LAPPEENRANTA UNIVERSITY OF TECHNOLOGY
LUT School of Engineering Science
Degree Program of Chemical Engineering

Master's Thesis
2018

Pavel Maksimov

APPLICATION OF COGNITIVE SOFTWARE FOR SPECIFICATION AND CHARACTERIZATION OF VALUABLE SIDE STREAMS IN METAL INDUSTRIES

Examiners: Professor Tuomas Koiranen
Tech. Lic. Esko Lahdenperä
Supervisor: Professor Tuomas Koiranen

ABSTRACT

Lappeenranta University of Technology
LUT School of Engineering Science
Degree Program of Chemical Engineering

Pavel Maksimov

Application of cognitive software for specification and characterization of valuable side streams in metal industries

Master's Thesis
2018

102 pages, 44 figures, 20 tables and 7 appendices

Examiners: Professor Tuomas Koiranen
Tech. Lic. Esko Lahdenperä
Supervisor: Professor Tuomas Koiranen

Key-words: data analysis, metal industries, products characterization, cognitive computing.

Currently, the rate of novel knowledge generation is ever increasing at an exponential rate, thus providing demand for new advanced data analysis methods. Therefore, large corporations along with smaller startups are trying to integrate new technologies for big data processing with the purpose of obtaining additional insights from seemingly useless datasets. At the same time, due to the recent widespread deployment of cloud-based technologies, software applications that require excessive computational capability can be accessed and operated from any device regardless of its processing power. In view of this, cloud-based software tools for information processing, that facilitate efficient data analysis without significant computational power consumption, have drawn increased attention from the modern scientific community. Thus far, these software utilities have found extensive application mostly in social studies or business and management areas, while their utilization for analysis of technical information has been relatively limited.

Within the limits of this work application of novel data processing software for metal industries related information analysis has been studied through the example of IBM Watson's analytical platform. A series of datasets containing relevant information about aluminum and copper production processes has been collected and initially refined to facilitate further analysis. These datasets have been studied with the purpose of revealing

hidden interdependencies between processing parameters and determination of the key statistical drivers affecting main products quality. Special consideration has been given to the side stream of copper production, namely slag, and main processing parameters that were associated with the most significant impact on its properties. Predictive analysis of main slag and matte constituents' content has been conducted in order to provide more insight into the process parameters and products properties. Additionally, a comparative analysis of the analytical platform with a conventional linear-regression based software application has been performed in order to assess and to highlight possibilities of Watson Analytics to group, classify and connect process related information.

The obtained results highlighted a crucial importance of quality and volume of the analyzed information set. More specifically, amount of available information had the most significant impact in terms of predictive analysis – increased quantity of observations resulted in generation of more detailed predictive models. Likewise, results of the product properties key drivers' determination were also associated with increased statistical accuracy in case of analysis of larger datasets. Furthermore, the conducted study revealed significant interdependence between main matte and slag constituents' concentrations. Among the most important processing parameters, cooling water temperature and oxygen concentration have been associated with the most substantial impact on the product properties.

Moreover, an additional study has been conducted by application of Watson Discovery service for analysis of a collection of scientific articles related to processing and reuse of waste and side streams in metal production industries. Data ingestion algorithms have been customized and the cognitive system has been trained to understand industry-specific terms and text context in order to facilitate natural language processing.

Contents

LITERATURE REVIEW

1. Introduction

Over the past years humanity generated an incredible amount of scientific knowledge – more specifically, data generated since 2015 is approximately equal to 80% of the data of all time. Owing to the fact, that computer storage related technologies experienced a significant improvement over the previous decade, and, consequently, data collection is getting cheaper every year, amount of accumulated novel research information increases almost at an exponential rate. Because of modern storage opportunities large companies and even small startups started to keep any potentially valuable data with the intention of further detailed analysis to acquire any useful meaningful information. Moreover, this tendency has been additionally supported by the recent revolution in data storage – cloud-based technology, that, empowered by improved computing resources along with increasing availability of fast internet connections, made the accumulation of large amounts of data even easier. However, 80% of the novel generated and collected information is unstructured, meaning that it does not have any inherent order or pattern in it, therefore, conventional computing algorithms are not capable to process this data and derive any helpful insights. Another issue to be considered is that, in general, unstructured data can contain plenty of so called data noise – more specifically, the information can be complex, confusing or even sometimes consist of contradictory statements. This "noisy" data is a common issue in science related materials that makes drawing a confident and well-reasoned inference difficult even for a specialist in the area. An algorithm capable of handling at least obvious conflicting indicators and resolving this unstructured information issues would offer a significant support for researchers and other specialists in terms of data analysis. In other words, traditional data analysis approaches are not anymore viable in the current context and another solution that can effectively deal with complexity of unstructured data is of crucial significance (Y. Chen, Argentinis & Weber 2016).

