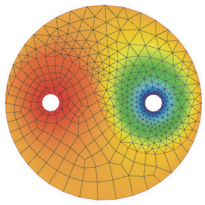
Publishing House
AKAPIT

**COMPUTER METHODS IN MATERIALS SCIENCE**

Informatyka w Technologii Materiałów

**Vol. 16, 2016, No. 4**



# APPLICATION OF SIGNIFICANCE ANALYSIS TO FINDING ROOT CAUSES OF PRODUCT DEFECTS IN CONTINUOUS CASTING OF STEEL

**A. RODZIEWICZ*, M. PERZYK**

*Institute of Manufacturing Technologies, Warsaw University of Technology, Narbutta 85, 02-524 Warszawa, Poland*
**Corresponding author: agnieszkarodziewicz88@gmail.com*

**Abstract**

The aim of this paper is to present an application of the input variable significance analysis to finding probable causes of product defects occurring in continuous casting (CC) of steel. The research was carried out using production data routinely recorded in one of Polish steel plants and basically referred to defective fraction of billets per heat as the process output. The data did not include the cases with zero defects which made the analysis difficult. The process inputs included eight parameters of different nature (physical, organizational and human). For determining which of the process input parameters are crucial for the output and which of them can be easily eliminated in further analyses two different approaches were applied and compared. The basic tool was an MLP-type Artificial Neural Network in which the relative significance was defined as the sum of the absolute weights of the connections from the given input node to all the nodes in the first hidden layer. As a complementary method the one-way analysis of variance (ANOVA) was utilized in which the value of the F-statistics is used as a measure of the input significance. It was found that the both methods indicate that the start-time of the CC process is the factor highly influencing the fraction of defective products. The process physical parameters which are expected to have a large influence on the billet quality, i.e. deviations from nominal casting temperature and deviation from nominal casting speed also appeared to be significant, moreover their variations also highly depend on the start-time of the CC process. The final conclusion is that the direct cause of the defective products are incorrect adjustments of the casting speed occurring mainly in the morning hours, however not correlated with particular operators. This finding can considerably facilitate the identification of the root cause of the defects by the plant engineers. Some recommendations concerning the future work are also given.

**Key words**: significance analysis, artificial neural networks, analysis of variance, continuous steel casting process

## 1. INTRODUCTION

The quality of manufacturing is a decisive factor in competing in a global market and it is of great interest of any production company to find root causes of defect in products, especially those appearing in the manufacturing process. In most companies large amounts of data are collected including those related to manufacturing, which can be a source of valuable knowledge, including failure and defect diagnosis. Extracting that knowledge from the data, using intelligent and partly automated techniques, is called Data Mining and is a multidisciplinary field including methodologies and tools from several disciplines such as database systems, visualization, statistics and especially advanced data-driven models called learning systems.

Detection of root causes of a deteriorating product quality is often a challenge for the engineering staff. A large variety of production factors can influence the quality, including those related to material, machines, operators, environmental conditions etc. It is important to point out that statistical methods which have been used extensively in manufacturing

 ISSN 1641-8581

industry for many years, such as Statistical Process Control tools, are not able to provide that kind of knowledge. They are 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.

The idea of the present study was to apply a significance analysis for the input process variables: those which would be found to be the most significant for a given quality parameter, e.g. percent of defective products, should be regarded as the first candidates for the causes of the quality decline. The considered production process was based on a key manufacturing technology utilized in steel industry, the continuous casting (CC).

## 2. METHODOLOGY OF THE RESEARCH

### 2.1. Industrial data

The data used in the present study have been collected in a Polish steel plant fitted with one continuous casting machine producing four strands. Each of them provides a semi-finished product called a billet which is a long square-profiled steel ingot. Defects of billets are formed during the solidification process. Due to the fact, that they can develop both internally and on the product surface, all possible defects can be divided into three types: surface, internal and shape - related.

The whole production data recorded in the plant refers to defective fraction obtained in the process and includes information regarding reason of the rejection. An application of the Pareto chart revealed that billets were classified as defective mostly due to heavy oscillation marks appearing on their surfaces, which can be easily detected as visible transverse depressions on the billet surface that are deeper than 4 mm. Although some types of defects can be reduced or repaired in the further stages of the manufacturing process, the heavy oscillation marks defect strictly determine the billet as a scrap which cannot be repaired and must be rejected. Therefore, the analysis carried out for this type of defect can bring a key information regarding improvement of the manufacturing economics.

