www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

DE GRUYTER OPEN

Arch. Min. Sci., Vol. 60 (2015), No 3, p. 847–857

Electronic version (in color) of this paper is available: http://mining.archives.pl

DOI 10.1515/amsc-2015-0055

KAROL FIREK*, JANUSZ RUSEK*, ALEKSANDER WODYŃSKI*

DECISION TREES IN THE ANALYSIS OF THE INTENSITY OF DAMAGE TO PORTAL FRAME BUILDINGS IN MINING AREAS

DRZEWA DECYZYJNE W ANALIZIE INTENSYWNOŚCI USZKODZEŃ BUDYNKÓW HALOWYCH NA TERENACH GÓRNICZYCH

The article presents a preliminary database analysis regarding the technical condition of 94 portal frame buildings located in the mining area of Legnica-Głogów Copper District (LGOM), using the methodology of *decision trees*. The scope of the analysis was divided into two stages. The first one included creating a *decision tree* by a standard *CART* method, and determining the importance of individual damage indices in the values of the technical wear of buildings. The second one was based on verification of the created *decision tree* and the importance of these indices in the technical wear of buildings by means of a simulation of individual dendritic models using the method of *random forest*. The obtained results confirmed the usefulness of *decision trees* in the early stage of data analysis. This methodology allows to build the initial model to describe the interaction between variables and to infer about the importance of individual input variables.

Keywords: damage to buildings, portal frame buildings, mining area, technical wear, decision trees

Celem prezentowanych w artykule badań było sprawdzenie możliwości pozyskiwania informacji na temat udziału uszkodzeń w zużyciu technicznym zabudowy terenu górniczego z wykorzystaniem metody *drzew decyzyjnych*.

Badania przeprowadzono na podstawie utworzonej przez autorów bazy danych o stanie technicznym i uszkodzeniach 94 budynków typu halowego, usytuowanych na terenie górniczym Legnicko-Głogowskiego Okręgu Miedziowego (LGOM).

Do analiz przyjęto metodę *drzew decyzyjnych CART – Classification & Regression Tree*, na bazie której utworzono model aproksymujący wartość zużycia technicznego budynków. W efekcie ustalono wpływ poszczególnych zmiennych na przebieg modelowanego procesu (Rys. 3 i 4). W drugim etapie, stosując metodę *losowych lasów* przeprowadzono weryfikację wyników uzyskanych dla modelu utworzonego metodą *CART* (Tab. 2).

Przeprowadzone badania pozwoliły na ustalenie udziałów wyspecyfikowanych kategorii uszkodzeń elementów badanych budynków w ich stopniu zużycia technicznego. Największy udział w zużyciu technicznym budynków stwierdzono w przypadku uszkodzeń ścian wypełniających i osłonowych, warstw

* AGH UNIVERSITY OF SCIENCE AND TECHNOLOGY, AL. A. MICKIEWICZA 30, 30-059 KRAKOW, POLAND

elewacyjnych oraz wewnętrznych elementów wykończeniowych. Z kolei najmniej istotny wpływ na stopień zużycia budynków mają uszkodzenia elementów stężających oraz zewnętrznych.

Z rezultatów przeprowadzonych badań wynika, że wykorzystanie metody *drzew decyzyjnych* może okazać się przydatne w początkowej fazie analizy danych. Pozwala ona na utworzenie wstępnego modelu oraz wnioskowanie o udziałach poszczególnych zmiennych wejściowych w zmienności zmiennej zależnej. Zaletę metody *drzew decyzyjnych* stanowi fakt, iż mimo niejawnej reprezentacji ostatecznego podziału przestrzeni wielowymiarowej, struktura drzewa jest w pełni przejrzysta i pozwala na interpretację powiązań przyczynowo-skutkowych w modelu. Drzewa decyzyjne, oprócz aproksymacji funkcji wielu zmiennych, pozwalają na analizę struktury utworzonego systemu z możliwością jego adaptacji do innych modeli wnioskowania (np. *sieci Bayesowskich* bądź *rozmytych systemów regułowych*), oraz ocenę istotności poszczególnych zmiennych wejściowych.

