Optimization of the aluminum electrolysis process by mathematical modeling of process algorithms on cleaned data

M S Chistyakov¹,6, A N Losev², S V Shamina³, A L Zolkin⁴ and G V Ryabkova⁵

¹ "Russian University of Cooperation", Vladimir branch, Vorovskogo Street 16, Vladimir 600000, Russia
² Department of Applied Informatics, Russian Timiryazev State Agrarian University, Timiryazevskaya Street 49, Moscow, 127550, Russia
³ Department Natural science disciplines, South Ural state agrarian University, Gagarin street 13, Troitsk 457100, Chelyabinsk region, Russia
⁴ Computer and Information Sciences Department, Povolzhskiy State University of Telecommunications and Informatics, L.Tolstogo Street 23, Samara 443010, Russia; Natural Sciences Department, Private institution of higher education "Medical University" Reaviz, Chapaevskaya Street 227, Samara, Russia
⁵ Department of Linguistics and Translation, Moscow Aviation Institute (National Research University), Volokolamskoe shosse 4, Moscow 125993, Russia

E-mail: shreyamax@mail.ru

Abstract. The article discusses the issues of construction of a digital model of aluminum electrolysis modes. Target hypotheses identified as priority ones have been worked out: recommendations for AlF₃ supply, forecast of the electrolyzer overhaul. For the selected hypotheses, models have been built and studies on an expanded data set have been carried out. The results of the work showed that the models have been trained and are ready for use with the required accuracy of obtaining the results. In addition, prototypes of management systems for influencing tools have been built, control parameters have been determined. The scenarios for use of models in real production have been identified and described. The results at this stage allow starting a full-fledged production experiment, which will make it possible to determine the parameters of models introduction into production processes.

1. Rationale
The main method for producing aluminum is the electrolysis of alumina dissolved in a cryolite melt at a temperature of 940-970 degrees Celsius (the Hall-Heroult method). The process is carried out in electrolyzers with baked or self-baked carbon anodes, which are burned and form carbon oxides during the anodic oxidation of oxygen ions. A discharge of aluminum ions occurs on a liquid aluminum cathode with the formation of a metal [1].

To achieve the objectives of the study, the following hypotheses have been put forward:
- anode wear forecast will increase the service life of the anodes and reduce the cost of their replacement;
• forecast of pots overhaul will extend the life of the pots, increase their productivity, improve product quality and reduce the cost of overhaul;
• identification of pots with abnormal electricity indicators will reduce electricity consumption and, thereby, increase the current efficiency;
• forecast and recommendations of AlF3 weighted portions will reduce the cost of expensive fluorides. As an additional, the hypothesis of determination of pots with abnormal ALF3 flow rates depending on the service life has been accepted;
• reducing of the amount of aluminum that is involved into the reverse reaction will allow to increase the current efficiency.
• the definition of controllable factors affecting the aluminum yield will allow to increase the current efficiency. In particular, it is necessary to investigate the factors that affect power consumption, that affect the artefact and waving on the metal and try to make an instability forecast;
• determination of controllable factors affecting the quality of aluminum will improve the quality of market products without current efficiency reduction and without costs increasing;
• prediction of the anode effect for a variety of parameters and determination of these parameters.

2. Data quality monitoring, data preparation and cleaning
The data of the electrolysis process, the state of electrolyzers, the qualitative composition of the electrolyte, raw materials and finished products at the electrolysis plant are not centralized and stored in a single database, are not normalized and essentially are scattered. Electrolyz telemetry data, for example, is stored in the Tenora database, process characteristics such as, for example, the weighed amount of fluorides, impurities in the aluminum obtained, data on the last overhaul of pots is stored in the Electrolyz database, data on the quality of the feedstock are stored in the 1С system. Pot overhaul data is stored in an Excel file. Many of the same data are generated from different sources, stored in different tables, and provide non-identical information about the same process or event [2].

Manually collected data (for example, melt temperature) are not validated for validity and relevance. As a consequence, a lot of work needs to be done to clean up the data.

The database of stage 1 electrolyzers (Tenora database) and the electrolysis process database (Electrolyz database) are physically different databases, therefore in order to obtain a reliable picture the data consolidation requires a preliminary data cleansing in addition to data processing.

In Tenora database, electrolyzers telemetry data are stored for each electrolyzer separately in its own table, thus the number of tables is equal to the number of electrolyzers. In such a situation, it is simply impossible to process data using standard SQL tools.

The difference in the number of eliminated records by features is explained by the variability of the unique values of this feature. So, there are very few unique values in stress, and it is impossible to weed it out at 5 and 95 percentile a time, since the same values lie outside the confidence interval as inside it. Power consumption, on the contrary, has a lot more variability, so about 10% of the entries have been screened out.

