ARCHIVES of FOUNDRY ENGINEERING

Published quarterly as the organ of the Foundry Commission of the Polish Academy of Sciences

ISSN (1897-3310)
Volume 7
Issue 3/2007

111 – 116

22/3

Statistical and Visualization Data Mining Tools for Foundry Production

M. Perzyk [a,*]
[a] Zakład Odlewnictwa ITMat, Politechnika Warszawska, ul. Narbutta 85, 02-524 Warszawa, Poland
* e-mail: M.Perzyk@acn.waw.pl

Received on: 15.04.2007; Approved for printing on: 27.04.2007

Abstract

In recent years a rapid development of a new, interdisciplinary knowledge area, called data mining, is observed. Its main task is extracting useful information from previously collected large amount of data. The main possibilities and potential applications of data mining in manufacturing industry are characterized. The main types of data mining techniques are briefly discussed, including statistical, artificial intelligence, data base and visualization tools. The statistical methods and visualization methods are presented in more detail, showing their general possibilities, advantages as well as characteristic examples of applications in foundry production. Results of the author's research are presented, aimed at validation of selected statistical tools which can be easily and effectively used in manufacturing industry. A performance analysis of ANOVA and contingency tables based methods, dedicated for determination of the most significant process parameters as well as for detection of possible interactions among them, has been made. Several numerical tests have been performed using simulated data sets, with assumed hidden relationships as well some real data, related to the strength of ductile cast iron, collected in a foundry. It is concluded that the statistical methods offer relatively easy and fairly reliable tools for extraction of that type of knowledge about foundry manufacturing processes. However, further research is needed, aimed at explanation of some imperfections of the investigated tools as well assessment of their validity for more complex tasks.

Keywords: Castings quality management, Data mining, Visualization tools, Statistical methods, Foundry process parameters

1. Introduction

In many manufacturing companies large amounts of data are collected and stored, related to designs, products, equipment, materials, manufacturing processes etc. This data can be a source of valuable information. The extracting useful knowledge from that data, using intelligent and partly automated techniques, is called data mining. Data mining is viewed as a multidisciplinary field, which includes methodologies and tools from several disciplines such as database systems, visualization, statistics and learning (AI) systems. It is important that data mining techniques can provide various types of information. Much work has been done to develop methods of automated knowledge extraction from the recorded past data, usually in the form of logic rules of the type "if … then…".

Until now, data mining has been primarily used in business area and social sciences. Application to manufacturing and design on a large scale are seldom [1-4]. In the foundry production area there are several types of important practical problems that can be solved through extracting knowledge from a recorded past data, such as:

- Prediction of results of manufacturing process changes, including indication of optimal or critical process parameters, e.g. combination of time and temperature for heat treatment, influence of variations of chemical composition of an alloy on its mechanical properties etc.
- Detection of causes of deteriorating product quality. This can apply to the final products, e.g. increasing percent of defective castings, or intermediate products, e.g. lowered strength of molding sand.
- Prediction of breakdowns of machines, furnaces etc. The reason of the failure can be a combination of circumstances

(operation parameters) which cannot be identified 'manually'.

- Establishing rules for design of casting processes, e.g. rigging systems, or for process operations, e.g. molding sand preparation, melting procedures etc.

It is worth noticing that many of the above problems, especially from the first three groups, can be solved by determination of the most significant input variables, including possible interactions between them. This can be done by means of various types of data mining tools, including relatively simple statistical methods, possibly assisted by some visualization tools.

It is important to point out that statistical methods which are already present in manufacturing industry, mainly in the form of Statistical Process Control tools, are not able to provide that kind of knowledge. They are very useful in detecting the appearance of abnormalities of the process in the form of excessive variations of process parameters, but they are unable to indicate the causes of the irregularities. This has to be done by the technical staff.

The purpose of the present work is first to demonstrate and to discuss some of the possibilities, advantages and problems related to statistical and visualization methods. In the second part of the paper some results of the author's own research are presented, aimed at analysis of performance of some statistical methods.

