LITERATURE REVIEW

This section briefly explains the components relevant to this research, including basic terms related to Bitcoin trading, the problem of Bitcoin price predicting and the predictions based on sentiment tracking.

2.1. Bitcoin

In 2008 a mysterious person, whose identity has yet to be discovered, named Satoshi Nakamoto released a paper “A Peer-to-Peer Electronic Cash System”. In addition, he also implemented open source blockchain software ushering in a cryptocurrency by the name of Bitcoin. The blockchain technology, on which the Bitcoin is based, has enabled a flow of money through online payments directly from one side to another without using a third party, in this case a financial institution (Nakamoto, 2008). This may be considered revolutionary, and it has had a great influence on the current financial system.

Since then, numerous cryptocurrencies and blockchain platforms have emerged. They brought a different perspective to financial markets, despite raising many questions and doubts. Numerous debates and discussions sprang up about the validity of cryptocurrencies and whether they should be banned or even criminalized and how the government could regulate them. Even though the rise of cryptocurrencies was followed by many negative effects and frauds, there are also a lot of benefits that they have brought. Some of them are:

- a faster settlement process,
- lower transaction fees,
- decentralization.

Nowadays, there are more than 1600 known cryptocurrencies. Even though they are different in nature and purpose and independently priced on the crypto market, their prices are mostly correlated to the price of Bitcoin. This fact allows the market to be easily influenced by the effects of “the herd” and “big players”. Gandal et al. (2018) analyzed the impact of suspicious trading activity on price manipulations in the Bitcoin market. The herd effect and market bubble creation were the subject covered by Gangwal and Longin (2018) and Kreuser and Sornette (2018).

2.2. Predictability and profitability

If one had invested in Bitcoin back in the 2010 it would have generated an astonishing return of more than 15.000.000% by the end of the year 2017. At the same time, investment in S&P500 stock index would have grown by a rate of 130% over the same period of time. Being this volatile is the main reason why people decide to either invest in Bitcoin or steer clear from it. Without the fundamental information (such as interest payments, book values, dividends etc.) investors are relying solely on price charts when it comes to investment decision making.

Numerous researches have been done to try to predict the Bitcoin price and to make those predictions profitable. Brauneis and Mestel (2018) tried to investigate the predictability and efficiency of several cryptocurrencies using the statistical testing methodology. They concluded the Bitcoin is the least predictable and most efficient in terms of the efficient market hypothesis. Prediction of fluctuations in cryptocurrencies has been researched by Kim et al. (2016) and Phillips and Gorse (2017), who showed great results and provided excellent trading guidelines. However, this topic needs to be further revised in the shortest possible time span because changes are occurring on an hourly/daily/weekly level.

2.3. Market sentiment analysis based on social media data

Using social media, people are generating a huge amount of information on a daily basis. If handled in a proper way, this information may be used to determine and understand the public opinion about a certain topic. One such topic is the price movement of cryptocurrencies. This topic can either be discussed directly by saying that the price is going to rise/fall for a certain reason; or indirectly by mentioning that you are going to invest in the near future, ask around about the cryptocurrency itself or show any kind of interest in it. All of the aforementioned can give us an indication of how the observed price could move in the future, which was underlined by Garcia et al. (2014).

Price predictions for Bitcoin cannot be obtained in an easy way. Contrary to the stock price predictions, which are based on company's business insights, the price of a certain cryptocurrency is formed based on the level of its usage in society, the level of the faith community has built, the mood of the people such as their skepticism, risk awareness and other feelings/emotions. By using these subjective factors, one can try to predict the future price movement like Bollen et al. (2011) did for American stock market using Twitter data.

3. METHODOLOGY

In this section, we introduce the methodology used to construct and test the proposed trading system based on artificial neural networks.

3.1. Trading system

In this paper, we build and test an automated trading system (ATS) for cryptocurrency trading using artificial neural networks as a machine learning component within the system. The proposed trading system is a continuation of the previous work by Rakićević, Milovanović and Aničić (2016), who build a similar ATS for stock market trading. The proposed system consists of two components (Figure 1): the pre-processing component and trading logic component. In pre-processing component there are three modules that work in parallel. Market analysis module gathers cryptocurrency market data and calculates different technical indicators. The second module is for social network data analysis. It crawls data from Reddit service and quantifies it using different indicators. Finally, there is a trading signal extraction module that uses the price data to extract trading signals used as target data for training neural networks. The second part of the system is reserved for trading logic. It consists of the learning part (implemented as a neural network) and a trading algorithm that uses neural network forecasts to decide whether to buy, sell or hold Bitcoin.

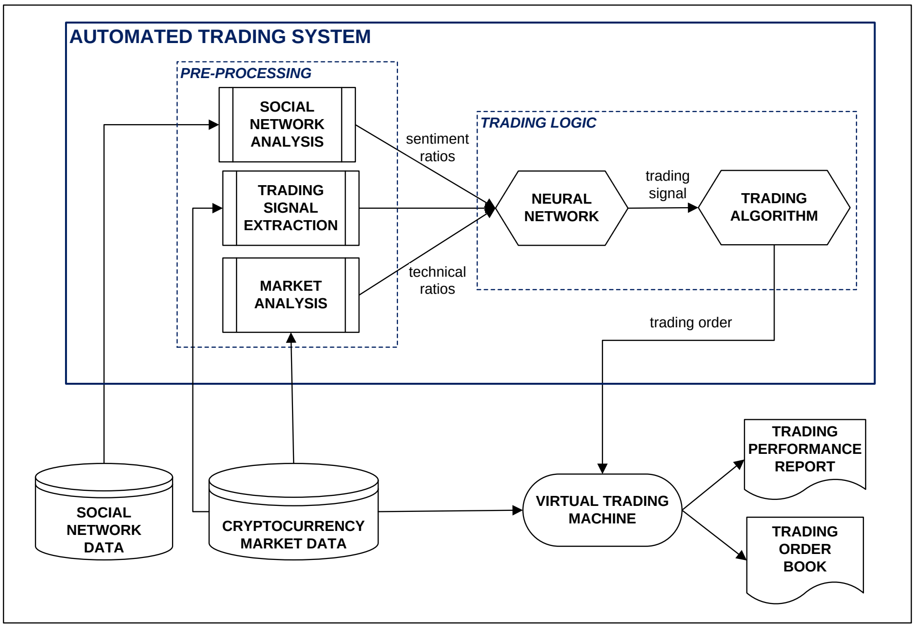


Figure 1: The structure of the proposed ATS

3.2. Artificial neural network

Our system uses a standard feed-forward neural network (FFNN) as a learning component to obtain intelligent system design. In this type of network, each network node aggregates all input variables as a weighted sum and then uses an activation function to obtain nonlinear output from the node. Our network model has three layers:

- input layer with four input variables,
- one hidden layer with 15 nodes,
- an output layer with one output variable.

The proposed FFNN is trained using aback-propagation algorithm. Further, log-sigmoid function is used as an activation function in all network nodes:

$$\sigma(x) = \frac{1}{1+e^{-x}}. \tag{1}$$

When the ATS starts operating, the first time periods of data *t* are used for training FFNN. The neural network uses technical and market sentiment indicators as input variables to learn how to predict the trading signal (extracted using a special pre-processing component). When the underlying model in the data is discovered, trading starts from the *t+1* time period. In this operating segment, the system is using a neural network to predict the trading signal based on input variables. Afterwards, the obtained trading signal prediction is used in a trading algorithm for the decision making process and creating the corresponding trading orders. Trading is done until the stopping criterion is met (the end of data series or another condition). If a stopping criterion is met, the ATS stops trading and starts training the neural network again (from the beginning of dataset until the stopping point). After the learning process is done, trading is resumed again (until the stopping criterion is met again).

3.3. Input variables

Our trading system uses four input variables obtained with two pre-processing modules. The market analysis module is based on technical analysis and uses market data to calculate indicator values. In this paper, we use two indicators: Relative Strength Index (RSI) and Volume-to-Market-Capitalization (Vol2MCap) indicator. The RSI indicator is one of the most common indicators in technical analysis used to analyze trend momentum. It is derived from the market price in the following way:

$$RSI = 100 - \frac{100}{1+RS}, \tag{2}$$

where RS is the ratio of average gain of up periods during the specified time frame to the average loss of down periods during the specified time frame.

The Vol2MCap indicator is used for market volume analysis, to provide us with an easy interpretation whether the current trading volume is high, low or average. We calculate this indicator with the following formula:

$$Vol2MCap = \frac{tradingVolume \cdot marketPrice}{marketCapilatization}. \tag{3}$$

To analyze the market sentiment through social networks, we use two additional indicators. The first one we call Sentiment Score (SS). It is calculated as a simple difference between positive reactions on the Reddit social network (ups) and negative reactions (downs) calculated on a daily basis:

$$SS = ups - downs. \tag{4}$$

The SS indicator should track investors’ feelings and give us an insight into psychology of cryptocurrency market. Another indicator that was used as an indicator of behavior on the social network is the number of comments on a daily basis. This indicator is aimed to detect “herd behavior” in the market. We assume that when the irrational euphoria hits the market, investors are tempted to talk more about markets on the web. These four indicators are input variables that are used by neural network to learn trading.

3.4. Trading signal extraction

In order to learn our system to trade on the market, we use FFNN as a learning mechanism. The output/target variable for completing that task is a trading signal. Trading signal is a value in [0,1] range that can be easily interpreted in the trading action (buy, sell, hold/do nothing). We use the well-known ZigZag technical indicator to extract values from price data.

The ZigZag indicator uses a general definition of a trend, a tendency in the data that creates a positive or a negative change in data level over a certain period of time, to construct the lines on charts that represent trends in data. The indicator tracks the change in price from the previous maximum/minimum value changes its direction should the current return cross the set threshold indicator. The smaller the threshold value is, the more sensitive the ZigZag indicator is to changes in price. In Figure 2 we show an example of trend identification and trading signal extraction.

To extract the trading signal from the ZigZag indicator we use a simple linear transformation which transforms ZigZag trend data into [0,1] value range. The values of 0 and 1 are assigned to be the reversal points in the trend. When the trading signal takes a value 0 it signals the market bottom to our system, while the value of 1 signals the market top.

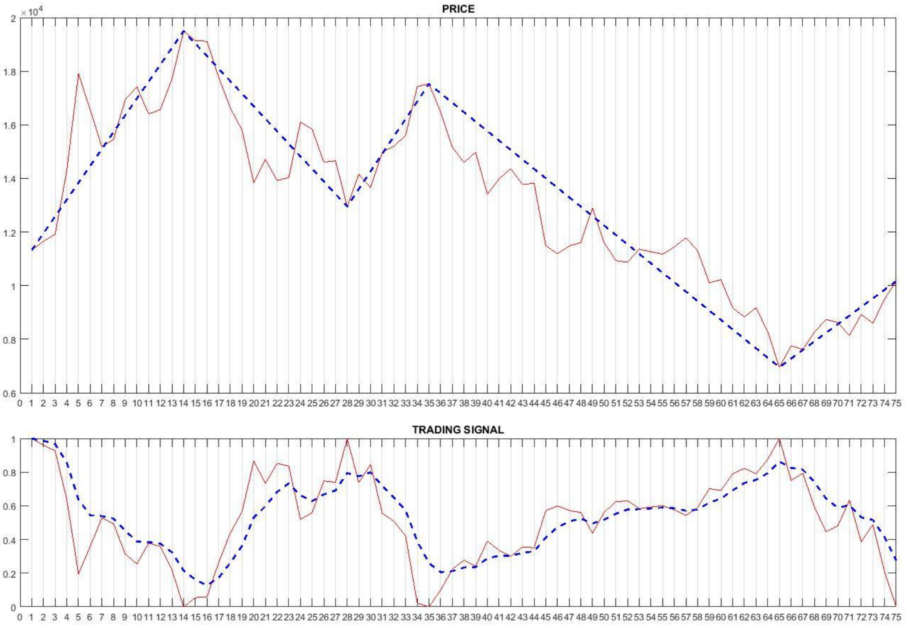


Figure 2: An example of trend identification and trading signal extraction

3.5. Trading algorithm and virtual trading machine

To enable automated trading based on neural network’s trading signal prediction we build a simple trading algorithm. We build three simple rules in order to convert trading signal into the corresponding trading action:

1. IF trading_signal > threshold_value THEN trading_action = buy
2. IF trading_signal < (1 - threshold_value) THEN trading_action = sell
3. IF (1 - threshold_value) < trading_signal < threshold_value THEN trading_action = hold

Furthermore, the trading algorithm splits the investment funds into trading lots, which are calculated as the certain percentage of the current account balance. Whenever the trading algorithm receives buy/sell trading signal it keeps purchasing/selling Bitcon until all the investment funds are spent.

Finally, to make trading simulation possible, we built a simple virtual trading machine (VTM) that uses real Bitcoin prices from the market. VTM manages trading orders, records transactions and does all the necessary accounting on trading accounts. It also delivers the trading report when the trading simulation is done.

4. RESULTS

In this section we preset trading simulation results followed by the corresponding discussion. Before results are shown, we describe our dataset, present specific parameter settings for our trading system and list the measures that we used to evaluate trading system performances.

4.1. Data

For this research, we collected data from two sources: cryptocurrency market and the social network Reddit. The market data includes standard prices (open, high, low and close), volume and market capitalization data. Social network data includes posts and comments on the subject of Bitcoin collected on the Reddit network. We chose this data source for two reasons: easily accessible data which we gained by scraping the “Daily discussion” posts about Bitcoin and because of almost one million subscribers of this network. We collected this data on a daily basis within the four-month time period (from December 2017 to April 2018) including 137 observations.

4.2. Parameter settings and performance measures

In Section 3 we presented a general methodology used for building our ATS. Here, we give detailed parameter specification which is used to obtain results presented in Section 4.3.

In Table 1 we summarize specific system parameters used in this study:

Table 1: Automated trading system parameters

Parameter	Value
Initial investment funds	$ 1.000.000
Lot size (as a percentage of current investment funds)	33%
ZigZag threshold	±10%
Trading signal threshold	0,7
Stopping criterion (for position)	5% loss
Stopping criterion (for trading)	3 consecutive losing trades

To measure the system performance, we use several standard performance measures: percentage of winning trades, return on investment (ROI), profit factor (PF) and maximal drawdown. To obtain a more detailed look at ATS results, we also calculate some of these measures for long and short positions separately.

The percentage of winning trades is often utilized to assess the performance of trading strategy. This indicator is focused only on the number of winning transactions and it does not take into account the amount that was earned/lost. ROI and PF are the most common indicators used to evaluate the efficiency of an investment. ROI is a ratio of the amount of return on an investment to the investment's cost, while PF indicates the number of monetary units earned over units lost. On the other hand, maximal drawdown is an indicator of the downside risk over a certain time period. It represents the relative difference between the highest and lowest account balance values.

4.3. Results and discussion

Experiments with neural networks showed that forecasting results of network can differ significantly when the experiment is repeated several times under the same conditions. This phenomenon is a direct consequence of how a neural network works. In order to obtain as accurate as possible approximations of network performances, researches repeat experiments several times and then calculate the average results. In this study, we did 10 simulations of trading with the same initial conditions. Simulation results are presented in Table 2.

As we can see from the results, our trading system proved to be quite a dynamic one. Considering all 10 simulations, it makes 13 daily trades on average during the four month period. Furthermore, the system was able to achieve an average ROI of 6% and average PF of 1,52 even though it has less than 50% of winning trades. It is worth noting that during the same period of time Bitcoin has lost 26% of its value. This can be used as a benchmark result to determine whether the proposed system is successful or not. Having all this in mind, we can conclude that we successfully beat the market.

If we look deeper and analyze long and short trades separately, we obtain interesting results. The results show that the proposed ATS performs much better when shorting (selling without previously buying the currency). This is probably the consequence of the predominant downward Bitcoin trend that happened after the market bubble burst in December of 2017. The following Figure 3 presents one of the simulated trading scenarios with the profit performance trajectory.

Table 2: Summary results

PERFORMANCE MEASURE	SIMULATIONS										AVG
	1	2	3	4	5	6	7	8	9	10	
Total trades	11	17	16	13	11	12	16	10	9	16	**13**
Percentage of win	45%	29%	44%	31%	27%	50%	25%	60%	44%	44%	**40%**
ROI	14%	-8%	14%	-3%	1%	14%	-17%	22%	23%	-5%	**6%**
PF	1,8	0,78	1,66	0,9	1,02	1,74	0,42	3,42	2,59	0,81	**1,52**
Max. drawdown	-13%	-9%	-9%	-8%	-20%	-9%	-8%	-13%	-9%	-8%	**-11%**
Long trades	8	11	11	11	9	10	10	5	6	10	**9**
Percentage of win	25	27	36	27	11	40	10	60	50	40	**33%**
Avg. ROI per long trade	0,7%	-0,2%	0,8%	-0,3%	-0,9%	0.7%	-1,4%	0,8%	1,1%	-0,3%	**0,1%**
Avg. PF per long trade	1,31	0,9	1,54	0,87	0,74	1,38	0,37	2,01	1,77	0,82	**1,17**
Short trades	3	6	5	2	2	2	6	5	3	6	**4**
Percentage of win	100%	33%	60%	50%	100%	100%	50%	60%	33%	50%	**64%**
Avg. ROI per short trade	2,9%	-0,9%	1%	0,4%	4,4%	3,5%	-0,6%	3,6%	5,5%	-0,3%	**2%**
Avg. PF per short trade	Inf.	0,58	2,03	1,32	Inf.	Inf.	0,54	4,47	3,82	0,8	**-**

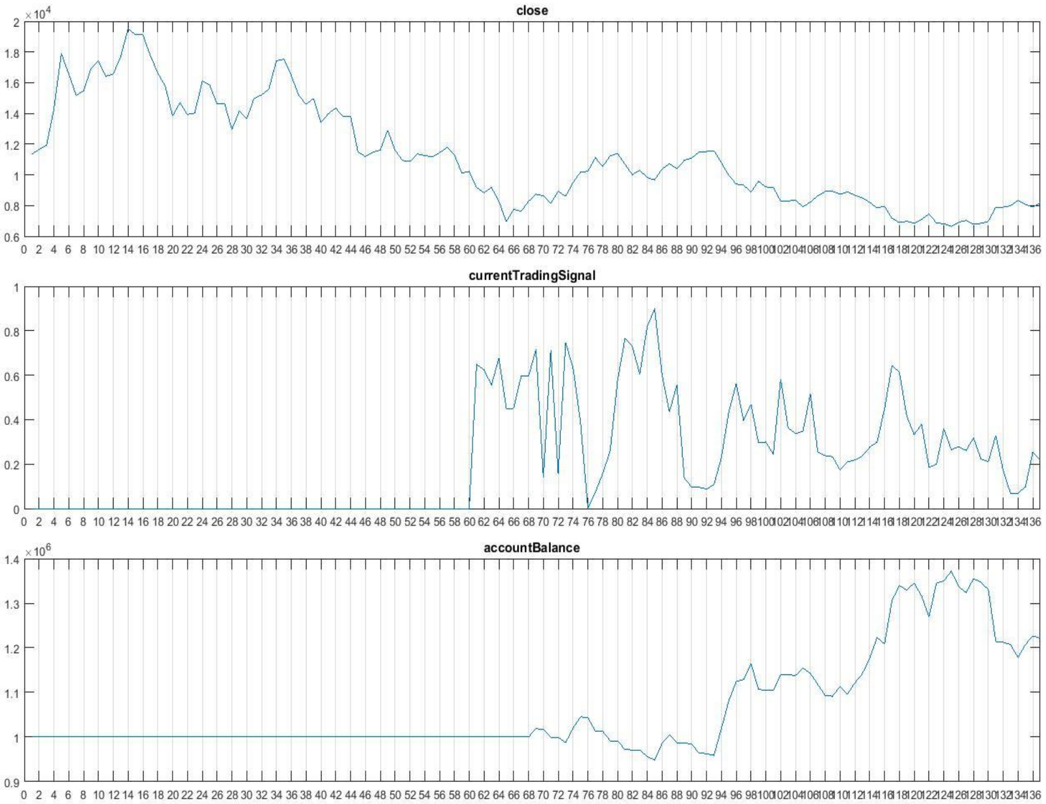


Figure 3: An example of trend identification and trading signal extraction

5. CONCLUSION

In this research, we presented an intelligent system for automated cryptocurrency trading based on neural networks. The system uses a neural network to analyze the market sentiment and learn to trade Bitcoin. For that purpose, it uses two technical indicators (Relative Strength Index and Volume-to-Market-Capitalization) in combination with two social network indicators (Sentiment Score and number of comments on the Reddit network). Another important aspect of the system is the way it extracts price trends from raw market data. In this paper, we use the well-known ZigZag indicator to extract trends. We then used a simple linear transformation to convert the obtained trend values into trading signals. Based on the trading signals, we trained the neural network to learn how to trade.

