Statistical Analysis of Aluminate Liquor Precipitation Process with Statistica: Classic and Modern Data Mining Methods

Vladimir Borovikov[1], Maksim Milkov[2]
1. Director
2. Technical Director
StatSoft Russia, Moscow, Russia
Correpsonding author: expert@statsoft.ru

Abstract

In the paper, a review of the application of modern and classic statistical methods for analysis of precipitation process of aluminate liquor is carried out. It is known that this process is highly inertial, characterized by significant variation and depends on many factors that affect the product quality. The review is using both classic and modern analysis methods of aluminate liquor decomposition on seed comprising neural networks and CART models (classification and regression trees), etc. Predictors in the studied models are: aluminate liquor flow, concentration of caustic soda, seed concentration, temperatures in a precipitator, etc. The target variables are content of aluminum hydroxide fractions - 5, - 20, - 45, +150 um after hydrocyclone. The paper compares the accuracy of models: classic regression, CART models, artificial neural networks. The viability of the methods for solving real production problems is demonstrated. The study was performed with Statistica 13 RU software.

Keywords: precipitation, aluminate liquor, statistical analysis, machine learning.

1. Introduction

One of the stages of alumina production is decomposition of a metastable aluminate liquor to produce aluminum hydroxide.

At Russian alumina refineries processing boehmite - diaspore raw materials, the process of crystallization is forcibly carried out at a high A/C ratio, reaching 1.65 - 1.71 units. Besides, the inability to significantly affect particle size distribution of crystals by controlled agglomeration leads to strong oscillatory fluctuations in PSD of the resulting product - the content of " - 45 μm" fraction can vary from 5 to 50 % within 3 - 4 months.

An example of the precipitation train used as the object of the study is presented in Figure 1. The flow of aluminate liquor and the seed flow are fed into the first tank of the precipitation train. As it passes through precipitators, the temperature of slurry is reduced resulting in crystallization of aluminum hydroxide. At the outlet of the last tank, crystals of aluminum hydroxide are classified in hydrocyclones. Spent liquor is separated from solid phase by filtration and removed. Coarse crystals after the hydrocyclone are removed from the process. The collected fine fraction is recycled to the first precipitator as a seed and reslurried with a flow of fresh undecomposed liquor.

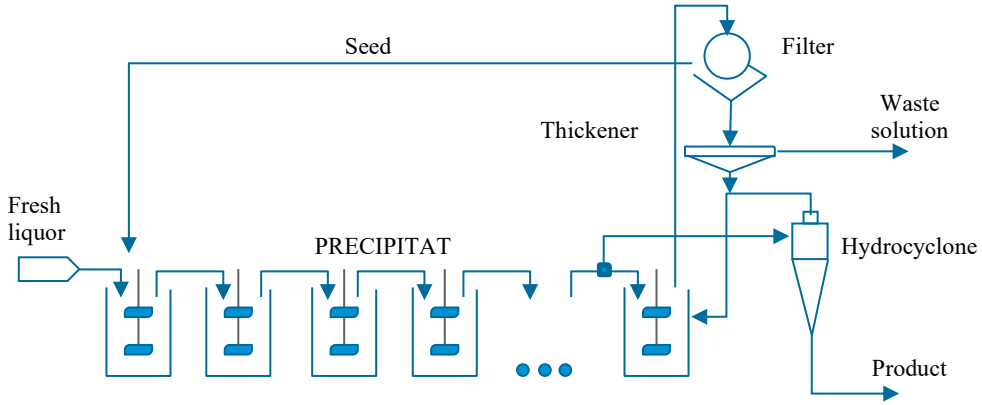

Figure 1. Flowsheet of precipitation process.

The process of precipitation is highly inertial, since the seed flow exceeds the flow of product by tens of times. However, at the same time, it is affected by numerous disturbing factors, such as: fluctuation in composition of aluminate liquor, ambient temperature, seed reactivity, etc. The process control can be implemented by changing aluminate liquor flow, seed concentration, temperature of the head and tail precipitators, profile of temperature variations by precipitators, pumping seed from other precipitation trains and sites.
So, along with the cyclic character of precipitation process, periods of instability is recorded with their amplitude of fluctuations (Figure 2).