On the other hand, simultaneous recent improvements in computing power, that to a large extent facilitates data processing, resulted in overall growth of interest towards artificial intelligence and cognitive based technologies.

One of the possible ways to achieve artificial intelligence is so called machine learning approach. Currently, the most common technique involved in machine learning is artificial neural network. Machine learning is based on algorithms that are capable of processing large amounts of so called training data by modeling complex relations between input and output

values to find patterns in the data and to learn from it in order to make viable and relevant predictions based on the acquired experience.

In the area of artificial intelligence technologies, most of the recent advances became possible because of such sub-technique as deep learning, that makes machine learning much faster and accurate. Basically, deep learning empowered algorithms employ parallel programming relying on various layers of neural networks and millions of instances of training data. Consequently, these machine learning algorithms extensively involved in the development of sophisticated artificial intelligence structures could be also applied for data analysis.

Furthermore, since big data mining driven by machine learning algorithms requires high computational capacity, previously mentioned cloud-based technology not only facilitates convenient data storage but also supports promotion and expansion of modern computing technologies. This way complex solutions can be available to common devices regardless of computing power. Therefore, cloud technology provides a unique opportunity for new computing product offerings to reach common masses, thus developing a solid foothold for the emergence of artificial intelligence based technologies.

Nowadays, these technical solutions are designed in the way, that even common users, without any experience in data science, are able to promptly analyze enormous amounts of information, clean data noise, find previously hidden patterns and insights and make well-reasoned conclusions.

2. Cognitive computing

2.1. Introduction to cognitive computing technologies

Cognitive based algorithms are a new computing paradigm empowered by the Internet that involves data collecting from different sources in order to find new solutions and improve existing technologies.

By itself, the term "cognitive computing" is hard to define due to its complexity and versatility, but in general, it describes the automated solutions that are able to learn complex tasks, interact with users by means of natural interfaces and make reasoned decisions, in a way, mimicking human cognitive abilities. This complex ability to think and analyze information is mostly achieved due to the combination of such computing tools as: image recognition, natural language processing, intelligent decision analysis and empowered search techniques. Integrated application of these techniques provides an opportunity to adapt underlying computing algorithms for processing and analysis of miscellaneous data (Tarafdar, Beath & Ross 2017).

This technology is at the forefront of novel generation of computing systems that enable efficient user-machine collaboration, where the system is not only able to understand commands in natural forms, but also self-sufficiently learns based on the previously processed information, iterating the operations in a sequential approach until the desired result is obtained. Therefore, by means of this feedback loop between computer and user, human operator and cognitive system are able to learn from each other.

Since, cognitive systems are equipped with machine learning algorithms, they are able to improve themselves as they collect and process data related to a certain topic by learning language, terminology and even certain semantic features of the domain. Gathering enough expertise in a certain field, these systems are capable of providing a decision support basis, thus helping human experts to create the most efficient solutions, based on the all available data and potentially undetected insights.

Currently, with rapid advancement of cloud services, cognitive computing technologies are switching to the cloud based functional deployment, which allows each network function to be operated through the effective cloud hardware. Thus, available computing power can be immediately scaled up according to the user requirements by allocating hardware resources for particularly demanding operations.

Cognitive computing facilitates plenty of data processing services, such as sensing, sharing, comparing, interpreting, correlating, etc. to enable and support human activities in various

fields of action. Consequently, these systems should be able to efficiently observe, filter and recognize valuable information among large amounts of empty data and process these observations to develop meaningful patterns.