To obtain and verify results, we performed multiple trading simulations. Summary results show positive average profit performance that significantly outperforms Bitcoin performance. These results may be perceived as a confirmation of the assumption that market psychology is an important factor in still emerging cryptocurrency market. This assumption is embedded in our system indirectly, i.e. through the choice of input variables. Another interesting fact is that our system performs significantly better when trading short position, which is in accordance with the downward trend that has dominated the Bitcoin market since the euphoria bubble burst in the middle of December 2017. These encouraging results give us a motivation to further improve the proposed methodology and to continue to investigate intelligent systems for automated trading in cryptocurrency markets. Some possible further improvements would involve the implementation of some other type of neural network (such as NARX), the development of some intelligent classifier to accurately separate between positive and negative comments on the Reddit network, and the development of a more complex trading algorithm to improve the investment decision making process within our system.

REFERENCES

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Brauneis, A., & Mestel, R. (2018). Price discovery of cryptocurrencies: Bitcoin and beyond. *Economics Letters*, *165*, 58-61.

Detzel, A. L., Liu, H., Strauss, J., Zhou, G., & Zhu, Y. (2018). Bitcoin: Learning, Predictability and Profitability via Technical Analysis. Doi:10.2139/ssrn.3115846

Gandal, N., Hamrick, J. T., Moore, T., & Oberman, T. (2018). Price manipulation in the Bitcoin ecosystem. *Journal of Monetary Economics*, in press. doi:10.1016/j.jmoneco.2017.12.004

Gangwal, S., & Longin, F. (2018). Extreme movements in Bitcoin prices: A study based on extreme value theory. Retrieved from https://www.longin.fr/Recherche_Publications/Resume_pdf/Gangwal_Longin_Extreme_movements_Bitcoin_prices.pdf

Garcia, D., & Schweitzer, F. (2015). Social signals and algorithmic trading of Bitcoin. *Royal Society open science*, *2*(9), 150288. Doi:10.1098/rsos.150288

Garcia, D., Tessone, C. J., Mavrodiev, P., & Perony, N. (2014). The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, *11*(99), 20140623. Doi:10.1098/rsif.2014.0623

Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J., & Kim, C. H. (2016). Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PloS one*, *11*(8), e0161197. Doi:10.1371/journal.pone.0161197

Kreuser, J. L., & Sornette, D. (2018). Bitcoin Bubble Trouble. *Swiss Finance Institute Research Paper*, 18-24.

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. Retrieved from http://www.bitcoin.org/bitcoin.pdf

Phillips, R. C., & Gorse, D. (2017, November). Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In *Computational Intelligence (SSCI), 2017 IEEE Symposium Series on* (pp. 1-7). IEEE.

Rakićević, A., Milovanović, A., & Aničić, R. (2016, June). An application of neural networks and fundamental analysis for automated trading: Belgrade stock exchange case. In *Symposium proceedings-XV International symposium Symorg 2016: Reshaping the Future Through Sustainable Business Development and Entepreneurship* (pp. 626-634). University of Belgrade, Faculty of Organizational Sciences.

EXPERIMENTAL COMPARISON OF MULTI-LABEL LEARNING METHODS

Dušica Stepić
Faculty of Organizational Sciences, University of Belgrade
e-mail: dusica.stepic@gmail.com

Abstract: *Multi-label learning has gotten noteworthy consideration in the research community over the recent years. This has led to the enhancement of diverse multi-label classification methods. This paper presented an experimental comparison of 5 multi-label classification techniques using 10 evaluation measures over 4 benchmark datasets from different application domains. Furthermore, additional efficiency analysis of the methods, in terms of time necessary to learn a classifier and time expected to predict a set of labels for an unseen example, was conducted to determine the most beneficial ones. The results of the analysis show that the best-performing multi-label learning methods are distinct RAndom k-labELsets (RAkELd) and Label power-set (LP).*

Keywords*: machine learning, classification, predictive modeling, multi-label, multi-label learning methods*

1. INTRODUCTION

In plenty application domains where single-label classification failed to solve a problem, multi-label classification succeeded. For example, single-label classification can classify a song as either “rap” or “pop” and not both, where in fact it could belong to both genres simultaneously. (Alazaidah & Ahmad, 2016). Similarly, in text classification issue, an article can belong to more than one conceptual class at the same time, for instance, a news article can be labeled with both “Politics” and “International Relations” (Nayak et al., 2013).

In the traditional single-label classification method, training instances are associated with a single label λ from a set of disjoint labels L, |L| > 1, where |L| represents the number of possible values of a label. However, genuinely more complex classification issues exist in the real world, where an instance can belong to more than one class at the same time. If |L| = 2, then the learning issue is called a binary classification issue, while if |L| > 2 it is called a multi-class classification issue. If the instances are associated with a previously predefined set of labels Y ⊆ L, then the issue is called multi-label classification. If it is possible that multiple target labels can be assigned to an example from the test set then it is concerning a multi-label classification issue, unlike the multi-class classification where only one label is assigned to each instance (Tsoumakas & Katakis, 2007).

The number of techniques accessible for solving the multi-label data issues is constantly on the rise. (Read & Hollmén, 2017). Latest application areas involve text, image and e-mail classification, functional genomics, music categorization into emotions and others (Madjarov et al., 2012). Many different approaches that were developed in order to solve multi-label problems can be grouped into three methods. On the one hand, existing methods, which attempt to adapt multi-class algorithms, in such manner to directly solve the multi-label issue, are called algorithm adaptation methods. On the other hand, exist problem transformation methods, which attempt to transform the problem of multi-label classification. They can transform it into multiple binary problems or multi-class classification issues (Probst et al., 2017). Furthermore, there are ensemble methods for multi-label learning. Ensembles are being used in this set of methods to make predictions, whereby classifiers they work with are part of problem transformation or algorithm adaptation techniques.

In this paper, five methods for multi-label learning were compared and evaluated. The multi-label methods comprise three different problem transformation methods, one algorithm adaptation and one ensemble method. Namely, those methods are Binary Relevance (BR), Classifier Chain (CC), Label power-set (LP), Multi-Label k-Nearest Neighbor (MLkNN) and RAndom k-labELsets (RAkEL), respectively. The main aim of the research is to compare these methods and give an estimation of their performance. By making a conclusion which method has achieved better performance it would be easier to decide which one to use in some future research. Without having any previous knowledge about utilized algorithms in the process of solving multi-label issues it would be possible to choose the best proven method or technique similar to it, just by using the results of this research as a guideline.

The remainder of the paper is organized as follows. Section 2 defines the main objective of multi-label classification and presents methods that were tested and evaluated in this study. Section 3 analyses research methodology and section 4 presents the results of the experiment. Finally, conclusion and future work are described in Section 5.

2. MULTI-LABEL LEARNING METHODS

Multi-label learning is instance-based, whereby each instance is associated with a subset of labels, which belong to an already predefined set of labels. The main objective of multi-label classification is to build a predictive model based on the training data. Next step is testing if the model is adequate. The result of this phase will be a list of labels that are relevant for a given, previously unseen instance (Madjarov et al., 2012).

Multi-label learning methods are divided into three categories: problem transformation, algorithm adaptation and ensemble methods. In the first category, whose methods are algorithm independent, any single-label learning algorithm can be used to solve multi-label classification issue. It consists of methods which transform the multi-label classification problem into either a few binary classification problems, such as the BR approach, or one multi-class classification issue, such as the LP approach. The second category is algorithm adaptation methods that consist of methods which extend existing learning algorithms to deal with multi-label data directly. The MLkNN algorithm used in this work belongs to this category (SpolaôR et al., 2013). The third group of methods is called ensemble methods. This group consists of methods that use ensembles to make multi-label predictions and their base classifiers belong to either problem transformation or algorithm adaptation methods, such as the RAkEL, which uses LP as a base classifier.

The multi-label learning methods used and evaluated in this experimental study are shown in Figure 1.

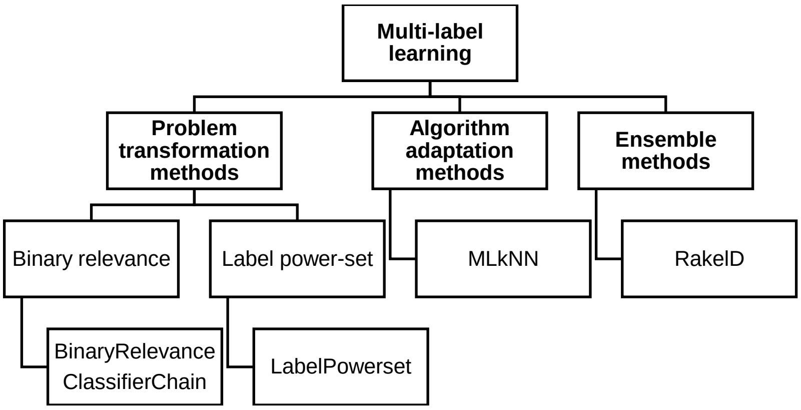


Figure 1: The multi-label learning methods used in this study

In this study five multi-label learning methods were used:

- Binary Relevance (BR) is the notable one-against-all strategy. It addresses the multi-label learning issue by learning one classifier for each class, using every one of the examples labeled with that class as positive cases and all remaining examples as negative. In order to make a set of labels that are relevant for a previously unseen example, each of the binary classifiers, when making a prediction, has to predict whether its label is important or not for that given example. BR technique has been actively analyzed and criticized because of its insufficiency of dealing with label dependency. In fact, the BR technique makes an assumption that every single label is independent of the others, which makes it easy to implement and generally effective, although unequipped of handling any sort of label relationship. (Cherman et al., 2011). During its transformation process, BR completely overlooks label relationships that exist in the training data. (Read et al., 2009)
- Classifier Chain (CC) is a technique firmly identified with the BR strategy. The fundamental idea of this algorithm is to transform the multi-label learning issue into a chain of binary classification issues, where consecutive binary classifiers in the chain are built upon the predictions of preceding ones (Zhang and Zhou, 2014). The CC model includes |L| binary classifiers as in BR, where |L| is the total number of labels. Classifiers are connected along a chain where every classifier handles the BR issue related with label $l_j \in L$, j = {1, ..., L}. (Read et al., 2009)
- Label power-set (LP) belongs to the group of the label power-set methods. The LP approach directly transforms the multi-label dataset into one single-label dataset, whereby each label is a unique combination of labels appearing on the training set, which was previously transformed to a single-label dataset, in which any method for the multi-class issue can be directly applied to. Moreover, this approach implicitly considers label dependence (SpolaôR et al., 2013). However, problem is that the space of

possible label subsets can be very large. If the number of meta-classes is large that can be an extremely difficult issue to resolve because it makes probability estimation challenging. Additionally, most LP-based method implementations essentially overlook label combinations that are not present in the training set. (Dembszynski et al., 2010)

- Multi-Label k-Nearest Neighbor (MLkNN) is the multi-label version of the well-known kNN algorithm. First, for each instance from the test set, the MLkNN determines its k nearest neighbors in the training set. At that point, according to statistical information gathered from the label sets of these nearest neighbors, namely the number of neighboring instances belonging to each of the possible labels, the maximum a posteriori principle is used to determine the relevant set of labels for an unseen instance from the test set. (Zhang & Zhou, 2007)
- Distinct RAndom k-labELsets multi-label classifier (RAkELd) is an ensemble technique for multi-label learning developed on top of the common problem transformation method (LP). RAkEL, as its name would suggest, constructs an ensemble of LP classifiers. It performs LP methods on M different, random subsets of size k, whereby subsets$\subset$ {1, ..., L}, and k is a small number, k < L. Those chosen subsets are called k-label sets. Each performed LP method, trains an LP classifier by using randomly chosen subsets. (Modi & Panchal, 2012). Thus, RAkEL takes in consideration label relationships and avoids the disadvantage of the LP method by applying single-label classifiers on k-label subsets with a satisfactory number of labels and a sufficient number of instances per label (Tsoumakas et al., 2011). Predictive power enhances by using ensembles of LP algorithms, in comparison to using only LP method.

3. RESEARCH METHODOLOGY

The basic idea of this research is to compare tested multi-label learning methods and identify the best ones based on their predictive performance and time complexity. The experiment was implemented in the Jupyter Notebook.

In this paper, an experimental assessment of methods for multi-label learning is presented. Five of the most popular methods for multi-label learning were evaluated using a wide range of evaluation measures. Datasets used in the experiment were acquired from UCI Machine Learning repository (Dua & Karra, 2017) and MULAN (Tsoumakas et al., 2011). List of datasets with basic details is shown in Table 1.

Table 1: Benchmark datasets from UCI repository and MULAN used in this study

Dataset	Classes	Attributes	Size
student performance	6	53	382
yeast	14	117	2417
emotions	6	78	593
scene	6	300	2407

First, five of the most popular multi-label methods that were recently proposed in the literature were selected. The chosen methods are divided into three main groups: problem transformation (three methods), algorithm adaptation (one method) and ensembles (one method). The methods use two types of basic algorithms for machine learning: SVM (problem transformation methods), random forest (an ensemble method) and k-nearest neighbors (an algorithm adaptation method). Moreover, 10 different evaluation measures that are typically used in the context of multi-label learning were considered. A better and more detailed view of the algorithm performance can be shown by using a variety of evaluation measures.

The evaluation measures are divided into three groups: example-based (Accuracy and Hamming loss), label-based (Macro-precision, Micro-precision, Macro-recall, Micro-recall, Macro-F1, Micro-F1 and Roc AUC score) and ranking-based (One error). Every multi-label method was trained on 70% of the data. Final testing and evaluation were performed on 30% of the data. Furthermore, the efficiency of the methods is evaluated by measuring the time necessary to learn the classifier and the time needed to create a prediction for an unseen example. Additionally, the methods were evaluated on four multi-label benchmark datasets from different application domains: social science, protein, image and music classification.

As far as the BR, LP and RAkEL techniques, they can use any learning algorithm for classifier training. The chosen methods were using two kinds of base classifiers for resolving the partial binary classification issues in all problem transformation methods and the ensemble method: support vector machines (SVM) and Random Forest (RF), respectively. The SVM with the Radial Basis Function kernel was used for training in all problem transformation methods and RF classifier was used for an ensemble method.

4. RESULTS

Evaluation measures (Accuracy, Macro-Precision, Micro-Precision, Macro-Recall, Micro-Recall, Macro-F1, Micro-F1, 0/1 Loss, Hamming Loss, Roc auc score) and efficiency measures of the tested methods (Train and Test time) are shown in Table 2 and Table 3. The best performance was bolded for each dataset, and the second best was presented in italic.

Further analyzing the predictive method performance across all ten evaluation measures and two efficiency measures: the very best performing methods on all measures are either RAkELd or LP. RAkELd is the best performing method, closely followed by LP. Performance of the BR was inferior over all evaluation measures, aside from precision, meaning that the labels predicted as relevant were truly relevant in the original examples (small number of false positives results in high precision). However, BR is leaving out some of the relevant labels when making predictions (larger number of false negatives results in low recall).