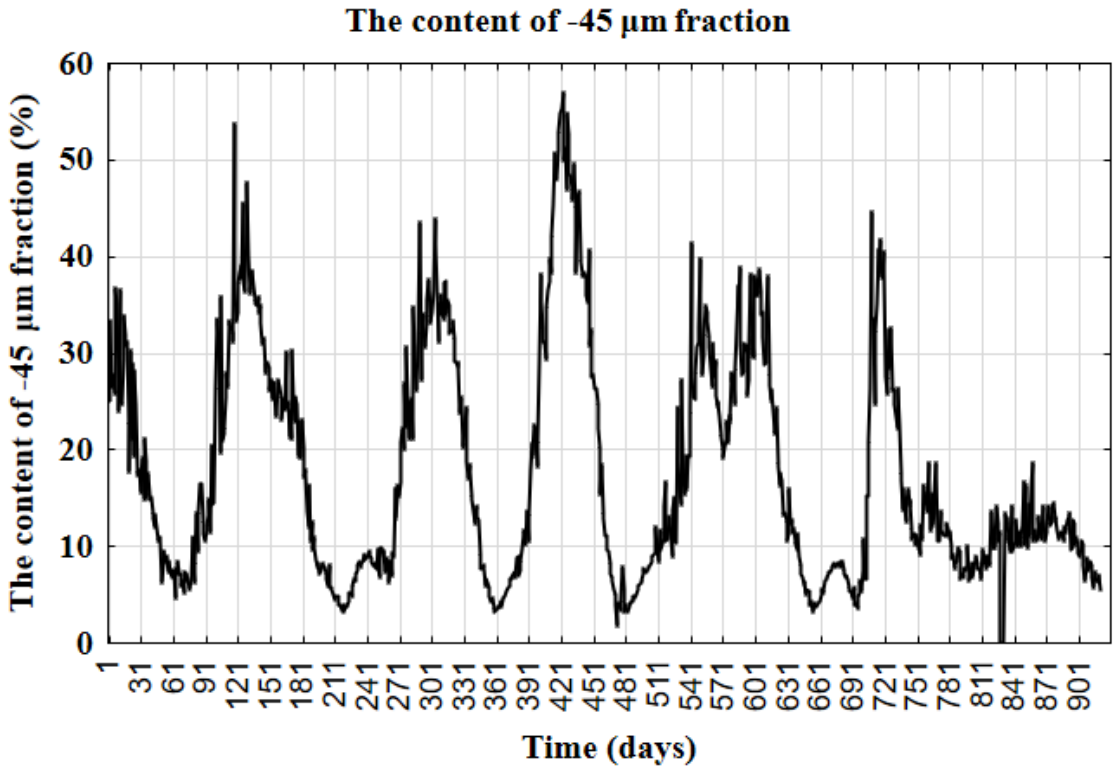

Figure 2. The content of -45 µm aluminum hydroxide fraction in the last precipitator.

To develop measures that allow to stabilize size distribution, it is necessary to have an idea of its change in the future. Today at Russian alumina refineries a forecast of this kind is carried out

manually. Due to this, a number of factors are taken into account when compiling data and the forecast horizon are very limited.
In this paper, we investigate the relationship between the factors of the process and its target parameter such as the content of the coarse fraction -45 μm, and assess the possibility of using different methods of data analysis for automatic forecasting based on current and historical data of APCS. The study of various predictive models to describe precipitation process is carried out, their accuracy and prospects for practical application are estimated.

2. Statistical Data Processing

The analysis is based on historical production data. Data discreteness - 1 day. Scope - 920 days.

2.1. Correlation Analysis

Existence of linear relationships between factors and target process parameter is checked by calculating the Pearson correlation.

Table 1. Pearson correlations between factors and target setting of the process.

Factor	Content of "-45 μm" fraction
Consumption of aluminate liquor, m^3/h	-0.16
Na_2O_k content in aluminate liquor, g/l	0.03
Al_2O_3 content in aluminate liquor, g/l	0.02
Al_2O_3 content in the 1st precipitator, g/l	-0.25
$Na_2O_к$ content in the 1st precipitator, g/l	0.12
Concentration of solids in the first precipitator, g/l	-0.32
Concentration of solids in the last precipitator, g/l	-0.41
Temperature in the first precipitator, °C	0.02
Temperature in the last precipitator, °C	0.19
A/C ratio of mother liquor	-0.02
Total volume of precipitators	-0.21
Input point for additional seed slurry (the first or the last precipitator)	0.09
Additional seed flow, m^3/h	-0.02
Additional seed concentration, g/ l	-0.13

The strongest linear relationship is the correlation of - 45 μm fraction content with concentration of solid phase in the last precipitator (r = -0.41).

