Agris on-line Papers in Economics and Informatics

Volume IX Number 4, 2017

Analytical System with Decision Tree for Economic Benefit

Jan Tyrychtr

Faculty of Economics and Management, Czech University of Life Sciences Prague, Czech Republic

Abstract

Data processing is an important aspect of business decision support systems (DSS). A good analytical system to process these data is essential to implement as a primary pillar for the development of complex expert systems. Businesses themselves are constantly confronted with deciding on investment opportunities to improve their performance. An important criterion for selecting investment is its profitability which cannot be easily determined when investing in analytical systems. Currently, there are two types of approaches to evaluating investments into information systems: normative and positive approaches. The simplest form of decisional analytical modeling is the decision tree (normative approach). The purpose of the article is to illustrate decision tree analysis as a component of an analytical system for evaluating two decision alternatives. The test case is demonstrated on an example of decision-making in agriculture.

Keywords

Analytical system, decision tree, decision rules, economic value, agriculture.

Tyrychtr, J. (2017) "Analytical System with Decision Tree for Economic Benefit", *AGRIS on-line Papers in Economics and Informatics*, Vol. 9, No. 4, pp. 123-129. ISSN 1804-1930. DOI 10.7160/aol.2017.090412.

Introduction

Farmers face a large number of day-to-day decisions about a number of activities linked to their production (animal changes, material shifting, land maintenance and planning, veterinary checks, etc.) and investment decisions (purchase and sale of animals or produce, modernization of buildings, stables or machines etc.). In connection with this necessary decision-making activity of the owners of agricultural companies, there is a need for the use of analytical systems, allowing for interactive and flexible data analysis (for example in relation to the evaluation of the current development of performance according to various aspects and time series) and on the other hand on agricultural activities.

The investment costs of analytical systems (i.e. hardware, software, and to a certain extent personnel costs) are represented by the market price. However, the actual benefits of the system (i.e. the effects of an analytical system on agricultural business performance) cannot be expressed in this way. The field of economic evaluation of systems is a non-trivial problem, which is well described in the study (Verstegen et al., 1995). In general, there are two approaches to assessing the economic value of information systems: a normative and positive approach. Normative approaches are based on decision making by means of theoretical (etc. decision tree analysis (Lahtinen et al., 2017), Bayesian information economics (Kleijnen, 1980)) or analytical approaches (etc. simulation or linear programming). Positive approaches are based primarily on experimental designs (time series, econometric modeling).

An objective approach to assessing the economic benefits of analytical systems is to use a measure that identifies the evolution of the revenues from the analytical system. Such information may be useful not only for farmers who are considering investing in a new analytical system but also for companies that design and sell these systems.

Analytical system

Analytical systems serve to support strategic decision making and to reveal hidden information to easily understand and anticipate user needs. The analytical system generally consists of three layers: the layer for data transformation (Extraction-Transformation-Loading (ETL) tools (Zekri et al., 2017)), the data storage layer (data warehouses, data markets and operational databases) and a layer for analytical data processing. Currently the most advanced types of analytical systems are systems for on-line analytical data processing (OLAP) (Wrembel, Koncilia, 2007), which are used in Business Intelligence (Rouhani et al., 2012, Tyrychtr and Vasilenko, 2015). Ways of storing data in analytical systems can be solved by designing so-called multidimensional databases. Multidimensional databases (Pedersen and Jensen, 2001) are suitable for storing (multidimensional) data of analytical type, over which analyzes and overviews are used most frequently for self-decision. The term multidimensional data represents the data of aggregate indicators generated by different grouping of relational data designed for OLAP. OLAP describes a decision support approach that aims to gain knowledge from a data warehouse or data markets (Abelló and Romero, 2009). The very way of organizing data in multidimensional databases is solved through a construction of a data cube. Data cube is a data structure for storing and analyzing large amounts of multidimensional data (Pedersen, 2009). The data cube represents an abstract structure, which, unlike the classical relational structure in the relational data model, is not defined unambiguously. There are many approaches to the formal definition of data cube operators (a comprehensive overview is available in the post (Vassiliadis and Sellis, 1999)). In general, the data cube consists of dimensions and measurements. Dimension is a hierarchically ordered set of dimensional values that provide categorical information that characterizes a particular aspect of data (Pedersen, 2009b). Measurements (monitored indicators) of the cube are primarily quantitative data that can be analyzed.