Słowa kluczowe: uszkodzenia budynków, budynki halowe, teren górniczy, zużycie techniczne, drzewa decyzyjne

1. Introduction

Due to the impacts of mining activities, risk assessment for a development located in a mining area requires, inter alia, a detailed inspection of its technical condition. This inventory is performed to assess the resistance to the effects of mining, and to determine the scope of necessary preventive security measures, as well as to assess the extent of possible mining damage. In mining areas, portal frame buildings are a technically different and less frequently analyzed group of structures. During the inspection of the technical condition of these structures, assessing the scope and intensity of damage, both in the context of determining their causes and ways of eliminating them, was of great difficulty (e.g. Wodyński, 2007; Ostrowski & Ćmiel, 2008; Barycz & Oruba, 2009).

The aim of the research presented in this article was to examine the possibilities of capturing information on the importance of damage in the technical wear of this type of development, using the method of *decision trees*. From the professional literature, it is apparent that *decision trees* are the recognized *data mining* method, allowing to solve both classification and regression problems (Breinman et al., 1984; Ćwik & Koronacki, 2005; Morzy, 2013). In addition to approximation of multiple variables function, they allow to analyze the structure of the created system and to assess the importance of individual input variables. In addition, *decision trees* in the regression approach, in contrast to, for example, the traditional method of multiple regression, allow to present the course of the approximated function, taking into account its local variation, performing approximation in separate partitions of the space of the input variables (Hastie et al., 2009).

The research studies were carried out based on the database prepared by the authors, regarding the technical condition and damage to 94 portal frame buildings located in the mining area of Legnica-Głogów Copper District (LGOM).

The method which was used for the analysis was the *decision tree* method using *CART - Classification & Regression Tree* (Statistica, 2011). This method formed the basis for creating a model approximating the value of technical wear of buildings, as well as individual input variable importance, resulting directly from the structure of the created model, were determined. The next stage of the study focused on the verification of the results obtained in the context of their stability. For this purpose, a method of *random forests* was used for verification (Freund & Shapire, 1997; Ćwik & Koronacki, 2005; Hastie et al., 2009; Morzy, 2013).

2. Research methodology

2.1. Introduction

Each *decision tree*, whether classification or regression one, takes the form of a coherent acyclic graph that represents the process of splitting the learning set into homogeneous subsets. The root of this tree comprises the entire learning data set (Stąpor, 2011), and the other elements are the interior nodes, or the places where the splitting of a given subset is performed, and the leaves, or the terminal place where the splitting of consecutive subsets does no longer proceed.

Graphical interpretation of the *decision tree* has been illustrated in Figure 1 (Hastie et al., 2009). There is (a), a graphical result of the modeled function approximation, and (b), the structure of the *decision tree*. In addition, there is (c), which is the manner of space division of the inputs (partitions), dictated by the structure of the tree.