2.1.1. Common algorithms in cognitive computing

Major part of reliable scientific information nowadays exists in a form of so-called "unstructured data". More specifically, this term can be described as an information set without any inner descriptive structure or pattern corresponding to the targeted task. Unstructured text typically represents large aggregate of various numbers, words, symbols and characters with a certain degree of organization, which is usually based on syntax and grammar rules for instance. Since, computing algorithms are intrinsically designed to understand and process data mostly in the forms of code or logic expressions, computers are adapted to processing of structured rather than unstructured form of information. From the perspective of a cognitive computing system in order to enable extraction of any meaningful insight, such dataset should be firstly transformed, adjusted and organized into suitable for analytic processing format. Furthermore, most of information sources contain tables, figures and cross-links, that also should be considered during data analysis process. This formation process usually involves several steps, such as:

- text normalization, where text is transformed into normalized and consistent form, with defined format and characters set;
- language identification, where basic contents and syntax features of the document's language are determined;
- tokenization, where different elements of a sentence (words, numbers, punctuation marks and other characters) are separated into different language-dependent blocks, to simplify further text processing;
- feature extraction, which involves patterns recognition in context to find and categorize information in accordance with pre-defined setup;
- entity recognition, where certain semantic text entities are identified according to their relation to the relevant topic;
- relationship extraction, when relations between two or more previously determined text entities are studied by the context analysis, and afterward identified and extracted.

Often, prior to text analysis by cognitive computing algorithms, it is profitable to develop a scheme of classification relevant to the studied knowledge area with a list of text entities to be discovered in order to optimize text processing.

Once the analyzed dataset is structured and transformed into suitable for the cognitive system format, a number of reasoning algorithms are applied to find the most relevant answer among the generated alternatives. Currently, major part of cognitive empowered systems operates according to the "alternative generation using selection" approach, where among the potentially suitable solutions the most relevant one is selected by means of comparison with the other alternatives in compliance with the pre-defined evaluation criteria. Furthermore, another reasoning engine, based on the extrapolation approach, can be applied. In this case, a new component is created by customization of the existing potentially most suitable alternative or overall configuration, so that the new proposed solution is modified to better meet the necessary criteria. Another frequently used reasoning engine involves configuration basis: the algorithm collects a set of components based on the user targets and configuration constraints to create a new alternative solution in reliance on existing ones (Sathi 2016).

In order to make emulation of cognitive activity possible, cognitive systems are typically designed with inherent training processes, which are quite similar to the human learning routine. More specifically, these systems are provided with a number of processes to implement certain activities in the specially simulated environment. One issue to be considered is that cognitive systems can remember everything they learn perfectly, but they are not able to apply so-called common sense or innovative reasoning when faced with a dead end. For this reason, sophisticated cases require plenty of initial pre-training and additional background to handle exceptions and other potential hindrances before decent data analysis can be conducted. Additionally, even though, cognitive systems are usually carefully designed to avoid this problem, as in case with any computer empowered smart systems, an endless loop of certain operations can be encountered, and human operator interference might be necessary to fix this issue.

Currently, there is a vast variety of different cognitive computing tools available in the market and new technologies are being developed with increasing frequency. However, not all of them are suitable for dealing with all kinds of objectives. More specifically, depending on the different tasks, in general, several common types of techniques are distinguished in relation to whether they are suitable for tackling narrow specific issues or broad problems that require large specter of available solutions.

For cases when specific tasks such as natural language recognition, text searching, or mining of unstructured data need to be tackled without any further processing, point solution approach is typically implemented. Generally, these tools are embedded with relatively

simple and uniform techniques, based upon vocabulary rules, syntax and grammar standards for a certain field of human activity.

Broad spectrum cognitive tools are typically developed with a set of general-purpose techniques which can be specifically adapted for certain applications. A good example of this kind of technologies can be drawn by looking at IBM's Watson® suite of applications, that comprises different techniques involving natural language processing, pattern recognition and information classification.

2.1.2. Data sources and interfaces

Successful application of cognitive computing based tools depends heavily on reliable and relevant data sources. Considering constantly growing rate of data and knowledge generation, the problem of choosing right information sources is especially acute, since the returned by a cognitive tool solution depends heavily on the processed data. A specific cognitive computing application might require additional external information for appropriate data processing and pattern recognition, such as, for example, scientific journals, legal documents, reports or libraries. Therefore, the process should be supervised by technical experts in the subject area, so that the intermediate steps and output responses can be validated and checked for accuracy, adequacy and relevance. Furthermore, information basis should be maintained and managed accordingly by the supervisor in such a manner that the knowledge quality is kept at an optimal level. Otherwise, there is a high possibility that over time, application's algorithms will lose their relevance and accuracy so that less and less information is processed in a meaningful way, thus providing incorrect and noncoherent results. With sufficient level of supervision by an experienced and well-trained expert, a cognitive computing software application can take on routine and tedious tasks, thus leaving complex and innovative solutions to human specialists. Taking into consideration constantly growing amount of different vast and disparate data sources, application of cognitive computing technologies in the industrial sector will help in addressing big data challenges by facilitating comprehensive analysis of unstructured information in a single solution, so that novel, previously hidden insights can be discovered. (Tarafdar et al. 2017, Y. Chen et al. 2016).

