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

Arch. Metall. Mater., Vol. 61 (2016), No 3, p. 1709–1714

DOI: 10.1515/amm-2016-0277

K. REGULSKI*, J. JAKUBSKI*, A. OPALIŃSKI*, M. BRZEZIŃSKI*,#, M. GŁOWACKI*,**

THE PREDICTION OF MOULDING SAND MOISTURE CONTENT BASED ON THE KNOWLEDGE ACQUIRED BY DATA MINING TECHNIQUES

The subject of the study is the improvement of the quality of moulding sand preparation. An exploration research performed on the data concerning moulding sand quality parameters was described. The aim of the research was to find relationships between various factors determining the properties of moulding sands and, based on the results obtained, build models predicting the sand moisture content with the induction of classification and regression trees. A two-match prediction approach was demonstrated and its effectiveness in evaluating the moulding sand moisture content was discussed. The knowledge in the form of rules acquired in this way can be used in the creation of knowledge bases for systems supporting decisions in the diagnostics of the moulding sand rebonding process. Formalized knowledge also facilitates further processing of the measurement data.

Streszczenie: Praca dotyczy poprawy jakości procesu przygotowania mas formierskich. Artykuł prezentuje badania ekploracyjne wykonane dla danych dotyczących parametrów jakości mas formierskich. Celem badań było określenie zależności pomiędzy czynnikami określającymi właściwości mas formierskich, by na ich podstawie zbudować modele predykcji wilgotności w oparciu o indukcję drzew regresyjnych i klasyfikacyjnych. Przedstawiono w artykule podejście dwustopniowego dopasowania predykcji i wykazano jego skuteczność w ocenie wilgotności masy formierskiej. Pozyskana w ten sposób wiedza w postaci reguł może być wykorzystana przy tworzeniu baz wiedzy dla systemów wspomagających decyzje w diagnostyce odświeżania mas formierskich. Sformalizowana wiedza ułatwia również dalsze przetwarzanie danych pomiarowych.

Keywords: Application of Information Technology to the Foundry Industry, Moulding sands, Decision trees, ANOVA, Data mining.

1. Introduction

The role of quality assurance in foundry production is growing at the present time in direct proportion to the requirements of ensuring competitiveness and respect for the environment. One way to ensure the required technical characteristics of castings is the strict control of production parameters. The most time-consuming part of the casting process is the preparation of moulding sand, foundry patterns and core boxes with the following manufacture of moulds and cores, all of which are the operations performed before the cycle of metal melting and treatment starts. Green moulding sands which are used to prepare moulds are heterogeneous mixture of quartz sand, binder and compounds to prevent overcooking.

In foundry practice, numerous parameters responsible for the quality of castings depend on one single parameter, namely on the content of moisture in the green moulding sand. The presence of water is indispensable for the binder to form bonds between the sand grains conferring adequate strength to the moulding sand. Water influences several green moulding sand properties, like apparent density, flowability, friability, permeability, plasticity, adhesion, mouldability and thermophysical behaviour. It is an essential component of the moulding mixture, but among all the benefits it also exerts a negative impact on the sand increasing the gas evolution rate and being largely responsible for the formation of moisture condensation zone. It also enters in the reaction with some of the liquid metal components [1-3].

The moulding sand preparation technology involves two-stage of water dosing, hence the problem of predicting the moulds moisture which is the subject of many research.

What assists the moulding mixture rebonding process is rapid cooling of the sand combined with pre-conditioning or pre-wetting. Pre-conditioning accelerates cooling of the sand mixture, reduces the rate of dust formation, mitigates the thermal destruction of binder, but changes the time taken by bentonite to absorb water. For bentonites currently used by the foundry plants, this time is very long and when the total amount of water is introduced to the moulding mixture as late as during the mixing process, the risk arises that the sand mixture supplied to the moulding shop will not have the required technological properties. Therefore the sand mixture preparation process should be conducted in such a way as to ensure that during the time of mixing water is added only in a small predetermined quantity selected from the specified narrow range of values. Prediction of water content in the moulding sand could ensure that the sand will require only a

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

** THE JAN KOCHANOWSKI UNIVERSITY (JKU), KIELCE

Corresponding author: mbr@agh.edu.pl

very low amount of the additional water before reaching the mixing plant.