Decision support system

From analytical systems it is possible to load aggregated or summarized data for further processing in decision support systems (DSS), (Burstein and Holsapple, 2008). Currently, DSS consists of a range of decision support applications or technologies such as model and data-oriented systems, multidimensional data analysis, query and reporting tools, online analytical processing (OLAP), Business Intelligence, document management, spatial DSSs, and executive information systems. All of these technologies and applications are designed to support decision making. However, the main characteristic of DSS is that it provides users with options for decision-making.

Expert system

Expert system (Wagner, 2017, Ugolnitskii and Usov, 2008) represents the knowledge base of control models that can be conditionally divided into two parts. In the first part is the known information from already existing control models (subsystems). This section includes a database of predictions of specific situations obtained using a pre-created scenario. The second part of the expert component uses information, models and expert-type data based on the knowledge, experience and intuition of experts. This section should constantly get new data.

Current state and motivation

Principles of decision trees are generally known from a number of areas. The way of representing knowledge in the form of decision trees is a clear and easy to interpret way of analyzing data. The goal of decision trees is to identify objects described by different attributes into classes. They do not require any special data preparation and process both categorical and numerical variables. Trees are relatively easily able to find non-linear relationships between input attributes. The result of the analysis is a graphically illustrated tree that can be, as a rule, easily interpreted. The algorithm of the decision trees method determines which attributes are key and which, on the other hand, do not matter and it is appropriate to drop them from the model. This property can be used to select dimensions when designing an OLAP database (Shmueli et al., 2017). Most commonly decision tree methods are used within Data Mining in Business Intelligence (Vercellis, 2011; Shmueli et al., 2017). However, these methods are currently neglected in some areas.

An example is the agricultural sector. For example, the survey of state of Business Intelligence in agriculture in the Czech Republic (Tyrychtr et al., 2015) shows that the level of use of analytical systems is rather marginal (1% of agricultural entities use an analytical system). At the same time, the results of the analysis of the current state of information needs in agriculture (Tyrychtr and Vostrovský, 2017) show that if the need for information on farms is higher, it also requires a higher level of ICT and DSS systems. Many of the farm problems that accompany these activities require timely and qualified decision-making. Given the high information needs in this sector, the farm management information systems (FMIS) must be able to use functions that are typical for expert and analytical systems and effectively support farmer's decision taking or their management. If a farmer decides to invest in analytical systems, it is essential that these systems support automated and easy-to-use analytical functions that are easy to interpret for his economic benefit.

The aim of the paper is to apply decision tree principles as a potential functionality of the analytical system to be used in the agricultural sector to support decision making. If the analytical system serves to support the decision-making of the farmer's main activities with an impact on economic benefits, such a system can represent a significant economic value for an enterprise.

Materials and methods

In order to calculate the economic value of the analytical system, the principles of the decision tree are used by the author. When making a decision tree, the followed method is used “*divide and conquer*”. The training data is gradually divided into smaller and smaller subsets (tree nodes) so that examples of one class predominate in these subsets. At the beginning, the whole training data consists of one set, at the end are subsets made up of examples of the same class (Quinlan, 1986). This principle is called *top-down induction of decision trees classifiers* (Rokach and Maimon, 2005).

For the choice of suitable attribute for the tree branching the attribute’s characteristics are used (Berka, 2005): entropy, information gain, relative information gain, χ^2 or Gini index.

Entropy is the degree of disorder of a system and is defined as follows:

$$H = -\sum_{t=1}^{T} (p_t \log_2 p_t),$$

where p_t is the probability of occurrence of class t and T is the number of classes. If $p = 1$ (all cases belong into the class) or $p = 0$ (no case belongs to the class), the entropy is zero. If both classes are represented by the same number of examples ($p = 0.5$), the entropy is at its maximum.

The calculation of entropy for one attribute is done in the following way. For each value v, which may be assumed by attribute A is calculated according to the entropy formula $H(A(v))$ on a group of examples that are covered by the category $A(v)$

$$H\big(A(v)\big) = -\sum_{t=1}^{T} \frac{n_t\big(A(v)\big)}{n\big(A(v)\big)} \log_2 \frac{n_t\big(A(v)\big)}{n\big(A(v)\big)}.$$

Medium entropy $H(A)$ is counted as a weighted sum of entropy $H(A(v))$, where the weights in sum are the relative frequencies of categories $A(v)$ in dat

$$H(A) = -\sum_{v \epsilon Val(A)} \frac{n\big(A(v)\big)}{n} H\big(A(v)\big).$$

The attribute with the smallest entropy is then selected for tree branching $H(A)$.

