# THE USAGE OF REGRESSION ANALYSIS AND ARTIFICIAL INTELLIGENCE TOOLS IN THE FIELD OF METALLURGY

TYKVA Tomáš[1,2], ŠPIČKA Ivo[1,2], ŠPIČKOVÁ Dagmar[1], ZIMNÝ Ondřej[1], BAHENSKÝ Ivo[1]

*[1]VSB - Technical University of Ostrava, Ostrava, Czech Republic, EU,*
*ivo.spicka@vsb.cz, spickova@bintell.cz, ondrej.zimny@vsb.cz, ivo.bahensky@vsb.cz*
*[2]University of Business and Law, Ostrava, Czech Republic, EU,*
*tomas.tykva@vspp.cz*

## Abstract

The presented article deals with the possibility of using artificial intelligence tools, specifically genetic algorithms, in the metallurgical industry. The genetic algorithms are used here to estimate coefficients of regression functions. In some cases, the standard regression analysis tools lead to incorrect results. And then the genetic algorithms can serve as an optimization tool searching for a certain state space to find functions or functionalities optimum. To verify the correctness of this premise we have used the genetic algorithms to find coefficients of two regression functions types, namely linear regression function and nonlinear regression function. By comparing the results, we have confirmed in the discussion the suitability of using genetic algorithms in the field of regression analysis as well.

Another result of this work is the determination of the general cost equation of the foundry furnaces. The data of selected foundries have been used both in classical regression analysis and in genetic algorithms applications.

**Keywords:** Cost analysis, genetic algorithms, metallurgical industry

## 1. INTRODUCTION

The main purpose of this article is to present the results of the partial cost analysis realized by using the regression analysis tools and by using the artificial intelligence - particularly by genetic algorithms - tools at the metallurgical industry field. In this particular case the mutual correlation of gas consumption per cycle on the charge weight, on the lag temperature and on the cycle length has been researched.

The regression analysis has been performed on the statistical sample. On the basis of the obtained coefficients of the regression function of three variables (charge weight, lag temperature and cycle length) the corresponding values of the estimated gas consumption as well as of the residual value have been computed.

Then for the same data the artificial intelligence tools - especially genetic algorithms - have been applied. Later the results obtained by the conventional procedure (regression analysis) use and those reached by the artificial intelligence tools (genetic algorithms) use have been compared.

## 2. REGRESSION ANALYSIS AND ARTIFICIAL INTELLIGENCE METHODS

As mentioned in the introduction, the same data from a particular normalizing furnace were analysed first by means of convectional statistical methods; with a certain degree of reliability the cost equations were defined and by using them the cost model of the heat treatment for the specific normalizing furnace was created. The used procedure is applicable also for the creation of the cost (calculation) model for furnaces in other factories that do not use it yet.

### 2.1. Regression analysis

In statistical modelling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis estimates the conditional expectation of the dependent variable given the independent variables - that is, the average value of the dependent variable when the independent variables are fixed.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation [1].

Considering that this is not a new method and its tools are generally known, it will not be more widely commented within this article; however, the use of genetic algorithms in this context, as evidenced by the search of available resources, is not applied so much and therefore more attention is paid to this matter in the following subchapter.

### 2.2. Artificial intelligence - Use of Genetic algorithms

Genetic algorithms (GAs) are powerful search algorithms that perform an exploration of the search space that evolves in analogy to the evolution in nature. The power of GAs consists in only needing objective function evaluations. So derivatives or other auxiliary knowledge are not used. Instead probabilistic transition rules of deterministic rules, and handle a population of candidate solutions (called individuals or chromosomes) that evolves iteratively are used. Each iteration of the algorithm is called generation. The evolution of the species is simulated through a fitness function and some genetic operators such as reproduction, crossover and mutation [2].

The main idea of genetic algorithms is that we look at the elements of a set of accessible solutions as at living organisms in an artificially created environment. Shortly: how the organisms thrive in that environment, i.e., their abilities to survive and to reproduce, correspond to the fact how "good" the solutions are. The actual search then consists in selecting a certain initial population of these organisms and then simulating of its development under the control of evolutionary mechanisms, including natural selection, reproduction, etc.

GAs are a subset of evolutionary algorithms. Biological motivation of GAs consists in working with the population of individuals and in guiding the search for space-based solutions by the means usual for the nature: sexual intercourse, mutation and selection of individuals with favouring more appropriate individuals (selective pressure). Mathematically, GAs are based on the so-called scheme of theorems and hypothesis of building blocks. Genetic algorithms are an universal stochastic search approach that is able to approach the global optimum within the bounded space of permissible solutions of the problem [3].