Table 2: Evaluation and efficiency measures for Student performance and Yeast dataset

Measures	**Student performance**					**Yeast**				
	BR	CC	LP	MLkNN	RAkELd	BR	CC	LP	MLkNN	RAkELd
Accuracy	*0.496*	**0.513**	*0.496*	0.4	0.487	0.011	0.014	0.134	*0.176*	**0.202**
Macro-Precision	0.786	*0.789*	0.765	0.779	**0.817**	**0.755**	*0.73*	0.581	0.717	0.623
Micro-Precision	0.781	*0.784*	0.764	0.776	**0.81**	0.243	0.18	0.222	**0.56**	*0.444*
Macro-Recall	0.976	*0.978*	**0.993**	0.969	0.953	0.146	0.169	0.298	*0.367*	**0.381**
Micro-Recall	0.979	*0.981*	**0.994**	0.973	0.96	0.356	0.378	0.543	**0.591**	*0.578*
Macro-F1	0.866	*0.869*	0.86	0.859	**0.875**	0.128	0.161	0.226	**0.397**	*0.391*
Micro-F1	0.872	*0.874*	0.865	0.865	**0.883**	0.484	0.498	0.561	**0.647**	*0.6*
0/1 loss	*0.504*	**0.487**	*0.504*	0.6	0.513	0.989	0.986	0.866	*0.824*	**0.798**
Hamming loss	0.217	*0.213*	0.235	0.229	**0.193**	*0.232*	0.233	0.259	**0.197**	0.236
Roc auc score	0.548	*0.554*	0.515	0.544	**0.635**	*0.501*	0.508	0.509	**0.589**	0.571
Train time	0.096	0.077	*0.076*	0.586	**0.07**	5.102	5.399	*2.774*	6.233	**1.475**
Test time	0.035	0.04	**0.023**	0.118	*0.027*	1.814	1.74	*1.481*	1.577	**0.156**

Table 3: Evaluation and efficiency measures for Emotions and Scene dataset

Measures	**Emotions**					**Scene**				
	BR	CC	LP	MLkNN	RAkELd	BR	CC	LP	MLkNN	RAkELd
Accuracy	0.157	0.157	**0.309**	0.135	*0.27*	0.27	0.335	**0.705**	0.633	*0.682*
Macro-Precision	**0.678**	**0.678**	0.581	*0.639*	0.61	**0.955**	*0.929*	0.765	0.771	0.736
Micro-Precision	**0.611**	**0.611**	0.554	0.577	*0.588*	0.645	**0.774**	*0.773*	0.769	0.745
Macro-Recall	0.398	0.405	*0.586*	0.403	**0.626**	0.286	0.351	**0.719**	0.696	*0.704*
Micro-Recall	0.423	0.429	*0.613*	0.435	**0.634**	0.276	0.339	**0.714**	0.69	*0.696*
Macro-F1	0.46	0.466	*0.567*	0.455	**0.602**	0.362	0.442	**0.743**	*0.728*	0.719
Micro-F1	0.521	0.526	*0.596*	0.518	**0.622**	0.429	0.496	**0.739**	*0.728*	0.716
0/1 loss	0.843	0.843	**0.691**	0.865	*0.73*	0.73	0.665	**0.295**	0.367	*0.318*
Hamming loss	0.243	*0.242*	0.258	0.253	**0.241**	0.131	0.123	**0.09**	*0.092*	0.099
Roc auc score	0.652	0.654	*0.691*	0.644	**0.719**	0.641	0.673	**0.835**	*0.826*	0.825
Train time	0.18	0.199	*0.161*	0.636	**0.066**	7.186	3.915	*2.866*	*6.455*	**0.513**
Test time	0.069	0.066	*0.045*	0.209	**0.031**	1.588	1.745	*0.862*	2.515	**0.125**

Table 4: Best and second-best performances of multi-label learning methods

	BR	**CC**	**LP**	**MLkNN**	**RAkELd**
First	4	5	13	6	22
Second	3	11	15	8	13

Table 4 shows number of best and second-best performances of each multi-label learning method through all four datasets. The best multi-label learning method was RAkELd, which performed best 22 times, and was second 13 times. LP has also shown great performances and BR performed worst of tested multi-label learning methods. It was first 4 times, and second only for 3 times. BR is the simplest problem transformation method. However, lately, it is sidelined on the grounds of assuming label independence. BR disregards label relationships that exist in the training data.

LP-based methods directly consider the label correlations. However, the main problem is the size of the set, which consists of possible label subsets, because that set can be quite large in LP. RAkELd is a tactic that creates a group of LP classifiers, or in other words an ensemble of them. During the classifier training process, each LP classifier is being trained on a different and random subset of the labels. This approach aims at considering label correlations and at the same time avoiding the problems of LP. Results of performance comparison against the BR and LP methods are in favor of the LP-based methods (RakelD and LP).

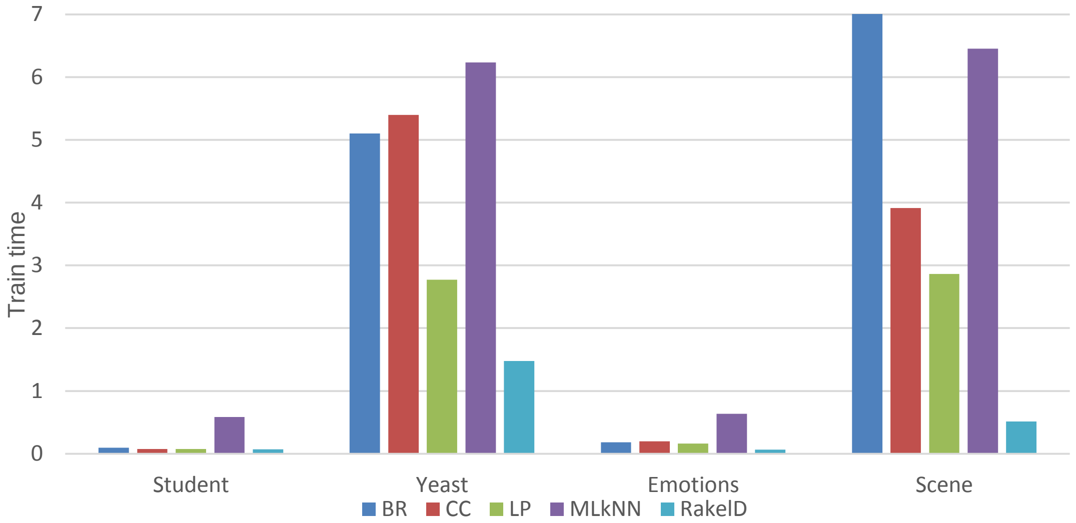


Figure 2: Multi-label learning methods - Train time (seconds)

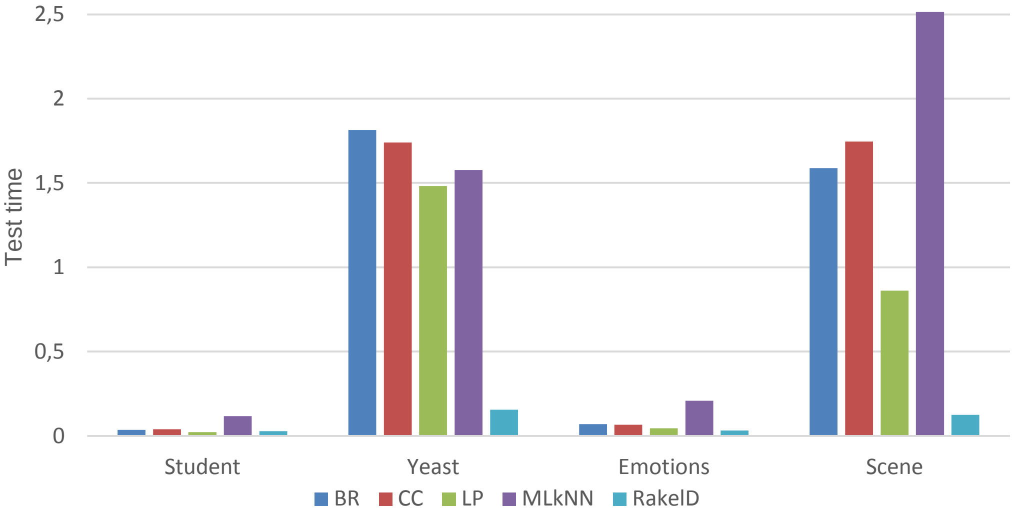


Figure 3: Multi-label learning methods - Test time (seconds)

As mentioned earlier, each dataset was randomly split into two sets: 70% for training and 30% for a test. For comparison, the charts in Fig. 2 and Fig. 3 show the efficiency of five tested methods, in terms of time complexity (test and train time of the classifiers). Considering efficiency, RAkELd (using the tree-based method as a base classifier) was generally faster to train a classifier and make a prediction for an unseen example than the SVM-based methods. The results show that RAkELd is faster than LP on testing time (on average 2.8 times) and on training time (on average 7.11 times). Furthermore, RAkELd is faster than MLkNN on testing time (on average 6.5 times) and on training time (on average 13 times).

When comparing only problem transformation methods among each other, LP performed better than BR and CC. The multi-label variant of k-nearest neighbors (ML-kNN) performed poor and had the worst train and test time. MLkNN is the most time-consuming algorithm.

5. CONCLUSION

This paper presented results of five different methods for multi-label classification. The goal was to show comparative experimental results of certain multi-label classification methods that were previously chosen from various research papers and to determine the greatest ones. This work also gave an organized presentation of those methods and provided empirical evaluation over four different multi-label datasets with a variety of evaluation and efficiency metrics.

All in all, considering the performance and the efficiency of the evaluated methods RAkELd, an ensemble method used in this paper, proved superior among other tested methods. Empirical evaluation demonstrates the competitiveness of RAkELd against other multi-label learning methods, which were tested, both in terms of predictive performance and time complexity.

Furthermore, RAkELd and LP proved to be the best in terms of predictive performance and demonstrated better competence than the remaining methods. Both of this method are LP-based, which prove that label dependency is important when it comes to predictive modeling in multi-label classification problems. On the other hand, it depends upon the concrete objective of the learning issue whether label dependencies can be used to enhance predictive performance (Dembszynski et al., 2010). Researchers have been working on the plan of designing multi-label methods competent for handling the diverse relationships between labels, particularly label dependency, correlation and co-occurrence. For example, in the LP method, inter-relationships between labels are mapped straightforwardly from the data, since all the existing combinations of single-labels present in the training examples are used as one of the possible labels in the given multi-class classification issue (Cherman et al., 2011).

Future work will include performing more extensive experiments with more datasets and extending this experimental study by including more multi-label learning methods, especially ensemble and algorithm adaptation methods. Moreover, including Grid search for hyperparameter optimization and investigating the issue of selecting appropriate parameters for all multi-label classifiers to improve performances.

REFERENCES

Alazaidah, R., & Ahmad, F. K. (2016). Trending Challenges in Multi Label Classification. IJACSA) International Journal of Advanced Computer Science and Applications, 7, 127-131.

Cherman, E. A., Monard, M. C., & Metz, J. (2011). Multi-label problem transformation methods: a case study. CLEI Electronic Journal, 14(1), 4-4.

Cortez, P. & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April 2008, EUROSIS, ISBN 978-9077381-39-7.

Dembszynski, K., Waegeman, W., Cheng, W., & Hüllermeier, E. (2010). On label dependence in multilabel classification. In LastCFP: ICML Workshop on Learning from Multi-label data. Ghent University, KERMIT, Department of Applied Mathematics, Biometrics and Process Control.

Dua, D. and Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern recognition, 45(9), 3084-3104.

Modi, H., & Panchal, M. (2012). Experimental comparison of different problem transformation methods for multi-label classification using MEKA. International Journal of Computer Applications, 59(15).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825-2830.

Probst, P., Au, Q., Casalicchio, G., Stachl, C., & Bischl, B. (2017). Multilabel classification with R package mlr. arXiv preprint arXiv:1703.08991.

Read, J., Martino, L., & Hollmén, J. (2017). Multi-label methods for prediction with sequential data. Pattern Recognition, 63, 45-55.

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009). Classifier chains for multi-label classification. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 254-269). Springer, Berlin, Heidelberg.

Read, J., Reutemann, P., Pfahringer, B., & Holmes, G. (2016). Meka: a multi-label/multi-target extension to weka. The Journal of Machine Learning Research, 17(1), 667-671.

Santos, A., Canuto, A., & Neto, A. F. (2011). A comparative analysis of classification methods to multi-label tasks in different application domains. Int. J. Comput. Inform. Syst. Indust. Manag. Appl, 3, 218-227.

SpolaôR, N., Cherman, E. A., Monard, M. C., & Lee, H. D. (2013). A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science, 292, 135-151.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. (2008, September). Multi-Label Classification of Music into Emotions. In *ISMIR* (Vol. 8, pp. 325-330).

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. International Journal of Data Warehousing and Mining, 3(3).

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011). Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering, 23(7), 1079-1089.

Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., Vlahavas, I. (2011) "Mulan: A Java Library for Multi-Label Learning", Journal of Machine Learning Research, 12, pp. 2411-2414.

Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.

Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8), 1819-1837.

DATA MINING USING ORACLE DATA MINER AND ANALYTIC FUNCTIONS WITH HADOOP

Ivan Rakić*[1], Željana Milošević[1], Slađan Babarogić[1]
[1]University of Belgrade, Faculty of Organizational Sciences, Serbia
*Corresponding author, e-mail: ivan.rakic@fon.bg.ac.rs

Abstract: *In this paper we have described how a large amount of data can be processed in Oracle Database using different algorithms and functions. After training data mining algorithms, their performance on new datasets can be measured and examined. We have focused on the comparison of specific algorithms in order to decide which of them gives more reliable data processing results. To make sure that more efficient algorithm is chosen, Oracle analytic functions can be used for data preparation before using the algorithms and for processing the results obtained after the execution of the algorithms. Hadoop framework and Hive, as part of it, have been used to improve access to data being processed.*

Keywords*: Data Mining, Oracle, Hadoop, classification, clustering, analytic functions*

1. INTRODUCTION

One of the biggest challenges that people who deal with information technology today face is how to handle large amount of data generated by users of different systems every day. (Thearling, 2017)

Lately, Data Mining is very important part of computer science that investigates data in order to discover and describe patterns. Because we live in a world where huge amounts of data are generated every day, it is imperative to find a way to extract useful information from this data, to classify it and to make certain conclusions (Thearling, 2017).

At the beginning, in the second chapter we described the problem of data analysis that we tried to solve in this paper.
In the third chapter is given a brief overview of what a Big Data and Data Mining represents, and after that is described the model which is usually used in Big data processing techniques as well as technologies and tools used for the realization of the demonstration case studies. Also, we make some review about the most useful Data Mining algorithms for classification and clustering, such as Decision tree, Naive Bayes, Hierarchical clustering and K-means algorithms. And finally, we present some of the most frequently used analytical functions, in this case used for data preparation and analysis of results of Data Mining algorithms. An example of using these algorithms and functions in Oracle Database is shown in the fourth chapter. In the end, in the fifth chapter we give our vision of further application of Data Mining algorithms.

2. PROBLEM STATEMENT

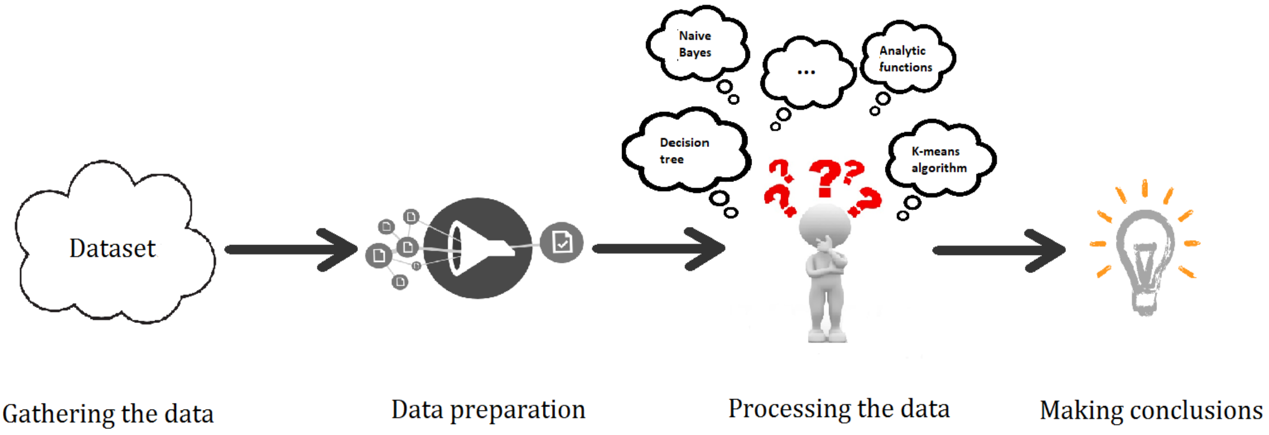


Figure 1: Problem statement

Nowadays, large amounts of data are piling up from various sources and the amount of this data has a constant tendency of growth. New ways to analyze this data need to be discovered, in order to find the best way to extract important information from it. Data in the original form represents only the noisy data that needs to be prepared for the application of various techniques and algorithms. Various techniques and algorithms can be applied to the data which is obtained as a result of prior data cleansing. In order to make a certain conclusion which of these techniques give the most reliable results depending on the context of the problem, it is necessary to apply several techniques separately or a combination of several techniques, functions and algorithms. If combination of techniques is applied, it is important to plan the sequence in which they will be used. Out of all these steps, choosing appropriate algorithm for given domain is still mostly unspecified, since there is no universal approach to this problem, and solutions vary from case to case.

3. THEORETICAL BACKGROUND

In this chapter are given main theoretical concepts concerning Big Data, Data Mining and corresponding algorithms, as well as required Oracle functionalities, such as analytic functions.