Information gain measures the reduction of entropy due to the choice of attribute A. It is defined as the entropy difference for the target attribute and for the considered attribute:

$$Gain(A) = H(C) - H(A),$$

where

$$H(C) = -\sum_{t=1}^{T} \frac{n_t}{n} \log_2 \frac{n_t}{n}.$$

Relative information gain also takes into account the number of attribute values and is defined as follows:

$$Gain\,ratio(A) = \frac{Gain(A)}{Furcation(A)},$$

where

$$Furcation(A) = \sum_{v \in Val(A)} \frac{n(A(v))}{n} \log_2 \frac{n(A(v))}{n}.$$

Data set

For the model example the data of a farmers in the Czech Republic is used. In the context of their agricultural activity they records the decision to sell or dispose of cattle in the Farmer’s Portal information system operated by the Ministry of Agriculture of the Czech Republic. All data about movements of animals in the farm are recorded in the Register of animal (IZR). Specific data values are not significant for design decision tree principles as a potential functionality of the analytical system. This dataset from IZR is simplified for the clarity of the decision tree induction algorithm. In table 1 the data is stated without numerical attributes.

Cow	Age	Weight	Sex	Disease	Sale/Transfer
c1	high	high	cow	no	yes
c2	high	high	bull	no	yes
c3	low	low	bull	no	no
c4	low	high	cow	yes	yes
c5	low	high	bull	yes	yes
c6	low	low	cow	yes	no
c7	high	low	bull	no	yes
c8	high	low	cow	yes	yes
c9	low	middle	bull	yes	no
c10	high	middle	cow	no	yes
c11	low	middle	cow	yes	no
c12	low	middle	bull	no	yes

Source: own work

Table 1: Data for creation of decision tree.

Results and discussion

In this section, the concept of the analytical system is designed by the author and a test example is created with the decision tree induction procedure. The result is the identification of the economic value of the analytical system for the enterprise.

Concept of analytical system

The relationship between the analytical system, the DSS and the expert system can be represented in the form of a pyramid (Figure 1). Within the analytical system, data is analyzed by means of summarizations, aggregations and filtrations. New data sets are created presenting important data from different points of view. This data enters the DSS to create various variants of monitored data - reports, dashboards, and various multidimensional reports to support decision-making. The last part of the pyramid is an expert system which, on the basis of the analyzed and processed data and information, will allow the user, based on already recorded knowledge, to provide expert evaluation of variants and prediction assessment.

Decision theoretical approaches

An analytical system that enables to efficiently analyze data through data mining methods can help increase the economic value of a business information system. In the next section, the author has made the entire decision tree induction process.

These calculations are made from the data listed in Table 1. Four-column tables are created from this table. Table 2 shows the values for Age and Sale/Transfer.

	Sale yes	Sale no
Age high	5	0
Age low	3	4

Source: own work

Table 2: Four-pole table for Age and Sale/Transfer.

1. step: Selection of attribute for the tree root

Initially, all examples are in one set. The attribute selection for the first branching is in all 12 examples selected based on the calculation of entropy for the individual attributes. Entropy for the Age attribute is calculated from the data in Table 2, i.e.

$$H(age) = \frac{5}{12}H\big(age(high)\big) + \frac{7}{12}H\big(age(low)\big),$$

where

$$H\big(age(high)\big) = -p_+\log_2 p_+ - p_-\log_2 p_- = -\frac{5}{5}\log_2\frac{5}{5} - \frac{0}{5}\log_2\frac{0}{5} = 0,$$

$$H\big(age(low)\big) = -p_+\log_2 p_+ - p_-\log_2 p_- = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985,$$

therefore

$$H(age) = \frac{5}{12}0 + \frac{7}{12}0.985 = 0.574.$$

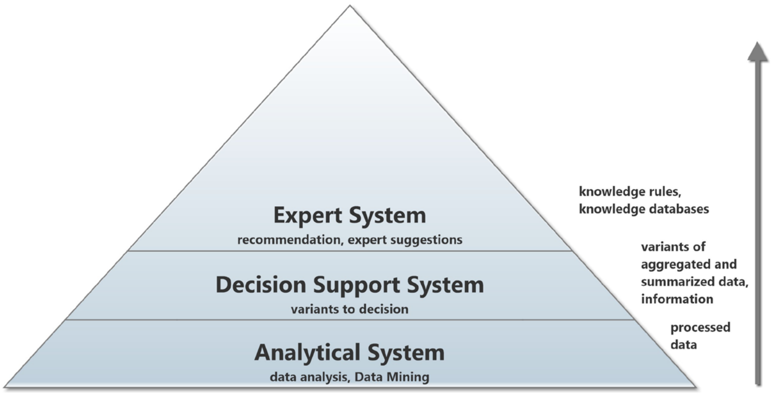
The entropy for other attributes is counted similarly:

H(weight) = 0.667,

H(sex) = 0.918,

H(disease) = 0.825.