The available industrial dataset was in an electronic format and consisted of a number of various process parameters for the heats in which defective billets were detected. The data were collected in the period of 6 years (from June 2008 till November 2014) and contained almost 600 records whereas the total number of heats, including those with no defects, was about 47000. The distribution of defective fractions in the defective heats is shown in figure 1. The fact that the heats with no defective billets were not included in the process data was an important disadvantage for the analysis which had to utilize the limited number of cases with (almost) defect-free production. For the heats with no defects only very limited information about process parameters was available, however, this data was also utilized in a complementary analysis.

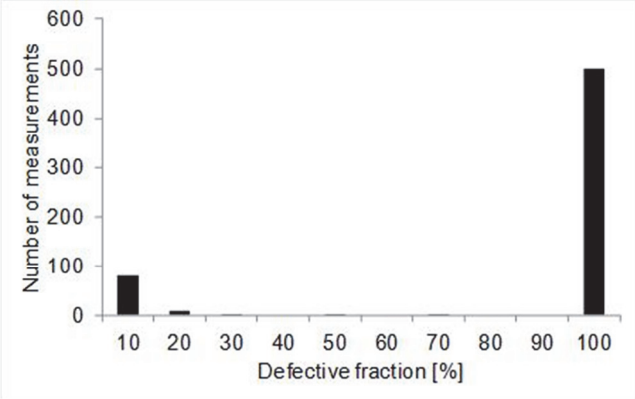



***Fig. 1.*** *Distribution of defective heats in the basic data set (with full information on process parameters) (Perzyk & Rodziewicz, 2016)*

The main process control parameter, influencing the quality of billets directly, is the casting speed. Its value should be adjusted according to the current temperature of the melt, i.e. the casting temperature. The nominal (optimal) values of the temperature are calculated for each steel grade from an empirical formula. However, in real conditions some deviations of the casting temperature appear which require appropriate corrections of the casting speed. The resulting deviations from the speed values corresponding the nominal casting temperature should therefore be precisely related to the temperature deviations. A quite obvious reason of the defects is therefore imprecise reaction of the operators on the current casting temperature, which is likely to be dependent on the magnitude of the temperature deviation. However, the simple correlation analysis revealed that the Pearson's coefficients for the relationships between the defective fraction of billets and the casting temperature deviations were very small (0,005 for the real values of the deviation and 0,03 for their absolute values). This prompted the authors to apply an advanced data-driven model in which all the available parameters of different nature which may be related to the occurrence of the defects are assumed as inputs. In the present study the following *input* parameters have been considered:



*Steel grade* - the research has been performed for a number of steel grades introducing both reinforcing and special quality steels,

*Billet dimension* – square billets 140, 160 and 220 mm,

*Working team ID* – there are 4 working teams in the steel plant,

*Shift ID* – there are 3 8-hour shifts per day,

*The last heat in the sequence* – heats are set in sequences depending on the order size and steel grade,

***Table 1.*** *Possible values of variables*

| Variable role | Variable name | Data type | Variable values |
|---|---|---|---|
| Input | Steel grade | Nominal | Free-cutting |
| | | | Carburizing |
| | | | Structural |
| | | | Bearing |
| | | | Micro-alloyed |
| | | | Spring |
| | | | Hardened and tempered |
| | | | Reinforcing |
| | Billet dimension | Nominal | 140 mm |
| | | | 160 mm |
| | | | 220 mm |
| | Working team ID | Nominal | 1 |
| | | | 2 |
| | | | 3 |
| | | | 4 |
| | Shift ID | Nominal | 1 |
| | | | 2 |
| | | | 3 |
| | The last heat in the sequence | Nominal | 0 (no) |
| | | | 1 (yes) |
| | Start-time of the CC process | Ordinal | 1 |
| | | | 2 |
| | | | 3 |
| | | | 4 |
| | | | 5 |
| Output | Defective fraction of billets in the heat | Real | (0; 100] % |
| Output or input | Deviation from nominal casting temperature (one of the two main process control parameters) | Real | [-21; 19.5] °C |
| | Deviation from nominal casting speed (one of the two main process control parameters) | Real | [-1.05; 0.38] m/min |

*Start-time of the CC process* – indicates part of the work-day when the continuous steel casting process starts. A workday has been divided into 5 equal parts so that '1' stands for initial part of the day and '5' means the final part of the day.

*Deviation from nominal casting temperature* – difference between averaged values of second and third recorded measurements (for practical reasons) and the nominal (desired) casting temperature indicated for each steel grade.