In large foundries expensive IT systems (eg. SAP R/3) are being implemented to handle finance and economic area, production planning, sales, purchasing and material management [22] . Expanding these systems for QM module (Quality management), which is important from quality control point of view, turns out very challenging to implement in foundries. The cheaper and easier solution would be to implement a simpler system customized to the actual needs of the foundry. In principle they should be more focused on the technologic data collection process. Such systems should verify the quality control in practice, and also should be based not only on the data recorded in real time but also on historical data [22].

In this regard, an interesting IT system named KonMas-final was presented in [21,23,24]. It facilitates making decision on the basis of the recorded data and also enables preparation of data as a base for analysis. This system represents the first of the developed quality management AQ support systems [22] fulfilling functions which are expected from the above-indicated module QM - [21]

Some interesting research on the properties of green molding sands were presented by Mahesh B. Parappagoudar, D.K. Pratihar, G. L. Datta [25]. They performed a study based on the properties of green moulding sands (degree of compaction, permeability, hardness of forms) in reference to moulding sand parameters (shape and grain size, type of binder, moisture). The authors present the possibility of using the backpropagation and generalized regression algorithms to predict properties of the moulds as a function of the parameters of moulding sands as well as to determine the input parameters of moulds based on the properties of the mould. Also some interesting studies using neural networks to predict the parameters of moulding sands are presented by M. Perzyk [26, 27]

These methods are helpfull to prepare the moulds with expected moisture and bentonite content, but there is a need to improve using data mining to suport rebonding of moulding sand for getting results closer to expected moisture.

The authors' intention is the attempt to use specific methods of prediction to support the proces of moulding sands rebonding.

2. Investigation methods

Study of relationships between various parameters of an industrial process has been the subject of many scientific publications i.e. [4-6, 11-15]. The statistical methods used for this purpose mainly aim at searching for correlations between the quantitative and qualitative factors, and building models for predicting the quantitative parameters and classification models for the dependent variables of quality.

From a range of the available classical statistical tools (Pearson's correlation coefficient, Chi-squared test χ^2, Test F, Sensitivity Analysis, Decision Trees, etc.) [6-8], in this study the analysis of variance (ANOVA) [4] was used. It allows for verification of the statistical significance of differences between means in groups determined by the quality factor which, in turn, allows determining the impact of this factor on the dependent variable examined. The significance test is carried out by variance analysis, i.e. splitting the total variance into components corresponding to the real random error and components that relate to the differences between means. The mean square deviation between groups (MS_{Ef}) and the mean square deviation within groups (MS_{Er}) are compared using Fisher's exact statistics with the distribution F [4]. If the hypothesis concerning the significance of differences between groups is true, then $MS_{Ef} > MS_{Er}$.

Another traditional research tool based on quantitative variables is the linear Pearson correlation coefficient [6]. The correlation coefficient is a measure of the linear relationship between variables. Calculated from the value of covariance, it is characterized by variability comprised in the range of -1 to 1. The absolute value of the coefficient determines the strength of the correlation between variables, where 0 indicates absence of the linear relationship. This factor is calculated as the ratio of the covariance of variables to their common total variance.

Using the traditional tools, i.e. the method of regression analysis, for forecasting an unknown quantitative value of the dependent variable, more complex algorithms for modelling of phenomena are built, such as the MARSplines (Multivariate Adaptive Regression Splines) algorithm building a regression model from the base functions which are a combination of a series of linear regression models, like in segmental regression [5].

When the examined phenomenon is of a non-linear nature, or underlying variable is a discrete signal, it is worthwhile to reach to the more modern statistical techniques, known as data mining, to mention even the induction of decision trees [6-8]. Data mining is a set of exploration techniques derived from the study of artificial intelligence aimed at finding dependencies, patterns and rules in large data sets operating in an automatic or semi-automatic mode.

Data Mining is a very broad term. The concept comprises both statistical tools like ANOVA or MARSplines and also rule induction algorithms using decision trees (CART, CHAID, and others) [9,10] and rough sets [11] or fuzzy logic [12], as well as artificial neural networks (ANN) [2, 13], support vector machines (SVM) and classifiers such as KNN or the Bayesian classifier [14,15].