For branching of the decision tree, the attribute *Age* is selected. This way were obtained two subsets of data. The first subset are examples included in category *age(high)* and belonging to the class *sale(yes)*, examples covered by the category *age(low)*, belong to other classes for which other attributes will be sought. Meaning that will be look for attributes which belong to the class of low age. The entropy is again calculated, this time for 7 examples – cattle with low age.


Expert System
recommendation, expert suggestions
Decision Support System
variants to decision
Analytical System
data analysis, Data Mining
knowledge rules, knowledge databases
variants of aggregated and summarized data, information
processed data

Source: own work

Figure 1: The visualization of analytical system concept.

2. step: Selection of attribute for various classes

$$H(weight) = \frac{2}{7}H\big(weight(high)\big) + \frac{3}{7}H\big(weight(middle)\big) + \frac{2}{7}H\big(weight(low)\big) = 0.394,$$

$$H(sex) = \frac{4}{7}H\big(sex(bull)\big) + \frac{3}{7}H\big(sex(cow)\big) = 0.965,$$

$$H(disease) = \frac{5}{7}H\big(disease(yes)\big) + \frac{2}{7}H\big(disease(no)\big) = 0.979.$$

Cattle with low age will be branched according to the weight. Examples covered by category *weight(high)* belong to the class *sale(yes)*, examples covered by category weight(low) belong to the class *sale(no)* and examples covered by category weight(middle) belong to various classes for which additional branching will be needed.

Entropy will be calculated again for the remaining attributes sex and *disease*:

$$H(sex) = \frac{2}{3}H\big(sex(bull)\big) + \frac{1}{3}H\big(sex(cow)\big) = 0.667,$$

$$H(disease) = \frac{2}{3}H\big(disease(yes)\big) + \frac{1}{3}H\big(disease(no)\big) = 0.$$

It is obvious from the results that the attribute disease is chosen and till cover the rest of the examples.

3. step: Creation of tree

Based on the above entropy calculations for individual attributes, a decision tree is created (Figure 2). The tree nodes have attributes used for branching, tree leaves are class assignment information, and edges of leaves match attribute values.

4. step: Transfer to knowledge rules

Each tree path from root to leaf corresponds to one rule. The attributes appear in the rule's prerequisitive and the leaf node will appear in the action rule (the action rule will appear in the leaf node). Decision tree from Figure 3 can be rewritten as follows:

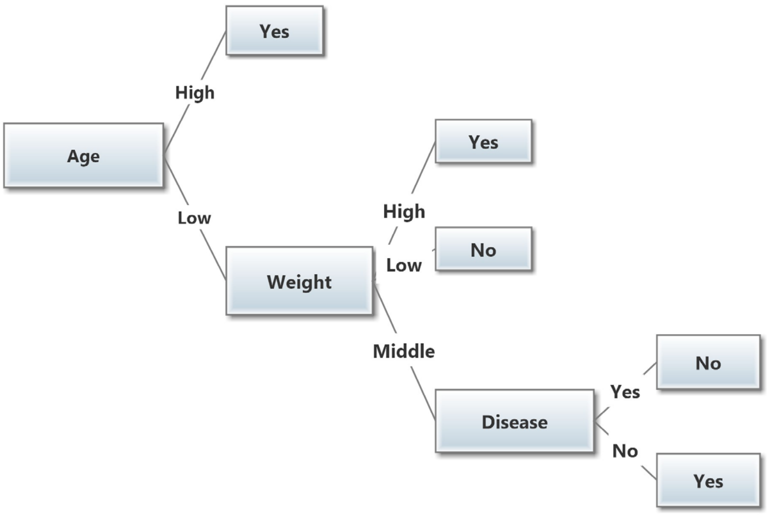
$IF\ age(high)\ THEN\ sale(yes)$
$IF\ age(low) \wedge weight(high)\ THEN\ sale(yes)$
$IF\ age(low) \wedge weight(low)\ THEN\ sale(no)$
$IF\ age(low) \wedge weight(middle) \wedge disease(no) THEN\ sale(yes)$
$IF\ age(low) \wedge weight(middle) \wedge disease(yes) THEN\ sale(no)$

Source: own work

Figure 3: From the decision tree to the rules.