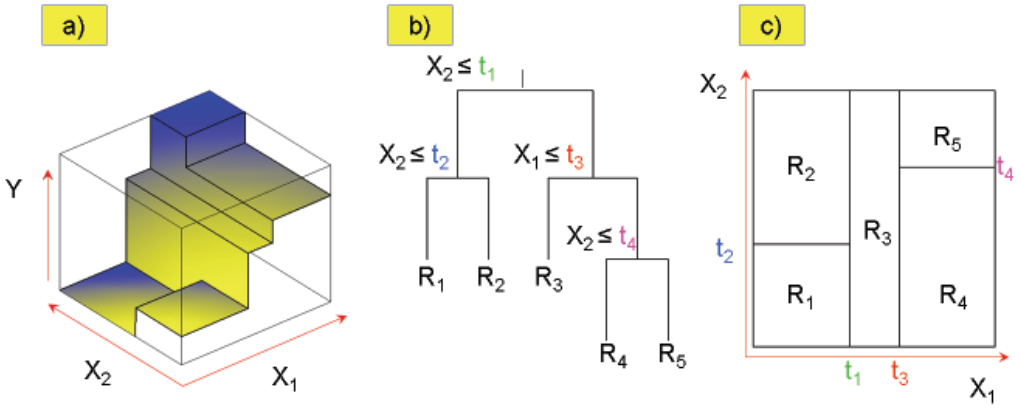

Fig. 1. Graphical interpretation of the *decision tree* structure (described in the text).
Source: Hastie et al., 2009

To solve similar problems, multiple regression is also used, the parameters of which are determined by means of the objective function minimization by the method of least squares. The advantage of *decision trees* over multiple regression is evident in at least two issues (Hastie et al., 2009).

The first one is the local estimation of the values of the degree of wear for the subsets determined during the constructing of the tree (Fig. 2), and identification of the interactions between input variables, possible to be interpreted by the user. The existence of the interactions between input variables during the stage of model construction disrupts the correctness of the results obtained by the method of least squares for multiple regression.

The second advantage is due to the dichotomy of splitting individual input variables, as imposed by the *decision trees* construction procedure. This allows to avoid a situation in which the approximated function abnormally fits to the subsets that differ in terms of data uniformity.

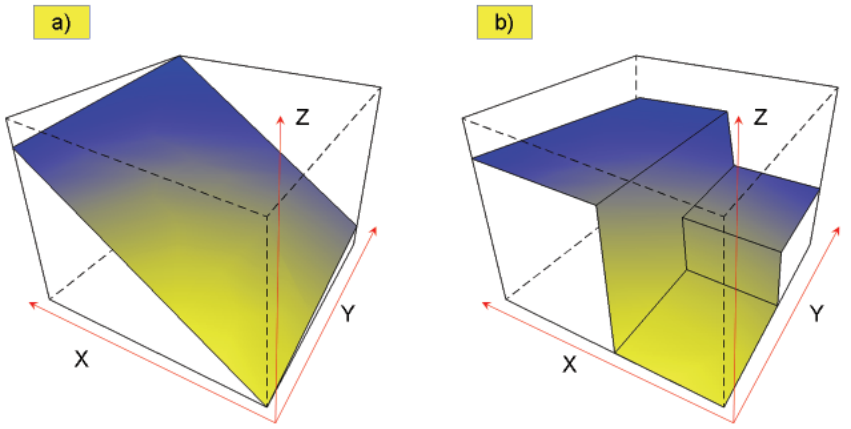

Fig. 2. Differences between approximation of a given function by: a) multiple regression, b) *decision tree*. *Source:* Carey et al., 2005

2.2. Mathematical bases for constructing and functioning of decision trees in regression terms of the *CART method*

The regression approach using *decision trees* and CART methodology (Breinman et al., 1984; Hastie et al., 2009; Morzy, 2013) is based on the division of the original space of input variables (features) into a set of the so-called partitions. In these areas, the local value of the approximated function is estimated, as a mean value of the observed data, attributable to a given partition.

The general form of the approximated function can be written as:

$$f(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\} \tag{1}$$

where:

$I\{\cdot\}$ — indicator function,

R_m — m-th partition obtained by dividing the space of input variables according to the procedure of the *CART* method,

c_m — the estimated value of the function, representative for the observed data, included in the partition R_m.

In the regression approach, the estimated value within a given partition R_m is the mean value of a set of learning data contained in it:

$$c_m = avg(y_i \mid x_i \in R_m) \tag{2}$$

The main problem in the *CART* method procedure is an appropriate division in the space of the features so that the approximate function had the best fitting to the raw data, and at the same time it did not exhibit overfitting (Borisov et al., 2009). The tree construction procedure is based on an adaptive approach and follows the minimization of the adopted objective function. In the regression approach, the objective function is a measure of *SSE* (*Sum of Squares Error*).