2.2. Available cognitive computing software tools

SparkCognition® – one of the global leaders in cognitive computing area. Company was founded in 2014 in Texas, USA. Augmented intelligence based products of the company are mostly utilized for cybersecurity improvement (SparkSecure™ software) and machine learning technology for identification and prevention of equipment potential failures (SparkPredict™ software) (SparkCognition Inc. 2018). Main feature of the SparkCognition technology lies in the ability of harnessing and processing big analog data generated by a vast variety of machines and improve itself based on it (Newstex 2015). Among successfully implemented projects of the company, a cognitive malware detection engine for Android based devices – DeepArmor™ should be mentioned. Additionally, SparkPredict™ software is widely applied for prevention of machinery malfunction of generators, actuators and turbines. The company's patented algorithm Artemis™ is developed for multi-variable anomaly identification, while machine learning technique Pythia™ generates a model for malfunction prediction.

Microsoft Cognitive Services® offers a wide range of possibilities for application development by utilizing natural language processing along with speech and image recognition algorithms. In terms of natural language understanding, a machine learning based algorithm LUIS™ is designed for text and speech recognition for building language models, which allow mobile/desktop apps or bots efficiently recognize commands. LUIS™ is powered by the logistic regression concept – modelling technique for prediction based on the processed information. (N. Pathak 2017). Microsoft Cognitive services pack is mainly focused on gathering and processing of unstructured information from the open sources to help web developers to understand users' requirements.

Numenta® machine learning algorithm powered by the neocortex theory (prediction, anomaly detection and pattern recognition by neurons in a human brain structure) was designed for anomaly detection in applications, servers and tracking data. Numenta software has been applied in such fields as: stock performance monitoring and prediction, human behavior anomaly detection and natural language processing. Main feature of the company's computing tool is ability to efficiently work with constantly updated streaming data and inherently structured data. Regarding the natural language processing issue, Numenta's technology has been utilized by Cortical.io® company for development of Natural Language Processing Solutions™ – semantic algorithm for text analysis in real time (Numenta Inc. 2018). Numenta's software has been involved in several scientific studies, among which are the following:

- Unsupervised Real-Time Anomaly Detection for Streaming Data;
- Multi-scale streaming anomalies detection for time series;
- An online prediction software toolbox based on cortical machine learning algorithm;
- Automatic detection of urban areas using the Hierarchical Temporal Memory of Numenta;
- Application of Numenta® Hierarchical Temporal Memory for land-use classification;

Expert System® provides cognitive computing solutions for accurate and automatic understanding of unstructured textual data, focusing on decision making improvement and cost savings. Main cognitive computing technology – Cogito™ is designed for natural language processing in such unstructured data sources as documents, e-mails and research articles. Among software products of the company, which are based on the mentioned cognitive computing concept, the most relevant one to the technology sector is CogitoStudio™ that combines semantic technology and machine learning algorithms for modelling of text analytics solutions to categorize and process information (Expert System Enterprise 2017a). Cogito™ technology has been utilized for unstructured data analysis in oil & gas industry by Expert System's prime partner – Shell® (Expert System Enterprise 2017b). External information sources such as scientific articles, blueprints and reports were used for improvement of information management in exploration and production phases of the industry. Processing of internal information allowed to minimize operating risks, protect production assets and verify economic viability of contracting organizations (Expert System Enterprise 2018).

CustomerMatrix® exploits novel cognitive computing approaches in the area of financial services to pinpoint and harness any potential possibility for economic growth. Main artificial intelligence based engine of the company – Opportunity Science™ is designed for improvement and acceleration of existing customer relationship management by applying cognitive computing and machine learning for processing of large quantities of unstructured data. The engine is capable of real-time processing of external and internal information sources linked to company's customers (PR Newswire Association LLC 2015b). Opportunity Science™ pinpoints at any signals that can have potential direct impact on revenue of a company and also develops the most suitable practices for the solution of the problem (PR Newswire Association LLC 2014). CustomerMatrix® software products found extensive support and interest in Asian banking systems and were implemented in tier 1 global banks (PR Newswire Association LLC 2015a).