All these methods have been tested and used in industry, and more detailed information can be found in the cited literature. In the problem under discussion only some of these tools have been used, the most important being the Classification and Regression Trees (CART).

CART algorithm is one of many algorithms for the induction of decision trees. It is based on hierarchical binary divisions of dataset aiming at a better separation of cases that are representatives of the dependent variable. The algorithm is heading to the ideal situation when in the resulting partition (leaf) will be placed instances of the same value of the dependent variable. CART is a universal algorithm when it comes to the type of dependent variable, for quantitative variables it builds regression trees, in which the criterion of division is based on the variance of quantitative trait, while for the discrete (qualitative) dependent variables it builds a classification tree based on the selected index of the purity of nodes (Gini index, G^2 - maximum likelihood statistical significance or Chi^2). Unfortunately, these methods are not as effective in

the prediction as artificial neural networks and support vector machines. The main reason why they do not achieve equally good results is the discretization of quantitative variables and generalizations forced thereby.

Decision trees are graphical representations of rules obtained by analyzing the data structures. These algorithms allows not only to create rules, but also to determine the validity of the variables in the model, which is sometimes as important as the model itself. The independent variable is defined as important in the classification process, or requesting information on classes, depending on its readiness to participate in the divisions of the dependent variable, which is measured during the construction of the tree. The determined validity enables creating a ranking of the independent variables in terms of their impact on the dependent variable. Validity is the degree of covariance with the dependent variable.

Certain indisputable advantages of classifiers based on the trees are: (1) their graphical representation - readable and easy to interpret and verify on the basis of domain knowledge; (2) ability to determine the validity of the predictors; (3) insensitivity to noise and outliers; (4) the result as a set of rules that can be used in other applications.

3. Investigation results

The results of testing moulding sand mixtures with different properties, different contents of bentonite and different moisture content levels are investigated in this study. The data set included 631 cases, but not all of the records contained the full feature vector. In individual studies, records with the lack of parameters were deleted in pairs in order to conduct the calculations. Previous studies [2, 18] indicated that the set of data containing 490, and even 370 values for each parameter is enough for precise results with ANN. To obtain more accurate results with ANOVA and CART we used 631 cases.

ANOVA revealed the lack of significance of the impact of the type of sand and bentonite content on the moulding sand moisture content (Fig. 1). Symbols ZS1, ZS2, MF1, MF4 and MF5 designates different kinds of moulding sands from different suppliers.

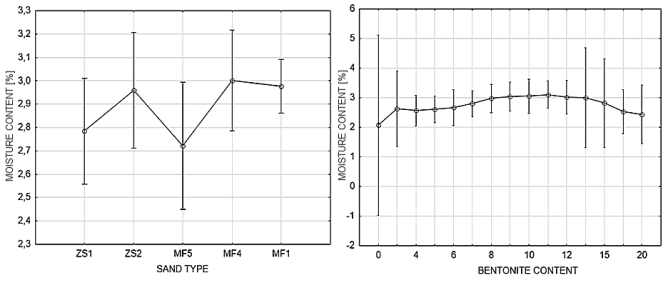

Fig. 1. The lack of significance of the impact of the type of sand and bentonite content on the moulding sand moisture content tested with ANOVA

Then the correlation was verified for quantitative variables. The correlation coefficients are given in Table 1. As can be seen, the strongest effect on the moulding sand moisture content was successively exerted by friability, compactability and permeability (although negligible), the weakest impact had the compressive strength, while density showed lack of any statistically significant effect on this parameter. (Symbols used in the table 1: r designates correlation and p stands for *p-value*.)

TABLE 1

The correlation matrix for the moisture content

	MOISTURE CONTENT	
	r	p
COMPRESSION STRENGTH	-0,21	p=,011
PERMEABILITY	-0,22	p=,008
COMPACTIBILITY	0,74	p=0,00
FRIABILITY	-0,84	p=0,00
DENSITY	0,11	p=,159

The type and form of the dependence can be seen in respective scatterplots (Fig. 2.). They confirm the already well-known relationship – moisture confers to moulding sand adequate strength and with increasing moisture content in the sand, the cohesiveness of the mixture increases to a maximum numerical value, and then decreases again. Water influences the following sand mixture properties: apparent density, flowability, friability, permeability, plasticity, adhesion, mouldability and thermophysical behaviour.