3.1. Big Data

Big data can be defined as “The storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to, NoSQL, MapReduce and machine learning”. (De Mauro, 2016) The term “Big Data” is used to identify amounts of data which size is bigger than software capacity that is usually used for storage, processing and data management. (Elgendy & Elragal, 2016)

3.1.1. Hadoop

MapReduce is one of the earliest and most popular programming models created by Google in 2004 and based on C++ language. This model is used for parallel processing and generating Big Data sets. (Yahya, Hegazy, & Ezat, 2012)

The MapReduce model is based on two basic functions: **Map** and **Reduce**. At first, the **Map function** is used to process the input records into a sequence of key-value pairs, and then generate a set of intermediate key-value pairs. The method key-value pairs, which are derived from the input data is defined by user for the Map function. The **Reduce function** merges all the key-value pairs with the same key into the same Reduce task. Reduce task processes one key at the time and combine all values associated with that key. The mode of combination of values is determined by the user for the Reduce function. (Jeffrey Dean, 2008)

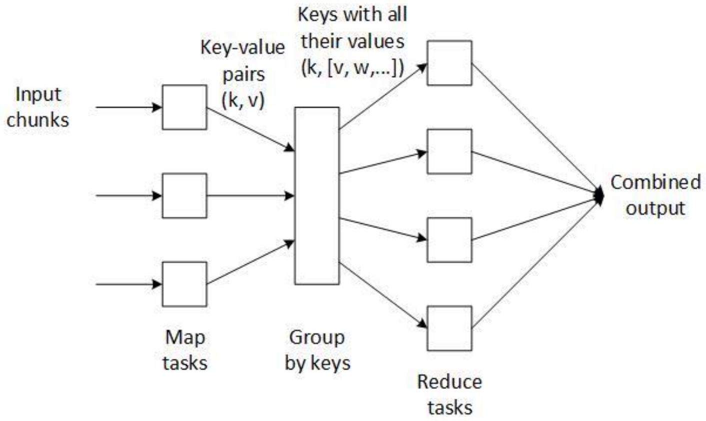


Figure 2: Schematic of MapReduce computation (Leskovec, Rajaraman, & Ullman, 2014)

There are many implementations of MapReduce model, but the most commonly used is a **Hadoop framework**. (Venner, 2009)

The big companies like Facebook, Google, Amazon and Yahoo were first to face rapid increase in the amount of data. They have to execute terabytes and petabytes of data by the day, in order to interpret the requirements, they receive from users and respond to them. The problem of handling and processing all of this data is resolved by Hadoop which has replaced scarce, existing tools. (Ghazi & Gangodkar, 2015)

3.1.2. Hive

One of the most common problems that Hadoop users encounter is creating queries over data with Hadoop as in traditional RDBMS infrastructure. **Hadoop Hive** is an open source SQL-based shared warehouse system which is suggested to solve mentioned problem by supplying an SQL-like abstraction above the Hadoop framework. Hive is translator which represents a combination of SQL language and MapReduce model. It uses HiveQL, in order to create queries over data stored in a cluster. Hive is used to bridge barriers between traditional applications which are implemented using SQL-based RDBMSs and Hadoop framework. (Dokeroglu, Ozal, Bayir, Cinar, & Cosar, 2014)

3.2. Data Mining and Data Mining algorithms

"**Data Mining** is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so they may be used in an enterprise's decision making." (Gupta, 2014)

Some of techniques which are used for Data Mining are very similar to the machine learning techniques, but on the other hand, a lot of them are closely related to techniques that are deployed for statistical analysis. (Gupta, 2014)

Data Mining uses advanced mathematical algorithms for automatically searching data warehouse in order to predict some future events based on past events. Although being a powerful tool, it is very important to remember that it is not able to recognize the value of the information which is derived from the data. (Data Mining Concepts, 2018)

Data Mining algorithms have various purposes today, ranging from fraud detection and bankruptcy prediction to target marketing and customer retention as, perhaps, its most commonly used purpose. However, results of such techniques depend largely on complex nature of data to be processed and choice of adequate Data Mining algorithm, since there is no established practice related to use of specific algorithm on specific problem domain. More traditional approach is represented by statistical algorithms, which rely heavily on mathematical models and assumptions such as data normality. Opposed to them, machine learning algorithms are assumption free, and tend to outperform above mentioned algorithms when mining business datasets. (Becker, 2001)

3.2.1. Classification

The aim of **classification** is to develop a classification rule, which can later be used to categorize given objects based only on their vector of features. Training dataset is necessary for classification, and it contains collection of object, their vector of features, as well as their correct classes. Most reliable way to determine correct classes is by consulting domain expert. In other words, the goal of classification is to distribute given objects into *m* distinct categories, with highest precision as possible. (Alpaydin, 2014) Classification is part of supervised learning scenario, in which "the learner receives a set of labeled examples as training data and makes predictions for all unseen points". (Mohri, Rostamizadeh, & Talwalkar, 2012) Some of the most used classification algorithms are:

- **Decision Tree -** a decision tree represents a binary tree in which leaves render decisions (classes) and interior nodes conditions on the object being classified. The classification process starts at the root of the tree where predicate is evaluated, and algorithm moves to one of the children based on the result of the evaluation. These steps are repeated for each interior node, until one of the leaves is reached. Reached leave represents object's class. Important task in this process is the choice of predicates for each of the interior nodes. (Leskovec, Rajaraman, & Ullman, 2014)

- **Naive Bayes** classifier is used to assign most likely class to a given object which is described by its feature vectors, with the assumption that given features are independent given class. The base of this classifier is Bayes' theorem, which describes probability of an event, based on conditions related to that event with "naive" assumptions about feature independency. These assumptions are not realistic, but despite that, Naive Bayes has shown exceptionally good results in practice. (Rish, 2001)

3.2.2. Clustering

Clustering is defined as process of examining a collection of certain “points” and grouping them into “clusters” according to some distance measure, in such manner that the distance between points in the same cluster is small and the distance between points in different clusters is large. The distance between two points can be defined in various ways, and most often it is traditional Euclidian distance. (Leskovec, Rajaraman, & Ullman, 2014)

Clustering is part of unsupervised learning scenario, where, based on an unlabeled dataset, learner makes predictions for all unseen points. Due to this, it is hard to evaluate the performance of a learner. (Leskovec, Rajaraman, & Ullman, 2014) Mostly used clustering algorithms are:

- **Hierarchical clustering -** each point starts as separate cluster, and in every iteration, new, larger clusters are built by merging two smaller clusters. The most common way to represent a cluster is by using centroids, and centroids to be merged are usually ones which have the shortest distance between each other. (Leskovec, Rajaraman, & Ullman, 2014)

- **K-means algorithms** require that the final number of clusters, *k*, is known in advance. In initial step *k* points which will represent clusters are selected. After that, each point, other than *k* selected points, are considered and assigned to the closest cluster. The distance between a point and a cluster is usually measured by distance between a point and the centroid of a cluster. There are several options for determining initial points, such as picking the points that are as far away from another as possible, clustering a sample of the data using some other algorithm, etc. (Leskovec, Rajaraman, & Ullman, 2014)

3.3. Oracle database support for data processing

Oracle Data Mining represents a powerful ability to analyze data inside of Oracle Database. It can be used for developing predictive Data Mining applications, adding smart possibilities in existing applications, or generating predictive queries to research data. Oracle Data Mining is a component of the Oracle Advanced Analytics Option of Oracle Database Enterprise Edition. One of the essential features of this component is the ability to work with Big Data sets in different forms. (Data Mining Concepts, 2018)

A development environment that can be used to execute Data Mining algorithms in Oracle databases is Oracle SQL Developer which has the appropriate extension called **Oracle Data Miner**. This tool is based on workflow paradigm which allows catching, documenting and automating the process of building, evaluating and using Data Mining algorithms. Inside of a workflow, it is possible to specify data transformations, build and evaluate more algorithms, and score more data sets. It is possible to create predictive models in Oracle Data Miner, which developers can integrate into applications to automate the discovery new business intelligence-predictions, patterns and discoveries-throughout the enterprise. (Data Mining Concepts, 2018)

3.3.1. Oracle analytic and aggregate functions

Data Miner can operate on data stored in tables and views in database, and results of tested algorithms can also be stored in such manner, which gives a lot of flexibility and potentiality by combining Data Mining process and big relational database, such as Oracle database. One of such advantages is the opportunity of using aggregate and analytic functions on datasets and results of Data Mining algorithms.

An aggregate function is used for aggregating data from several rows into a single result row. Analytic functions also operate on subsets of rows, but they do not decrease the number of rows returned by the query. Analytic functions calculate an aggregate value based on a group of rows. They differ from aggregate functions in that they return multiple rows for each group. (Analytic Functions, 2018)

Some of the most used aggregate and analytic functions are **MAX**, **MIN**, **VARIANCE**, **STDDEV**, etc. (Analytic Functions, 2018)

4. SELECTING DATA MINING ALGORITHMS USING ORACLE DATA MINER, HADOOP AND ORACLE ANALYTIC FUNCTIONS

In this chapter will be described the example of using Oracle Data Miner extension and Oracle analytic functions for classification and clustering of datasets stored in Hadoop file system. This example was implemented using Oracle Big Data Appliance, which contains all necessary services and software.

The process of selecting the most appropriate data mining algorithm for given dataset is described using following BPMN 2.0 diagram.

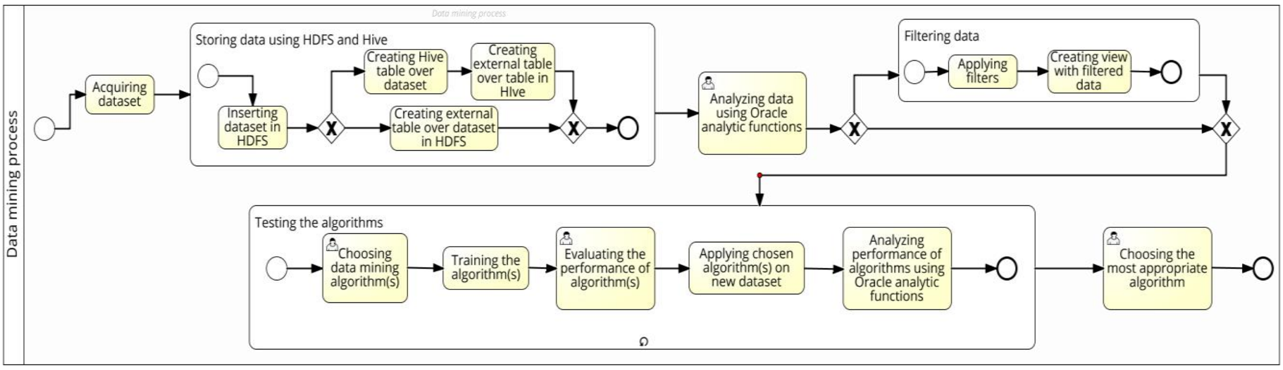

Data mining process
Data mining process
Acquiring dataset
Storing data using HDFS and Hive
Inserting dataset in HDFS
Creating Hive table over dataset
Creating external table over table in HIve
Creating external table over dataset in HDFS
Analyzing data using Oracle analytic functions
Filtering data
Applying filters
Creating view with filtered data
Testing the algorithms
Choosing data mining algorithm(s)
Training the algorithm(s)
Evaluating the performance of algorithm(s)
Applying chosen algorithm(s) on new dataset
Analyzing performance of algorithms using Oracle analytic functions
Choosing the most appropriate algorithm

Figure 3: BPMN diagram describing the selection of data mining algorithm

Process shown in diagram above starts with finding suitable dataset for desired mining area. After that it is necessary to store acquired dataset in some manner. One of the most popular ways to do that today is by using Hadoop and HDFS. After inserting dataset in HDFS, Hive can be used to access imported data, or data can be accessed directly in HDFS. Since this example shows data mining in Oracle database, this step also includes creating external tables over previously imported data. This is followed by data analysis, which includes detecting and eliminating inconsistent and “dirty” data, possibly done with Oracle analytic functions. If inconsistencies are found, data should be filtered and result of that filtering should be stored in some way, possibly by creating a view. Next step is about testing data mining algorithms, and this includes choosing algorithms to be tested, training as well as evaluating results of that training. After that, chosen algorithms can be applied on new dataset and their performance on this dataset can be analyzed, again with Oracle analytic functions. Last few steps are repeated until a decision can be made, choosing the right algorithm.

4.1. CLASSIFICATION EXAMPLE

Dataset used for classification is about detection of credit card frauds. First step described in Figure 3 is acquiring dataset and **storing it using HDFS and Hive**. After starting all necessary services, dataset file is imported in Hadoop file system via terminal commands. Since the format in which the files are stored is .csv format, it is stated that all fields are separated by comma. Next step is about creating external table in Oracle database over the previously created Hive table.

No missing or “dirty” data were found after **analyzing the data**, so there is no need for **filtering**.

Next step is about **testing and evaluating algorithms**. Before all, Data Miner connection must be created, and that connection must be created over the same schema that was used for creating external tables in previous steps. After that, workflows can be added.

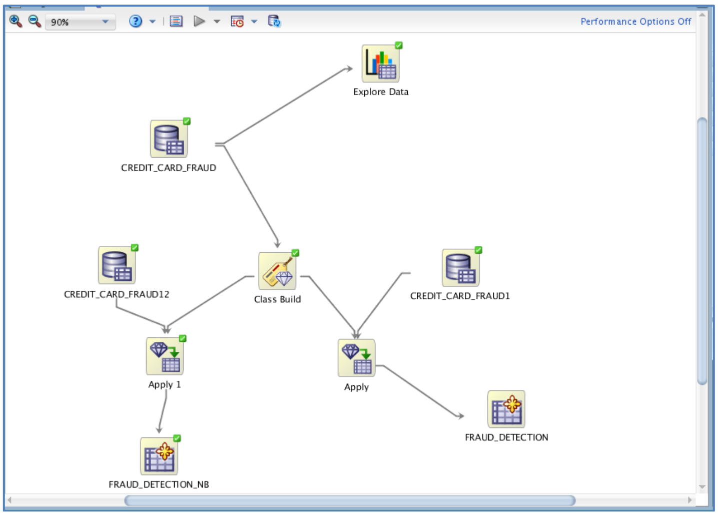
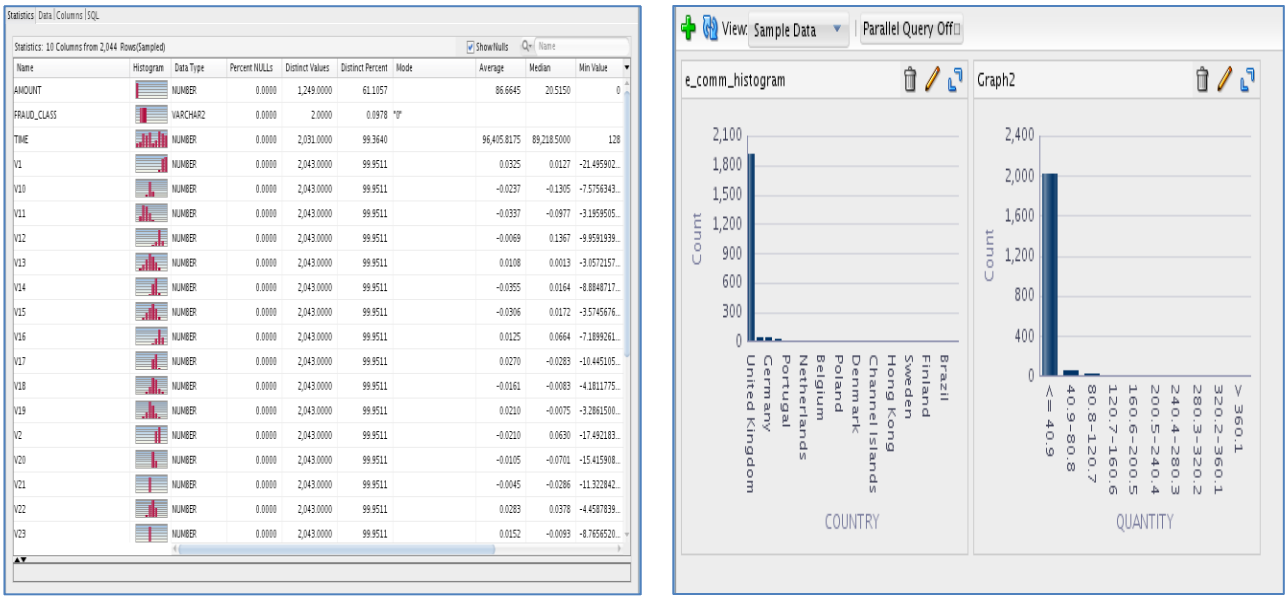


Figure 4: Workflow for classification in Data Miner

After adding node that represents data source and connecting it with appropriate table, different operations can be performed with it. Wide range of statistics is at disposal, such as minimum and maximum value, median, variance, standard deviation, etc. Data can be also visualized, using many different graphs, as shown in pictures below.

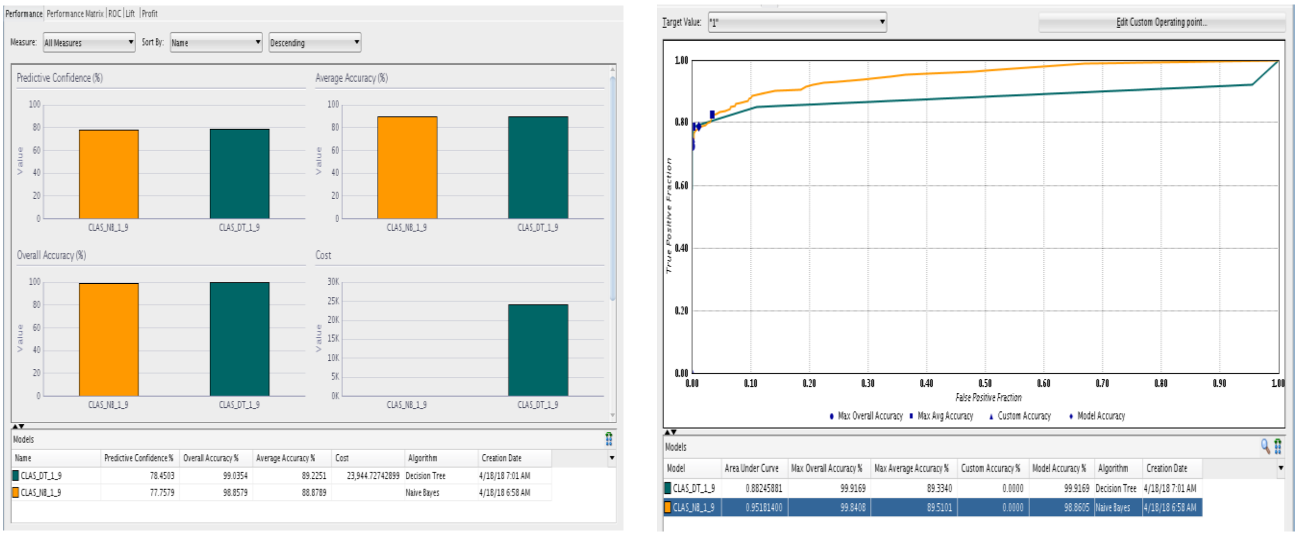


Statistics: 10 Columns from 2,044 Rows(Sampled)

Name	Histogram	Data Type	Percent NULLs	Distinct Values	Distinct Percent	Mode	Average	Median	Min Value
AMOUNT		NUMBER	0.0000	1,249.0000	61.1057		86.6645	20.5150	0
FRAUD_CLASS		VARCHAR2	0.0000	2.0000	0.0978	"0"			
TIME		NUMBER	0.0000	2,031.0000	99.3640		96,405.8175	89,218.5000	128
V1		NUMBER	0.0000	2,043.0000	99.9511		0.0325	0.0127	-21.495902...
V10		NUMBER	0.0000	2,043.0000	99.9511		-0.0237	-0.1305	-7.5756343...
V11		NUMBER	0.0000	2,043.0000	99.9511		-0.0337	-0.0977	-3.1959505...
V12		NUMBER	0.0000	2,043.0000	99.9511		-0.0069	0.1367	-9.9591939...
V13		NUMBER	0.0000	2,043.0000	99.9511		0.0108	0.0013	-3.0572157...
V14		NUMBER	0.0000	2,043.0000	99.9511		-0.0355	0.0164	-8.8848717...
V15		NUMBER	0.0000	2,043.0000	99.9511		-0.0306	0.0172	-3.5745676...
V16		NUMBER	0.0000	2,043.0000	99.9511		0.0125	0.0664	-7.1899261...
V17		NUMBER	0.0000	2,043.0000	99.9511		0.0270	-0.0283	-10.445105...
V18		NUMBER	0.0000	2,043.0000	99.9511		-0.0161	-0.0083	-4.1811775...
V19		NUMBER	0.0000	2,043.0000	99.9511		0.0210	-0.0075	-3.2861500...
V2		NUMBER	0.0000	2,043.0000	99.9511		-0.0210	0.0630	-17.492183...
V20		NUMBER	0.0000	2,043.0000	99.9511		-0.0105	-0.0701	-15.415908...
V21		NUMBER	0.0000	2,043.0000	99.9511		-0.0045	-0.0286	-11.322842...
V22		NUMBER	0.0000	2,043.0000	99.9511		0.0283	0.0378	-4.4587839...
V23		NUMBER	0.0000	2,043.0000	99.9511		0.0152	-0.0093	-8.7656520...

