THE POSSIBILITY OF USING DATA MINING IN THE RESEARCH OF AGRICULTURAL HOLDINGS

Milan Milunović[1] , Radovan Damnjanović[2] , Nedžad Imamović[3] , Radan Kostić[4] , Mihailo Ćurčić[5] , Vladimir Ristić[6] , Dragan Bojanić[7]

**Corresponding author E-mail: radovandam78@gmail.com*

ARTICLE INFO

Review Article

Received: *06 June 2018*

Accepted: *17 August 2018*

doi:10.5937/ekoPolj1803139M

UDC 631:[004.62+519.816]

Keywords:

agricultural, data mining, unsupervised discriminant analysis, decision tree.

JEL*:* Q12, C82

ABSTRACT

Purpose. The aim of this study was to examine the usefulness and accuracy of Data Mining techniques on the example of testing the presence of impact evaluation of the quality of the land on the level of income of agricultural holdings on the basis of test samples. Methodology. The study was analysis conducted on a random sample for identifying key factors in the research of impact evaluation of the quality of the land on the level of income of agricultural holdings, on a data set of 179 examples, where the input consists of various variables: factor of erosivity, the power of the land, reducing the pH value, presence of organic matter, then target discrete variables with two descriptive values: at a expected yield and real yield. Results and Conclusions. The results obtained from the experiments agree confirmed a physical and chemical factors properties largely determines the classification results.

Introduction

Land quality is usually defined as "the capacity of a specific type of land that functions within the natural or farmland boundaries of the ecosystem, to maintain the plant and plant animal productivity, preserves or increases the quality of water and air and supports

1 Milan Milunović, PhD., Assistant Professor, Budget Department, Ministry of Defense, Birčaninova 5, 11000 Belgrade, Serbia, milunmil68@gmail.com

2 Radovan Damnjanović, PhD. Assistant Professor, General Staff of the Serbian Armed Forces, Belgrade, Gardijska 7, 11000 Belgrade, Serbia, E-mail: radovandam78@gmail.com.

3 Nedžad Imamović, PhD, Ministry of Defense, Gardijska 7, 11000 Belgrade, Serbia, nedzimam66@gmail.com

4 Radan Kostić, PhD, Assistant Professor, Military Academy, University of Defence, Pavla Jurišića Šturma br. 33, 11000 Belgrade, Serbia, kosrad74@gmail.com

5 Mihailo Ćurčić, M.A., Military Academy, University of Defence, Pavla Jurišića Šturma br. 33, 11000 Belgrade, Serbia, e-mail: curcicmihailo@gmail.com

6 Vladimir Ristić, M.A., University of Defence, Pavla Jurišića Šturma br. 1, 11000 Belgrade, Serbia, vladirist72@gmail.com

7 Dragan Bojanić, M.A., University of Defence, Pavla Jurišića Šturma br. 33, 11000 Belgrade, Serbia, draganboj74@gmail.com

health and standard of human beings "(Karlen et al., 1997). Inherent soil quality can be assessed based on the study land within the network of national monitoring. The dynamic quality of the land includes those land properties that can be change in a short period of time under the influence of usage measures and land management. Land quality (Trgovčević Prokić, Počuča, 2016) is initially presented as an approach that makes it better use of land for different land functions, thus putting the accent on the live and the dynamic nature of the land. The quality of the land is assessed in relation to its functions (Karlen et al. 1997). The ability of the soil to perform any of the many functions depends on its physical, biological and chemical properties.(Vukoje, 2013). The aim of this study was to examine the usefulness and accuracy of Data Mining techniques on the example of testing the presence of impact evaluation of the quality of the land on the level of income of agricultural holdings on the basis of test samples.