The fittest individuals will survive generation after generation while also reproducing and generating offspring’s that might be stronger and stronger. At the same time, the weakest individuals disappear from each generation. Individuals must be encoded in some alphabet, like binary strings, real numbers, and vectors and others. In a practical application of genetic algorithms, a population pool of chromosomes has to be installed and they can be randomly set in the beginning. In each cycle of genetic evolution, a subsequent generation is created from the chromosomes in the current population. The cycle of evolution is repeated until a termination criterion is reached. The number of evolution cycles, or a predefined fitness value can set this criterion [4].

As being stated in the previous chapter, a relatively limited number of authors deals with the issue of genetic algorithms specifically in the metallurgical industry application. Except those ones already mentioned before, the work of the Feliks, Lenort and Besta team and their article "Model of multilayer artificial neural network for prediction of iron ore demand" [5] or the article "Application of Artificial Intelligence Methods for Prediction of Steel Mechanical Properties" [6] of prof. Koštialová Jančíková can be quoted.

## 3. PROBLEM SOLVING

This chapter defines how the authors have proceeded to solve a particular assignment; methodological apparatus, the results of the work and the discussion about the obtained results are mentioned there, too.

### 3.1. Solution process

As first a regression analysis was performed by the method of minimization of deviations which includes also the terms that are the multiple of individual independent variables. The next step was to determine the coefficients of the general regression function using genetic algorithms

In case of genetic algorithms the comparison of the practically measured data with the predicted data was also performed, depending on the value of the individual variables. The relationship of the variables mentioned below can be written mathematically as follows:

$$y \sim 1 + x_1 + x_2 + x_3 \qquad (1)$$

The meaning of the following variables is as follows:

- dependent „**consumption ($y$)**“ indicates the gaseous media consumption in normal m$^3$,
- „**constant value**“ (in **Table 1** labelled as intercept, this is a point of intersection with an x-axis for zero values of all variables),
- „**charge weight ($x_1$)**“ indicates the weight of the charge in tonnes,
- „**cycle time ($x_2$)**“ indicates the length of the whole cycle in hours,
- „**temperature of delays ($x_3$)**“ indicates the normalization temperature in °C.

### 3.2. The comparison of results obtained by regression analysis and by GAs using

**Table 1** The comparison of regression analysis with using of GAs and conventional procedures

| | GAs | Linear regression model obtained by conventional procedure | | | |
|---|---|---|---|---|---|
| | Estimate | Estimate | SE | tStat | pValue |
| (Intercept) | -1212.7705 | -1212.8 | 41.216 | -29.424 | 2.1279e-147 |
| $x_1$ | 11.4007 | 11.401 | 0.7197 | 15.841 | 5.8725e-52 |
| $x_2$ | 17.5686 | 17.569 | 0.9403 | 18.684 | 1.8521e-69 |
| $x_3$ | 3.3945 | 3.3945 | 0.0486 | 69.835 | 0 |

The list of parameters below expresses the statistical significance of each item of the regression function calculated by using of the conventional procedure. When using Gas, the statistical regression parameters must be determined by a separate calculation.

Parameters of linear regression:

- Number of observations: 1352;
- Error degrees of freedom: 1348;

- Root Mean Squared Error: 391;
- R-squared: 0.834;
- Adjusted R-Squared: 0.833;
- F-statistic vs. constant model: 2.25e+03. p-value = 0.

### 3.3. The results of regression analysis with GAs using

The GAs best practice and GAs approach shows the selected graphs that characterize the gradual search for the final solution and the speed (i.e. the number of generations) with which GAs will find the final result.