The scatterplots show another important aspect of this relationship, namely its non-linearity, which prevents the use of linear regression models. In the case of non-linear phenomena, whose form is not known a priori, models of regression trees are applicable. In this case, a model of the tree was built basing on the interactive CART algorithm using STATISTICA package (http://www.statsoft.com, 06-07-2016,).

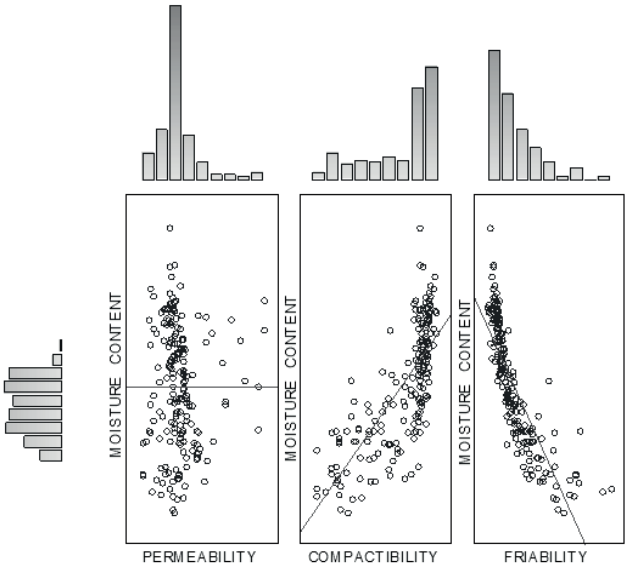

Fig. 2. Analysis of the impact of various parameters of sand on the moisture content

Three predictors of the variable called moisture content were used, namely friability, compactability and permeability.

The algorithm enabled the induction of 9 rules allowing for the prediction of moisture content based on predictors:

if ("FRIABILITY" <= 3.065) and ("PERMEABILITY" <= 205) then (PREDVAL = 5.788, VARIVAL = 0.36)
if ("FRIABILITY" <= 3.065) and ("PERMEABILITY" > 205) then (PREDVAL = 5.006, VARIVAL = 0.357)
if ("FRIABILITY" > 3.065) and ("FRIABILITY" <= 14,735) then (PREDVAL = 3.67, VARIVAL = 0.179)
if ("FRIABILITY" > 14.735) and ("FRIABILITY" <= 23.55) and ("PERMEABILITY" <= 335) then (PREDVAL = 2.426, VARIVAL = 0.366)
if ("FRIABILITY" > 14.735) and ("FRIABILITY" <= 23.55) and ("PERMEABILITY" > 335) then (PREDVAL = 2.911, VARIVAL = 0.0919)
if ("FRIABILITY" > 23.55) and ("COMPACTIBILITY" <= 8) then (PREDVAL = 3.156, VARIVAL = 0.092)
if ("FRIABILITY" > 23.55) and ("FRIABILITY" <= 50.5) and ("COMPACTIBILITY" <= 47) then (PREDVAL = 2.36, VARIVAL = 0.17)
if ("FRIABILITY" > 23.55) and ("FRIABILITY" <= 50.5) and ("COMPACTIBILITY" > 47) then (PREDVAL = 1.855, VARIVAL = 0.22)
if ("FRIABILITY" > 50.5) and ("COMPACTIBILITY" > 8) then (PREDVAL = 1.979, VARIVAL = 0.12)

In the above rules PREDVAL means 'predicted value of moisture content', while VARIVAL means 'the value of the variance in the class'. It can therefore be noted that the variance of the variable MOISTURE CONTENT has significantly decreased from the level of 1.3 for the full set to an average of 0.22 for each class. It means that the tree to a large extent explains the variability of the variable MOISTURE, as confirmed by matching of the results (determination coefficient) at a level of $R^2 = 0.85$ (Fig. 3a.). This result is only slightly inferior to the result obtained while using the artificial neural networks [2]. Some attention certainly deserves the fact that the study was performed on a large sample of high variability, which quite naturally gave the results of a much more generalized nature.