Figure 5: An example of statistical functions and graphs applied on data sets

Four classification algorithms are available: Decision Tree, Naive Bayes, Generalized Linear Model and Support Vector Machine. After the selection of the algorithms (in this example Decision Tree and Naive Bayes algorithms are used), connecting classification node with data source and setup of parameters (such as percent of dataset to be used for training and testing), classification process can be started. When this is complete, performance of chosen algorithm can be examined, and different algorithms can be compared using many criteria, such as predictive confidence, average accuracy, overall accuracy, cost, or using visual representation, as shown on pictures below.



Measure: All Measures Sort By: Name Descending

Models

Name	Predictive Confidence %	Overall Accuracy %	Average Accuracy %	Cost	Algorithm	Creation Date
🟩 CLAS_DT_1_9	78.4503	99.0354	89.2251	23,944.72742899	Decision Tree	4/18/18 7:01 AM
🟧 CLAS_NB_1_9	77.7579	98.8579	88.8789		Naive Bayes	4/18/18 6:58 AM

Target Value: "1"

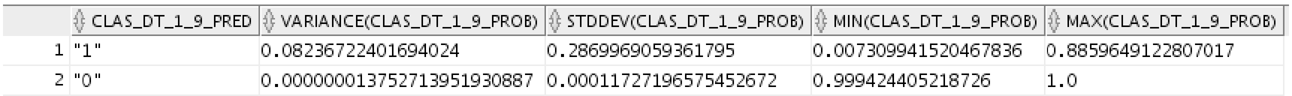
Models

Model	Area Under Curve	Max Overall Accuracy %	Max Average Accuracy %	Custom Accuracy %	Model Accuracy %	Algorithm	Creation Date
🟩 CLAS_DT_1_9	0.88245881	99.9169	89.3340	0.0000	99.9169	Decision Tree	4/18/18 7:01 AM
🟧 CLAS_NB_1_9	0.95181400	99.8408	89.5101	0.0000	98.8605	Naive Bayes	4/18/18 6:58 AM

Figure 6: Compared performances of algorithms and ROC curve

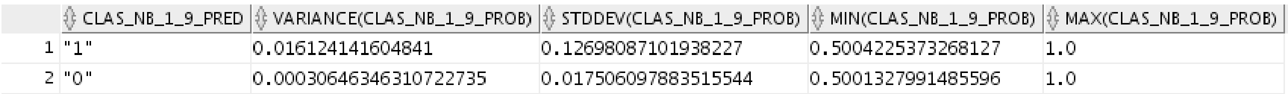
In this example, Naive Bayes algorithm (orange bar) and Decision Tree algorithm (green bar) are almost equally precise, with less than a percent of difference. Because of that, it is hard to tell which of these two algorithms will fare better in classification of new data. This is where Oracle analytic functions come in handy. Each of trained algorithms can be separately applied on test data set (as shown in Figure 3), and results of each appliance can be stored in Oracle tables. These tables by default contain information about predicted class for each row and probability with which the algorithm designated class for a given row. Now, wide range of queries can be applied on created tables. For example, we can compare variance and standard deviation for prediction probabilities of both of algorithms, and minimal and maximal probability can also be of interest.

In last step described in Figure 3, a **choice** must be made which algorithm suits better for a given problem. After using VARIANCE, STDDEV, MIN and MAX functions to examine mentioned probabilities, we can conclude that that Naive Bayes algorithm, although a percent less accurate than Decision Tree algorithm, has significantly less variance and standard deviation when it comes to predicting positive fraud cases (class with label “1”). We can also note that Naive Bayes algorithm have minimal probability of approximately 0.5 against approximately 0.007 for Decision Tree algorithm. Based on these analyses, conclusion can be made that Naive Bayes algorithm is far more confident when it comes to credit card fraud detection.



	CLAS_DT_1_9_PRED	VARIANCE(CLAS_DT_1_9_PROB)	STDDEV(CLAS_DT_1_9_PROB)	MIN(CLAS_DT_1_9_PROB)	MAX(CLAS_DT_1_9_PROB)
1	"1"	0.08236722401694024	0.2869969059361795	0.007309941520467836	0.8859649122807017
2	"0"	0.000000013752713951930887	0.00011727196575452672	0.999424405218726	1.0

Figure 7: Analyses of prediction probability for Decision Tree algorithm



	CLAS_NB_1_9_PRED	VARIANCE(CLAS_NB_1_9_PROB)	STDDEV(CLAS_NB_1_9_PROB)	MIN(CLAS_NB_1_9_PROB)	MAX(CLAS_NB_1_9_PROB)
1	"1"	0.016124141604841	0.12698087101938227	0.5004225373268127	1.0
2	"0"	0.00030646346310722735	0.017506097883515544	0.5001327991485596	1.0

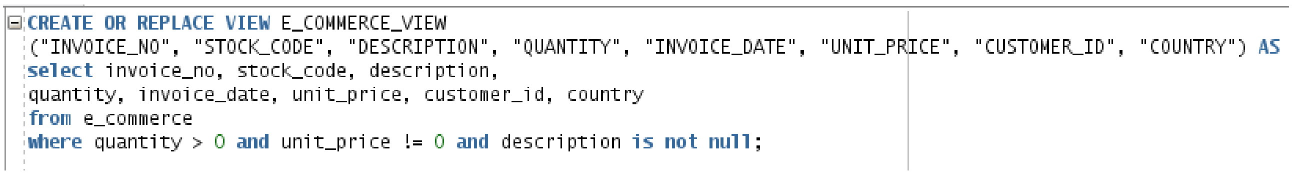
Figure 8: Analyses of prediction probability for Naive Bayes algorithm

4.2. CLUSTERING EXAMPLE

Among other models, Data Miner also supports clustering. In this section will be described process of **filtering the data,** since the rest of the steps are the same as in classification example.

Acquired datasets usually contain inconsistent data, which should be eliminated before data mining algorithms are trained on them. Examples of inconsistent data include negative quantities of products, negative prices, missing values for identifiers and other important fields and so on. This is where Oracle analytic functions might become useful again. Using these functions, anomalies in datasets can be easily identified and eliminated. Another advantage of using Oracle relational database management system is that there is no need to delete inconsistent rows from the original table. Simply, a view can be created with filtered data, and this view will be later used for clustering. That way, original data can be kept, and data mining algorithms can be applied on filtered data.

Dataset used for clustering contains e-commerce data, such as data about products bought, their prices, quantity of bought products, date of purchase, and so on. Analyzing the data with MIN and MAX functions, we can detect negative values for quantity and price of purchased products and also missing values for description of products. Since these rows make only a small percentage of the entire dataset (about 0.03%), they can be eliminated without fear that this will jeopardize the training process. Now, a view can be created that will contain only valid data, and it will be used for clustering algorithms.



```
CREATE OR REPLACE VIEW E_COMMERCE_VIEW
("INVOICE_NO", "STOCK_CODE", "DESCRIPTION", "QUANTITY", "INVOICE_DATE", "UNIT_PRICE", "CUSTOMER_ID", "COUNTRY") AS
select invoice_no, stock_code, description,
quantity, invoice_date, unit_price, customer_id, country
from e_commerce
where quantity > 0 and unit_price != 0 and description is not null;
```

Figure 9: A view containing filtered data

Clustering node also requires data source for its execution, and three algorithms are available: Expectation Maximization, K-Means and O-Cluster, an Oracle proprietary algorithm similar to hierarchical clustering. In this example K-Means and O-Cluster will be compared. After the selection of algorithms, desired number of clusters and other parameters, clustering process can be started. When it is done, received clusters from both algorithms can be examined with information such as number of instances in each cluster, information about centroid, percent of all instances in given cluster, etc. Both algorithms divided rows in five clusters, as can be seen in pictures.

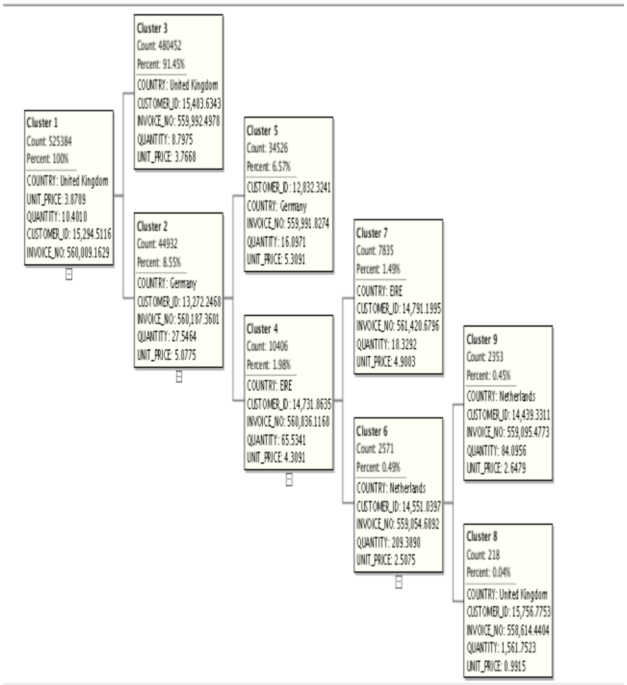


Figure 11: Clusters obtained using O-Cluster algorithm

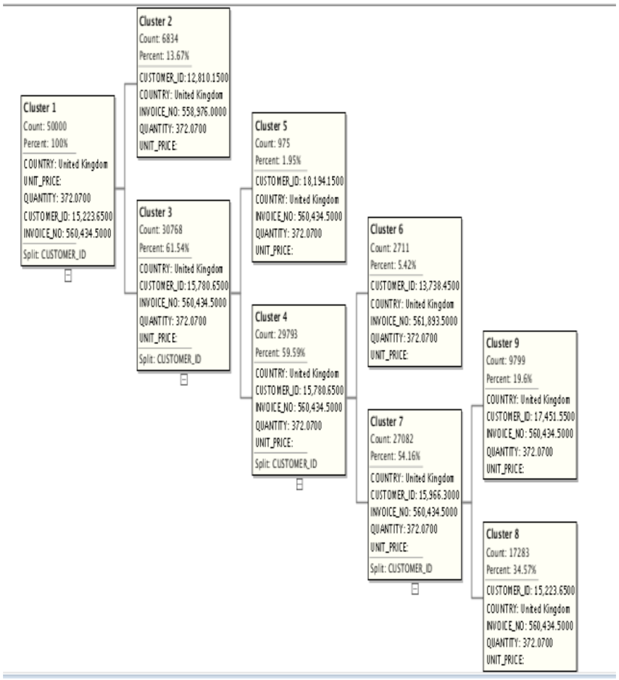


Figure 10: Clusters obtained using K-Means algorithm

One of, or both algorithms, can be used for clustering of a new data source, and results of that operation can be saved in separate table as described in classification.

5. CONCLUSION

There has been a rapid growth in Data Mining usage in different areas like business, education, medicine, science, etc. It is not possible to examine all the data that users generate through various applications in these areas, therefore it is crucial to use different algorithms and techniques for their processing and making certain conclusions from them.

In this paper, the accent was on the very process of Data Mining and selection of the most appropriate algorithm for given domain. Being one of the most rapidly developing disciplines, Data Mining has been integrated in all aspects of IT world, using their advantages to produce more effective solutions, generated faster. One of such examples is using Data Mining algorithms in combination with relational databases and all the benefits they provide. These benefits can be used to complement and aid mentioned algorithms and eliminate eventual flaws. There are numerous ways, perhaps yet undiscovered, to achieve this, and using analytic functions to additionally analyze Data Mining algorithms is just one of them.

The process of applying certain algorithm on dataset can be divided in steps, and largest part of these steps is independent of the type of algorithm applied. This implies that it can be abstracted to some degree, which could lead to its automation. This could further lead to acceleration of Data Mining process, and its standardization. This also represent course of further development of the subject depicted in this paper.

REFERENCES

Alpaydin, E. (2014). *Introduction to Machine Learning.* MIT Press.

Analytic Functions. (2018, April 20). Retrieved from Oracle-base: https://oracle-base.com

Becker, S. (2001). *Data Warehousing and Web Engineering.* IGI Global.

Data Mining Concepts. (2018, April 11). Retrieved from Oracle: http://www.oracle.com

De Mauro, A. G. (2016). A formal definition of Big Data based on its essential features. *Library Review, 65(3)*, 122-135.

Dokeroglu, T., Ozal, S., Bayir, M. A., Cinar, M. S., & Cosar, A. (2014). Improving the performance of Hadoop Hive by sharing scan and computation tasks. *Journal of Cloud Computing, 3(1), 12*.

Elgendy, N., & Elragal, A. (2016). Big data analytics in support of the decision making process. *Procedia Computer Science, 100*, 1071-1084.

Ghazi, M. R., & Gangodkar, D. (2015). Hadoop, MapReduce and HDFS: a developers perspective. *Procedia Computer Science, 48*, 45-50.

Gupta, G. K. (2014). *Introduction to data mining with case studies.* PHI Learning Pvt. Ltd.

Jeffrey Dean, S. G. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM, 51(1)*, 107-113.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets.* Cambridge university press.

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2012). *Foundations of Machine Learning.* MIT Press.

Rish, I. (2001). *An empirical study of the naive Bayes classifier.* New York.

Thearling, K. (2017). *An introduction to data mining.*

Venner, J. (2009). *Pro hadoop.* Apress.

Yahya, O., Hegazy, O., & Ezat, E. (2012). An efficient implementation of a-priori algorithm based on hadoop-MapReduce model. *International sjournal of Reviews in Computing, 12*.

BIG DATA ANALYSIS IN SOCIAL MEDIA

Jelena Ljubenović*, Ognjen Pantelić, Ana Pajić Simović[1]
[1]University of Belgrade, Faculty of organizational sciences, Serbia
*Corresponding author, e-mail: jelena.ljubenovic@fon.bg.ac.rs

Abstract: *The Internet community has changed a lot in the past few years. Social media has changed the way of communication and connected millions of people. With the increase of information that is being generated in these communication channels, there is a new problem: storing all the data. Relational databases are no longer most suitable storage for social media data. Two technical entities have come together. First is big data for massive amounts of data. Second, there's advanced analytics, which is a collection of different tool types for getting meaningful information from the data. Together they make big data analytics, a new practice in BI today.*

This paper presents big data, social media, and social media data infrastructure. It also shows tools for analyzing big data. The contribution of this paper is to provide an overview of tools and techniques for analyzing social media data either for research or business purposes.

Keywords*: big data, big data analysis, social media, social media data, social media data infrastructure*

1. INTRODUCTION

Information overload and the limits of human cognition in dealing with information are much older than new technologies and Internet network. As hardware technology has improved, software technology has improved as well. It's never been that easy to communicate and stay connected with other people. But, the information and data overload have also increased. A brand new way of communication and new technologies have brought massive amounts of various data and a new problem: How to store it? The answer is big data. It is not regular, structured data and can be explained through three dimensions: volumes, variety, and velocity.

Companies are recognizing the potential value of this data and they are putting great efforts in order to extract some useful and meaningful information from it. The only way to derive value from big data is the use of analytics. It must be analyzed and the results should be used by decision makers and organizational processes in order to generate value. The type of information extracted from social media is constantly in motion. Analyzing social media data can contribute to better understanding people's behavior, their needs and habits. New data sources bring new challenges and new challenging questions.

2. BIG DATA

Big data is a very popular buzzword used by IT companies to increase their sales. However, it is more than just a word. It is real and extremely important technology trend with huge business potential. Big data can often be associated with social media. Rather, it is a combination of data-management technologies that have evolved over time. Big data enables organizations to store, manage and manipulate vast amounts of data. Over time, size and type of data that needs to be stored have changed and new ways of storing data appeared. Big data does not represent regular, structured data. It can be explained through three dimensions: volumes, velocity, and variety, so-called three Vs (Xiang, Du, Ma, and Fan, 2017).

Volume is also known as the big part. It represents the amount of data that is being generated in everyday use. As we generate more and more data, this point of big data will keep growing. Some experts consider that the size of memory that is used for storing data is measured in Petabytes.

Velocity represents the speed of generating new data or the frequency of delivering data. It shows how fast that data is processed. Big data improves real-time data analyzing and enables fast access to a large amount of data.

Variety is about the different type of data and file types available in everyday use. For example there are tweets (500 million per day), video files on YouTube (3.25 billion hours of watching video each month), posts on Facebook (510,000 per second), profiles on LinkedIn (over 5 millions) and then there are also all the different log files and other data produced by any computer system we use. So the explanation of big data could be that it is dealing with high-velocity and high-volume data streams while data types are highly diverse.

Big data requires methods and tools that can be used for analyzing and extracting patterns from large-scale data. The large Volume of data poses a challenge to conventional computing environments. It requires scalable storage and a strategy to data querying and analysis (Choi, Chan, and Yue, 2017). The Volume of data is also a major advantage of big data. Dealing with the Variety among different data is a unique challenge for Big Data, which requires preprocessing of unstructured data in order to extract structured information. The importance of dealing with Velocity in big data is the quickness of the feedback, translating data input into usable information (Bello-Orgaz, Jung, and Camacho, 2016).