*Deviation from nominal casting speed* is similarly calculated, and is a measure of the difference between an average value calculated using all recorded measurements and its nominal value, i.e. resulting from the nominal casting temperature for each steel grade.

The fundamental *output* parameter was *Defective fraction of billets in the heat*, i.e. percentage fraction of defective billets in a heat, calculated by weight. The last two input parameters have been also considered as outputs in a part of the analysis. The summary of all variables together with their possible values are shown in table 1.

### 2.2. Methodology of significance analysis

Finding relative significances of process input variables can be done using different approaches, which assume different definitions of the significance, utilize different data-driven models and can be applied to different types of data. Some published studies on this subject (Wieczorek & Golak, 2004; Perzyk et al., 2008; Perzyk et al., 2014) have shown that there is no single procedure which can be recommended in any particular case. In the present study two methodologies were used, briefly described below.

Artificial Neural Network (ANN), being one of the leading advanced data-driven models and has proved their excellent characteristics in a huge number of industrial applications, was utilized as a basic tool for the significance analysis of input variables. There are several methodologies of extracting this kind of information from a trained ANN (Wieczorek & Golak, 2004; Perzyk et al., 2008; Perzyk et al., 2014), however, a natural and frequently offered by the software providers is the one based on the network weights. The authors' intention was to apply this simple method at first, in order to check whether some inputs can be identified as considerably standing out from the others. In the present study the MLP-type ANNs were constructed with the use of the software package EasyNN plus in which the relative significance is defined as the sum of the absolute weights of the connections from the given



input node to all the nodes in the first hidden layer. Thus obtained values were then normalized by dividing by the maximum value obtained for all input variables.

In accordance with the general principle for input variables of nominal type (Masters, 1993), individual input nodes of the network were assigned to each of the values appearing in the data. For example for the variable 'Steel grade' there were 8 input nodes, each denoting a given grade, with only two values: 1 (yes) or 0 (no). Hence, the total number of input nodes in the network constructed for the output 'Defective fraction' was 29 and for the networks for the outputs 'Deviation from nominal casting temperature' and 'Deviation from nominal casting speed' there were 27 input nodes. Based on the previous experience and many recommendations the ANN consisted of one hidden layer only. A number of preliminary tests led to a neural model with 27 hidden neurons trained in 3000 cycles (backpropagation algorithm). The training data set utilized all available records, without isolating any testing or validating subsets.

predictions for new input values. The consequence is the approach applied in the present study in which all the available data were utilized 'as is', i.e. without testing the model using separated fraction of the data in order to avoid overfitting. Similarly, no elimination of possible outliers was applied, which is always considered a controversial practice and the original data was not balanced by rejecting the overbalancing records (with 100% defective fraction). This kind of approach to ANN testing and elimination of outliers in data was applied e.g. in (Perzyk et al., 2008).

As a complementary tool for the significance analysis the one-way ANOVA (Analysis of Variance) was used, which provides a statistical, non-parametric method dedicated for output variables of numerical type. The input variables have to be of a discrete type as ANOVA utilizes variances of output variable in subgroups of records including the same values of the given input. As shown in (Perzyk et al., 2008), the one-way ANOVA based significance factor can be defined as the F-statistics values calculated for dependency between a given input variable