Application of Data Mining in the last decade has brought about a significant methodological shift in the field of scientific research in agricultural. The classical method which assumes normative-descriptive methods supported by classical multivariate statistical methods has become the basis for establishment of machine learning as productive and more accurate scientific methods in agricultural research (Mihajlović, 2014). Study Hira and Deshpande (2015) has been to investigate analysis approach helps to build model and apply advance techniques like multidimensional data analysis, statistical mining and data mining to extract knowledge for to analyze agriculture productivity using various agriculture related parameters. The methods employed were Unsupervised Linear Discriminant Analysis and Decision Trees (Bengio et al., 2004). The present study contributes to agricultural research by examining the suggested variables in order to identify those that can best discriminate cases variables which are dominant in validation in case at a expected yield and real yield.

Materials and methods

The research process is based on well-defined and grouped data, determining the variables (target variable) and applying selected methods with analysing and interpreting the results. In this paper we used the following research method Decision trees and rules, as well as teaching methods with a combined classifier and methods of accuracy estimation of learned classifiers: test set methods, cross-validation method (Kotsiantis, Zaharakis & Pintelas, 2007).

We used the following software tools in our research: - system for knowledge exploration – WEKA, a large number of implemented inductive learning algorithms, analysis of reducing the dimensionality of space attributes; Random Forests algorithm (Breiman, 2001; Pamučar, Ćirović, 2018); Tanagra and Sipina Data Mining software for academic and research purposes. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and data base area (Xuab Z., Lee J., Parka D., Chunga Y., 2017).

Our research is related to the results of identifying key factors of the agricultural

research of a physical and chemical factors on 179 cases, where the key factors are: factor of erosivity from 0 to 31, where average the is 10.13 and the power of the land from 0 to 33, with average value is 11.12. The same principles are used in the case of reducing the pH value from 0 to 19 with the average value of 5.99 and presence of organic matter from 0 to 20 with the average of 6.65.

From the sample we identified 89 cases of expected yield, and 90 with real yield.

Our research is related to the results of identifying the key factors impacting whether at a expected yield or real yield. Data was grouped into five categories (variables), namely: X1 = fcsex (soil types): 1 = Black soil , 2 = Vertisol; X2 = faggr (factor of erosivity); X3 = maggr (the power of the land); X4 = fanxdepr (reducing the pH value); X5 = manxdepr (presence of organic matter), X6 = clin (yield): 1 = expected yield , 2 = real yield and it based on sample of 179 items.

Identifying the key variables of X1, X2, X3, X4, and X5, (excluding X6 as a target variable), can be identified as a typical classification problem, and takes place in two procedural stages. In the first phase, the model is trained to use the soil sample. The sample is organized in rows and columns. One of the attributes is the class attribute, which predominantly affects the key factor for defining real yield. In the second step, the model is trying to classify an object that does not belong in the soil sample.

The authors used supervised linear discriminant function with validation accuracy processing of the classification method, namely: cross-validation. The method of cross-validation randomly divided a set of examples D on k mutually exclusive subsets D1, D2,...,Dk of approximately the same size. The sampling process and assessment is repeated k times, each time using one subset of Di as a test set. The bootstrap method is a family of methods for the estimation of prediction accuracy. For a given set of examples, a bootstrap sample is formed randomly taking n examples uniformly from a set of examples, with a replacement (Kohavi, 1995). Other than LDA method, we used the Decision Trees (Breiman et al., 1984) in our research by algorithm C 3.4 with validation methods, namely Random Forests (Breiman, 2001). The main goal of these statistical methods is to determine the useful variables for the purpose of classification. The first step uses the method of supervised learning by linear discriminant function with continual variables, X1, X3, X4, X5, and X6, as the target - predictor discrete variable category: expected yield or real yield.

The results indicate that there is a substitution error of 0.18. The detailed results of the sampling process imply that several variables do not seem relevant in the classification model. Summary of the LDA model are designed FCSEX, MAGGR and FANXDEPR as significant. Since the first function is standardized, these coefficient can be used to make judgments about relative importance of each variable. Since FCSEX makes the largest contribution to the first discrimination function followed by MAGGR- the power of the land and FANXDEPR - reducing the pH value.

$$Z = 6.3X_1 + 0.07X_2 + 0.17X_3 + 0.06X_4 + 0.03X_5 - 6.34$$

But the question that remains is whether all variables or only some are relevant in the resulting model? To answer this question we used the selection feature, the process of selecting a subset of relevant features for use in the construction model. The main assumption when using a selection technique feature is that the data contains many redundant or irrelevant features. We realized that the re-substitution error rate is not accurate and requires the use resampling method (Bootstrap) for obtaining honest error estimate. We observe that true error rate is about 0.195.