3. SOCIAL MEDIA

Social media was originally intended for people who want to interact with friends and family but later many companies started using it as a very powerful tool for communication with customers. Now we can define it as a technology based on computers and Internet that is meant for facilitating the sharing of information and ideas. A huge advantage that social media has is the ability to connect with anyone and share information as long as they also use social media. Some of the most popular social media services are Facebook, Twitter, LinkedIn, Instagram, Pinterest, YouTube etc.

Twitter is a real-time service that allows users to post short messages of 140 or fewer characters. Messages posted on Twitter are called tweets. The network infrastructure of “friends“ and “followers“ on Twitter is asymmetric. User’s friends are the accounts that user is following and user’s followers are the accounts that are following the user. Very important dimension of Twitter is the ability to analyze and track user activity. Retweet function is a form of data sharing. It is a way for Twitter users to participate in discussion without starting it. Twitter offers Apps for various mobile phones and tablets. Interactions or integrations with Twitter are done using the Twitter APIs (Newman, 2017).

Facebook is social networking site. Originally it was designed by Mark Zuckerberg for college students at Harward University. Today, Facebook is the largest social network with more than 1 billion users all over the world. Facebook allows users to send messages, post pictures, videos, links, share information and like content posted by other users.

LinkedIn is a social network for business people. LinkedIn started in 2003 and had only 2,700 members the first week. Today, it has more than 350 million members. It provides a way to connect with other professionals. LinkedIn is used for exchanging knowledge, ideas, as well as finding a job.

4. SOCIAL MEDIA DATA INFRASTRUCTURE

Next three chapters show data analytics infrastructure at Facebook, LinkedIn, and Twitter.

4.1. DATA INFRASTRUCTURE AT FACEBOOK

There are two sources that Facebook collects data from. These are federated MySQL tier and web servers tier.Web servers generate event-based log data and the data is collected to Scribe servers which are executed in Hadoop clusters. The Scribe servers aggregate log data, which is written to Hadoop Distributed File System (HDFS) (Kamar Sahu, 2015). Data from HDFS is compressed and transferred to Production Hive-Hadoop clusters for processing. Federated MySQL tier contains user data.

There are two different clusters for data analysis. Tasks that have a higher priority are executed in the Production Hive-Hadoop cluster. Tasks with lower priority and ad hoc analysis tasks are executed in Ad-hoc Hive-Hadoop cluster. Data from Production cluster is replicated to the Ad hoc cluster. Data analysis results are saved in Hive-Hadoop cluster or to MySQL tier. The graphical user interface (HiPal) or Hive command-line interface (Hive CLI) are used for specifying queries for Ad hoc analysis.

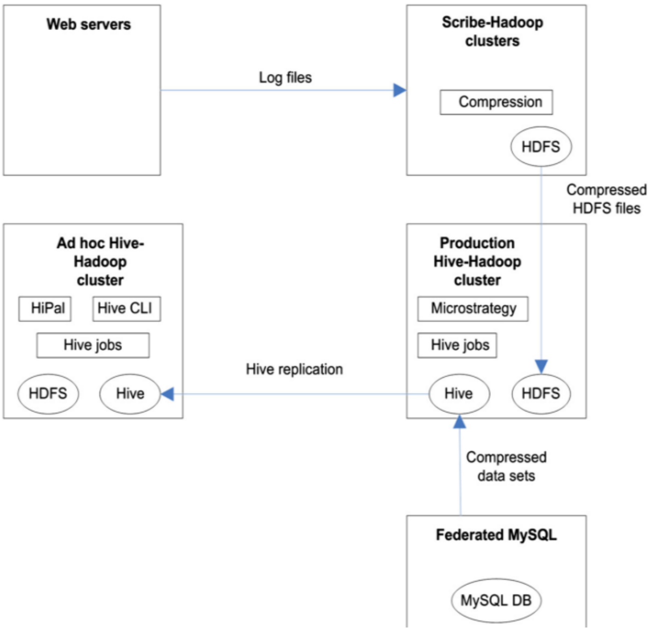


Figure 1. Data analytics infrastructure at Facebook
(Source: Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems)

4.2. DATA INFRASTRUCTURE AT LINKEDIN

LinkedIn started as a single monolithic application called Leo. As the site began to grow, it was necessary to "Kill Leo" and break it up into many services. As the amount of data that needed to be collected also grew, LinkedIn developed many custom data pipelines for streaming and querying data. The site needed to scale and each pipeline needed to scale. The result was the development of Kafka, distributed messaging platform which is used for collection of streaming events.

Data is collected from two sources: activity data that includes streaming events based on usage of LinkedIn's services and database snapshots. Kafka producers report events to topics at a Kafka broker, and Kafka consumers read data at their own pace. Kafka's event data is transferred to Hadoop ETL cluster for further processing. Data from the ETL cluster is copied into production cluster and development cluster.

Azkaban is a workload scheduler for supporting various types of tasks. It is realized as MapReduce, Pig, Hive jobs or shell script. Workloads are tested in the development cluster and transferred to production after testing. Analysis results are moved to an offline debugging database or to an online database. It can also be back to Kafka cluster.

Avatara is used for the preparation of OLAP data. Analysed data is read from the Voldemort database, pre-processed, and aggregated for OLAP, and saved to another Voldemort read-only database (Clemm J. 2015).

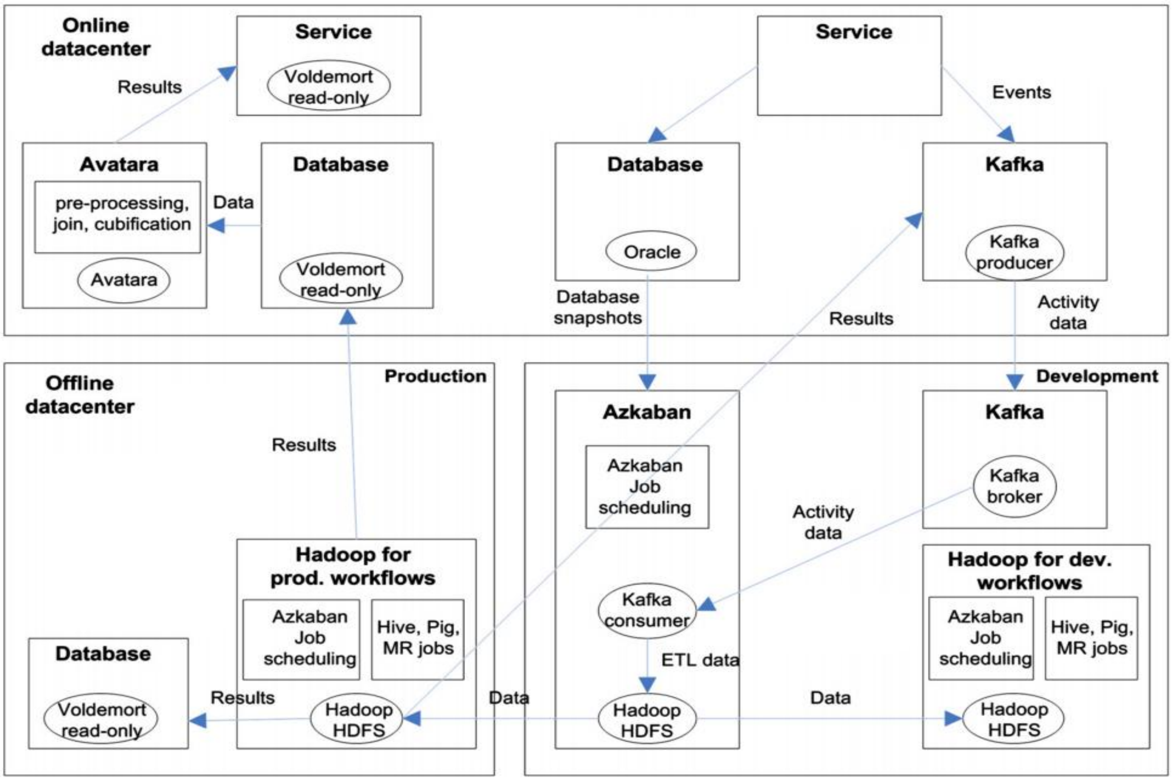


Figure 2. Data analytics infrastructure on LinkedIn
(Source: Pääkkönen, P., & Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems)

4.3. DATA INFRASTRUCTURE AT TWITTER

Twitter has published two different implementation architectures of their infrastructure: one based on Hadoop batch processing which processes streaming data and the other custom-made solution that meets real-time requirements.

In infrastructure for real-time services, all requests (searching for tweets or user accounts via a QueryHose service) coming from Twitter are managed by Blender. The EarlyBird is a real-time engine designed for providing low latency and high throughput for search queries. FireHose service uses tweets as an input. After receiving tweets, EarlyBird servers filter data, personalize and index data. The EarlyBird servers also serve incoming requests from the QueryHose/Blender.
When tweet or a query is served, search assistance engine saves statistics into one of three in-memory stores. Session store saves user sessions, Query statistics store saves individual queries information, and Query co-occurrence store saves information about pairs of co-occurring queries. The results of the analysis are saved into Hadoop HDFS. In the end, Front-end cache takes results of analysis from the HDFS and presents them to Twitter users.

Twitter has three streaming sources from which data is extracted: tweets, Updater, and queries. REST API transforms tweets and queries in JSON format. Data format from Updater is not known. Blender and ingestion pipeline work as Stream temp data stores. The data is being processed. EarlyBird servers contain processed stream-based data (Stream data store). Hadoop HDFS storing the analysis results is modeled as a Stream analysis data store. Front-end cache (Serving data store) provides data to a Twitter app.

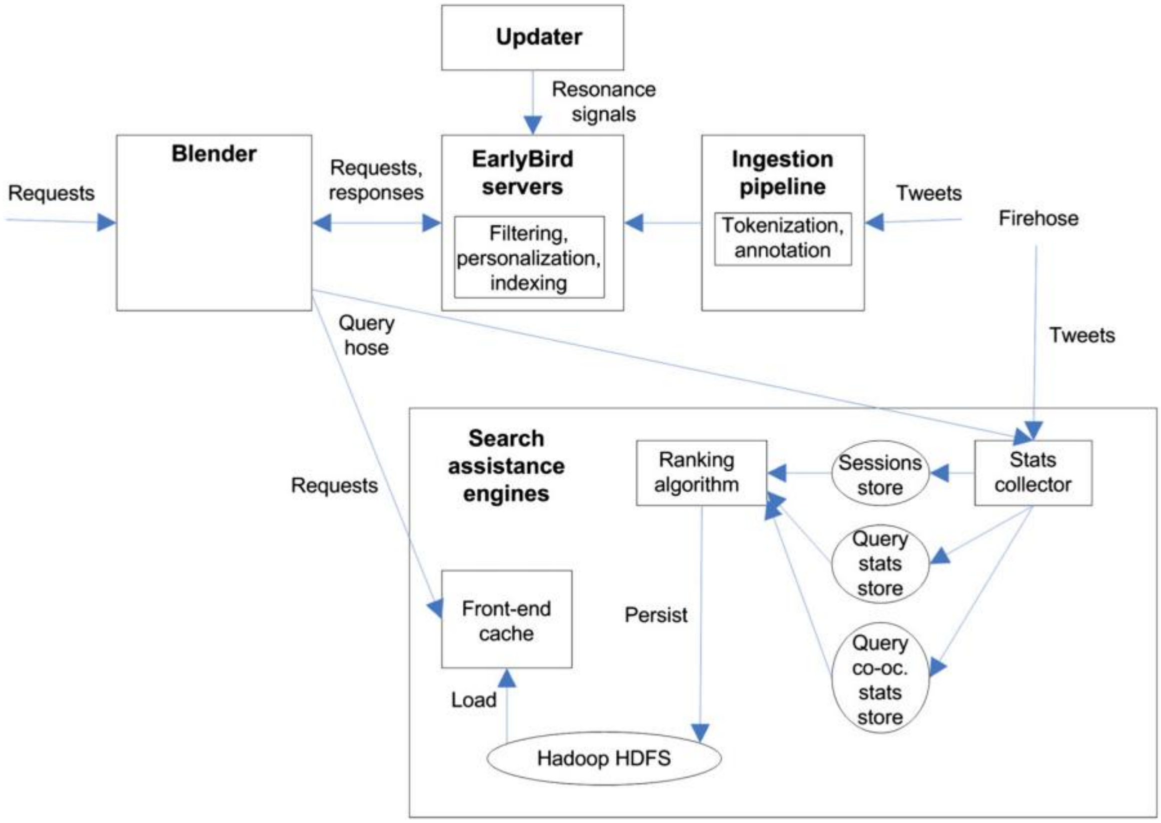


Figure 3. Data analytics infrastructure on Twitter
(Source: Pääkkönen, P., &Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems)

5. ANALYSING BIG DATA IN SOCIAL MEDIA

Social media store a great amount of data. Underneath the surface of a Facebook profile, Facebook page, Instagram business profile or just tweets on Twitter, there is a huge information potential that is still just mostly unstructured data. Social media analysis is the analysis of structured and unstructured data from its channels.

At the moment, data in social media is available either via simple routines or require analysis in some research programs that use programming languages such as MATLAB, Java or Python. Researchers require:
• **Analytics dashboards** — non-programming interfaces for ‘deep’ access to ‘raw’ data.
• **Holistic data analysis** — tools required for combining multiple social media and other data sets.
• **Data visualization**—researchers also require visualization tools so information can be visualized in some schematic form.

There is a number of tools for big data text analyzing that can be used for analyzing social media data. Attensity’s software is one of them. Attensity is one of the original text analytics companies. It uses a Hadoop framework (MapReduce, HDFS, and HBase) to store data. Another text analytics vendor is Calarbridge. Clarabridge CX Social is its tool specialized for exploring social media data.IBM offers several solutions such as Watson, SPSS, IBM Content Analytics with Enterprise Search (ICAES). IBM BigInsights for Apache Hadoop is an industry standard Hadoop. It is a data management platform. SAS is also solving big data problems. The SAS High Performance Analytics Server is an in-memory solution that allows a user to analyze complete data. It uses Hadoop Distributed File System (HDFS).

There are also custom applications for big data analysis. The purpose of these applications is to increase the speed of decision making or action taking. For example, the “R” environment is based on the “S” statistics and analysis language. It is an integrated suite of software tools and technologies designed to create applications that are used to ease data manipulation and analysis. It supports effective data-handling, operations for arrays and other ordered data types and tools to a wide variety of data analyses.

The Google Prediction API is an example of an emerging class of big data analysis application tools. It functions by looking for patterns and matching proscriptive, prescriptive or the other existing patterns. While doing this matching, it also learns. The source can be postings from Facebook, Twitter, Amazon, LinkedIn and the goal finding specific patterns of behavior. That can be very useful in business for a consumer products company. Based on the gathered information, the company might launch a new product or upgrade the old one. It is implemented as a RESTful API and supports .NET, Java, PHP, JavaScript, Python, Ruby, and many other programming languages. It also provides scripts for accessing the API as well as a client library for R.

Hadoop is an open source framework which is designed to solve problems associated with distributed data storage, analysis, and retrieval of big data. A distributed file system (HDFS) stores data and replicates it so it is always available. MapReduce is a distributed processing system for parallelizable problems. The idea is to design map functions that are used for generating a set of key/value pairs after which the reduce function will merge all the values associated with the same key. In the first step (Map), a problem is divided into many small problems and sent to servers for processing. In the second step (Reduce) the results of the previous step are combined to create the final results of the problem. Reduce can't begin until all the mapping is done, and it isn't finished until all instances are complete. The output of mapping and reduce are key-value pairs. Hive is a data warehouse system in which the user can specify instructions convert them to MapReduce tasks. Pig is another member of Hadoop family. It has similar functions like Hive, but it uses a programming language called Pig Latin, which is more oriented to data flows. HBase is another component of the Hadoop ecosystem, which implements Google's BigTable data store. Bigtable is a multidimensional sorted map. Elements in the map are an array of bytes. They are indexed by a row key, a column key, and a timestamp. The official Hadoop project is Apache. Hadoop solutions are also offered by companies such as Cloudera and Hortonworks. There is also a company MapR. It offers a commercial implementation of Hadoop (Zadrozny and Kodali, 2013).

6. COMPARISON OF SOCIAL MEDIA DATA INFRASTRUCTURE

With large incoming data and the fact that more and more data needs to be properly stored in order to allow data analysis, many social media platforms use Hadoop tools. Hive-Hadoop is an ad hoc cluster. It is a data warehousing framework built on Hadoop. It was created by Facebook and then given to Hadoop as a subproject of Hadoop ecosystem. Hive allows users to query and analyze data using SQL which makes data processing easier. In LinkedIn event data from Kafka cluster is transferred to Hadoop ELT (Extract – Transform - Load). ELT on Hadoop is a data integration process that provides flexibility in data processing environment. Also, Azkaban, a workload scheduler for supporting various types of tasks, is realized as MapReduce, Pig, Hive jobs or shell script. All the three social media networks, Facebook, Twitter and LinkedIn use Hadoop HDFS typically for storing structured data or storing the results of the analysis.

LinkedIn data is collected from two sources: activity data and database snapshots. Facebook also collects data from two sources. From federated MySQL tier it collects structured and stream-based user data. From web servers tier it generates event–based log data. The data is applied for batch-based data analysis. Every time user sends a Facebook message, add a contact on LinkedIn or tweets on Twitter digital data is being generated. These platforms are able to gather real-time data. All these events are captured, streamed and processed in digital form as they occur (Pigni, Piccoli, and Watson, 2016). Twitter uses a new stream processing system called Heron. It provides significant performance improvements and other advantages such as debugging-ability, manageability, and scalability. If something goes wrong, the design of Heron makes it transparent as to which part is failing. Each Heron Instance is executing a single task, so it is easier to debug that instance (Kulkarni, Bhagat, Fu, Kedigehalli, Kellogg, Mittal, and Taneja, 2015).

Relational databases are still applied for storing important user data as MySQL for Facebook and Oracle for LinkedIn. For storing data analysis results NoSQL databases or in-memory stores are used (LinkedIn's Voldemort). Log facilities are used for storing stream-based data. For Facebook it is Scribe, For LinkedIn Kafka. Technologies for data processing can be classified as a batch and stream processing. Real-time jobs require special technologies and algorithms (Twitter's ranking algorithm and the EarlyBird architecture). Batch processing is used for jobs with less strict timing requirements (Facebook's and LinkedIn's MapReduce, Hive and Pig scripts). Jobs for batch processing in LinkedIn are scheduled with Azkaban, in Facebook's with Databee. Processed data can also be presented with commercial Business Intelligence tools. Facebook uses MicroStrategy (Pääkkönen and Pakkala, 2015).