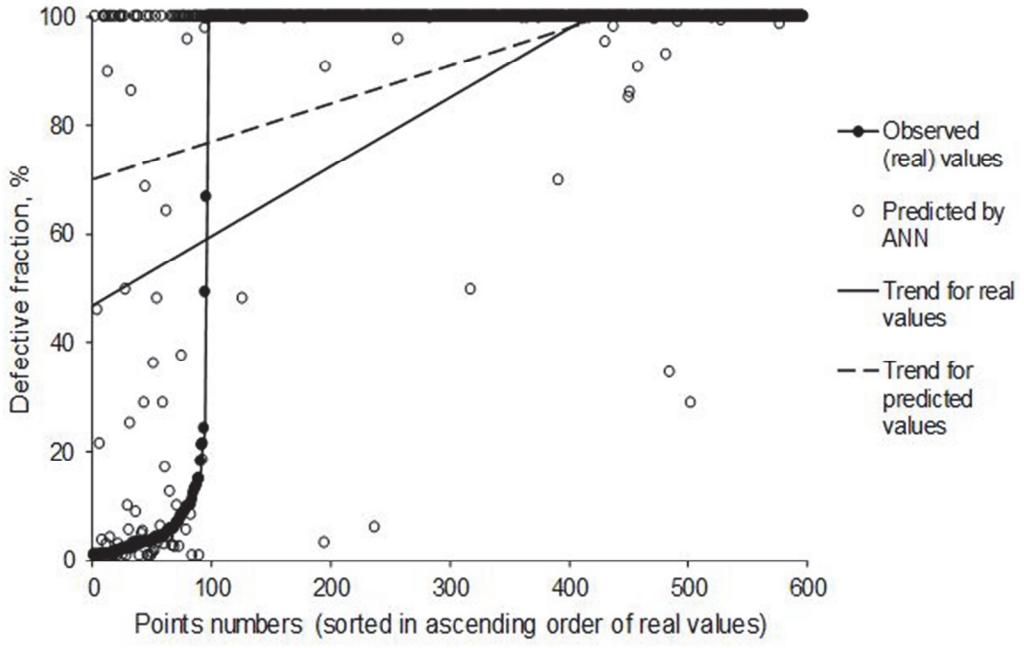



***Fig. 2.*** *Comparison of real and predicted by neural model defective fraction of steel billets*

The present and many other authors' experience of using neural models for extraction of useful information from real industrial data is that the information obtained is often not distinct and clear enough. The anxiety of a wrong indication is less serious than a loss of some important information, especially when the neural model is not used for

and the dependent variable. However, in the present work the basic approach using p-values was found to be applicable. All ANOVA computations were made using the Statistica v. 9 software from StatSoft, Inc., USA.



## 3. RESULTS

### 3.1. Significance analysis based on neural model

In figure 2 a comparison between real and predicted by the neural model fractions of defective billets is shown. It can be seen that although the differences are large, the values obtained from the model reflect the general trend in the recorded data in spite of the large scatter, inherent in that kind of data (the Pearson's correlation coefficient between real and predicted values is 0,64). This implies that the significance analysis based on the ANN can be fairly reliable.

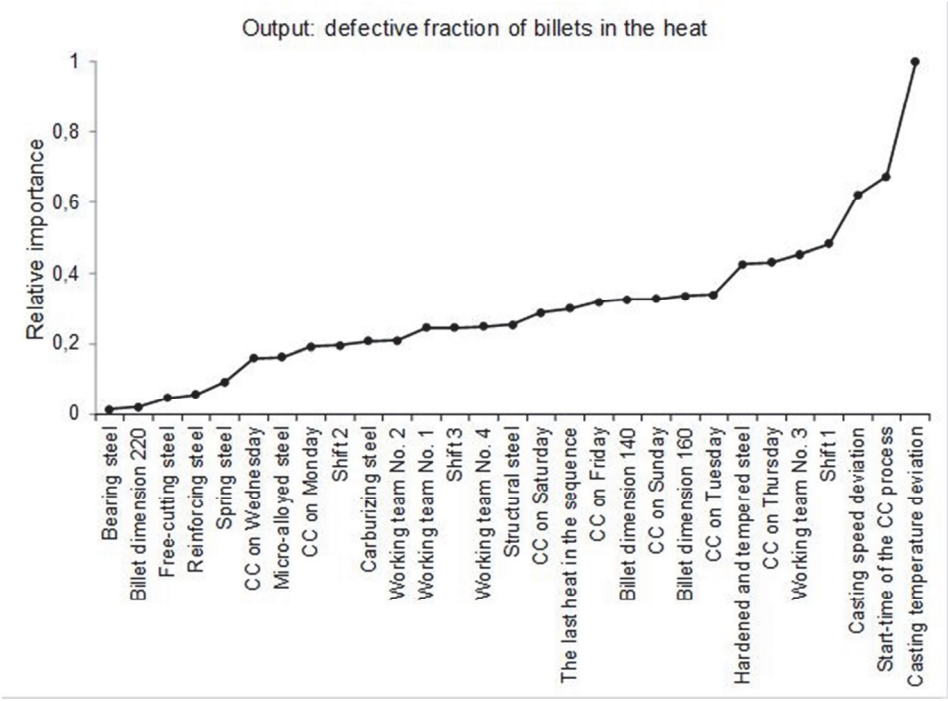



***Fig. 3.*** *Relative significances of input variables for the CC process influencing the defective fraction of billets obtained from ANN, sorted in ascending order (the separate inputs for each value of the nominal-type of variables result from neural networks specificity as described in Chapter 2)*

In figure 3 the relative significances of the input variables defined as described in Chapter 2 are shown. It should be noticed that although the total number of independent variables for the problem was 8 (see table 1), the number of input variables (nodes) in the neural model was 29 due to the treatment of the nominal type variables described in Section 2.2. This may have led to some blurring of the significance of those variables. Some comments on this problem are made in Chapter 4.