2.2. Multiple Linear Regression

Let's build a dependence linear model of the kind:

$$Y = b_0 + b_1x_1 + \ldots + b_kx_k + \varepsilon \quad (1)$$

where:
Y target characteristic
$x_{...}$ factor (predictor)
k number of factors in the model
$b_{...}$ estimated parameter (coefficient) of the model
ε model error

When building the model, we use an algorithm for step-by-step inclusion of factors. This allows to select the strongest (of greater impact) predictors from statistical point of view.

The results of model building are shown in Table 2.

Table 2. Coefficients of a multiple linear regression model.

Factor	***b***
Free term	-120.648
Concentration of solids in the last precipitator, g/l	-0.021
Al_2O_3 content in the 1st precipitator, g/l	-1.711
$Na_2O_к$ content in 1st precipitator, g/l	1.199
Temperature in the last precipitator, °C	1.061
Concentration of solids in the first precipitator, g/l	-0.041
Al_2O_3 content in 1st precipitator, g/l	0.871
Concentration of additional seed, g / l	-0.009
A/C ratio of mother liquor	7.876

All model coefficients are statistically significant. Determination coefficient of the resulting model is $R^2 = 0.41$. 50 % of residuals, quartile range is [-6.40, - 6.16].

2.3. Decision Trees CART

The decision tree model makes a forecast by means of sequential logical “if - then” operations. Such models have a demonstrated visualization in the form of a tree having a hierarchical structure.

Branching starts from the top of the tree (the root node), which contains all observations of the original data table (sample). Next, the algorithm selects an optimal stratification of the sample into 2 subsamples. To this aid, the division condition is determined on the basis of independent variables.

In the resulting model, the first branching occurs according to the condition:

Seed concentration (tail decomposer) <= 586,5 g/l

If for observation (measurement in a particular moment of time) this condition is true, then the observation falls into the left node of the tree, otherwise into the right one.

The average value of the target variable "fraction -45 μm content " throughout the sample (i.e. in

the root node) is 17.8 %. After branching, 70 observations fell into the left node. The average value of -45 μm fraction content for them was 33.7 %. The remaining 830 observations fell into the right node. The average value of -45 μm fraction content for them was 16.4%. This process proceeds iteratively until the node is recognized as terminal. From here on, the terminal node is not divided into subsamples.

In the branching described earlier, the left node with 70 observations is considered terminal and the node is no longer "branching". The right node with 830 observations is not terminal. For this node, the following iteration of branching occurs according to the condition:

The total volume of decomposers <= 42 969 m^3

The resulting tree model CART is presented in Figure 3.