HPE Haven OnDemand® provides augmented intelligence solutions, powered by Microsoft Azure® platform, designed for enterprise application developers (S.Yegulalp 2016). Haven OnDemand® approach is designed for processing and extracting valuable insights from multiple text formats, visual and audio data. The company provides vast variety of relatively simple algorithms for text analysis: concept extraction based on statistical methods; documents categorization; entity extraction from unstructured data; language identification; sentiment analysis; text tokenization designed for terms specification (Hewlett Packard Enterprise Development LP 2017).

CognitiveScale® is yet another company involved in cognitive computing industry, that combines machine learning approaches and big data processing for acceleration and improvement of decision making. The company's software is mostly designed for application in the area of enterprise applications development and internal business processes improvement. Augmented intelligence platform CognitiveScale Cortex™ has been trained to understand certain features and nuances in financial services, healthcare and digital commerce industries (CognitiveScale Inc. 2018). Another feature of CognitiveScale Cortex™ that should be highlighted is its transparency – for every decision it provides detailed reasoning and explanation.

DeepMind® – a wholly owned sub-company of the Google® conglomerate, that provides cognitive computing based software solutions for various fields and for health industry especially. In partnership with National Health Service of the Great Britain, DeepMind® was set to process and analyze millions of patients' personal records in order to assist in diagnosis of kidney related diseases. It goes without saying that all the collected information about patients is highly encrypted and protected. Currently, DeepMind® plans to significantly broaden its area of health expertise and get involved into solving other disease related problems (J. Powles 2017). Regarding more industry-oriented solutions provided by the company, a good example can be drawn by looking at successful application of DeepMind® in terms of energy consumption reduction in Google Data Centre. Machine learning software has been successfully applied to process and analyze collected unstructured data from thousands of sensors in the cooling system (temperature, power, flowrates, etc.). The obtained results were used to create a new improved framework for the cooling system, with 40% reduced energy consumption (R. Evans 2016). Additionally, DeepMind® current research projects include speech recognition algorithm WaveNet™ and famous AlphaGO™ computer program that managed to defeat a professional human Go player (A. Oord, T. Walters, T. Strohman 2017).

Atomian® – a technology-based R&D company that specializes on applying cognitive based approach to provide enterprises with valuable insights among tons of internal unstructured data through efficient natural language processing algorithms. Atomian® divides unstructured data into "knowledge atoms" and collects them in different memory levels: episodic, semantic, declarative, logical-mathematic, etc. This way, Atomian® has been designed to rearticulate language data into bundles of computable units. By processing unstructured or structured data (.txt, .pdf, .xls formats) Atomian® is able to create its own cognitive architecture which can help organizations to reveal previously undetected opportunities for improvement and reduce amount of time necessary to find certain information (Atomian 2016). The company's software was designed for application in such areas as: healthcare industry – for processing of medical records (Atomian Healthcare™); banking industry – for risk analysis optimization (Atomian Focus™); travelling industry – for travel request process management (Atomian GO™); production industry – for processes automation and utilities savings (Atomian Enterprise Projects™).

Coseer® focuses on tactical cognitive computing (compilation of cognitive calibrating and natural language processing) for solving complex and tedious problems in business industries. An augmented intelligence powered self-learning software of the company for natural language processing is based upon the Calibrated Quantum Mesh™ – approach designed to mimic the human reasoning process. This algorithm is designed based on the notion that any symbol, word or other entity can be recognized differently according to the conditions so that these "quantum states" are all taken into account with least possible ambiguity to find the most accurate solution. Additionally, the Calibrated Quantum Mesh™ approach assumes, that everything is related to each other with different extents of influence, so these interdependencies are collected in the single mesh to increase accuracy (Coseer Inc. 2016). Cognitive computing module Longseer™ has been applied in hardware production industry for collecting and processing information from the Internet, relevant industries and individual reports in order to provide the most valuable insights from any possible information source (Arbot Solutions Inc 2017).

ROSS Intelligence® – company that features utilization of augmented intelligence based solutions in legal services industry. Proprietary software solutions of the company are combined with IBM's cognitive technology to utilize machine learning approach and natural language algorithms in order to provide legal authorities with relevant specific insights among big data. The company's program – ROSS™ is provided with semantic analysis abilities to understand information written in plain unstructured language. On average,

implementation of the ROSS™ software resulted in the 22% reduction of time required for necessary information searching and also relevancy of the retrieved authorities increased by 40% (Blue Hill Research 2017).