It can be seen that the following three input variables are standing out, having the largest influence on the defective fraction of billets: the casting temperature deviation, the start-time of the casting process and the casting speed deviation. The other variables are evidently much less significant, only the 'Shift 1' could be possible considered as a more significant parameter.

An interesting insight give the results of significance analysis in which the both deviations are treated as outputs, as shown in figures 4 and 5. The start time of CC is the most significant parameter influencing the deviation of the casting temperature and the casting speed.

All the above findings indicate that during certain time of a day something happens which considerably influences the process. The defects occurrence seems to be directly related with the magnitude of the casting speed and casting temperature deviations, in spite of the low Pearson's coefficient values.

In figure 6 the correlation between the casting temperature and casting speed is shown. This graph shows the results of the control actions of the operators adjusting the casting speed to the current casting temperature.

The general trend reflects the binding principle: the larger temperature values require more time to solidify the steel. However, the large scatter reveals significant imperfections of this control which may be the direct cause of the defects. In particular, for the same values of the casting temperature, much differentiated values of the casting speed are observed.



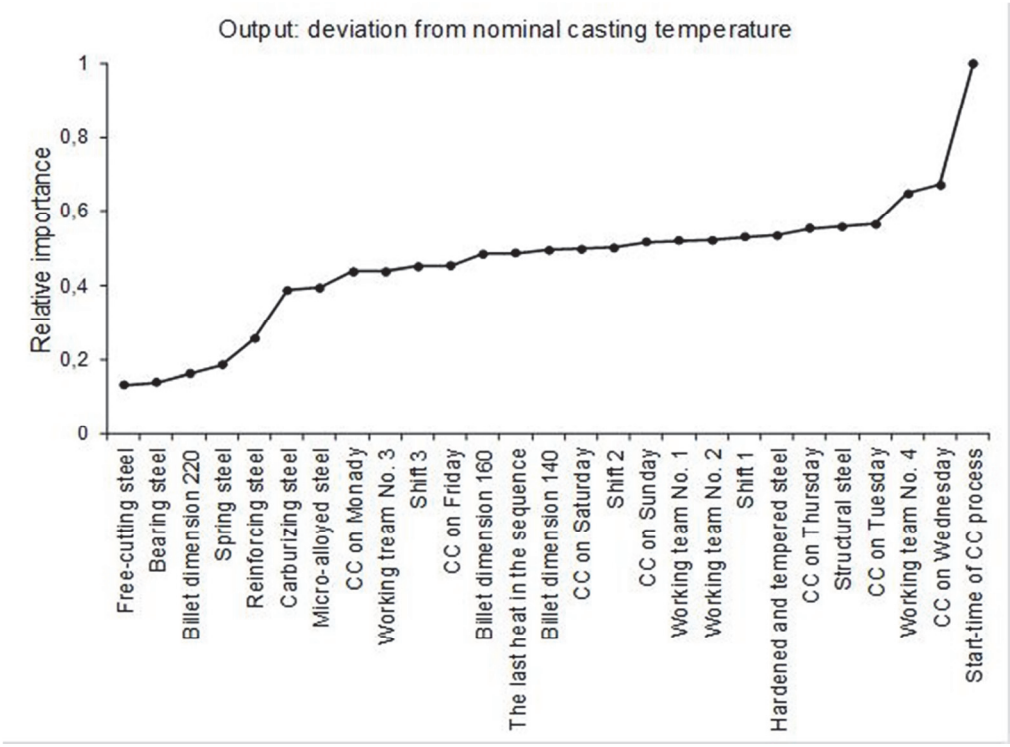



***Fig. 4.*** *Relative significances of input variables for the CC process influencing the deviation from nominal casting temperature obtained from ANN (see caption of figure 3 for further details)*