2. Characterization of selected data mining tools for foundry applications

2.1. Visualization tools

Visualization tools are often treated as supplementary methods, providing better understanding and easier to interpret the knowledge discovered by the models [1]. Examples of such tools are flow charts, run charts, Pareto charts, Ishikawa diagrams, histograms, scatter plots, identification of outliers and others. However, some of those methods can be also extremely valuable for initial analysis of the problem, aimed at the right choice of the mode's variables, i.e. identification of potential process parameters and interdependencies between them, which could play important role in the process. Below, examples of two methods, probably less recognized by foundry technical staff, will be given.

The Pareto principle states that „not all of the causes of a particular phenomenon occur with the same frequency or with the same impact". Such characteristics can be presented using Pareto charts, which show the most frequently occurring factors and help to make best use of limited resources by pointing at the most important problems to tackle. From the exemplary chart shown below it can be concluded that the foundry staff should concentrate on reducing 2 defects: 'sand inclusions ' and 'gas holes ', which make up 72% of all defects. The Pareto diagrams can therefore be particularly useful in defining the targets of the whole data mining system.

Cause-and-effect diagram is a kind of putting together of factors affecting a process. Because of its shape sometimes it is also called fishbone diagram, or Ishikawa diagram (due to the name of its author, professor Kaoru Ishikawa from Tokyo University). They can reveal important relationships among various process variables and possible causes of faults as well as provide additional insight into process behavior. Construction of the diagram requires the following consecutively taken actions:

- make up a flow chart of the process,
- define the problem to be solved,
- find all possible causes of the problem (brain storm technique can be used),
- group these causes into categories,
- build the diagram which illustrates the relationships between the causes.

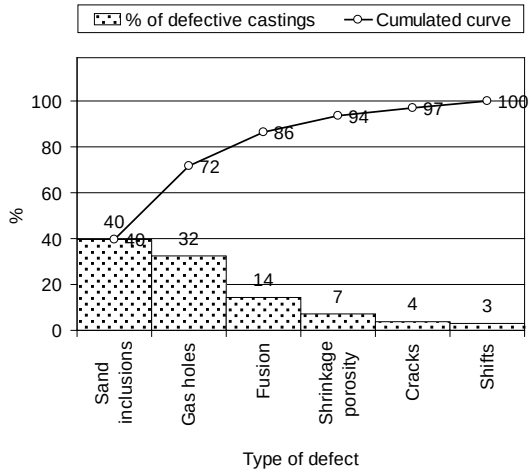

Fig. 1. Example of Pareto chart

In Fig. 2 an example of Ishikawa diagram used in a USA foundry [5] is shown; some more examples can be found in [6, 7].

2.2. Statistical methods

Statistical methods can be of several different types, characterized below.

Statistical regression models are probably the most popular in continuous type data analysis and generalization, often used in the form of so called empirical relationships. It is important that a particular form of the function must be assumed which requires a certain amount of knowledge about the modeled process. Choice of the type of function is usually carried out after plotting the experimental data. Significant difficulties occur for multivariable functions, which will be discussed later, in section 3.2. Linear and non-linear functions can be used (linear and non-linear regression tasks). The latter include polynomials (of arbitrary order) as well as other functions (e.g. power, exponential etc.). Analytical (unique) methods of determination of the parameters are available only for linear or polynomial types of functions. For other types, one can employ linearization of the function, or other optimization methods of the function parameters.

Statistical ANOVA (analysis of variance) methods as well as *contingency tables* techniques can be applied to detect and measure strength of dependence between variables. They can be also used to determine relative significance of each of the input variables on the output variable as well interactions existing between two of them. This can be particularly useful for solving many of the foundry problems mentioned in chapter 1.