In the course of the work, data for the period from January 2017 to May 2019 have been used. This volume was enough to train the model. The expansion of the time range did not improve the accuracy, therefore, to save time and resources, it has been decided to use a sample from 2017-2019 data. It has 57677 records. During the data cleaning procedures described above, 23657 records have been screened out and 34020 staid. Since if there is a missing day in the data (a record line for a specific date), then the previous three days and the next three days are not included in the sample for training the recommender model (t-3 (the day before yesterday), t-2 (yesterday), t-1 (today), t (tomorrow)), then the training set consisted of 14295 records. Thus, if there is a missing day in the data (a record line for a specific date), then the previous three days and the next three days are not included in the sample for training the recommender model.

In the course of the work, the features of the data have been identified (in particular, the different
discreteness of obtaining data in different sources). Table 1 shows the data elements used in datasets, the frequency of data acquisition, and the method for converting the data to a single frequency.

**Table 1. Data elements.**

| Table name | Data discreteness | Method |
|------------|-------------------|--------|
| Tenora.cxt.tenrpt* | Every two minutes | Average value per day |
| ElectrolizMD.dbo.El_StatusVann | As the status of the pot changes (under repair or in operation) | The calculated attribute "Days since pot overhaul" has been created |
| ElectrolizMD.dbo.EL_CZL_ALCEA | From one day to 2 weeks | The last available value is used. |
| ElectrolizMD.dbo.EL_TehnEl | Daily, but there are breaks from one day to 2 weeks | The last available value is used. |
| ElectrolizMD.dbo.EL_CZL_Electrolit | From one day to 2 weeks | The last available value is used. |
| ElectrolizMD.dbo.EL_Vilivka_ME | Daily, but there are breaks from one day to 2 weeks | The last available value is used. |
| ElectrolizMD.dbo.EL_Naveski | Daily, but there are breaks from one day to 2 weeks | The last available value is used. |

In the course of work, the tables ElectrolizMD.dbo.EL_Naveski and ElectrolizMD.dbo.EL_TehnEl were combined by the fields EL_Naveski.id_T_TehnEl and EL_TehnEl.TehnEl_cod, the gaps were filled with the last values.

The data for the datasets was selected based on the principle of maximum impact:

- for the model based on the recommendation of AlF3 weighed portions - on the temperature change when the cryolite ratio changes depending on the aluminum fluoride weighed portion;
- for the overhaul model - on the likelihood of the need to take the pot out of service and putting it into overhaul, as well as the likelihood of a critical event for each of the event submodels

For each submodel, its own feature importance (distribution of influencing parameters) has been compiled. It determines the top of the most important features that have an impact on changing the target parameter of the model.

A dataset can be reviewed as a data lake. It is used in various combinations in the overhaul forecast model submodels:

- board temperature forecast model;
- model for forecasting the temperature of the bottom;
- flow outlet model;
- iron content prediction model.

For the dataset, the following data grouped by date and cell number in the period from the start of the plant in 2007 to August 2019 have been selected:

- shift report (bbrpt of Tenora database);
- the chemical composition of the electrolyte (ElectrolizMd.dbo.EL_CZL_Electrolit);
- the chemical composition of aluminum (ElectrolizMd.dbo.EL_CZL_ALCEA);
- tapping data (ElectrolizMd.dbo.EL_Vilivka_ME);
- weighed portions of fluoride salts (ElectrolizMd.dbo.EL_Naveski);
- process parameters (ElectrolizMd.dbo.EL_TehnEl);
• overhaul data (ElectrolizMd.dbo.El_StatusVann).
• For the dataset, electrolyzers that are in operation but not under overhaul (and at least 4 years have passed since the date of the last overhaul) have been taken.

Data are collected for the following observation period:

• the beginning of the observation period is the current date minus 90 days (3 months).
• the end of the observation period is the current date.

The data are filtered by the date field starting from the beginning of the observation period (period_start) and ending with the end of the observation period (period_finish), i.e. those lines for which period_start<= date <= period_finish are selected.

As part of the preparation of the dataset, features have been selected. All signs have been taken for the submodels of prediction of the probability of exceeding of the iron content limit. For the submodels of prediction of the probability of exceeding of the threshold of the side temperature and the bottom temperature, combinations of features have been enumerated to find the best result on the same model [3]. As a result, 15 features have been selected for each submodel. A similar approach has been applied in the submodel of prediction of the probability of a decrease in the average current efficiency, which uses 16 features.