7. CONCLUSION

Big data analysis is a very powerful way to get right insights, information, and meaningful data from a huge volume of various data on social media platforms. It is now easier to collect data than it has ever been before. But extracting and utilizing useful information is not easy at all. From the business viewpoint, online reviews, photos, and the reviewer's personal information are very important in order to understand and influence user behavior. Social media isn't just used by marketers to find out about the position of their brand on the market or about customers opinion. It is used by healthcare institutions to discover health threats all around the world. The government is using it to predict terrorist attacks. Business intelligence can benefit from big data analytics. Social media data hides great informational potential. With the right tools and the right techniques, data can be transformed into valuable information.

As big data evolves, software vendors are competing to make the best tool for all sorts of analysis. There are numbers of programs designed to meet the needs of data analyzing. It's up to analyzer to choose the one that is most suitable for the given problem.

In this paper, various definitions of big data and social networks have been presented. A primary focus has been on big data analytics tools and social media data infrastructure for three different social media platforms: Facebook, LinkedIn, and Twitter. The paper can be a good starting point for further social media data exploring.

REFERENCES

Bello-Orgaz, G., Jung, J. J., & Camacho, D. (2016). Social big data: Recent achievements and new challenges. *Information Fusion*, *28*, 45-59.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51-65.

Newman, T. P. (2017). Tracking the release of IPCC AR5 on Twitter: Users, comments, and sources following the release of the Working Group I Summary for Policymakers. *Public Understanding of Science*, *26*(7), 815-825.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., &Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Zadrozny, P., &Kodali, R. (2013). *Big data Analytics Using Splunk: Deriving Operational Intelligence from Social Media, Machine Data, Existing Data Warehouses, and Other Real-Time Streaming Sources*. Apress.

Russell, M. A. (2013). *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*." O'Reilly Media, Inc.".

Clemm, J. (2015), A Brief History of Scaling LinkedIn, retrieved from https://engineering.linkedin.com/architecture/brief-history-scaling-linkedin

Kumar Sahu, B. (2015), Big data Analytics Reference Architectures - Big data on Facebook, LinkedIn and Twitter, retrieved from https://www.linkedin.com/pulse/big-data-analytics-reference-architectures-facebook-sahu.

Thusoo, A., Shao, Z., Anthony, S., Borthakur, D., Jain, N., SenSarma, J., ...& Liu, H. (2010, June). Data warehousing and analytics infrastructure at facebook.In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 1013-1020).ACM.

Pääkkönen, P., &Pakkala, D. (2015).Reference architecture and classification of technologies, products and services for big data systems. *Big data Research*, *2*(4), 166-186.

Gandomi, A., &Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144.

Batrinca, B., &Treleaven, P. C. (2015). Social media analytics: a survey of techniques, tools and platforms. *Ai & Society*, *30*(1), 89-116.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter*, *19*(4), 1-34.

Borthakur, D., Gray, J., Sarma, J. S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., ...& Schmidt, R. (2011, June). Apache Hadoop goes realtime at Facebook. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (pp. 1071-1080).ACM.

Watson, H. J. (2014). Tutorial: Big data analytics: Concepts, technologies, and applications. *CAIS*, *34*, 65.

Das, T. K., & Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. *International Journal of Engineering Science & Technology*, *5*(1), 153.

Choi, T. M., Chan, H. K., &Yue, X. (2017).Recent development in big data analytics for business operations and risk management. *IEEE transactions on cybernetics*, *47*(1), 81-92.

Kulkarni, S., Bhagat, N., Fu, M., Kedigehalli, V., Kellogg, C., Mittal, S., ...&Taneja, S. (2015, May). Twitter heron: Stream processing at scale. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data* (pp. 239-250).ACM.

Pigni, F., Piccoli, G., & Watson, R. (2016). Digital Data Streams: Creating value from the real-time flow of big data. *California Management Review*, *58*(3), 5-25.

QUERY PROCESSING ASPECT IN HETEROGENEOUS DBMS

Sofija Prokić*, Jelena Ljubenović[1]
Univerzitet u Beogradu, Fakultet organizacionih nauka[1]
*Corresponding author, e-mail: sofija.prokic@fon.bg.ac.rs

Abstract: *This paper presents the method of realization of query processing in two heterogeneous database management systems. There is NoSQL and SQL query processing aspect that are integrated into different ways depending on the architecture. Integrated results are displayed to the user. In HQPS, Resource Description Framework (RDF) is used for data integration from heterogeneous databases. The CloudMdsQL uses a functional SQL-language, capable of querying multiple heterogeneous data stores.*

Keywords: *NoSQL database, relational database, HQPS system, CloudMdsQL system*

1. INTRODUCTION

Until a few years ago relational databases played a leading role in storing data. They provide great precision, consistency, and availability of data. Relational databases store data using SQL query language (Bergamaschi, Guerra, Interlandi, Trillo-Lado, & Velegrakis, 2016). Characteristic for the SQL database is that they are based on ACID properties (Đurđevac, 2012). The ACID properties are the acronym for:

- **A**tomicity
- **C**onsistency
- **I**solation
- **D**urability

The ACID model is in the database layer, which means that all operations that are performed must be executed in one transaction and all operations are performed at the database level. However, relational databases are not the best solution for storing a large amount of data or when it is necessary to manipulate data at high speed. Then NoSQL databases appeared. NoSQL do not use the SQL query language, but support it. NoSQL databases are much more flexible than relational because they do not have a strictly defined schema in data storage and consume much fewer resources. Also, the NoSQL databases sacrifice the consistency that is the property of ACID in order to provide better research performance. NoSQL is short for Not Only SQL (Oluwafemi, Sahalu, & Abdullahi, 2016).
These databases support a set of BASE features that are the acronym for:

- **B**asically **A**vailable
- **S**oft state
- **E**ventually consistent

The BASE property model is in the application layer and it is good for use in storing data on the web, where it is necessary not to block transactions, while others are still not executed. NoSQL databases work with structured, semi-structured and unstructured data (Nikolaus, 2015). Eric Brewer (Brewer, 2012) has devised CAP theorem in response to the fact that there is a conflict in the availability of data in distributed systems. The CAP theorem shows that in distributed systems cannot be simultaneously available more than two of the following properties: (Nayak, Poriya, & Poojary, 2013)

- **C**onsistency
- **A**vailability
- **P**artition tolerance

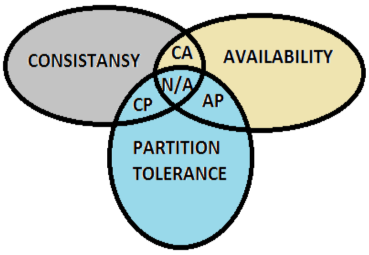


Figure 1: *CAP theorem*

The second chapter gives a brief overview of the types of NoSQL databases. The third chapter deals with the integration of SQL and NoSQL data; the architecture view shows how the data is transformed from one form to another and through which layers they pass through the transformation. The fourth chapter presents two HQPS and CloudMdsQL architectures that analyze the query and give a plan for its execution. Mapping functions between different data formats are also given. In the fifth chapter, we give our vision for further research on heterogeneous systems, in order to improve and maximize their performance.

2. TYPES OF NOSQL DATABASES

1. Key/value database

The key/value do not have a strictly defined schema. This model can be presented as a single table in a relational model with two columns, where one column represents a key and the other value. Three operations that can be performed over the database are: PUT, GET and DELETE (Luković, 2015).

2. Document oriented databases

The central part of these databases is a document. With such databases, each JSON document represents a separate object.

3. Column oriented database

With this database, the data is arranged by columns, unlike RDBMS, where data is stored in rows. Such warehouses do not require a complete schema, but pre-defined column families, which are mutually independent. Apart from the concept of a family of columns, there is a notion of super columns. A super column may contain other columns, but not other super columns.

4. Graphic databases

Graphic databases are based on graph theory, using graphic structures with nodes (entities), branches (connecting nodes), and properties (attributes) for displaying and storing data. Each node can have an unlimited number of attributes that describe nodes (Veljković, 2013).

3. PROBLEM OF INTEGRATION OF RELATIONAL AND NOSQL DATABASES

Although in recent years there is rapid growth and development, as well as the popularization of NoSQL databases, there are still situations when it is more convenient to use relational databases. For example, when it is necessary to provide ACID properties that reduce anomalies and protect the integrity of the database by defining the way in which transactions communicate with the database. NoSQL databases sacrifice these properties for flexibility and speed of execution. In NoSQL databases, there is also no inner join or transaction (Varughese & Rajeswari, 2009).

NoSQL databases prevent the occurrence of bottlenecks or congestion due to volume. They also reduce storage costs and easily deploy data to multiple servers, allowing simultaneous access to a large number of different users. The big problem is the selection of relational databases that ensure consistency and usefulness in combination with NoSQL databases that provide benefits in terms of scalability and partition tolerance. The problem occurs when one software product needs to provide data storage, where one part is ideally stored in the NoSQL database, while the other data is ideally stored in a relational database (Adeyi, Abdullahi, & Junaidu, 2013). Indian scientist Ali (Ali, 2009) proposed a solution that creates a global scheme instead of multiple local schemes. One access to all components of a heterogeneous database, or heterogeneous system, is formed. Ali believes that in this way there is an integration of several different local systems, and it can be called MDBS (Multidatabase system). To the end user, this system is presented as a single database that they can use, while on the lower layers it consists of several local databases (Haber, et al., 2015).

The problem that occurs when storing data is that one part is stored in the SQL database, while the other is stored in the NoSQL database, so the user needs to extract data from different sources. To avoid writing a different code for each type of database in which the data is located, the Roijackers suggested using a hybrid database. It requires the construction of an abstraction layer at the top of SQL and the NoSQL database. The abstraction layer is responsible for translating data from NoSQL into a triple format and incorporating the resulting format into an SQL database as a virtual relationship. In this way, using one query language, data can be downloaded regardless of the database, because this layer is responsible for collecting relevant data, but also combining received data in one query (Katsov, 2012).

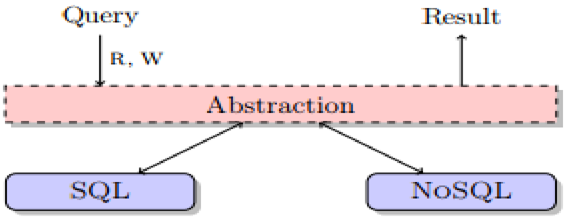


Figure 2: An example of the architecture of the hybrid base proposed by Roijdzakers (Roijackers & Fletcher, 2012).

To virtualize data from different databases (MySQL and MongoDB), RDF is used as a data format for relational and NoSQL data. RDF (*Resource Description Framework*) is a language for displaying resource information and metadata on the World Wide Web. The RDF is designed to store and model information as well as to provide interoperability between applications that exchange information that is understandable for machines on the web. Distributed NoSQL data is in the form of a triple format in the RDF (subject, predicate, object), so that they can be embedded in the SQL database (Thant & Naing, 2014). Triple notation represents arbitrary data that facilitate the work of the hybrid base. The problem that arises with NoSQL data is that there is no standard that can be compared with SQL data from a relational database. All SQL databases use the same SQL query language, while in the NoSQL database, there are different NoSQL query languages that can be used (Roijackers & Fletcher, 2012).

The picture shows that the triple format is represented by F (id, key, value), where different data can be described in triple format. The data to be found in the F relation are derived from the outside world, or from the NoSQL database. All nodes that actually represent NoSQL data must have a unique ID value according to a set of key/value pairs. All keys at the same level of nesting, with the same "parent", must be unique. The combination of id and key is unique, so it refers to one value of NoSQL data (Roijackers & Fletcher, 2012).

Table 1: *Relational representation*

ID	**Name**	**Age**
1	Ana	18
2	Boba	35

Table 2: *Representation of triple format*

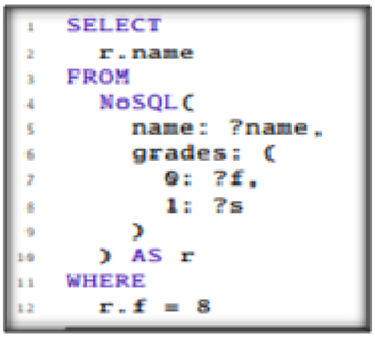
ID	**Name**	**Value**
i1	id	1
i2	Name	Ana
i3	Age	18
i4	Id	2
i5	Name	Boba
i6	age	35

Table 3: Nested data

ID	**Name**	**Value**
i1	name	Boba
i1	grades	i2
i2	0	8
i2	1	6

If NoSQL data is nested within another NoSQL data, a transformation function must be performed so that all nested elements in the triple format can have the same id value. SPARQL is used to represent the NoSQL query. It is the W3C standardized query language for RDF. The most important part of the SPARQL query is the basic graphic representation (Roijackers & Fletcher, 2012). The picture shows the nesting object, the SQL query and the nested NoSQL query within it. The process of translating NoSQL into the corresponding

SQL fragment is shown in the section of the architecture in the following image, which is colored in yellow. The translation should be executed automatically, before executing the query itself, but this transformation still does not represent a final SQL query, it is yet to come (Mijalkovic, 2013).



```
SELECT
  r.name
FROM
  NoSQL(
    name: ?name,
    grades: (
      0: ?f,
      1: ?s
    )
  ) AS r
WHERE
  r.f = 8
```

Figure 3: *SQL query and nested NoSQL query within it.*

Each transformation is unique depending on the type of the NoSQL database, which means that for different NoSQL solutions it is necessary to make new transformations in triple RDF formats. This part of the architecture is shown in the green part of the image. Only after this transformation, the obtained RDF triple format is embedded in the SQL base and is merged with the SQL query set by the user, if the user wrote the SQL query at all. The last blue section represents SQL query that is a result of user's query and NoSQL query which was previously transformed into a triple format. In the end, this new query is sent to the relational database and executed as a normal SQL query. The result is an output made from SQL query as s response to a user's request.

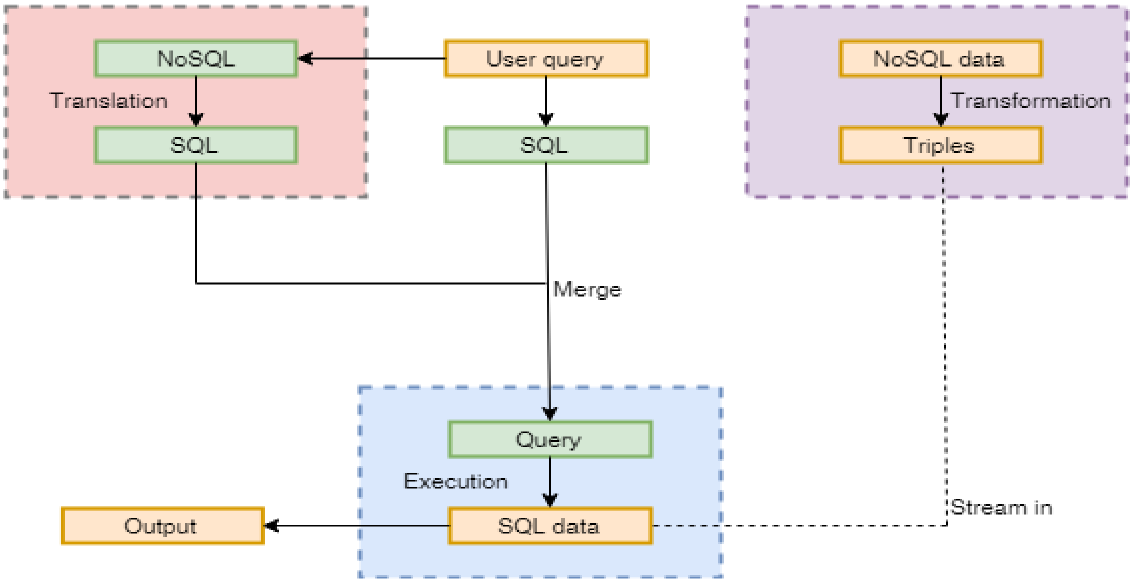


Figure 4: Architecture that illustrates execution and query transformation (Roijackers & Fletcher, 2012).

4. ARCHITECTURE OF INTEGRATED SOLUTION

4.1. Hybrid Query Processing System (HQPS)

Hybrid Query Processing System–HQPS was created to use and manage data from the SQL database, as well as data from the NoSQL database, or to handle both types of user queries.
HQPS consists of two components, which are:

1) Query Analyzer &Generator (Control SQL and NoSQL Queries)
2) Result Preparation (Preparation of results where results are prepared and presented to users)

In the picture below is given a graphical representation of the HQPS system. The user writes a query in Query Analyzer &Generator, then the same query is sent to the SQL or NoSQL query processing unit, depending on the type of query. In query processing units, queries are executed with the help of the database the query belongs to (SQL or NoSQL), and in the end the result is sent to Result Preparation. In Result Preparation integration of data from heterogeneous databases, SQL or NoSQL is performed. Integration is done using the RDF format, and if the data is from the NoSQL database it is translated into a triple RDF format. Then the data is integrated with the data from the SQL database, and the result of integration is sent to the user as a response to a user's query. (Thant & Naing, 2014).

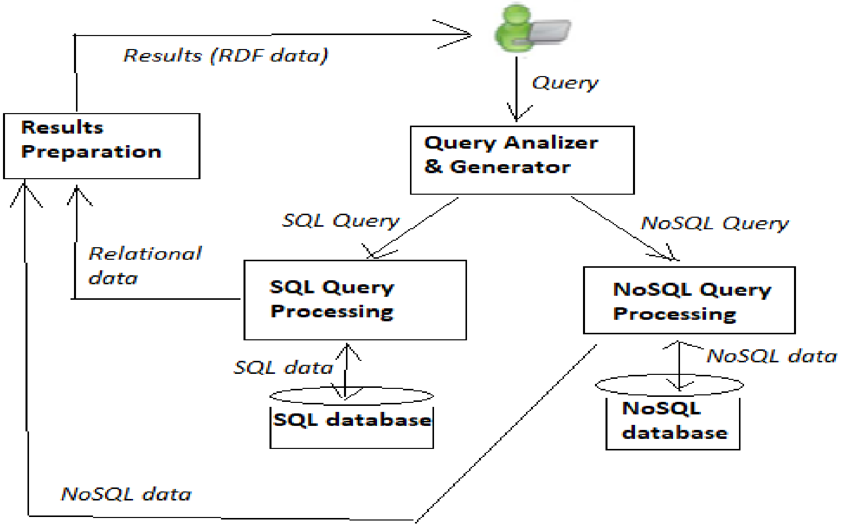


Figure 5: Architecture of the HQPS system

When a query is sent to Result Preparation, it first enters the query layer and then into the virtual database that is in the target layer. From a virtual database, the data is sent to a Relational or NoSQL database depending on the type of query. Sending the response backward goes through the same layers only in the opposite order. In the picture below is given a summary of the Result Preparation architecture.