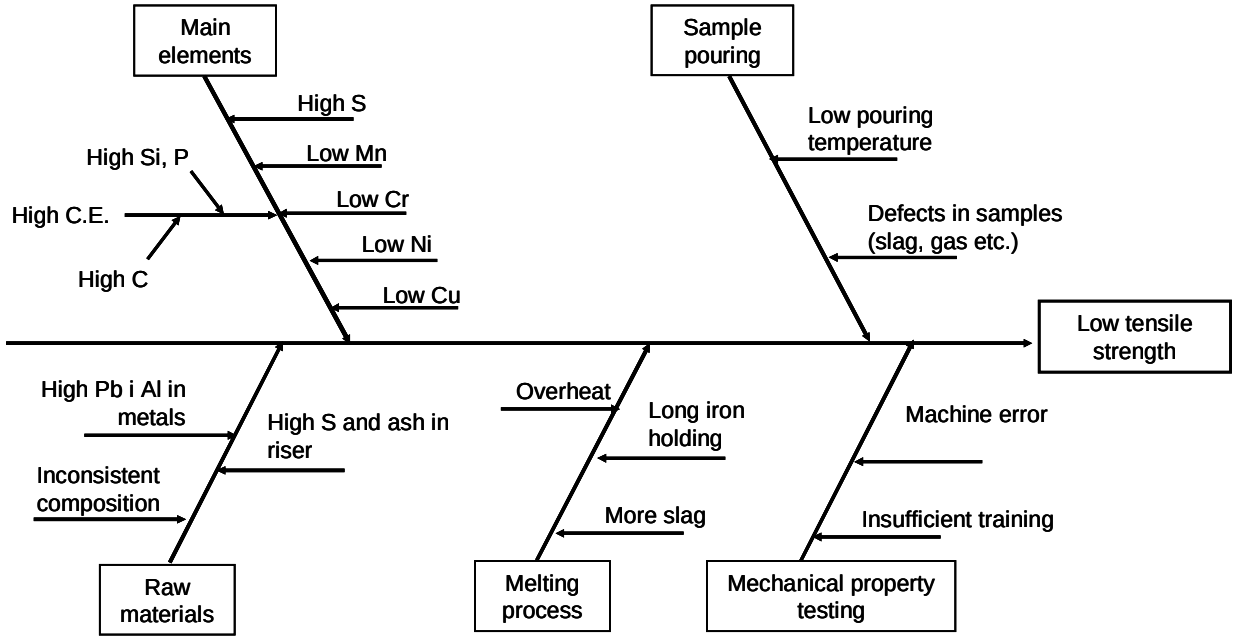

Fig. 2. Example of cause-and-effect (Ishikawa) diagram, used in Crane Valves foundry, USA [5]

It is important that these statistical methods do not require any assumption about the form of the dependency between input and output variables, which is the main disadvantage of the statistical regression models (see section 3.2).

It is worth noticing that the variables for ANOVA and contingency tables analyses must be of discrete type, i.e. nominal or ordinal. The data of continuous type must be converted to the ordinal type data before processing. This can be done by various algorithms, dependent on the characteristics of the data.

Statistical clustering methods and multidimensional scaling are the methods which apply to situations, when there is no dependent variable. They are used for grouping (clustering) of variables exhibiting similar characteristics. This kind of analysis can be also useful in analysis of manufacturing processes. Examples of potential applications are briefly characterized below.

- If a group, in which certain combinations of process parameters are included, is characterized also by a larger defectiveness of products, it could mean that this combination is a source of a lower quality.
- If a group, in which extreme values of parameters (close to the specified limits) are included, is associated with a particular operator, it is likely that he or she does not his work properly.
- If the clustering algorithm tends to group the process parameters in a number of significantly distinct groups which is greater from the number of different product types, it could indicate, that the process suffers from some severe and undesired variations.

Applications of the clustering methods to analysis and knowledge extraction in the manufacturing environment, particularly foundry production, are seldom.

3. Performance analysis of selected statistical data mining methods

3.1. Methodology

The general methodology employed in this research is based on utilization of simulated data sets containing assumed, but hidden relationships between variables. The data records were generated in the following way. First, an analytical formula of the type Y = f(X1, X2, …) was assumed. Then for random values of independent variables X1, X2, … the dependent variable Y was calculated. Finally, a Gaussian type noise with maximum deviations ±20% was imposed on the independent variables. Usually 1000 records for each data set were generated in that way. All the values were normalized within 0 – 1 interval.