The next step is applying a stepwise discriminant analysis (STEPDISC) (Rikalović, Soares, Ignjatić, 2018) approach in the process of sampling, with the purpose of finding how many variables are sufficient for the classification variables that determine membership in the expected yield or real yield?

We applied FORWARD strategy and set the comparison of the F statistics as a stopping rule. We saw that two variables are selected out of 5, and according to the analysis STEPDISK, the only relevant attributes (variables) are MAGGR - the power of the land and FANXDEPR - reducing the pH value.

The next step in analysis is establishing the control of efficiency. In this sense, analysis is performed through supervised LDA and bootstrap components. (Mc Farlanea et al., 2016) Classification performance measured by the rate of error is the same this time 0.179, but now with a new LD function of only one variable, which is of decisive importance for the determination of classification in the expected yield and real yield, is as follows:

$$Z = 0.15MAGGR + 0.14FAXDEPR - 1.4$$

with reduced bootstrap error of 0.189.

Results

In addition to discrimination analysis, a study was carried through the decision tree, through the C 4.5 algorithm, which is based on a tree structure, where each leaf node represents a test attribute, and each branch represents the results of the test. The goodness of a split is based on the selection of attributes that are better separated in the sample. Identifying target variable can be regarded as a typical classification problem. (Sabarina, Priya, 2015)Classification is a two-step procedure. In the first step, a model is trained by using a soil sample. The sample is organized in tuples (rows) and variables (columns). One of the attributes, the class label attribute, contains values indicating the predefined class to which each tuple belongs. This step is also known as supervised sampling. In the second step, the model attempts to classify objects which do not belong to the training sample and form the validation sample. In this study we employed the well-known ID3 algorithm. ID3 uses an entropy-based measure, known as information gain, in order to select the splitting attribute (Han & Kamber, (2000). The successive division of the sample may produce a large tree. Some of the tree's branches may reflect anomalies in the soil set, like false values or outliers. For that reason tree pruning is

required. Tree pruning involves the removal of splitting nodes in a way that does not significantly affect the model's accuracy rate. In order to classify a previously unseen object, the variable or attribute values of the object are tested against the splitting nodes of the Decision Tree.(Matei et al., 2017) According to this test, a path is traced that will conclude with the object's class prediction. Main advantages of Decision Trees are that they provide a meaningful way of representing acquired knowledge and make it easy to extract IF–THEN classification rules.

Supervised learning produced decision tree, with classifier performance: error rate 0.146 and decision tree, as follows:

- *MANXDEPR < 9.5000 then CLIN =* ***EXPECTED YIELD****(84.09 % of 88 examples)*
- *MANXDEPR >= 9.5000 then CLIN =* ***REAL-YIELD*** *(80.00 % of 5 examples)*

- *MAGGR >= 9.5000*
  - *FANXDEPR < 8.5000*
    - *MAGGR < 15.8250*
      - *FANXDEPR < 5.1900*
        - *MAGGR < 13.5000 then CLIN =* ***EXPECTED YIELD*** *(63.64 % of 11 examples)*
        - *MAGGR >= 13.5000 then CLIN =* ***REAL-YIELD*** *(66.67 % of 6 examples)*
      - *FANXDEPR >= 5.1900 then CLIN =* ***EXPECTED YIELD*** *(75.00 % of 8 examples)*
    - *MAGGR >= 15.8250 then CLIN =* ***EXPECTED YIELD*** *(88.00 % of 25 examples)*
  - *FANXDEPR >= 8.5000 then CLIN =* ***EXPECTED YIELD*** *(100.00 % of 35 examples)*

We use C 4.5 and cross-validation in order to evaluate the accuracy of a standard (individual) decision tree algorithm. The error rate is 0.247.