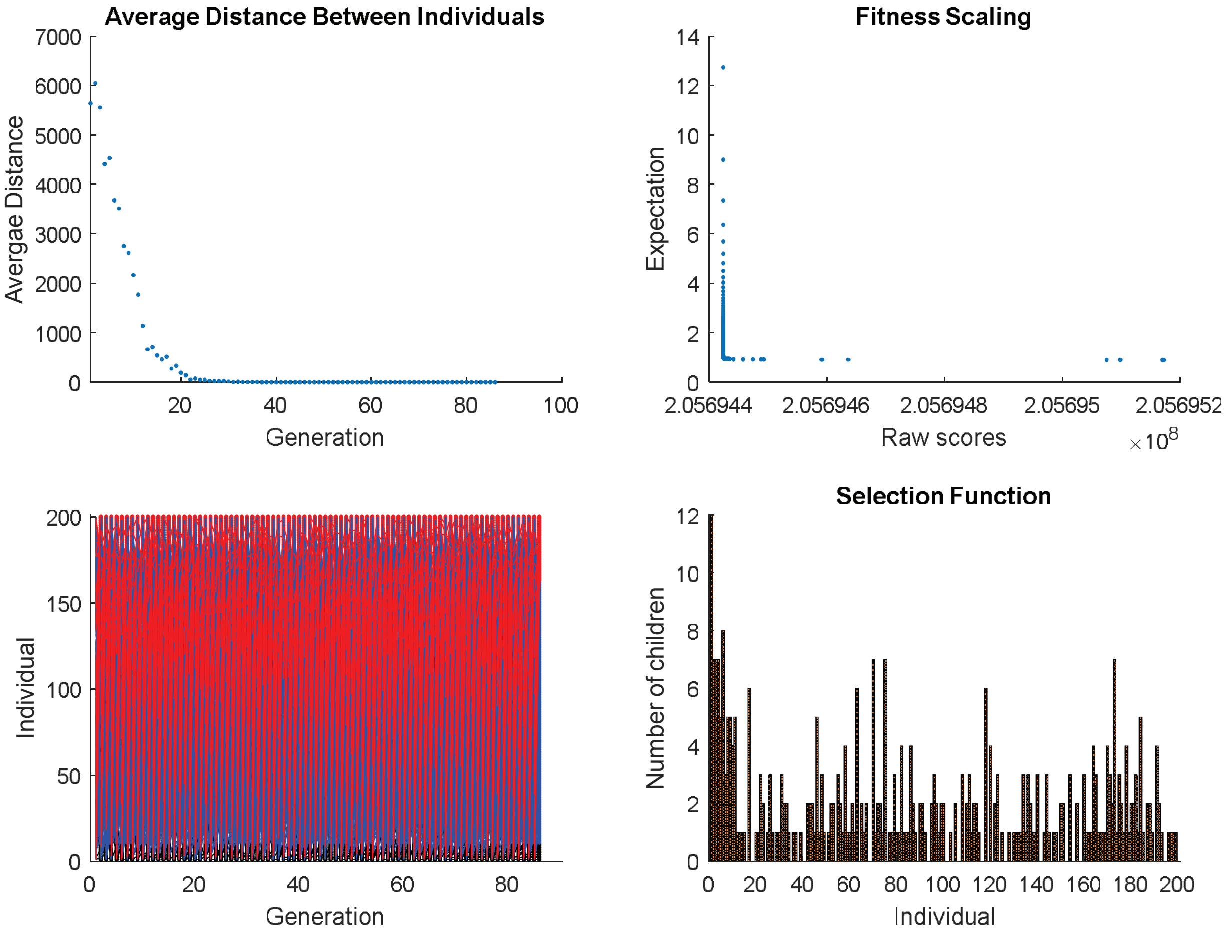

**Figure 1** Graphic depiction of GAs usage

The first graph in **Figure 1** (Average distance between individuals) shows the average difference of individuals in the respective generation. It can be visible in the graph that the distance (which expresses the difference of individuals in the separate generations) has not changed significantly after the 25th generation.

The second graph in **Figure 1** (Fitness Scaling) shows that in the most individuals in the generation the score lies at a very narrow interval and therefore the results of the crossing will no longer be very different from previous generations. Raw scores are returned by the fitness function to values in a range that is suitable for the selection function. The selection function uses the scaled fitness values to select the parents of the next generation. The selection function assigns a higher probability of selection to individuals with higher scaled values.

The third graph in **Figure 1** shows the crossing procedure of the separate generations. The fourth chart shows the number of off springs of the last generation particular members.

For the linear regression the GAs using is not needed as common regression analysis methods find a global optimum and do not stay in a local optimum (this is due to the use of linear functions). For nonlinear regression, in particular by using of more complex relationships, it can lead - in case of classical regression - to a local extreme lockup, especially in case of inappropriately determined initial estimation of function parameters. Just GAs that are used to search large state spaces, due to crossing and mutation properties, have the ability to find the global optimum [7].

### 3.4. Nonlinear regression with GAs using

The previous chapter, at the simple example, shows the use of GAs to create a linear regression model (cost model).

The coefficient $k_0$ (Intercept), which is negative for linear regression relation, does not give any direct physical sense. One of the possible interpretations would be that it means the overheads of the furnace, but in that case the coefficient should be positive. Therefore, the efforts were made to find non-linear regression functions for which this coefficient would be positive. The general form of the regression function is shown by the following relationship:

$$y = k_0 + k_1x_1 + k_2x_2 + k_3x_3 + k_4x_1.x_2 + k_5x_2.x_3 + k_6x_1.x_2 + k_7x_1.x_2.x_3 + k_{10}.\frac{x_1}{x_2} + k_{11}.\frac{x_2}{x_1} + k_{12}.\frac{x_2}{x_3} + k_{13}.\frac{x_3}{x_2} + k_{14}.\frac{x_1}{x_3} + k_{15}.\frac{x_3}{x_1} \quad (2)$$

The genetic algorithm has found a minimum of deviations between the actual and predicted gas consumption but using of another algorithm shows that when looking for coefficients of this complex function by standard procedures the values of some parameters are poorly conditioned (it means that the found values of individual coefficients are only one of many variants for which regression functions will predict almost identical results in terms of the sum of deviations).