Most of the data sets were generated repeatedly 5 times and the results of significance and interaction coefficients were presented in the form of their averages and 95% confidence intervals. Employment of that procedure allowed to evaluate the sensitivity of the analyzed coefficients to the noise present in the data.

For the statistical methods which require discrete types of data the number of intervals used for conversion real (continuous) values to categories was 10.

Additionally, the previously used industrial data set was utilized which relates chemical composition of ductile cast iron with its tensile ultimate strength (for details, see [8]).

For determination of the relative significance factors of independent variables two statistical methods were used: single-factor analysis of variance (ANOVA) and contingency tables. For

the industrial data set a 2nd order polynomial approximation (with mixed terms) was also applied, for comparison.

The ANOVA based significance factors were defined as the F-statistics values calculated for dependency between given independent and dependent variables, normalized by dividing them by the maximum value obtained for all independent variables. Similarly, the contingency tables based significance factors were defined as the normalized V-Cramer measures.

The significance factors calculated from the polynomial, also calculated for the industrial data set, were simple the normalized sums of the 2nd order and linear terms coefficients for the given variable, ignoring the mixed terms which included that variable.

The definition of interaction coefficient is based on the multi-factor analysis of variance and is defined as:

$$\frac{F_{i,j}}{\left(F_i + F_j\right)/2} \cdot 100 \qquad (1)$$

where F_i and F_j are test statistics for single variables(factors) i and j, respectively, and $F_{i,j}$ is the test statistics for the interaction of those variables.

The software used for ANOVA computations was Statistica version 7 package (by StatSoft). For the contingency tables and polynomial calculations a software developed by the present author, using VBA for Excel as the programming language, was applied.

3.2. Results

Exemplary results of the relative significance factors of independent variables obtained for the simulated data sets are shown in Fig. 3 and Fig. 4. It can be seen that the both investigated statistical methods reflect the general expected tendencies, however, their values are not accurate.

The ANOVA based factors essentially underestimate the significance of less important variables while the contingency based factors overestimate them. This observation is valid not only for the input variables without interactions, like those presented in Fig. 3, but also for the strongly interacting variables, like X1 and X2 in Fig. 4. The ANOVA based factors also exhibit much higher sensitivity to the noise present in the data. Some more results of that type can be found in [9].

In Fig. 5 the relative significance coefficients for the industrial data sets are presented. Here, the general tendency of higher significance predictions of ANOVA based method, compared to contingency tables, does not hold. However, the distinct role of copper, which was the main component used by the foundry to obtain high strength grades of SG iron, was identified, as the most important one, by both ANOVA and contingency tables methods.

The significance factor for copper obtained from the polynomial terms is different. It is quite likely, that the 2nd order polynomial was not able to correctly approximate complex relationships evidently existing in this case. Of course, one reason could be the assumed definition of the significance of single variables, ignoring the mixed terms in which that variable possibly also appears. This example illustrates evident disadvantages of the statistical regression methods. If the number of independent variables is large then the polynomial which would be capable of reflecting all possible interactions between variables should contain all the possible mixed terms and thus it would become extremely complex. This leads to problems related to finding of the model's constants as well as the analysis of the variables' significances.

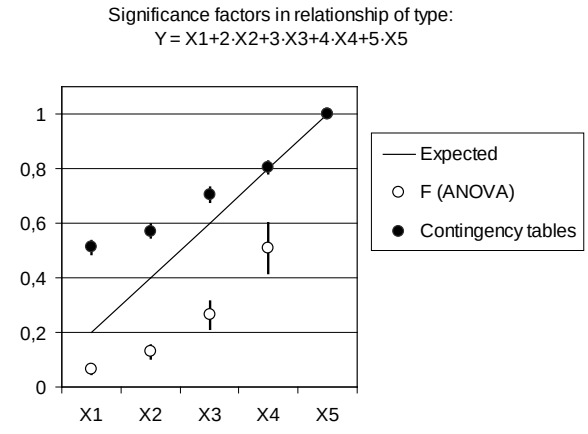

Fig. 3. Comparison of relative significance factors of independent variables without interactions, obtained by two statistical methods from 5 generations of simulated data sets; vertical bars denote 95% confidence intervals