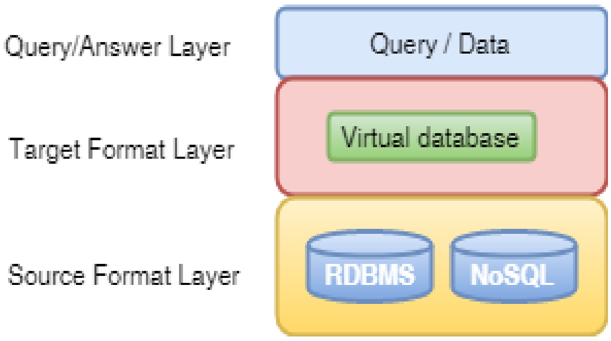


Figure 6: Architecture of Result Preparation (Thant & Naing, 2014).

To integrate data, NoSQL data must be transformed into an RDF triple format, which is easy when the NoSQL data is normalized. However, when the data is nested one within another, a more complex mapping is required. In order to transform the nested key-value structure into a triple format, it is important that all nested elements have the same ID value that will define them and which will be found in triple format.

4.1.1. Mapping between NoSQL and RDF triple format is given in the following lines:

Let

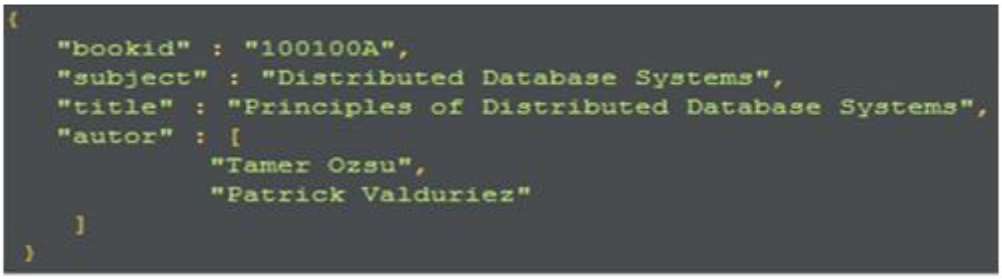
α = transformation function for a key value pair

δ = transformation function for set of key value pairs

i = subscript for given equal id values

$$\propto i(N) = \left\{ \begin{array}{l} \{(i,\ Nk,\ Nv)\},\ if\ \ Nv\ is\ not\ a\ set \\ \{(i,\ Nk,\ j)\cup\delta j(Nv),\ if\ \ Nv\ is\ a\ set\} \end{array} \right\}$$

$$\delta i(S) = \ U\ n \in s\ \ \alpha i(N)$$



```
{
   "bookid" : "100100A",
   "subject" : "Distributed Database Systems",
   "title" : "Principles of Distributed Database Systems",
   "autor" : [
            "Tamer Ozsu",
            "Patrick Valduriez"
    ]
 }
```

$$\delta i\ \ (S)\ \ = \cup n \in s\ \ \alpha i\ \ (N)$$

$$= \ \{(i1,\ BookId,\ 1000100A)\} \cup \{(i1,\ Subject,\ Distributed\ \ Database\ \ Systems)\}$$

$$\cup\{(i1,\ Title,\ Principles\ of\ \ Distributed\ \ Database\ \ Systems)\}$$

$$\cup\{(i1,\ Author,\ i2)\}$$

$$\cup\{(i2,\ 0,\ M.\ Tamer\ \ Ozsu)\} \cup \{(i2,\ 1,\ Patrick\ \ Valduriez)\}$$

Figure 7: An example of a transformation function from NoSQL to a RFD triple format with a nesting document Author (Thant & Naing, 2014).

4.1.2. Mapping between Relation Model and RDF Triple Format:

Let

β = transformation function for a relational record

φ = transformation function for relational records

i = subscript for given equal id values

$$\beta i(R)\ \ = \ \{(\ i,\ Ak,\ Vj\ \ )\forall j\ ,Vj \in Nj\ \}\quad \varnothing i(S) = \cup \beta i(R)$$

Table 4: Example of mapping between Relation Model and RDF triple format

Member ID	Name	Password	Email	Major
000001	MonMon	345mon	mon@gmail.com	Computer Science

$$\begin{gathered}
\varnothing i\ (S)\ = \cup r \in s\ \beta i(R) \\
= \beta i2\ (MemberID,\ 000001) \cup \beta i1\ (Name,\ Mon\ Mon) \\
\cup \beta i1(Password,\ 345mon) \cup \beta i1\ (Email,\ mon@gmail.com) \\
\cup \beta i1(Major,\ Computer\ Science) \\
= \{(i1,\ MemberID,\ 000001)\} \cup \{(i1,\ Name,\ Mon\ Mon)\} \\
\cup \{(i1,\ Password,\ 345mon)\} \cup \{(i1,\ Email,\ mon@gmail.com)\} \\
\cup \{(i1,\ Major,\ Computer\ Science)\}
\end{gathered}$$

Figure 8: An example of a transformation function from SQL to an RFD triple format. *(Source: Hybrid Query Processing System (HQPS) for Heterogeneous Database (Relational and NoSQL))*

4.1.3. Mapping algorithm from NoSQL to RDF and from SQL to RDF format

Using the mapping algorithm, you can see what the inputs are and what is the outcome, as well as all the steps that are executed until an RDF format is obtained. The mapping algorithm from NoSQL to RDF receives the query and db properties as inputs, and as the output has an RDF format. In the first step, a database is loaded, then a query is executed as long as the cursor reports that there are still queries or entities for execution and converting to objects. Converting is done in RDF format (key, name, value) and in the end, the RDF format is obtained.

The mapping algorithm from the relational database to the RDF format is similar. The data from the database is loaded, the query executed and a set of results is obtained. As long as there are results or data in the list, the data from the relational base is converted to RDF format (key, name, value) and the result in RDF format is printed. When it is necessary to return to the user the results of a query from two databases, one relational (MySQL) and one non-relational (MongoDB), the system itself integrates data from MySQL and JSON data from MongoDB and then shows integrated data to the user. The implementation of relational data in RDF mapping is done using a free database called Sakila, a product of MySQL Corporation. For mapping operations, the HQPS system needs to provide SQL expression, while the result will be displayed in the RDF format (Thant & Naing, 2014).

4.2. CloudMdsQL System

CloudMdsQL is a specific DBMS that is categorized as web-based data. CloudMdsQL uses different sizes for displaying different data and databases. It uses relational properties for data exchange, such as ACID, transaction independence, SQL query language. It provides scalability, schemaless databases, and good performance. CloudMdsQL has provided a solution to the problem of data storage in three different databases. It is necessary to write a program that will have access to three different databases through their APIs and which will integrate their solutions.

The query engine is part of the platform that allows deployment over one or more data centers. The picture below shows the architecture of a single data center. The architecture of the query engine is fully distributed, so that query engine nodes can directly communicate with each other. The query engine does not follow the traditional mediator/wrapper architectural model where mediator and wrappers are centralized. Each query node consists of two parts of the master and the workers, and they are located on each data hub in a computer cluster. Every master or worker has a communication processor that supports the sending and receiving of data exchange operators and commands between nodes.

If there are multiple masters, the client chooses one of them to send a query to. A master takes as input a query and produces a query plan, which it sends to one chosen query engine node for execution. The query planner performs query analysis and optimization and produces a query plan serialized in a JSON form. All operations in the plan have the identifier of the query engine node that is in charge of performing it. The operation determines the first worker, which the master should send the query plan to. Query execution controller controls the execution of a query plan by interacting with the operator engine for local execution with one or more workers. The query execution controller will synchronize the execution of the operators that require the intermediate results produced by the distant workers, once they are received back.

Operator engine executes the query plan operators on data retrieved from the wrapper, from another worker, or from the table storage. These operators include CloudMdsQL operators to execute table expressions in the query and communication operators to exchange data with other workers. Table storage provides efficient, uniform storage for result data in the form of tables. Wrapper with its data store interacts through its API to retrieve data, transforms the result in the form of a table, and writes the result in table storage or delivers it to the operator engine.

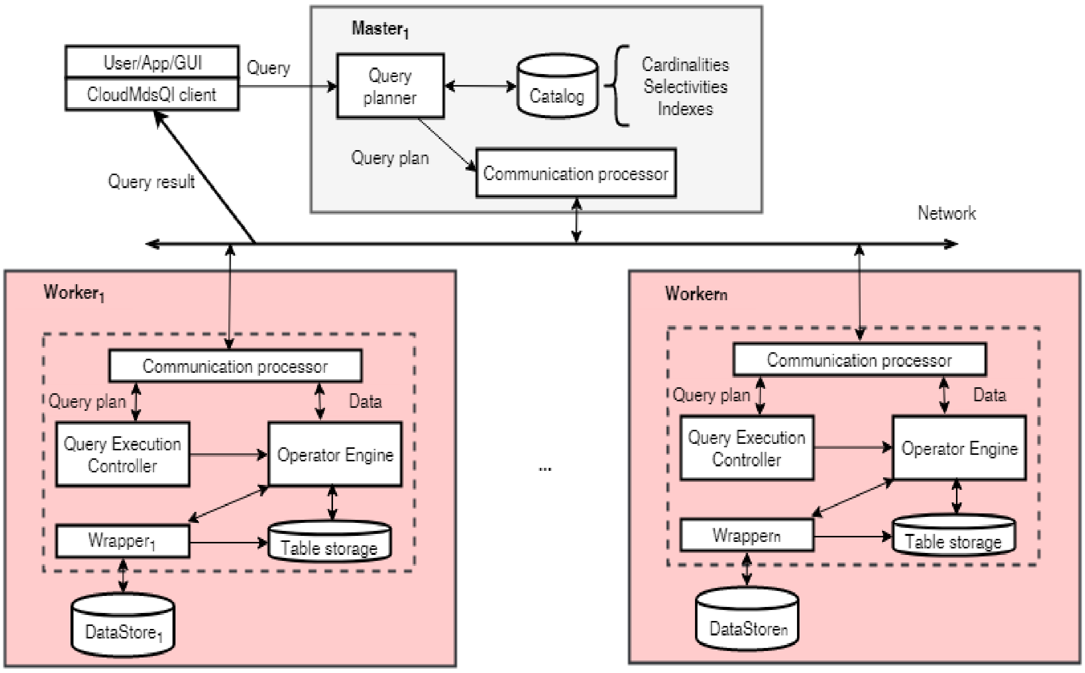


Figure 9: Architecture of the query engine (Kolev, Valduriez, Bondiombouy, Jiménez, Pau, & Pereira, 2016).

The architecture that uses the CloudMdsQL language consists of a mediator and a wrapper and it is given in the picture below (Kolev, Valduriez, Bondiombouy, Jiménez, Pau, & Pereira, 2016). The mediator collects the information found in the global schema. Transforms queries set in common language for the wrapper and integrates query results. The wrapper displays information about source schemes and provides mapping functions that translate data between source schemas and schema mediators. The wrapper transforms queries written in a common language into queries for different DBs. Transforms query results into a common data model. The characteristic of this architecture is that one query that represents output from one database can be input to another database (Kolev, Valduriez, Bondiombouy, Jiménez, Pau, & Pereira, 2016).

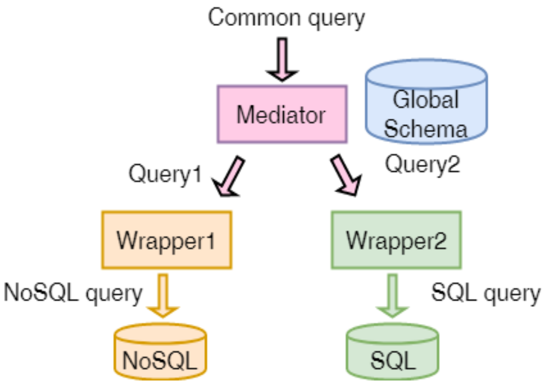


Figure 10: Mediator/wrapper Architecture (Kolev, Valduriez, Bondiombouy, Jiménez, Pau, & Pereira, 2016).

CloudMdsQL is not built to integrate web data. CloudMdsQL sticks to the relational data model for data representation, integration of data and the benefits of relational algebras operations.

Below are examples of relational, non-relational and common queries:

*1**/ show from document db all reviews made in 2018*

```
reviews_2018 ( pub_id int, reviewer string )@DB2 = {*
  db.reviews.find(  {'date': {'$gte': '2018-01-01', '$lte': '2018-12-31'} }, {'pub_id': 1, 'reviewer': 1, '_id': 0} )
                                                                          *}
```

*2**/ show from relational db the publications of scientists from Africa*

```
pubs_Africa ( id int, title string, author string )@DB1 = (
 SELECT pubs.id, pubs.title, pubs.author
 FROM pubs   JOIN scientists ON (pubs.author = scientists.name)
 WHERE scientists.affiliation = 'AFRICA' )
```

*3**/ joining the previous two queries*

```
SELECT pubs_ Africa.id, pubs_ Africa.title, pubs_ Africa.author, reviews_2018.reviewer
FROM   pubs_ Africa   JOIN reviews_2018 ON (pubs_ Africa.id = reviews_2018.pub_id) ;
```

Table 5: Joining the previous two queries

ID	Title	Author	Reviewer
6	Principles…	Mark	Peter
6	Principles…	Mark	Rui

5. CONCLUSION

Researching and analyzing HQPS system and CloudMdsQL system reveals some differences between them, as differences in data transformation. In the HQPS system, a user prints a query that is sent to the Query Analizer&Generator, which checks if the last query is relational or not, and sends it to the corresponding database (SQL or NoSQL). When a query is executed in the database, it is sent in the Result Preparation for the integration of the results from different databases. In the Result Preparation, the transformation of the NoSQL data into the RDF triple format is executed and that result is shown to the user.

In CloudMdsQL architectures that are fully distributed, nodes can communicate with each other, which enables optimization, or minimization of data transfer between nodes. CloudMdsQL sticks to the relational data model, because of its intuitive data representation and integration datasets by applying joins, unions and other relational algebra operations (Kolev, Valduriez, Bondiombouy, Jiménez, Pau, & Pereira, 2016). In CloudMdsQL System nesting is allowed in both SQL and native expressions. CloudMdsQL allows for optimizing the query execution by rewriting queries according to bind joins and, planning optimal join execution orders, and the delivery of intermediate data is optimal. CloudMdsQL query can exploit the full power of local data stores.

Architecture based on CloudMdsQL has little advantage over HQPS architecture if the criteria of comparison are the speed of transforming data. The conclusion is based on the CloudMdsQL architecture which allows nodes to communicate with each other which makes transforming data faster. We examine factors that influence query result performance, including databases with different sizes and different table sizes. The results show size affects the time of executing queriess. In the future, other criteria for comparing different databases should be considered, in order to better identify the benefits of both with the optimization of query execution.

REFERENCES

Adeyi, T. S., Abdullahi, S. E., & Junaidu, S. B. (2013). DualFetchQL System: A Platform for Integrating Relational and NoSQL Databases. *International Journal of Engineering, 2(12)* .

Ali, M. G. (2009). Object Oriented Approach for integration of heterogeneous databases in a multidatabase system and local schemas modifications propagation. *arXiv preprint arXiv:0912.0603.*

Bergamaschi, S., Guerra, F., Interlandi, M., Trillo-Lado, R., & Velegrakis, Y. (2016). Combining user and database perspective for solving keyword queries over relational databases. *Information Systems, 55* , 1-19.

Brewer, E. (2012). CAP twelve years later: How the" rules" have changed. *Computer, 45(2)* , 23-29.

Đurđevac, I. (2012, may 23.). *Da li MySQL odlazi u istoriju?* Retrieved june 1., 2017, from Đukijev blog: https://ivandjurdjevac.me/category/kompijuteri-i-it/nosql/

Haber, A., Look, M., Perez, A. N., Nazari, P. M., Rumpe, B., Wortmann, A., et al. (2015). Integration of heterogeneous modeling languages via extensible and composable language components. *In Model-Driven Engineering and Software Development (MODELSWARD)* , 19-31.

Katsov, I. (2012). *NOSQL DATA MODELING TECHNIQUES.* Ontario, Canada: wordpress.com.

Kolev, B., Valduriez, P., Bondiombouy, C., Jiménez, R., Pau, R., & Pereira, J. (2016). CloudMdsQL: querying heterogeneous cloud data stores with a common language. *Distributed and parallel databases, 34(4)* , 463-503.

Luković, I. (2015, may 10). *Alternativni pristupi u izgradnji sistema baza podataka.* Retrieved octobar 1, 2017, from Univerzitet u Novom Sadu, Fakultet tehničkih nauka: http://www.acs.uns.ac.rs/sites/default/files/06_BP_Alternativni_Pristupi_Izgradnji_SBP.pdf

Mijalkovic, S. (2013). *NoSQL Baze podataka.* Beograd: Matematicki fakultet,Studentski trg 16.

Nayak, A., Poriya, A., & Poojary, D. (2013). Type of NOSQL databases and its comparison with relational databases. *International Journal of Applied Information Systems, 5(4)* , 16-19.

Nikolaus, M. (2015). *Map Reduce kao tehnika za obradu velikih količina polustrukturiranih podataka.* Varaţdin: SVEUČILIŠTE U ZAGREBU FAKULTET ORGANIZACIJE I INFORMATIKE.

Oluwafemi, O. E., Sahalu, J. B., & Abdullahi, S. E. (2016). TripleFetchQL: A Platform for Integrating Relational and NoSQL Databases. *International Journal of Applied Information Systems (IJAIS)* , 54-57.

Roijackers, J., & Fletcher, G. (2012). Bridging sql and nosql. *Master's thesis, Eindhoven University of Technology* .

Thant, P. T., & Naing, T. T. (2014). Hybrid Query Processing System (HQPS) for Heterogeneous Database (Relational and NoSQL). *In Proceeding of the International Conference on Computer Networks and Information Technology* , 53-58.

Varughese, D. K., & Rajeswari, V. (2009). Heterogeneous database integration for web applications. *International Journal of Computer Science and Engineering* , 1-8.

Veljković, N. (2013). Nerelacione baze podataka. *Univerzitet u Beogradu Matematicki fakultet* , 1-11.