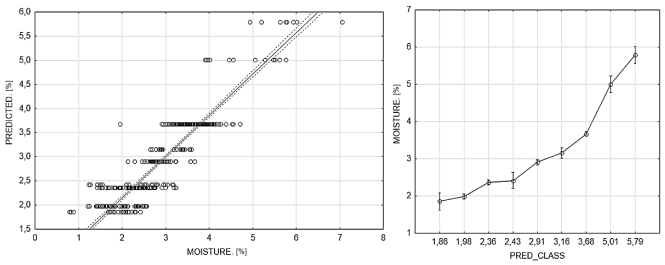

Fig. 3. (a) Output data distribution in relation to the experimental data for regression tree; (b) Average moisture content in each of predicted classes

The regression tree makes the prediction of moisture content using 9 classes of variability. The ANOVA significance tests showed that the first three pairs of classes did not differ in any more significant way, as can be seen on both the results of prediction (Fig. 3a) and diagrams of interaction between the variable MOISTURE and factor of classes (Fig. 3b). It was therefore decided to combine classes with the closest values, obtaining in this way 6 classes of the moisture content variations. Using this procedure, the discretization of the variable MOISTURE was done using a regression tree. The newly created variable MOIST_DIS was qualitative and could be used in further steps to build a classification tree (Fig. 4). The whole process can be called a two-match prediction.

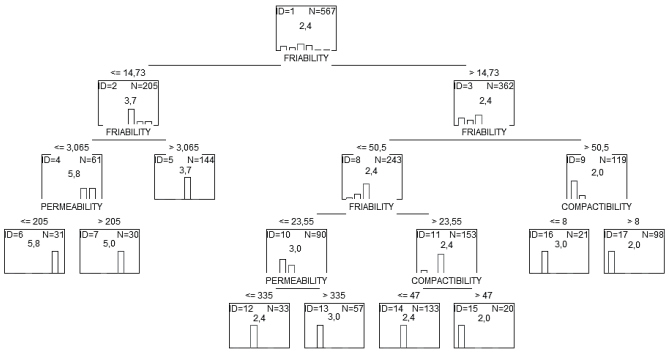

Fig. 4. Classification tree for the variable MOIST_DIS

The CART algorithm was used once again with the G^2 measure of the goodness of fit (maximum likelihood statistical significance). The tree allows for error-free classification, which is confirmed by the graph of matrix classification (Fig. 5a). The bar chart indicates which classes the tree has predicted the result. The bars on the diagonal indicate that the decision of the tree is consistent with the value of the observation. The bars outside the diagonal represent wrong decisions of the tree, which do not exist in the case of the tree built with the help of two-match prediction.

To verify the effectiveness of the discretization approach using the regression tree, the results were compared with a classification tree based on the dependent variable MOISTURE subjected to the traditional discretization using intervals of constant width. The result was obtained in the form of the error of 22% of wrong classifications (Fig. 5b).

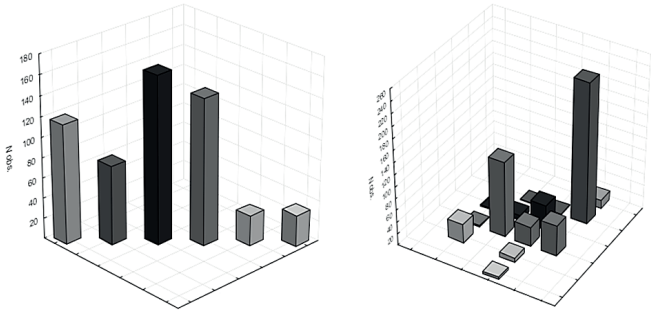

Fig. 5. (a) Classification matrix for the tree MOIST_DIS; (b) classification matrix for the tree built for MOISTURE with classical discretization

4. Conclusions

The presented data mining tools allowed for the analysis of relationships existing between moulding sand parameters and helped to build models for the moisture content prediction based on the sand properties. The quality of prediction measured by the coefficient of determination was fully

satisfactory for the regression trees, although slightly inferior to the models made by ANN described in previous publications of the authors [2].

Decision trees, however, have a very important advantage, which ANN do not have, namely allow for the extraction of knowledge in the form of rules, while neural networks operate on the principle of a black box, which means that they return the result, but do not allow for exploration of the phenomenon.