In the Tenora database, the data of the electrolyser operation parameters are stored in such a way that a separate table is created in the database for each pot. The corresponding table is determined by the number of the electrolyser, and the records corresponding to the sampling period by date are retrieved from it. The BAID (wave) and ZHENZH (artefact) fields are selected. The following features are calculated for them:

• average value for the entire period;
• average value for the last day;
• average value for the last week.

Thus, the following set is obtained:

• BAID_mean;
• BAID_last_day_mean;
• BAID_week_mean;
• ZHENZH_mean;
• ZHENZH_last_day_mean;
• ZHENZH_week_mean.

The features have been chosen in accordance with the most significant feature importance. In total, there are 7 features in the submodel.

3. Elaboration of target hypotheses and identification of patterns
As a result of testing of the put forward hypotheses, two promising hypotheses have been identified for building of predictive and recommendatory models to achieve the research objectives.

Models:

• model recommendation for AlF3 weighted portions;
• model for forecasting of the pot overhaul.

Recommendations for AlF3 weighted portions are based on the prediction of the electrolyte melt
temperature. For the normal course of the electrolysis process and in order to obtain the maximum yield of the purest aluminum, it is necessary to maintain the temperature in the electrolytic cell at 953 degrees Celsius. The temperature is regulated by AlF3 weighed portions. Currently, at the electrolysis plant, in order to determine the size of the fluoride samples based on the current temperature, either standard values are used, or the sample size is determined based on the expert opinion of the process engineer. But at the same time, forecasts of how the temperature could change in the coming day with the current state of the electrolyte are completely ignored [4]. The use of such a forecast will make it possible to obtain more accurate weighed amounts of fluoride, to optimize the production process, to increase the current yield of aluminum, and to give a noticeable saving in expensive AlF3.

The model uses the RandomForest algorithm (RF - use of a committee of decision trees). It gives comparable results to the XGBoost (Extreme Gradient Boosting) algorithm, but works and learns faster than the second one.

The following dataset has been used. First, all parameters that, in principle, could affect the temperature have been collected. Then, using RFECV the optimal number of parameters (41 for RF) has been determined. Subsequently, the sum of the weighted portions for three days has been used to recommend the weight of the portion, instead of three separate weighed portions for each day [5,6]. The number of parameters has decreased to 38. Feature importance has been built for these parameters.

As a result, it has been determined that the temperature for the next day is mostly affected by: interval, pot life, energy consumption, weight of portion for three days.

To split the dataset, the train_test_split random split method from the model_selection library of the sklearn framework has been used. The dataset has been divided into 2 parts (training sample - 80% of the dataset, test sample - 20%). The method splits the values so that the distribution of the target variable value is maintained for each subsample. In this case, the lines are selected in random order. Rows from subsamples do not overlap. Each row of the dataset is used once and only in one of the subsamples).

4. Influence tools management and introduction of models into production processes

To obtain a real picture of the work of models, models specification, creation of influence tools and embed models in production processes, it is necessary to conduct a full-fledged production experiment with obtaining actual data on the use of models. The production experiment will make it possible to assess the use of models in comparison with the indicators of real production, econduct the xpert assessment of specialists in production, as well as evaluate the effectiveness of the use of models [7].

For a production experiment, the following model scenarios are proposed:

- for each pot process engineer enters the temperature measurement value in the model interface and launches the model to receive recommendations on the weighted portion;
- the model requests the pot state data for the last 3 days in the Tenora database;
- the model requests the electrolysis process data for the last 3 days into the electrolysis database;
- the model requests the date of the last start-up of the pot after a major overhaul into the overhaul database;
- the model calculates a range of temperature values for the next day when using weights from 0 to 70 kg. Than the value that is the closest to 953 degrees is selected;
- the model calculates the required amount of weighted portion of AlF3 per day (in kg) in order to maintain the regulated temperature;
- the model displays weighted portion recommendations to the process engineer;
- the process engineer sets the value of the sample in the automated process control system.

If at any step of the script a failure occurs or the database contains insufficient data for an exact recommendation, then an empty value is written to the table of hints recommendations, the reason for
the error or the reason for the impossibility of obtaining a recommendation is written as a comment[8,9]. The process engineer must make a decision, either to supplement the data for correct working off and restart the calculation, or to set the regulatory value of the weighted portion [10].

If at any step of the script a failure occurs or the database contains insufficient data for an accurate recommendation, then an empty value is written to the forecast table, an error log or the reason for the impossibility of obtaining a forecast is written with a comment, the decision to start overhaul is made on the basis of the regulations.

5. Findings

Thus, in the work, hypotheses have been put forward and tested on the factors that allow influencing the processes of optimizing aluminum production, increasing the volume of commercial products, reducing production costs and improving product quality.