The next step is implementing the Random Forests algorithm. There are two steps in order to insert the Random Forest method as controlling methods in the diagram: First we applied BAGGING. Bagging shall generate multiple versions of the classifier used as a unified whole, through the mechanism of voting. More classifiers are generated by the soils set as an example is bootstrap-Travels. The sampling each set is an independent pattern of examples and some examples have been omitted, while some are repetitive. As with other methods of ensembles, the procedure is suitable for aggregation of

results of "unstable" algorithms, algorithms relationship in which small changes in the training of its rally caused major changes in the learned set of rules. Embed method and then Fandom Forests methods (Saeys, Inza & Larrañaga, 2007) and techniques as controlling method by resampling data and by selection subsets of attributes in process of resampling by induction trees (Touw, Bayjanov, Overmars, Backus, Boekhorst, Wels & Hijum, 2012). Finally we evaluate sampling accuracy with cross- validation components. Estimation of the accuracy of the classification using a method of cross-validation gets the error rate of 18%.

Discussions

Three alternative models were built, each based on a different method. First, the Decision Tree model was constructed using the Sipina Research Edition software (Kamilaris et. al.,2017). The model was built with confidence level 0.05. We used the whole sample as a soil set. The model was tested against the sampling and managed to correctly classify 81% expected yield and 82% real yield.

Interpreting the IF- THAN is as follows: IF MANXDEP - presence of organic matter IS GREATER OR EQUAL 2.5 TNEN 37% is expected yield and 63% is real yield AND IF MAGGR - the power of the land IS LESS 9.5 THEN 71% is expected yield and 29% is real yield AND IF FAGGR -factor of erosivity IS LESS 7.5 THEN 83% is expected yield and 17% is real yield.

We noticed that IF MAGGR - the power of the land IS GREATER OR EQUAL 9.5 THAN expected yield case is 16% and 84% is real yield, and in last case. IF FAGGR - factor of erosivity IS GREATER 7.5 THEN 40% is expected yield and 60% is real yield.

The use of the proposed methodological framework could be of assistance to agricultural holdings. In terms of performance, we apply the supervised discriminant analysis (SPV LDA) in order to identify specific variables that influence expected yield and real yield for both with the methods of classification accuracy of validation influence variables and identifying the key variables which are MAGGR, and FAXDEPR in some weight (0.15* MAGGR and 0.14*FAXDEPR) with accuracy at a level of 82%. Decision Tree method was used, which gave results that are more accurate in terms of determining the extent and logical connection between variables of impact of individual variables. Obtained logical connection between variables and its weight of impact and unlike traditional multivariate study (Hiironen J., Riekkinen K., 2016), this is where the research through classical statistical methods worked out three variables that influence the presence of impact on evaluation for factor of erosivity, the power of the land, reducing the pH value, presence of organic matter tree variables are not decisive for qualification. All the research into the foreground has been achieved and this is a great degree of accuracy studies. The present study contributes to agricultural research by examining the suggested variables in order to identify those that can best discriminate cases variables which are dominant in validation in case at a expected yield and real yield.

Conclusions

The application of Data Mining would significantly help researchers in the field of agricultural, particularly because of the possibility that the research work sparse datasets, which have a large number of attributes and a very small number of examples. In our example we recognized the problem of sparse data and sparse data set (data with small relations between number of observation and number of variables). Implicitly it prevents sufficient good differentiation of new examples. In agricultural it is necessary to achieve the same or greater accuracy of prediction, which is very subjective evaluation, that the development of agricultural changes.

The present study contributes to agricultural research by examining the suggested variables in order to identify those that can best discriminate cases variables which are dominant in validation in case expected yield and real yield properties.