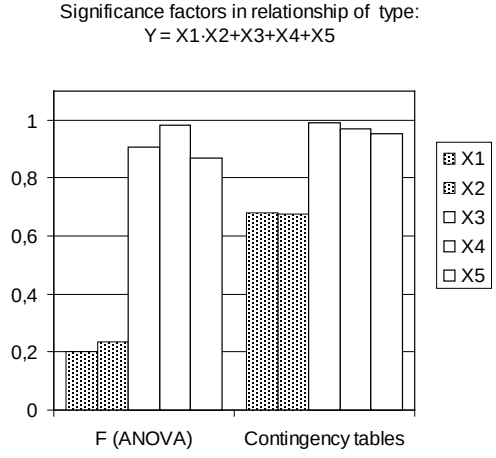

Fig. 4. Comparison of relative significance factors of independent variables with, and without, interactions, obtained by two statistical methods as averages from 5 generations of simulated data sets

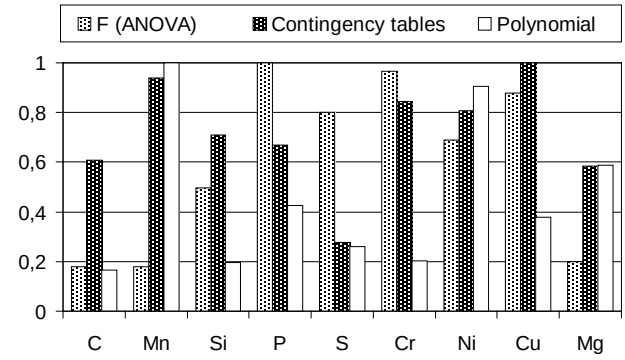

Fig. 5. Significance factors of alloying components calculated for tensile strength of ductile cast iron, based on an industrial data

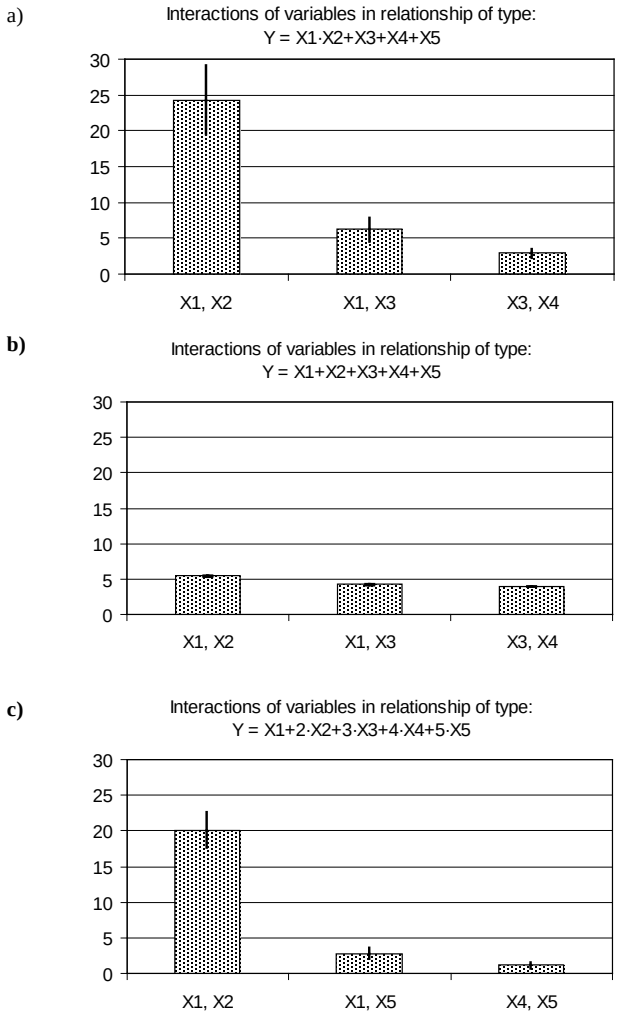

Fig. 6. Interaction coefficients between two input variables obtained for three simulated data sets by the multifactor ANOVA based procedure; a – data with strong interactions between two variables X1, X2 and with no other interactions; b and c – data with no interactions

In Fig. 6 exemplary results of the interaction coefficients calculated according formula (1) for several pairs of input variables are shown.