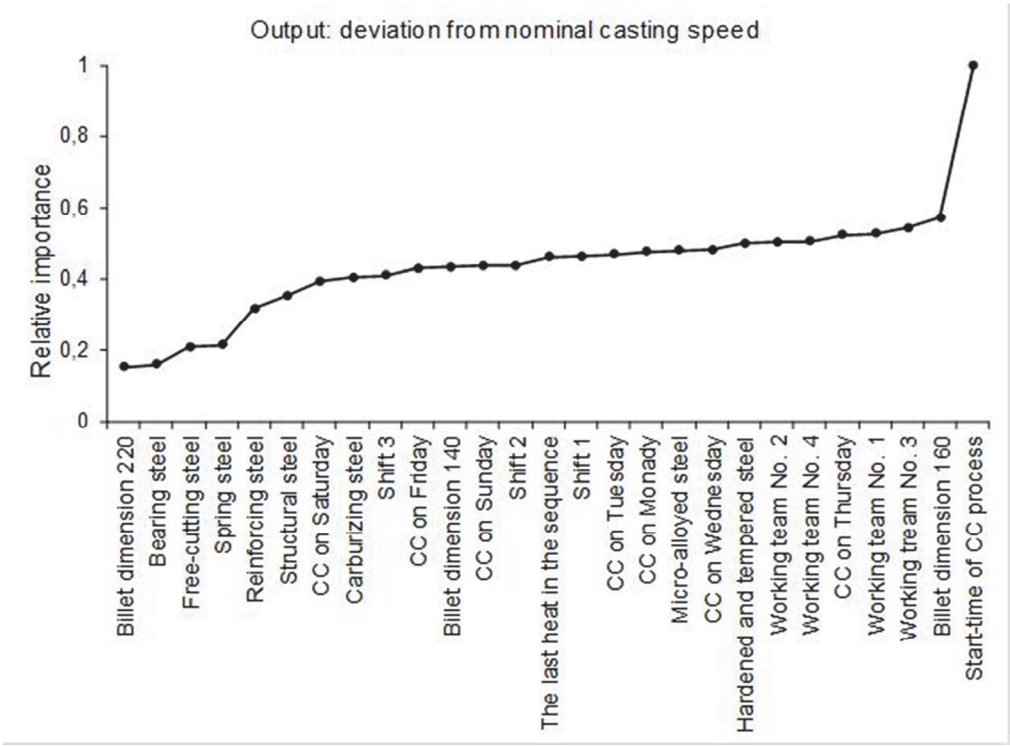



***Fig. 5.*** *Relative significances of input variables for the CC process influencing the deviation from nominal casting speed obtained from ANN (see caption of figure 3 for further details)*



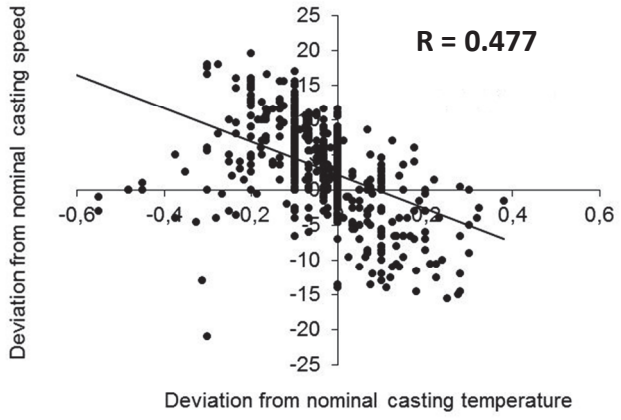



*Fig. 6. Deviation from nominal casting speed vs. deviation from nominal casting temperature*

The observed large significance of the start-time of the CC process requires a detailed discussion which will be done in the next section, as this finding is also confirmed by the ANOVA based analysis.

### 3.2. Significance analysis based on ANOVA

The two main ANOVA parameters, i.e. F-statistics and p-value, for the input variables are shown in table 2.

*Table 2. F-statistics and p-values for input variables calculated for the output 'Defective fraction of billets in the heat'*

| Input variable | F | p |
|---|---|---|
| Steel grade | 1,25 | 0,27 |
| Billet dimension | 0,39 | 0,67 |
| Working team ID | 14,03 | 0,00 |
| Shift ID | 2,16 | 0,12 |
| The last heat in the sequence | 2,76 | 0,10 |
| Start-time of the CC process | 1,38 | 0,24 |
| Deviation from nominal casting temperature | 0,36 | 0,83 |
| Deviation from nominal casting speed | 0,97 | 0,41 |

The main observation is that assuming typical levels of p-values (0.05 – 0.01) only the 'Working team ID' input variable may be considered as significant for the defective fraction of billets. All the three variables indicated in the ANN-based analysis, i.e. 'Start-time of CC process', 'Deviation from nominal casting temperature' and 'Deviation from nominal casting speed' appeared to be insignificant. In the authors' opinion, these results should not be regarded as much surprising because of the completely different nature of the ANN-based and ANOVA-based analyses, large scatter in the data and much diversified representativeness of the cases belonging to individual input categories.