Discussion

In this work was introduced the concept of an analytical system in the context of an expert system and a DSS type system. Emphasis was placed on presenting the potential of current analytical systems, which can be considered an important component for data processing in enterprises and organizations. Currently, these systems are characterized by principles based on OLAP technology. This has the potential to directly enrich these OLAP approaches via methods such as decision trees, Bayesian classification, and other statistical methods. The advantage of such a solution would be a complex analytical system, without the need to own more sophisticated tools for Data Mining such as statistical software, tools for working with neural networks etc. In this article was applied



Source: own work

Figure 1: Figure 2: The decision tree.

a simple method for classifying data - decision trees. Below author of this article assess the validity of the results achieved in this work:

The concept of the analytical system is based on the categorization and the possibility of differentiating the DSS system, expert systems and analytical systems. In literature, it is no exception that these systems are interchanged or form a common entity but mostly the DSS is usually discussed.

The example of the data model is selected from real business situations, but it is very simplified. The purpose was not to model the decision tree for a specific activity in organizations but to demonstrate how such a model could be useful in the analytical system.

For the development of the analytical system, the procedure was not translated into the algorithm. Algorithms of decision trees are generally known. However, it would be necessary to directly link them to the OLAP functional principles so that data from OLAP can be categorized directly through the decision tree.

Conclusion

In this article, author has been working on approaches that could improve the value of analytical systems in a business. The concept of the analytical system was designed by the author and a decision tree was created from the example data of agricultural decisions. The result was the identification of the economic value of the analytical system for the enterprise.

These systems can easily help company management interpret data analysis from which it is possible to gain relevant knowledge about their economic performance. Through an analytical system using a simple decision tree method, the user can choose to sell (in this case livestock) or some other action important to the business or organization. A correct decision e.g. concerning sales, represents an economic benefit for the users of the analytical system in the form of a per piece payment.

The example presented in this article demonstrates the importance of introducing these methods into analytical systems solutions. The purpose is to make decision trees and other methods directly part of these systems. If analytical systems directly incorporate functionality for classification and further data processing, it is possible to clearly define the economic benefits of such a system for an enterprise. The proposed approach provides system engineers with a methodological framework for designing the OLAP system, respectively structures of multidimensional databases. Because similar research has not yet been carried out, the results and benefits of this article offer new insights into the development of analytical systems.

Acknowledgements

The results and knowledge included herein have been obtained owing to support from the Internal Grant Agency of the Faculty of Economics and Management, Czech University of Life Sciences in Prague, grant no. 20161019, “Economic value of analytical systems in agriculture“ and within project DEPIES – Decision Processes in Intelligent Environments funded through the Czech Science Foundation, Czech Republic, grant no. 15-11724S.

Corresponding author:
Ing. Jan Tyrychtr, Ph.D.
Department of Information Technologies, Faculty of Economics and Management
Czech University of Life Sciences in Prague, Kamýcká 129, Prague 6 – Suchdol, Czech Republic
E-mail: tyrychtr@pef.czu.cz

References

[1] Abelló, A. and Romero, O. (2009) “On-Line Analytical Processing”, *Encyclopedia of Database Systems*, Ling L., Özsu, T. M. (ed.), USA: Springer US, pp. 836-836. ISBN 978-0-387-35544-3.

[2] Berka, P. (2005) "*Dobývání znalostí z databází*" (in Czech), 1st ed., Prague: Academia, 386 p. ISBN 80-200-1062-9.

[3] Burstein, F. and Holsapple, C. W. (2008) “*Handbook on Decision Support Systems 1 : Basic Themes*”, 1st ed., Springer-Verlag Berlin Heidelberg, 854 p. ISBN 978-3-540-48713-5. DOI 10.1007/978-3-540-48713-5.

[4] Kleijnen, J. P. C. (1980) “Information Systems in Management Science - Bayesian Information Economics: An Evaluation”, *Interfaces*, Vol. 10, No. 3, pp. 93-97. DOI 10.1287/inte.10.3.93.