It can be seen that the ANOVA based method correctly predicts distinct interaction between variables X1 and X2 appearing in the first type of relationship (Fig. 6a) as well as the absence of other interactions in that data. Also for all pairs of variables appearing in the relationship presented in Fig. 6b it correctly indicates lack of interactions (small values of the coefficients). However, in the presence of other, more significant variables, as in the data generated according relationship shown in Fig. 6c, the behavior of the interaction coefficients is not entirely acceptable. In particular, unexpectedly large interaction is predicted by the ANOVA for the variables X1 and X2 which are obviously also independent on each other in that data set.

Another case of unexpected values of the multifactor ANOVA based interaction coefficients is presented in Fig. 7. In both relationships: Y = X1·X2 and Y = X1·X2+X3+X4+X5 the interactions between variables X1 and X2 should be this same. However, the formula (1) gives essentially lower interaction coefficient values for the second relationship, i.e. for the case where also other input variables contribute to the output variable.

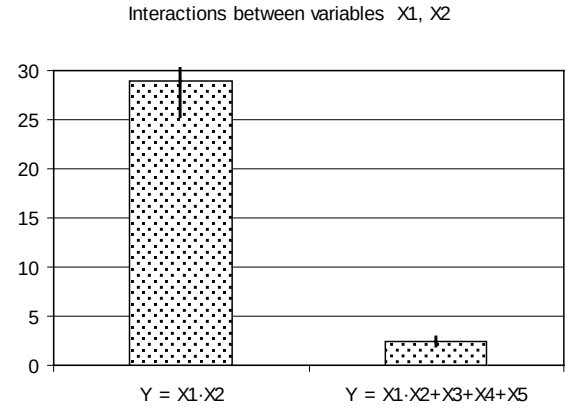

Fig. 7. Interaction coefficients between two variables with strong interactions obtained for two simulated data sets by the multifactor ANOVA based procedure

4. Discussion of results and conclusions

Application of data mining techniques in foundry industry creates new chances for achieving a better quality of products (final and intermediate) and higher production effectiveness. This can be accomplished by extraction and visualization of the knowledge hidden in the recorded past data. Although most of the advanced modeling methods concentrated until now on design and control of production processes (for foundry industry see [10-24]), their

application to the knowledge extraction seems to be a natural step forward now.

The performance of statistical techniques considered in the present research is only partly acceptable and a vast further work is needed. In general, it should include further analysis of behavior of various data mining methods and development of improved definitions of significance and interaction of variables (e.g. detection of synergetic action of several variables) as well as development of the software oriented at manufacturing problems.

In particular, the unexpected behavior of the multifactor ANOVA based interaction coefficients has to be investigated and explained. Also, the large and ambiguous differences between significance factors obtained form ANOVA and contingency tables for more complex relationships existing in the data require further exploration and clarification.

It is worth noticing that the methods of evaluation of relative significance of input variables based on single factor ANOVA and the contingency tables are utilized also in some commercial statistical software packages. They are recommended as important tools for preliminary identification of the less significant variables, which could be possibly ignored in application of advanced data mining models, such as artificial neural networks or classification and regression trees. The present results clearly indicate that this should be done with caution.

References

[1] D. Braha (ed.): Data Mining for Design and Manufacturing - Methods an Applications, Kluwer Academic Publ., Dordrecht, Boston, London, 2001.

[2] H. Sadoyan, A. Zakarian, P. Mohanty, Data mining algorithm for manufacturing process control, Int. J. Advanced Manuf. Technol. vol. 28, No. 3–4 (2006) 342–350.

[3] T. Demski, Statistics and data mining in practice, StatSoft, Warszawa – Kraków, 2004 (in Polish).

[4] M. Perzyk, Data mining in foundry production, Research in Polish Metallurgy at the Beginning of XXI Century, Committee of Metallurgy of the Polish Academy of Sciences, ed. K. Świątkowski, Kraków, 2006.