Depending on the type of a task, which could be a study of regression or classification, objective functions can take other forms, together classified as the so-called impurity measures (Borisov et al., 2009, Breinman et al., 1984; Stąpor, 2011).

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}^{N}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}^{N}(y_i-c_2)^2\right] \tag{3}$$

where:

$\{x_i,y_i\}$, $i=1\ldots N$, $x_i=[x_{i1},x_{i2},\ldots,x_{ip}]$ — a set of learning data,

$R_1(j,s)=\{X|X_j\le s\}$; $R_2(j,s)=\{X|X_j>s\}$ — areas of conditional belonging of data, resulting from the splitting of j-th predictor,

$c_1=avg(y_i|x_i\in R_1(j,s))$; $c_2=avg(y_i|x_i\in R_2(j,s))$ — mean values of the data in a given node to a dichotomously separated subsets R_1 and R_2.

In this paper, in addition to the constructing of a *decision tree* for the approximation of multi-dimensional function describing the degree of technical wear, also an attempt was made to assess the importance of individual input variables in the course of the modeled phenomenon. The assessment is made possible by the construction procedure of a *decision tree* itself. During the construction of a tree, at each stage of the division, these variables are selected whose splitting causes the greatest reduction in the error of the estimated value of the function of the degree of technical wear.

For a single *decision tree* (T), in (Breinman et al., 1984) a measure was proposed, defining the importance of individual variables (*VI-Variable Importance*) that have the greatest influence during node division in a given tree. This measure takes the general form (Borisov et al., 2009):

$$VI(X_i,T)=\sum_{t\in T}\Delta I(X_i,t) \tag{4}$$

where:

$VI(X_i,T)$ — a measure of *Variable Importance* determined for all nodes of the *decision tree,* in which the division is made with respect to the variable X_i,

$\Delta I(X_i,t)$ — reducing the value of the "impurity" function, obtained with respect to variable X_i at the node t,

$I(X_i, t)$ — "impurity" function for the regression X_i at the node t, expressed as the value of $SSE=\sum_{i\in t}^{N}(y_i-c_t)^2$,

T — the number of all nodes in the *decision tree* outside the root.

It should be noted that this measure is quantitative, and more specifically, it describes the importance of a given input variable in clarifying the variability of the dependent variable by the structure of the tree T. In addition, according to the implementation of such a procedure in the Statistica, the importance analysis in all nodes takes into consideration the rest of the variables that are potentially competing with the target variable, at which there is a division at a given level. Therefore, the ranking of importance of input variables can take into account these variables, which are not present in the final structure of the tree (Statistica, 2011).

2.3. The method of *random forests*

A very stable approach, applied only to *decision trees*, is the extension of the model construction called *random forests*. In contrast to the *CART* method, in the method of *random forests* the model is a group of trees having a simpler structure, and the prediction value is obtained by averaging the results from each tree component (Ćwik et al., 2005).

The accuracy of prediction in *random forests* is improved thanks to the weakening of the correlation between the component trees (Hastie et al., 2009; Morzy, 2013). In general, the reduction of correlation is possible due to the iterative procedure, during which individual input attributes are selected at random.

Due to the fact that there is no possibility of a direct comparison of *CART* model structures with respect to *random forests,* it was decided that this extension, as a more stable one, will be used only to verify the results obtained by *CART,* regarding the influence of individual input variables on the value of technical wear of building structures.

3. Description of database

3.1. Technical characteristics of the study group of portal frame buildings

The study used information collected during the inventories carried out with the participation of the authors in the years 2002-2010 (Inventory sheets, 2002-2010). A database of 94 portal frame buildings, located in the mining area of Legnica-Głogów Copper District (LGOM), was created. These structures were built in the years 1960-2005, and their average age is 30 years. They perform a variety of functions: production, commercial, office or warehouse.