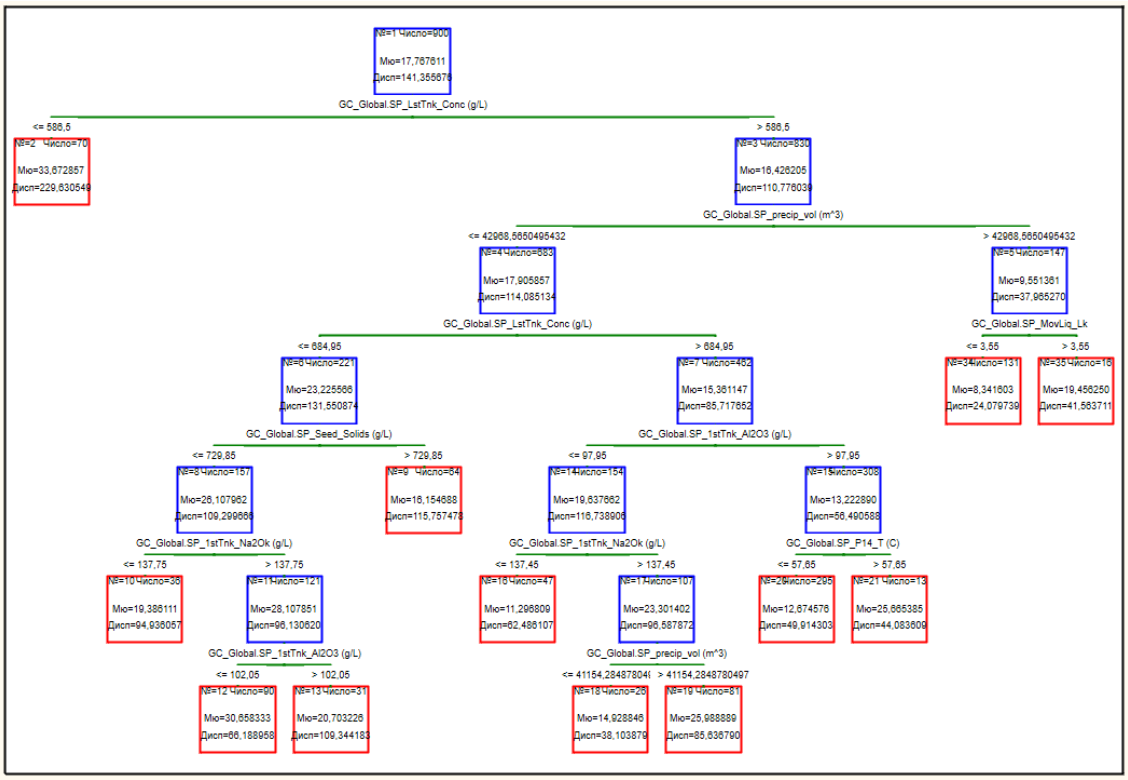

Figure 3. CART Model.

The tree contains 12 terminal nodes and 11 branch nodes.

The value Mu inside the node is the average value of the target variable from the observations contained in this node. This value is predictive.

Thus, the forecast tree is discrete.

The determination coefficient of the resulting model is $R^2 = 0.48$. 50% of the residuals, quartile range varies is [-4.59, - 4.76].

2.4. Artificial Neural Networks

Neural networks are a modeling method that can reproduce extremely complex dependencies. A

number of principles of neural networks are based on crude low-level models of biological neural information-analyzing systems, their study led to development of artificial intelligent computer systems used in various problems of data analysis. Neural networks have arisen in the study of the basics of artificial intelligence, and mainly in attempts to reproduce a fault-tolerant and "learning-capable" biological neural system by simulation of a low-level structure of the brain.

Neural networks are non-linear in nature. They are able to find and extract useful information (rules and trends) from complex, noisy and inaccurate data. They can be used to extract patterns and detect trends that are defined by complex mathematical functions, and, if possible, to build models using analytical or parametric methods. Neural networks are able to predict accurately target values for data not involved in the learning process. This ability is also called a generalizing ability of a neural network.

When building the model, we chose the most commonly used type of architecture - Multilayer perceptron, with one hidden layer, including 13 neurons with logistic activation function (Figure 4).

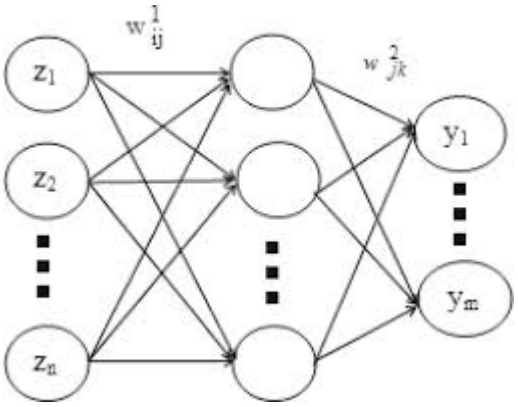

Figure 4. Architecture of Multilayer perceptron with one hidden layer.

Determination coefficient obtained by the neural network model is R^2 = 0.82. 50 % of the residuals, quartile range is [-3.30, -2.92].

2.5. Smoothing

The studied characteristics are subject to strong local variation, are noisy. For greater generalizing ability of the models, we smooth the series by the moving average method with a window of 14 days.

All models were rebuilt for smoothed data.

3. Analysis of the Results

The figure below represents a joint linear graph of the actual and forecast model values of the target variable for the entire observation period, as well as for one cycle period for unsmoothed data:

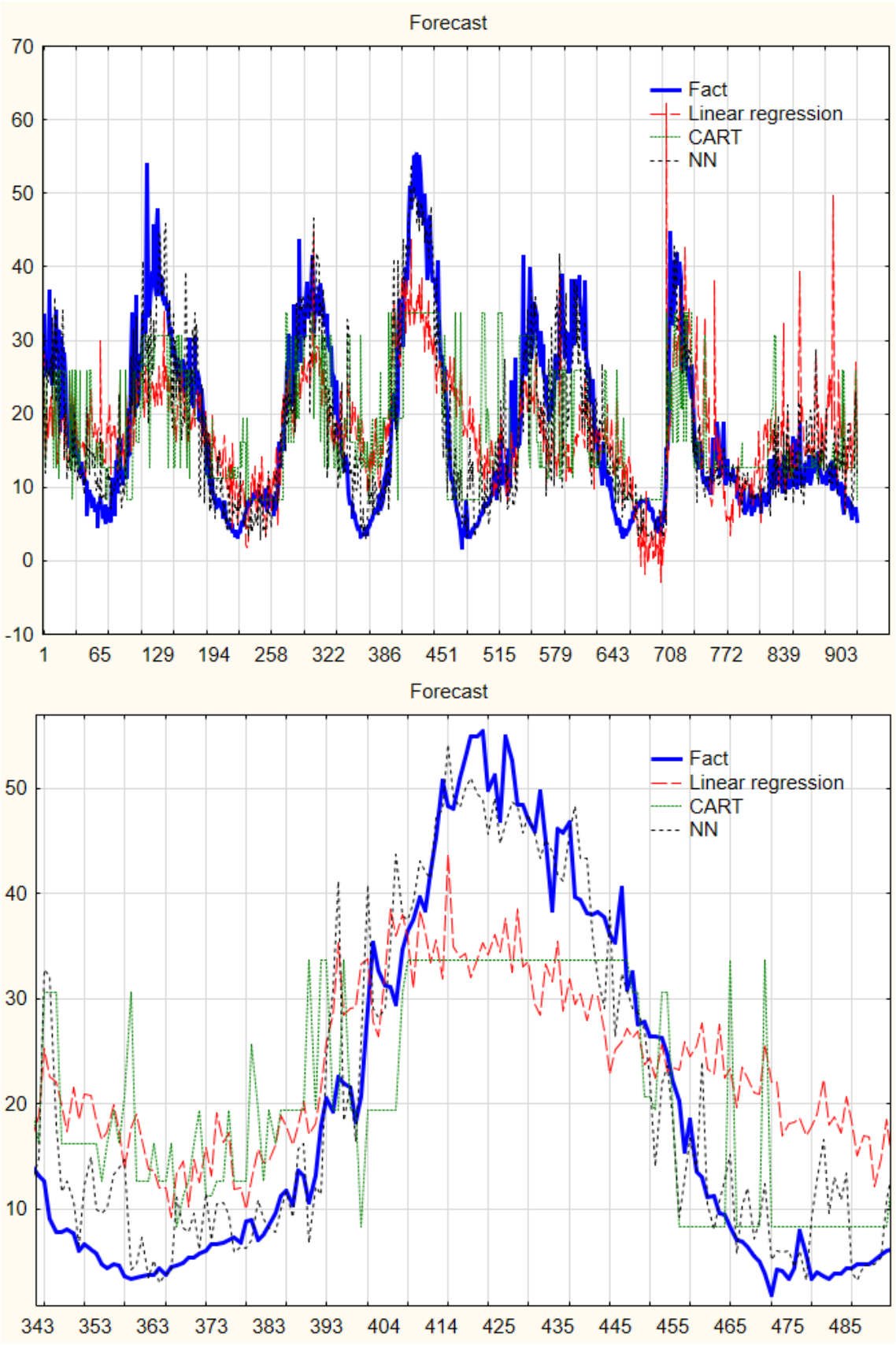

Figure 5. Graph of actual and forecast values.

The box plot for model residuals (dot — median, "box" - quartile range, segment — range) for unsmoothed data:

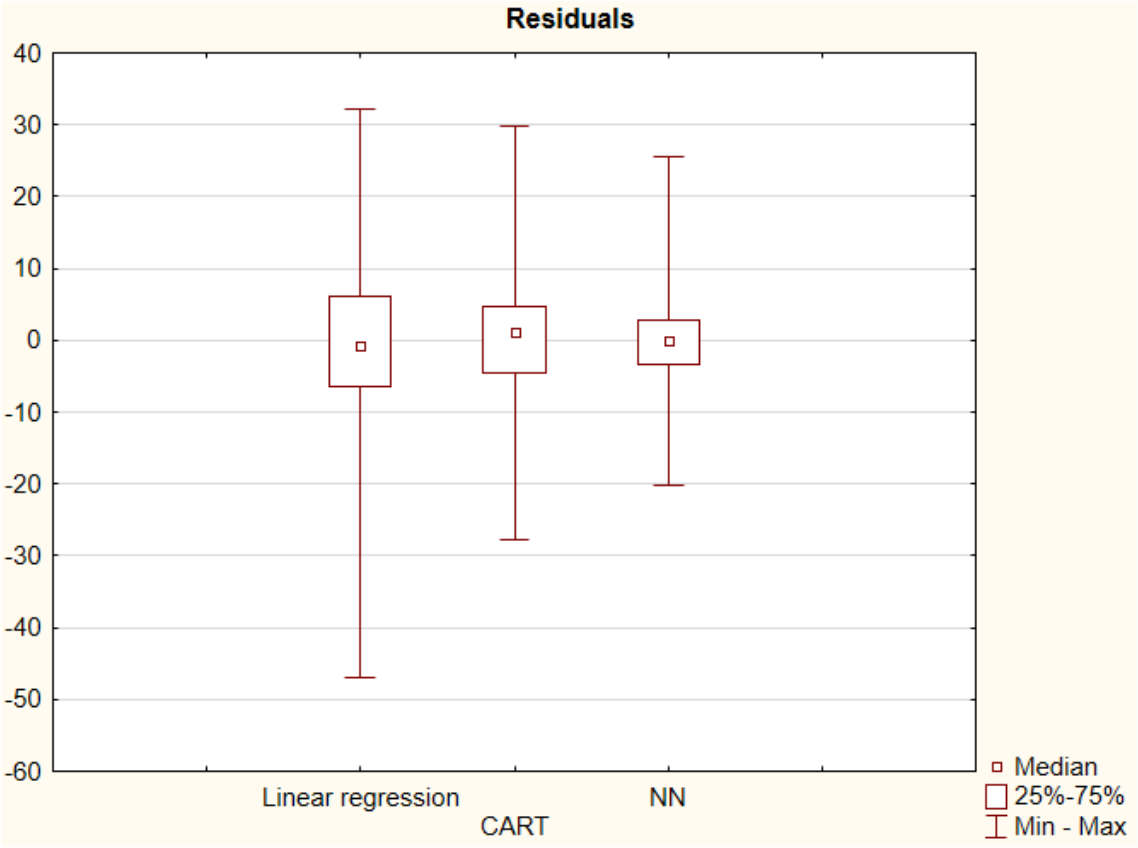

Figure 6. Box plot for residuals.

The summary table below displays the accuracy of the predictive models obtained using the coefficient of determination metric R^2:

Table 3. The coefficients of determination of built models.

Data	Linear regression	CART	Neural networks
Raw data	0.41	0.48	0.82
Smoothed data	0.74	0.82	0.99

In this study, the most accurate model is the neural network one. The conventional disadvantage of the model is the complexity of interpreting relationships between the target and independent variables. Often a neural network model is considered as a "black box" due to the complexity and non-linearity of the model.

The tree CART method gives an intermediate in accuracy forecast (it is more accurate as compared to linear models, but less accurate than neural networks). This method has an advantage - visibility and ability to interpret the resulting model by an expert.

Linear models demonstrated the least forecast accuracy.

4. Conclusion

The study examined both linear and non-linear dependency models:

- Multiple linear regression,
- Tree CART,
- Artificial neural networks.

An analysis of the factory data of precipitation stage showed that non-linear Data Mining methods allow building models with high accuracy to predict the target characteristic - the content of - 45 μm aluminum hydroxide fraction.

Thus, these methods can be considered promising for solving the problem of control and stability of size distribution of aluminum hydroxide product: using the obtained forecast model, having solved the optimization problem, it is necessary to calculate preliminary the volume of hydrate output per day and the values of other controlled factors of the process to obtain the desired response.

Data Mining methods have also proven to be good in solving the problems of various industries, both non-production and production. We give examples of such problems from ferrous metallurgy:

- Prediction of mechanical properties and optimization of chemical composition of metal,
- Identification of causes and prevention of defects of various origin,
- Statistical product quality management.

5. References

1. V.P. Borovikov. Statistica: the art of data analysis on computer, St. Petersburg, Peter 2003
2. V.P. Borovikov. A popular introduction to modern data analysis and machine learning with Statistica, Moscow, Hot Line – Telecom, 2018
3. I.V. Loginova, A. V. Kurchikov, N. P. Penyugalova Alumina Production Technology, Yekaterinburg, Publishing House of the Ural University 2015
4. I.A. Troitskiy, V.A. Zheleznov, Metallurgy of Aluminum, Moscow, Metallurgy, 1984