[5] X. Guo, Implementing Six Sigma in Foundry Industry, AFS Transactions, vol. 110 (2002), 199–210.

[6] S. Kannan, J. E. Thixton, System Approach to Casting Defect Analyses and Reduction: Hydrogen Gas Defect in Iron Castings, AFS Transactions, vol. 112 (2004), 115–119.

[7] P.L. Barker, B. Bidassie, Using Statistical Tools to Detect and Improve Core Shift: A Case Study, AFS Transactions, vol. 112 (2004), 121–130.

[8] M. Perzyk, A. Kochański, Prediction of ductile cast iron quality by artificial neural networks, Journal of Materials Processing Technology, 109/3 (2001), 305–307.

[9] M. Perzyk, J. Kozłowski, Comparison of statistical and neural networks-based methods in analysis of significance and interaction of manufacturing processes parameters, Computer Methods in Materials Science, vol. 6, No. 2 (2006), 81–93.

[10] K. Hatanaka, T. Tanaka, H. Kominami, Breakout forecasting system based on multiple neural networks for continuous casting in steel production, Fuijtsu Scientific and Technical Journal, vol. 29 (1993), 265–270.

[11] K. Terashima, Y. Maesa, H. Namura, Learning–control of mould hardness in blow molding, Proc. 60th World Foundry Congress, Hague 1993.

[12] W. Chen, G. Duan, C. Ou, Neural network applied to predicting molten steel temperature profile from converter to continuous casting, Iron and Steel (China), vol. 32 (1997), 30–32.

[13] Y. Otsuka, M. Konishi, Neural network models and its applications to iron and steel making processes, Journal of the Iron and Steel Institute of Japan, vol. 77 (1991), 1539–1543.

[14] P.F. Bartelt, N.G. Bliss, J.S. Moberley, Application of artificial intelligence to power input control in the modern foundry. AFS Transactions, vol. 103 (1995), 221–225.

[15] M.B. Brady, N.G. Bliss, H.D. Phillips, Integrated computer controls for the modern foundry meltshop. Proc. 51st Electric Furnace Conference, Washington, 1993, 117–120.

[16] E.D. Larsen, D.E. Clark, H.B. Smart, K.L. Moore, Intelligent control of cupola melting. AFS Transactions, vol. 103 (1995), 215–219.

[17] G. Wang, T.Y. Huang, Application of artificial neural networks in foundry industry. Proc. Third Asian Foundry Congress, Kyongju, South Korea, 1995, 424–431.

[18] P.F. Bartelt, M.R. Grady, D. Dibble, Application of intelligent techniques for green sand control. AFS Transactions vol. 104 (1996), 635–642.

[19] T. Watanabe, K. Omura, M. Konishi, K. Watanabe, K. Furukawa, Mold level control in continuous caster by neural network model, ISIJ International, vol. 39 (1999), 1053–1060.

[20] S. Calcaterra, G. Campana, L. Tomesani, Prediction of Mechanical properties in spheroidal cast iron by neural networks, J. Mater. Proc. Technol., vol. 104 (2000), 74–80.

[21] A. Faessler, M. Loher, Quality control in die casting with neural networks, Proc. Int. Symp. On Neuro-Fuzzy Systems, IEEE USA, Laussanne, 1996, 147–153.

[22] P.K.D.V. Yarlagadda, E.C.W. Chiang, Neural network system for prediction of process parameters in pressure die casting, J. Mater. Proc. Technol., vol. 89–90 (1999), 583–590.

[23] J.H. Zietsman, S. Kumar, J.A. Meech, I.V. Samarsekera, J.K. Brimacombe, Taper design in continuous billet casting using artificial neural networks, Ironmaking and Steelmaking, vol. 25 (1998), 476–483.

[24] M. Perzyk, R. Biernacki, Modeling of manufacturing processes by learning systems: The naive Bayesian classifier versus artificial neural networks, J. Mater. Proc. Technol., vol. 164–165 (2005), 1430–1435.