[5] Lahtinen, T. J., Hämäläinen, R. P. and Liesiö, J. (2017) “Portfolio decision analysis methods in environmental decision making”, *Environmental Modelling & Software*, Vol. 94, pp. 73-86. ISSN 1364-8152. DOI 10.1016/j.envsoft.2017.04.001.

[6] Pedersen, T. B. and Jensen, C. S. (2001) “Multidimensional database technology”, *Computer*, Vol. 34, No. 12, pp. 40-46. ISSN 0018-9162. DOI 10.1109/2.970558.

[7] Pedersen, T. B. (2009a) "*Multidimensional Modeling*“, Encyclopedia of Database Systems, Ling L., Özsu, T. M. (ed.), USA: Springer US, pp. 1777-1784. ISBN 978-0-387-35544-3.

[8] Pedersen, T. B. (2009b) "*Dimension*“, Encyclopedia of Database Systems, Ling L., Özsu, T. M. (ed.), USA: Springer US, pp. 836-836. ISBN 978-0-387-35544-3.

[9] Quinlan, J. R. (1986) "Induction of decision trees“, *Machine Learning*, Vol. 1, No. 1, pp. 81-106. ISSN 1573-0565. DOI 10.1007/BF00116251.

[10] Rokach, L. and Maimon, O. (2005) "Top-down induction of decision trees classifiers-a survey“, *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 35, No. 4, pp. 476-487. ISSN 1558-2442. DOI 10.1109/TSMCC.2004.843247.

[11] Rouhani, S., Asgari, S. and Mirhosseini, S. V. (2012) “Review study: business intelligence concepts and approaches”, *American Journal of Scientific Research*, Vol. 50, No. 1, pp. 62-75. ISSN 1450-223X.

[12] Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R. and Lichtendahl Jr. K. C. (2017) “*Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*”, John Wiley & Sons, p. 576. ISBN 978-1-118-87936-8.

[13] Tyrychtr, J. and Vasilenko, A. (2015) “*Business Intelligence in Agriculture: Fundamental Concepts and Research*”, 1st ed. Brno: Konvoj, p. 80. ISBN 978-80-7302-170-2.

[14] Tyrychtr, J., Ulman, M. and Vostrovský, V. (2015) “Evaluation of the state of the Business Intelligence among small Czech farms”, *Agricultural Economics*, Vol. 61, No. 2, pp. 63-71. ISSN 0139-570X. DOI 10.17221/108/2014-AGRICECON.

[15] Tyrychtr, J. and Vostrovský, V. (2017) “The current state of the issue of information needs and dispositions among small Czech farms”, *Agricultural Economics*, Vol. 63, No. 4, pp. 164-174. ISSN 0139-570X. DOI 10.17221/321/2015-AGRICECON.

[16] Ugolnitskii, G. A and Usov, A. B. (2008) “Information-Analytical System for Control of Ecological-Economic Objects”, *Journal of Computer and Systems Sciences International*, Vol. 47, No. 2, pp. 321-328. ISSN 1064-2307.

[17] Vassiliadis, P. and Sellis, T. (1999) “A Survey of Logical Models for OLAP Databases”, *ACM Sigmod Record*, Vol. 28, No. 4, pp. 64-69. ISSN 0163-5808.

[18] Vercellis, C. (2011) “*Business intelligence: data mining and optimization for decision making*”, John Wiley & Sons, 436 p. ISBN 978-0-470-51138-1.

[19] Verstegen, J. A., Huirne, R. B., Dijkhuizen, A. A. and Kleijnen, J. P. (1995) “Economic value of management information systems in agriculture: a review of evaluation approaches”, *Computers and electronics in agriculture*, Vol. 13, No. 4, pp. 273-288. ISSN 01681699.

[20] Wagner, W. P. (2017) “Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies”, *Expert Systems with Applications*, Vol. 76, pp. 85-96. ISSN 0957-4174. DOI 10.1016/j.eswa.2017.01.028.

[21] Wrembel, R. and Koncilia, C. (Eds.) (2007) “*Data warehouses and OLAP: concepts, architectures, and solutions*”, Igi Global, p. 332. ISBN 1-59904-364-5.

[22] Zekri, A., Massaâbi, M., Layouni, O. and Akaichi, J. (2017) “Trajectory ETL Modeling”, In: *International Conference on Intelligent Interactive Multimedia Systems and Services*, Springer, Cham, pp. 380-389. ISBN 978-331959479-8. DOI 10.1007/978-3-319-59480-4_38.