The study group includes single-story portal frame buildings, mostly one-segment structures (69%), and in 28 cases multi-segmented structures, divided by expansion joints 3-80 mm wide. In 49% of cases, the horizontal projection was qualified as simple with compact shape, and in 37% as elongated. A poorly fragmented projection is less frequently occurring (12%). The studied buildings mostly have constant height.

In terms of design, buildings with steel structures are the most numerous in the study group (29%), with masonry load-bearing walls (28%) and with a reinforced concrete prefabricated structure (26%).

Mostly, these are portal frame buildings with a double frame static scheme, and less often with a rigid framework.

The foundations were made as monolithic reinforced concrete (85%) or concrete, founded at a constant level. Foundation walls were made of concrete (78%), of brick or concrete blocks. As far as load-bearing walls or infill walls of the superstructure are concerned, those of brick masonry and cellular concrete blocks dominate (53%), as well as those built of concrete blocks. The lintels were made as monolithic reinforced concrete, prefabricated or on steel beams. Most flat roofs are full slab roofs, with a classic arrangement of insulating layers, laid on prefabricated reinforced concrete roofing beams, or in the case of a steel structure – on the roof structure of trapezoidal steel sheet.

3.2. Technical condition and the extent of damage in the studied buildings

The measure of technical condition of buildings is the degree of their wear. As part of the described research, the degree of technical wear was determined for individual buildings by the method of weighted average, taking into account individual construction and technological solutions (e.g. Wodyński, 2007). Most of the studied portal frame buildings have a degree of wear within the range of 20-50%.

In order to examine the importance of damage in the technical wear of each building, qualitative damage intensity index w_{ui} was determined for the individual structural and non-structural components (Table 1). This index was defined in (Firek, 2009) in a 6-point scale, in which $w_{ui} = 0$ means that the damage does not occur, $w_{ui} = 1$ – slight damage (0-10%⟩, $w_{ui} = 2$ - moderate damage (10-30%⟩, $w_{ui} = 3$ – heavy damage (30-50%⟩, $w_{ui} = 4$ (and 5) – very intensive damage (over 50%).

Preliminary analysis of the value of the damage intensity index w_{ui} in the study group of buildings proved that most of the objects were damaged slightly or moderately.

TABLE 1

Damage intensity indices of building components.
Source: own study

Designation	Index description
Components of load-bearing structure	
w_{u1}	intensity of damage to components of load-bearing structure (columns, transoms, or load-bearing walls)
w_{u2}	intensity of damage to roofing structure (flat roofs or ceilings)
Secondary components (finishing elements)	
w_{u3}	intensity of damage to bracings (elements ensuring spatial rigidity)
w_{u4}	intensity of damage to infill and curtain walls
w_{u5}	intensity of damage to layers of a facade
w_{u6}	intensity of damage to interior finishing elements (partition walls, plasters wall and floor coverings)
w_{u7}	intensity of damage to damp-proof insulation, roofing, flashings, gutters and downpipes
w_{u8}	intensity of damage to exterior elements (roofing at the entrance to the building, platforms, wall trims)

For example, the most common damage to the reinforced concrete load-bearing elements of portal frame buildings (damage intensity index w_{u1}) are their scratches and cracks of concrete, as well as reinforcing steel corrosion and deformation and, in the case of steel columns and transoms – corrosion of steel. In the study group, the elements of the load-bearing structure (columns and transoms) do not exhibit damage threatening the safety of the structure.

4. The scope and results

The research studies described in this paper were divided into two stages. During the first stage, a model was created by the *CART* method, and inference was conducted about the influence of individual input variables.

During the second stage, using the method of *random forests*, the results obtained for the model created by the *CART* method were verified.

4.1. Building a model using the *CART* method and assessment of input variable importance

The database was divided in a ratio of 0.7/0.3 to obtain a learning (training) set and a test set. Using the prepared test set, *cross-validation test* was performed to reduce the risk of model overfitting. For the calculations, the Statistica module was used "General models of Classification and Regression Trees" (Statistica, 2011).