A careful analysis of the plant data records has enabled the authors to extract one additional information for the heats with no defects, which was particularly important for the present study: the start-time of the casting process. The calculated ANOVA parameters for this variable obtained from the data including all heats (about 47000 records) are: F=12.757 and p< 0.00. These values evidently confirm a large influence of the start-time of the casting process on the appearance of defects obtained from the significance analysis based on ANN. In figure 7 a reproduction of a graph related to the ANOVA results obtained from the *Statistica* software is shown.

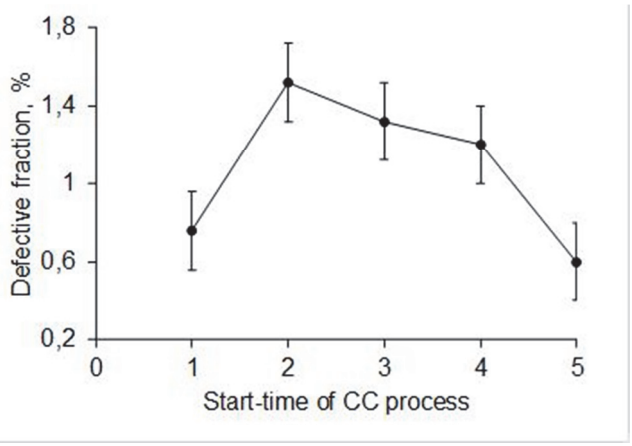



*Fig. 7. Graph illustrating ANOVA results obtained for the data including all heats*

The graph in figure 7 also illustrates the dependency of defect occurrence on the start-time of CC process. In particular, a strong increase between the time interval 1 (i.e. from midnight to 5 AM) and 2 (i.e. form 5 AM till 9.30 AM) is observed. This observation also agrees with a fairly significant role of the input variable 'Shift 1' indicated by the analysis based on ANN, as shown in figure 3.

## 4. SUMMARY, CONCLUSIONS AND FURTHER WORK

The significance analysis of the CC process input parameters was carried out using the existing plant data, basically collected only for the cases in which the product defects were observed. This was a significant difficulty in finding possible root causes of the defects. However, the neural model was able to point out process parameters which were clearly standing out from the remaining variables consid-



ered as influencing the defects appearance. The large significance of one of them, the start-time of the casting process, was also confirmed by one-way ANOVA applied to extended data, extracted by the authors from the existing plant database and covering values of the start-time of CC for all heats in the analyzed period.

There are two important findings of the present study which can significantly facilitate identification and elimination of the root causes of the defects:

1) the defects are highly influenced by the start-time of CC process and tend to appear mostly in the morning;
2) the actual physical cause of the defect are imperfections in adjusting the casting speed to the current casting temperature. The variations of these main process physical parameters also depend significantly on the start-time of CC.

The present results cannot answer the question why the above imperfections in the process control occur mainly in the morning hours but, as a result of the present study, the plant staff can focus on the search for the defect root causes in a much narrower range. In particular, it should be taken into account that the large variations of the process control parameters and the imperfections in their adjustment which take place in the morning are probably not associated with particular operators (e.g. poor training or carelessness). This conclusion is fairly well-founded because the variables defining working teams are among less significant compared to the above mentioned variables connected with time and the teams are assigned to the shifts on a rotation basis. It is important that after identification the potential factors a systematic recording of the relevant data should start for all the heats. This would enable to obtain a reliable and exhaustive database for further, similar analyses as presented in the present study.

The above findings are mostly based on the results of significance analysis carried out using a neural model. In this context, some comments concerning the methodology of finding the relative significances of the nominal-type variables are necessary. As mentioned in Section 2.2 each variable of that kind is 'decomposed' to a number of 'component' variables, equal to the number of values of the original variable appearing in the data. A question arises how to calculate the overall significance of the original variable based on significances of the 'component' variables? In the authors' opinion an appropriate methodology requires development and testing.

It would be also advisable to apply Data Mining methods alternative to ANN, both to the historical and the new data. In particular, the methods oriented at classification tasks, such as those based on the Rough Sets Theory, Classification Trees or Bayesian classification could be helpful. It is also worth noticing that the dependency of defects appearance on time implies that the time-series analysis could be also a useful tool in finding their root causes or predicting their occurrence. That kind of approach has been applied to the continuous casting of steel process and the results are described in (Camisani-Calzolari et al., 2003; Perzyk & Rodziewicz, 2016).