The course of analysis presented in this paper (Fig. 6.) further comprises the process of discretization of the dependent variable (MOISTURE CONTENT) to develop tree / rules for classes of moisture content.

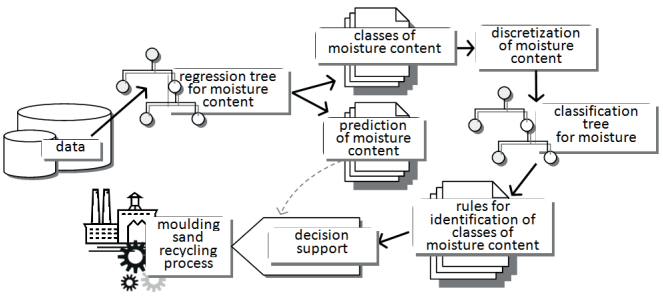

Fig. 6. Practical application of two-match prediction

The discretization by classes derived of a regression tree enabled a 100% effective classification, which could not be achieved in the case of a classification based on the variable discretized with intervals of constant width. It should be noted that the error of misclassification can not be compared directly, because in the first case the error was transferred to the fit of regression tree. However, the quality of prediction using two-match prediction clearly improves the quality of classification, which confirms the validity of this approach.

Obtained results are time-independent. Taking under consideration three parameters we can predict potential moisture of the sand. If the model could be a part of the industrial process with the mixing devices, time should be involved in the prediction as the feedback from the process, but it implies the need for new research and experiments.

Knowledge in the form of rules acquired in this process is applicable in decision- supporting systems regarding moulding sand rebonding. The possibility is also available to expand the exploratory analyses, and thus the acquired scope of knowledge, to include data from the Internet. First attempts of this type have already been made for other metallurgical processes [16,17]. Interesting seems also application of the developed methodology to data on the impact of other factors on the moulding sand [18]. Algorithmically acquired knowledge can assist technology choices and could be used in industrial systems for knowledge and quality management [19] or even by moulding sand mixers and other devices used in moulding sand preparation [20]. Acquired results apply to the tested samples of sands. Trying to use in a particular plant, presented methodology of calculation could be used.

It also seems reasonable to apply in the case of models for phenomena not fully determined a hybrid approach (neural networks, regression trees), which will provide the material for further studies.

Acknowledgements

The work has been supported by the Polish Ministry of Science and Higher Education – AGH University of Science and Technology Funds No. 11.11.110.300.

REFERENCES

[1] D. Hartmann, Process management and virtual engineering In foundries, Foundry – Science and Practice, 50, COCAFTEC, 2006, Foundry Research Institute, Cracow.

[2] J. Jakubski, Artificial neural networks (ANN) as a tool for predicting the moulding sands properties in terms of supporting green moulding sand quality control, PhD thesis, AGH University of Science and Technology, Cracow 2012, Archives of Foundry Engineering, Katowice-Gliwice (2013).

[3] J. Jakubski, St.M. Dobosz, Selected parameters of moulding sands for designing quality control systems, Archives of Foundry Engineering **10**(3), 11–16 (2010)

[4] P. Lewicki, T. Hill, Statistics: methods and applications, 2006, Tulsa, OK. Statsoft.

[5] J.H. Friedman, Multivariate Adaptive Regression Splines, Ann. Statist. **19**(1), 1-67 (1991).

[6] K. Regulski, D. Szeliga, J. Kusiak, Data Exploration Approach Versus Sensitivity Analysis for Optimization of Metal Forming Processes, Key Engineering Materials **611–612**, 1390–1395 (2014).

[7] J.R. Quinlan, Induction of decision trees, Machine Learning, **1**(1), 81-106 (1986).

[8] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and regression trees. 1984, CRC press.

[9] N. Speybroeck, Classification and regression trees, International Journal of Public Health **57**(1), 243-246 (2012).

[10] G.V. Kass, An exploratory technique for investigating large quantities of categorical data, Applied Statistics, 119-127, (1980).