In the *CART* method, it can be inferred about the importance of individual variables in the course of the modeled value of the degree of wear. The inference can be performed based on the determined values of the parameter $VI(X_i,T)$, in accordance with the implementation of the developed procedure which is used in the program Statistica 10.

The resulting importance of damage intensity indices of individual structural and secondary components in the approximated value of the degree of technical wear of a building have been illustrated in Figure 3. The structure of a *decision tree* for the degree of technical wear s_z of the studied buildings has been presented in Figure 4.

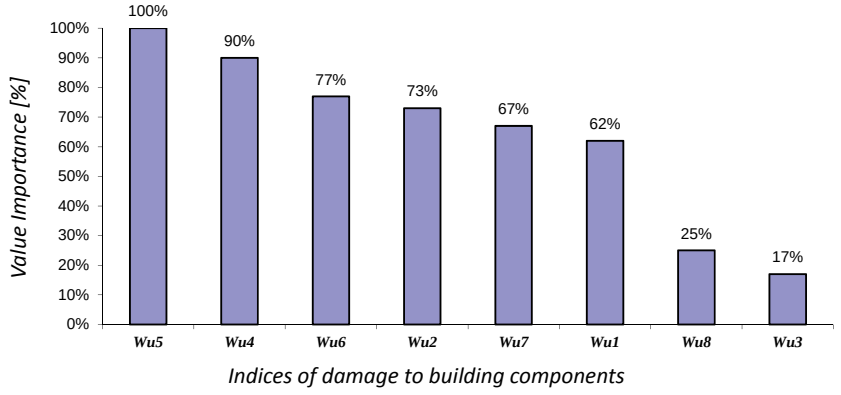

Fig. 3. Importance values of indices of damage to individual structural and secondary components in the degree of technical wear of buildings according to the adopted measure $VI(X_i,T)$
Source: own study

The structure of the created *decision tree* (Fig. 4) identifies the input variables that were included in the final model. For these variables, a splitting of data set is performed to minimize the SSE obtained at the output. Although some variables were not included in the final dendrogram (indices w_{u1} i w_{u8}), the assessment of their importance is also taken into account. It results from the process of selecting the variables optimal for splitting at every level of the tree, used in Statistica (Statistica, 2011). In contrast to the original method of determining input variable importance proposed by Breinman (Breinman et al., 1984), the program took into consideration not only those variables which eventually occurred in the dendrogram. The construction procedure