The hypotheses for predicting the wear of the anodes and reducing the amount of aluminum reacting were rejected by the process engineers during the discussion of the hypotheses as poorly implemented.

Determination of controllable factors affecting the quality of aluminum have been suspended in development, due to the fact that they are not a priority.

Work on a number of descriptive hypotheses has been carried out: to identify pots with abnormal electricity rates, determination of pots with abnormal ALF3 consumption rates depending on the service life, determination of factors affecting power consumption, determination of factors affecting waves on the metal. For each of these hypotheses, data as well as statistics and underlying causal factors for the problems have been collected [11].

The hypothesis for prediction of pots overhaul is one of the most promising hypotheses. To confirm this hypothesis, a model that recommends the pots overhaul will be built. Extensive data preprocessing has been performed for this model.

The hypothesis for forecasting of the aluminum yield also includes the determination of controllable factors affecting the aluminum yield. Work to collect basic statistics on the current efficiency of the pots of the first stage has been carried out, a model that predicts the current efficiency has been built. This model made it possible to identify the main influencing factors.

On the recommendation of fluorides supply, it has been decided to proceed step by step: first, a model for prediction of the temperature for the next day shall be build up and then a model for recommendation of weighted portion shall be build up on its basis. Work to collect and preprocess data for the model has been carried out. A model for prediction of the next day temperature as well as a model for prediction of weighted portion have been built up. The last one needs serious revision at the subsequent stages.

The following hypotheses have been selected as priority ones:

- forecast of the pot overhaul;
- recommendations for AlF3 supply.

References

[1] Simakov D A 2006 Study of the features of electrolysis of alumina suspensions in fluoride melts in order to improve the Hall-Heroult process: thesis abstract ... PhD in Engineering scince (Krasnoyarsk: Krasnoyarsk State University of Non-Ferrous Metals and Gold) 26 p

[2] Makushin D V 2007 Increasing the efficiency of aluminum electrolysis based on optimizing the distribution of trace impurities in the "electrolyte - cathode aluminum" system: abstract for Ph.D. thesis in Engineering Science (St. Petersburg: Saint-Petersburg Mining University) 22 p

[3] Utyuzh A S, Yumashev A V and Mikhailova M V 2016 Spectrographic analysis of titanium alloys in prosthetic dentistry Journal of Global Pharma Technology 8(12) pp 7-11

[4] Yumashev A and Mikhaylov A 2020 Development of polymer film coatings with high adhesion
to steel alloys and high wear resistance *Polymer Composites* 41(7) pp 2875-880 doi: 10.1002/pc.25583

[5] Yumashev A, Semenycheva I, Rakhadilov B and Tsymbal A 2019 Development of biocompatible coatings for dental implants based on transition metal nitrides *Journal of Global Pharma Technology* 11(5) pp 22-8

[6] Perminova O M, Lobanova G A and Faizullin R V 2019 Raw materials inventory model applied by regional enterprises of the industrial cluster *Acta Logistica* 6(2) pp 35-41

[7] Lavrov E, Krivodub A, Tolbatov A and Pasko N 2016 Mathematical models for the distribution of functions between the operators of the computer-integrated flexible manufacturing systems *Modern Problems of Radio Engineering, Telecommunications and Computer Science, Proceedings of the 13th International Conference on TCSET* 13 pp 72-5

[8] Lavrov E, Volosiuk A, Pasko N, Gonchar V and Kozhevnikov G 2018 Computer simulation of discrete human-machine interaction for providing reliability and cybersecurity of critical systems *Proceedings of the 3rd International Conference Ergo-2018: Human Factors in Complex Technical Systems and Environments, Ergo* 67-70 DOI: 10.1109/ERGO.2018.8443846

[9] Lavrov E, Pasko N, Siryk O, Burov O and Natalia M 2020 Mathematical models for reducing functional networks to ensure the reliability and cybersecurity of ergotic control systems *Proceedings - 15th International Conference on Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering, TCSET* 179-84 DOI: 10.1109/TCSET49122.2020.235418

[10] Burov O, Lytvynova S, Lavrov E, Krylova-Grek Y, Orliyk O, Petrenko S, Shevchenko S and Tkachenko O 2020 Cybersecurity in educational networks *Advances in Intelligent Systems and Computing* 1131 AISC 359-364 DOI: 10.1007/978-3-030-39512-4_56

[11] Lavrov E, Paderno P, Volosiuk A, Pasko N and Kyzenko V 2019 Automation of functional reliability evaluation for critical human-machine control systems *Proceedings of 2019 3rd International Conference on Control in Technical Systems, CTS* 144-7 DOI: 10.1109/CTS48763.2019.8973294