[11] S.Kluska-Nawarecka, D.Wilk-Kołodziejczyk, K.Regulski, G.Dobrowolski, Rough sets applied to the RoughCast system for steel castings , Intelligent Information and Database Systems. Part II, Third International Conference, ACIIDS 2011, Daegu, Korea, April 20-22, 2011, Proceedings, Part II, Series: Springer Lecture Notes in Computer Science, Volume 6592/2011, 52-61, DOI: 10.1007/978-3-642-20042-7_6, Subseries: Lecture Notes in Artificial Intelligence, Nguyen, Ngoc Thanh; Kim, Chong-Gun; Janiak, Adam (Eds.), 1st Edition., 2011, XXVII, 580 p.

[12] M. Warmuzek, K. Regulski, A Procedure for in situ Identification of the Intermetallic AITMSi Phase Precipitates in the Microstructure of the Aluminum Alloys, Praktische Metallographie-Practical Metallography **48**, 12, 660-683 (2011).

[13] J. David, P. Svec, R. Frischer, R. Garzinova, The Computer Support of Diagnostics of Circle Crystallizers; Metalurgija **53** (2):193-196; APR-JUN (2014).

[14] J. Jakubski, St.M. Dobosz, The usage of data mining tools for green moulding sands quality control, Archives of Metallurgy and Materials **55**(3), 843-849 (2010)

[15] A. Glowacz, A. Glowacz, Z. Glowacz, Recognition of thermal images of direct current motor with application of area

perimeter vector and bayes classifier, Measurement Science Review **15** (3), 119-126 (2015). DOI: 10.1515/msr-2015-0018

[16] I. Olejarczyk-Wożeńska, A. Adrian, H. Adrian, B. Mrzygłód, Parametric representation of TTT diagrams of ADI cast iron, Archives of Metallurgy and Materials **57**, 981-986 (2012), DOI: 10.2478/v10172-012-0065-9

[17] A. Opalinski, W. Turek, K. Cetnarowicz, Scalable web monitoring system, in: Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on. IEEE, 2013.

[18] J. Jakubski, P. Malinowski, St.M. Dobosz, K. Major-Gabryś, ANN Modelling For The Analysis Of The Green Moulding Sands Properties, Archives of Metallurgy and Materials **58**(3), 961-964 (2013)

[19] P. Malinowski, J.S. Suchy, J. Jakubski, Technological knowledge management system for foundry industry, Archives of Metallurgy and Materials **58**(3), 965-968 (2013)

[20] K. Smyksy, E. Ziółkowski, R. Wrona, M. Brzeziński, Performance evaluation of rotary mixers through monitoring of power energy parameters, Archives of Metallurgy and Materials **58**(3), 911-914 (2013)

[21] Z. Ignaszak, R. Sika, System do eksploracji wybranych danych produkcyjnych oraz jego testowanie w odlewni. Archiwum Technologii Maszyn i Automatyzacji, 28(1), 61 – 72 (2008)

[22] K. Bramczewski, M. Idee, S. Szwajkowski, System pomiaru i rejestracji temperatury zalewania form, instrukcja obsługi programu RTO PC Soft s.c, Piła 1996.

[23] R. Sika, Studium nad strukturą systemu SAP R/3 i możliwości jego dostosowania do zarządzania oraz sterowania jakością w Odlewni Żeliwa ŚREM S.A., praca dyplomowa pod kierunkiem Z. Ignaszaka, Politechnika Poznańska, Wydział Budowy Maszyn i Zarządzania 2006.

[24] R. Sika, Z. Ignaszak, Po wdrożeniu programu KonMas-final -jego wykorzystanie do analizy procesu produkcji odlewów na wydziale W6 - Odlewni Żeliwa ŚREM S.A., w: XI International Symposium - Modeling of casting and foundry processes, 26™ 27 October 2006, Poznań-Śrem (Poland).

[25] B. Mahesh Parappagoudar, D.K. Pratihar, G.L. Datta, Forward and reverse mappings in green sand mould system using neural networks. Applied Soft Computing **8**, 239-260 (2008)

[26] M. Perzyk, A.W. Kochański, Prediction of ductile cast iron quality by artificial neural networks. Journal of Material Processing Technology **109**, 305-307 (2001)

[27] M. Perzyk, R. Biernacki, A. Kochański, Modeling of manufacturing processes by learning systems: The naïve Bayesian classifier versus artificial neural networks. Journal of Material Processing Technology **164–165**, 430–1435 (2005)