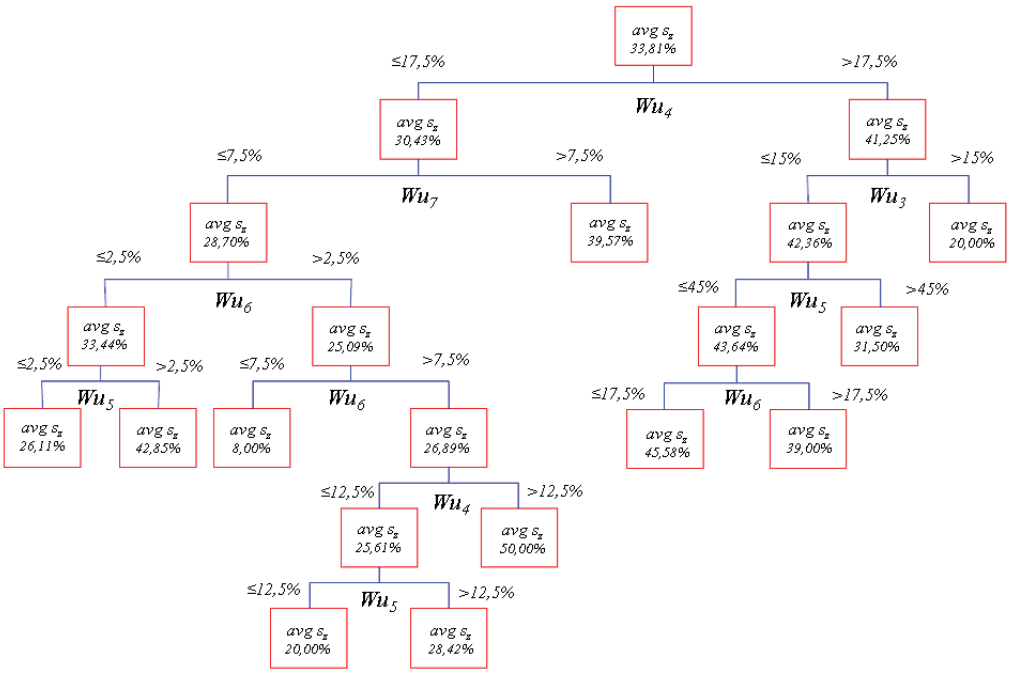

Fig. 4. Schematic *decision tree* structure created to predict values of the degree of technical wear of portal frame buildings. *Source:* own study

also took into account those variables that were to be split, but their importance did not prove to be significant enough at each stage.

Basing on the obtained model, it can be concluded that:

- the greatest importance in the modeled value of the degree of technical wear of buildings have variables describing the intensity of damage to elevation layers (w_{u5} index) and infill and curtain walls (w_{u4} index),
- the smallest influence on the value of technical wear of buildings have variables describing the intensity of damage to bracings (w_{u3}) and external elements (w_{u8}), characterized by much lower importance in the value of technical wear than other factors,
- other variables, including the variable describing the intensity of damage to load-bearing structure (w_{u1}) exhibit a relatively similar importance in the value of technical wear.

4.2. Analyzing the stability of results using the method of *random forest*

Analyses were performed for a specified test set in the same proportion as for the construction of a *decision tree* using *CART* method (0.7/0.3 = a training/learning set). A total of 30 independent experiments for learning subsets of variables, generated during the training were performed. As part of these analyses, attention was primarily focused on the stability of the obtained importance value of each category of damage generated for each subsequent experiment.

Summary of the obtained importance values (ranks) of intensity indices of damage to individual structural and secondary components in the degree of technical wear of a building for the model of a *decision tree* created using the *CART* method, and simulation using *random forests* have been presented in Table 2. The yellow color marks the highest ranks of indices, blue – intermediate, and green – the lowest.

TABLE 2

Ranking of damage intensity indices of specific elements included in the degree of technical wear of a building structure. Source: own study

Method	Ranking of damage intensity indices							
	w_{u1}	w_{u2}	w_{u3}	w_{u4}	w_{u5}	w_{u6}		w_{u8}
CART	6	4	8	2	1	3	5	7
random forests	5	6	8	1	3	2	4	7

Table 2 illustrates that assessment of importance of indices of damage to individual structural and secondary components using the *CART* method and *random forests* are comparable. Assuming that the *random forests* approach is a stable method, the results obtained by the *CART* method were confirmed regarding the assessment of influence of individual variables on the course of the modeled process.

It should be emphasized that in the problem of assessing importance of indices of damage to individual structural and secondary components in the course of the modeled process, the method of *random forests* may be both a complement to, and a basis for, the verification of a selected *decision tree* obtained by the standard *CART* method. This is important because the result of the *CART* method, unlike of the *random forests,* is one *decision tree* structure (Ćwik et al., 2005; Freund et al., 1997; Hastie et al., 2009; Morzy, 2013), which in turn can be the basis for extended analyses (see Fig. 4).

5. Conclusions

The article presents a preliminary analysis of the database regarding the technical condition of 94 portal frame buildings located in the mining area of Legnica-Głogów Copper District (LGOM).