The equation was created by a direct transcription of the MatLab program code which did not contain members with coefficients $k_8$ and $k_9$ and therefore also in the **Table 2** there are the limits set to 0. The coefficient $k_7$ value was set by the genetic algorithm.

**Table 2** Coefficients of nonlinear regression function obtained by using GAs

| | $k_0$ | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ | $k_6$ | $k_7$ | $k_8$ | $k_9$ | $k_{10}$ | $k_{11}$ | $k_{12}$ | $k_{13}$ | $k_{14}$ | $k_{15}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| limits | ± 1000 | ± 1000 | ± 1000 | ± 1000 | ± 1000 | ± 1000 | ± 1000 | ± 1000 | 0 | 0 | ± 5000 | ± 5000 | ± 5000 | ± 5000 | ± 5000 | ± 5000 |
| coeff. value | -594.22 | 24.65 | -7.79 | 2.15 | -717.1 | 0.02 | 715.06 | 0 | 0 | 0 | -256.5 | 37.64 | 5000 | 4.84 | 5000 | -1.50 |

Once the values of the coefficients from the **Table 2** have been set, the function is following:

$$y = -594.22 + 24.65x_1 - 7.79x_2 + 2.15x_3 - 717.1x_1.x_2 + 0.02x_2.x_3 + 715.06x_1.x_2 - 256.5.\frac{x_1}{x_2} + 37.64.\frac{x_2}{x_1} + 5000.\frac{x_2}{x_3} + 4.84.\frac{x_3}{x_2} + 5000.\frac{x_1}{x_3} - 1.5.\frac{x_3}{x_1} \quad (3)$$

It is shown that even the deviation coefficient $R^2$ (in this case $R^2 = 0.8672$) is higher than in the case of linear regression, this function is not too much suitable for prediction of the actual furnace operation considering the complexity as well as the variability of the coefficients.

Genetic algorithms seem to be an appropriate tool for determining the coefficients of such complex dependencies. Due to the high variability of individual coefficients we assume that it will be appropriate to extend the cost function by other independent variables, such as: preheated air temperature, lag time, permitted heating and cooling temperature gradient, or others.

## 4. CONCLUSION

The purpose of the study, the results of which are presented in this article, was to compare one of the foundry cost items, namely gas consumption for normalizing for different furnaces and different types of technological processes. The result shows that all examined furnaces showed a compliance in the structure of the regression function (the same type). Different furnace designs led to different values of regression function parameters. The differences did not, however, exceed the size of one degree, from which it can be deduced that the regression formula is suitable as a universal cost estimation tool, however it is necessary to set these parameters individually for each furnace. The model will be further elaborated to explain better the individual regression function parameters. This will be done by selecting of further regression function variables.

In the frame of other research activities it would be appropriate to create a complex cost model for the foundries. It means a model that will not only involve the heat treatment process, but all previous and subsequent activities, from the preparation of moulding compounds, mould and core production, through steel melting, resp. casting, shaping, cleaning, diagnostics and defects removal, to the expedition.

## ACKNOWLEDGEMENTS

***At this point we would like to thank the Faculty of Metallurgy and Material Engineering of VŠB - TU Ostrava, especially the Department of Automation and Computer Engineering in Metallurgy to support the creation of this article within the research projects SP2017/63 and SP2017/38.***

## REFERENCES

[1] ARMSTRONG. J. S. Illusions in regression analysis. *International Journal of Forecasting* [online], 2012, vol. 28, no. 3, pp. 689-694.

[2] BENISIS, A. *Business Process Management: a data cube To analyze business process simulation data for decision making*. Saarbrücken: VDM Verlag Dr. Müller Aktiengesellschaft & Co. KG, 2010. 204 p.

[3] ZELINKA, I. *Evoluční výpočetní techniky: principy a aplikace*. Praha: BEN - technická literatura, 2009. 534 p.

[4] HAUPT, R. L., HAUPT, S. E. *Practical genetic algorithms*. 2nd ed. Hoboken. N.J: John Wiley, 2004. 261 p.

[5] FELIKS, J., LENORT, R., BESTA, P. Model of multilayer artificial neural network for prediction of iron ore demand. In: *METAL 2011*: *Proceedings of the 20th International Metallurgical & Materials Conference METAL*. Brno: Tanger, 2011, pp. 65 - 70.

[6] JANČÍKOVÁ, Z., ROUBÍČEK, V., JUCHELKOVÁ, D. Application of Artificial Intelligence Methods for Prediction of Steel Mechanical Properties. *Metalurgija*, 2008, vol. 47, no. 4, pp. 339 - 342.

LASÁK, R. Návrh nákladového modelu tepelného zpracování s využitím statistických metod. In: *Oceláři 2017*. Rožnov pod Radhoštěm: Tanger, 2017, pp. N/A.