However, it is possible to achieve the results of the minor increase of the number of variables and the extent of sample to the level of about 93% to 100% accuracy. Also we assigned a problem of scarcity associated with the assessment of the severity of quality of the land problems (task difficulty), which is solved in the domain of Data Mining by reducing the number of attributes (variables). Such approaches allow detection methodically of so far hidden knowledge in agricultural, and especially the causes that determine the decisive variables and attributes and factors for solving research problems in these and other research areas.

Conflict of interests

The authors declare no conflict of interest.

References

1. Bengio, Y., Delalleau, O., Le Roux, N., Paiement, J. F., Vincent, P., & Ouimet, M. (2004). Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Computation*, 16(10), 2197-2219.
2. Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
3. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984).*Classification and regression trees*. CRC press, Chapman and Hall, London.
4. Han, J., & Kamber, M. (2000). *Data Mining: Concepts and Techniques*, Elsevier, London.
5. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th international joint conference on Artificial intelligence*, Montreal, Quebec, Canada — August 20 - 25, 1995, 14(2), 1137-1145.
6. Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3), 249-268.
7. Mihajlovic, M. (2016). Relationship between corporate management and corporate governance. *ODITOR*, Center for Economic and Financial Research, 2(1), 4-10.

8. Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), 2507-2517.
9. Touw, W. G., Bayjanov, J. R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., & van Hijum, S. A. (2012). Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?. *Briefings in bioinformatics*, 14(3), 315-326.
10. Hiironen, J., Riekkinen, K. (2016). Agricultural impacts and profitability of land consolidations. *Land Use Policy*, Elsevier, 55(9), 309-317.
11. Hira, S., Deshpande, P.S. (2015). Data Analysis using Multidimensional Modeling, Statistical Analysis and Data Mining on Agriculture Parameters, *Procedia Computer Science*, Elsevier, 55(3), 431-439.
12. Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F.X. (2017). A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture*, Elsevier, 143(9), 23-37.
13. Karlen, D. L., Mausbach, M. J., Doran, J. W., Cline, R. G., Harris, R. F., Schuman, G. E. (1997). Soil Quality: A Concept, Definition, and Framework for Evaluation, *Soil Sci. Soc. Am. J.,* 61(6), 4-10.
14. Mc Farlanea, J.A., Blackwellab, B.D., Mountera, S.W., Grantc, B.J. (2016). From agriculture to mining: The changing economic base of a rural economy and implications for development. *Economic Analysis and Policy*, Elsevier, 49(3), 56-65.
15. Xuab, Z., Lee, J., Parka, D., Chunga, Y. (2017). Multidimensional analysis model for highly pathogenic avian influenza using data cube and data mining techniques, *Biosystems Engineering,* Elsevier, 157(10), 109-121.
16. Matei, O., Rusu, T., Petrovan, A., Mihuțc G. (2017). A Data Mining System for Real Time Soil Moisture Prediction. *Procedia Engineering*, Elsevier, 181(4), 837-844.
17. Pamučar, D., Ćirović, G. (2018). Vehicle route selection with an adaptive neuro fuzzy inference system in uncertainty conditions. *Decision Making: Applications in Management and Engineering*, 1(1), 13-37.
18. Sabarina, K., Priya, N. (2015). Lowering Data Dimensionality in Big Data for the Benefit of Precision Agriculture. Procedia Computer Science, Elsevier, 48(3), 548-554.
19. Rikalović, A., Soares, G.A., Ignjatić, J. (2018). Spatial analysis of logistics center location:A comprehensive approach. *Decision Making: Applications in Management and Engineering,* 1(1), 38-50.
20. Trgovčević Prokić, M., Počuča, M. (2016). Acquisition of agricultural land, *Economics of Agriculture*, 63(4), 1281-1296.
21. Vukoje, A. (2013)., Factors of existence as a condition of creating a market position of the company. ODITOR, Center for Economic and Financial Research 1(5), 27-37. [*in Serbian:* Faktori egzistencije kao uslov stvaranja tržišne pozicije preduzeća. *ODITOR*, Centar za ekonomska i finansijska istraživanja]