The study used the methodology of *decision trees*. The scope of the analysis was divided into two stages. The first one included creating a *decision tree* by a standard *CART* method, and interpreting the obtained importance of individual damage indices in the values of the technical wear of buildings. The second one, on the other hand, was based on verification of the created *decision tree* and the determined importance values by means of a simulation of individual dendritic models using the method of *random forest*.

The conducted research allowed to assess the importance of specified categories of damage to the elements of the studied buildings in the value of their degree of technical wear. And so, the greatest importance in the technical wear of buildings was identified for damage to infill and curtain walls, elevation layers and internal finishing elements, represented in the analysis by the indices w_{u4}, w_{u5} and w_{u6}. The least important in the technical wear of buildings proved to be damage to bracing elements (w_{u3}) and external elements (w_{u8}).

From the results of the performed analyses it is apparent that the approach using *decision trees* may prove useful in the initial phase of database analysis. It allows to create an initial model and to infer about the importance of individual input variables in the variance of the dependent output variable.

The conducted study confirmed that the advantage of *decision trees* is that despite the implicit representation of the final division of multidimensional space, the structure of the tree is fully transparent and allows for the interpretation of cause and effect linkages in the model. Decision trees, in addition to the approximation of functions of multiple variables, allow for the analysis of the structure of the created system (with a possibility of its adaptation to other models of inference, for example *Bayesian networks* or *fuzzy rule-based systems*) and for the assessment of the importance of various input variables. Moreover, an advantage of this method, both in regression and classification tasks, is that there are no restrictive requirements imposed as to the quality of data, such as, for example, preserving normal distribution or independence of input variables.

The article was prepared within the scope of the AGH Statutory Research no. 11.11.150.005.

References

Inventory sheets, 2002-2010. *Arkusze inwentaryzacyjne zabudowy miasta Polkowice i Lubin wraz z oceną odporności na wpływy górnicze, wykonane w Katedrze Geodezji Inżynieryjnej i Budownictwa AGH w latach 2002-2010*. Prace niepublikowane.

Barycz S., Oruba R., 2009. *Industrial reinforced concrete chimneys resistance to the influence of mining activities*. Arch. Min. Sci., Vol. 54, Iss. 4.

Borisov A., Runger G., Torkkola K., Tuv E., 2009. *Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination*. Journal of Machine Learning Research.

Breinman L., Friedman J.H., Olshen R.A., Stone C.I., 1984. *Classification and regression trees*. Belmont, Calif.: Wadsworth.

Carey V., Dudoit S., Gentleman R., Huber W., Irizarry R., 2005. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor Statistics for Biology and Health*. Springer Science+Business Media Inc., New York.

Ćwik J., Koronacki J., 2005. *Statystyczne systemy uczące się*. Wydawnictwo Naukowo-Techniczne, Warszawa.

Firek K., 2009. *Proposal for classification of prefabricated panel building damage intensity rate in mining areas*. Arch. Min. Sci., Vol. 54, Iss. 3.

Freund R., Shapire R., 1997. *A decision-theorethic generalization of online learning and an application to boosting*. Journal of Computer and System Science, 55.

Hastie T., Tibshirani R., Friedman J., 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics.

Morzy T., 2013. *Eksploracja danych. Metody i algorytmy*. Wydawnictwo Naukowe PWN, Warszawa.

Ostrowski, J., Ćmiel A., 2008. *The use of a logit model to predict the probability of damage to building structures in mining terrains*. Arch. Min. Sci., Vol. 53, Iss. 2.

Stąpor K., 2011. *Metody klasyfikacji obiektów w wizji komputerowej*. Wydawnictwo Naukowe PWN, Warszawa.

Statistica 2011. Data analysis software system, version 10, www.statsoft.com, StatSoft, Inc.

Wodyński A., 2007. *Zużycie techniczne budynków na terenach górniczych (The process of technical wear of buildings in mining areas)*. Uczelniane Wydawnictwa Naukowo Dydaktyczne AGH, Kraków.

Received: 04 June 2014