## REFERENCES

Camisani-Calzolari, F.R., Craig, I.K. & Pistorius, P.C., 2003, Quality prediction in continuous casting of stainless steel slabs, *Journal of The South African Institute of Mining and Metallurgy,* 103, 651-665.

Masters, T., 1993, *Practical neural network recipes in C++*, Academic Press, San Diego.

Perzyk, M., Biernacki, R., Kozłowski, J., 2008, Data mining in manufacturing: Significance analysis of process parameters, Journal of Engineering Manufacture, *Proceedings of the Institution of Mechanical Engineers*, part B, 222, 1503-1516.

Perzyk, M., Kochański, A., Kozłowski, J., Soroczyński, A., Biernacki, R., 2014, Comparison of data mining tools for significance analysis of process parameters in applications to process fault diagnosis, *Information Sciences*, 259, 380-392.

Perzyk, M., Rodziewicz, A., 2016, Application of time-series analysis for predicting defects in continuous steel casting process, *Archives of Foundry Engineering*, 16, 4, 125-130.

Wieczorek, T., Golak, S., 2004, An algorithm of knowledge extraction from trained neural networks, *Advances in soft computing*, eds, Klopotek, M.A, Wierzchon, S.T., Trojanowski, K. Springer Verlag, Berlin Heidelberg, New York, 470-475.

### ZASTOSOWANIE ANALIZY ISTOTNOŚCI DO ZNAJDOWANIA PRZYCZYN POWSTAWANIA WAD WYROBÓW W PROCESIE CIĄGŁEGO ODLEWANIA STALI

Streszczenie

Celem niniejszej pracy jest przedstawienie analizy istotności zmiennych wejściowych jako narzędzia pozwalającego na znalezienie prawdopodobnych przyczyn wad wyrobów w procesie ciągłego odlewania stali. Do analizy wykorzystano dane przemysłowe zebrane w jednej z polskich hut stali, dotyczące produkcji kęsów. Skorzystano z zarejestrowanych danych związanych z przebiegiem i parametrami procesu produkcyjnego, według obowiązujących w zakładzie procedur. Podstawowym



źródłem danych była baza tzw. wybraku technologicznego rejestrowanego przez zakład produkcyjny, informująca o fakcie pojawienia się braków oraz ich ilości wyrażonej w procentach dla każdego wytopu. Baza ta nie zawierała jednak pełnej informacji o tych wytopach, w których braki nie występowały, co powodowało istotne trudności i ograniczenia prowadzonych badań. Analizie poddano osiem parametrów wejściowych o różnym charakterze: fizycznym, organizacyjnym i ludzkim. W celu określenia, które z parametrów procesu mają największe znaczenie z punktu widzenia pojawiania się braków, a które z nich mogą być pominięte, zastosowano i porównano dwie metody. Podstawowym narzędziem były sztuczne sieci neuronowe typu MLP, w których istotność względna jest definiowana jako bezwzględna suma wag połączeń między danym wejściem a wszystkimi węzłami w pierwszej warstwie ukrytej. Drugą metoda była jednoczynnikowa analiza wariancji (ANOVA), badająca wpływ poziomu jednego czynnika klasyfikującego na wartości badanej zmiennej zależnej typu rzeczywistego, określany wartością statystyki F. Otrzymane wyniki pozwalają jednoznacznie stwierdzić, że parametr wejściowy 'pora spustu' ma znaczący wpływ na kształtowanie się parametru wyjściowego 'udział braków'. Podstawowe fizyczne parametry procesu, tj. odchyłka od nominalnej temperatury odlewania i odchyłka od szybkości odlewania, zgodnie z oczekiwaniami wykazały również duże znaczenie, a ponadto duże zróżnicowanie w zależności od wartości parametru 'pora spustu'. Końcowym wnioskiem wynikającym z przeprowadzonych analiz jest zidentyfikowanie bezpośredniej przyczyny powstawania wadliwych produktów, którą jest niepoprawne dostosowywanie prędkości odlewania do aktualnej temperatury stali, występujące głównie w godzinach porannych, lecz niezwiązane z konkretnymi pracownikami. Poczynione spostrzeżenia mogą w znaczący sposób ułatwić dokonanie ostatecznej identyfikacji źródeł powstawania wad w wyrobach przez personel inżynieryjny zakładu. W pracy zawarto również zalecenia dotyczące przyszłych badań, zarówno przemysłowych jak i związanych z metodyką prowadzenia podobnych badań.

*Received: August 30, 2016*
*Received in a revised form: October 27, 2016*
*Accepted: November 12, 2016*