Digital Reasoning® is another leading company involved in cognitive computing industry. The company provides solutions powered by machine learning to understand the context of natural language and infer valuable insights or make accurate predictions. One key distinction is that the company's algorithm is developed to automatically create a graph-based data model, that is used for analysis and predictions. So far Digital Reasoning® software solutions have found application in such industries as: healthcare industry, financial and governmental services. The cognitive powered platform Synthesis™ has been applied in financial industry in terms of employee surveillance for potential human risk reduction. Additionally, the platform can be used to analyze any information about company's customers to reveal valuable insights. Similar to ROSS™ platform, Synthesis™ has also been applied in the legal services industry to use machine learning and cognitive approach for data analysis acceleration (Digital Reasoning Systems 2017, T. Zerucha 2016).

Perpetuuiti TechnoSoft® – is software developing company that combines cognitive computing approach with automation systems, helping enterprises maximize efficiency. Av3ar™ is the company's main software cognitive empowered platform, that works on deep learning technology for natural language processing among large quantities of unstructured data. For Av3ar™ typical information sources are Internet and internal information about potential customers. Av3ar™ was designed for application in accounting (audit data analysis and cash flow management), marketing (customer related information processing) and recruiting (pre-screening of applicants) (PerpetuuitiTechnoSoft Inc 2017).

Beyond Limits® – a cognitive augmented intelligence company famous for providing efficient solutions for NASA's Deep Space program®. Beyond Limits® cognitive empowered software solutions were applied for such famous missions as: "Pathfinder", "Voyager" and "Curiosity". Currently, the company decided to shift its area of expertise and provides software that can be applied in such areas as: finance, energy and healthcare industries. More specifically, BP® is involved in strong partnership with Beyond Limits®. Machine learning cognitive approach has been applied to improve the drilling process in oil & gas industry by collecting and analyzing such data as: seismic activity, vibrations, temperature gradients, pressure values etc. The sub-company Beyond Life Sciences® is currently involved in the healthcare industry for gathering and processing information about

patients to help doctors design the best possible treatment procedure (Bindi 2017, Law 2017, E. M. Yang 2017).

Headai® – is a Finnish company founded in 2015 that combines semantic neurocomputing and machine learning for developing of self-teaching software bots, that can potentially improve intellectual labor by deep data analysis. The company's software can work with unstructured data from various sources and organize data according to mutual interconnections recognized within these sources. Headai® "robots" operate as a separate cloud services helping in data mining, thereby minimizing repetitive working processes (Headai Ltd 2017).

IBM Watson® – perhaps one of the best examples of contemporary artificial intelligence product. Initially, the system was intended as a computer-based competitor to real human players on the American tele-show "Jeopardy", that eventually won 2 human champions in 2011. This victory was achieved because of novel IBM Watson's processor – TrueNorth, which is based on the neuromorphic chips, that consist of millions of so-called "computer neurons", that work in parallel with interneuron connections, making processing operations much faster (N. Pathak 2017). IBW Watson® cognitive technology has been specifically designed to understand technical, industry related unstructured data and apply advanced reasoning, predictive modeling and machine learning algorithms to this content in order to facilitate scientific investigations. Furthermore, Watson technology has been developed with ability to comprehend specific scientific terminology, so it is capable of making reasonable connections among millions of pages of scientific text. Current version of the platform involves more than a hundred of various natural language processing algorithms (Shariyar Murtaza et al. 2016). Since, one of the major areas of Watson technology applications is to facilitate life sciences research, current version of this application contains some typical information regarding data related to pharmaceutical and chemical industries, patents, relevant literature and terms. The Watson artificial intelligence platform has found extensive applicability in big data processing for scientific research. More specifically, it was vastly applied in life science research area for studies in drug identification and repurposing, oncology treatment, heart failure early detection and many other cases (Guidi et al. 2016). Most of the time, IBM Watson proved to be exceedingly useful by pulling all relevant available data sources together and extracting meaningful connections, thus unlocking novel evidence driven insights (Y. Chen et al. 2016, Contractor, Telang 2017). Apart from healthcare industry research IBM has been applied in such areas as banking, legal services,

tourism etc. Additionally, IBM Watson's analytic platform supports cognitive text analysis that enables extraction of information by submitting a typed in natural language question.

2.3. Technology behind IBM Watson

2.3.1. Basic principles description

In terms of cognitive computing, the data observation algorithm is the cornerstone of the system. More specifically, this term refers to various data processing techniques, which involve aggregation, integration and examination of the information. In order to make data observation efficient, the cognitive system requires access to large volumes of reliable and relevant information, therefore, data collection and normalization also play an important role in the whole process. The obtained at this step results provide the foundation for further analysis and evaluation.

The IBM Watson data collection technique is based on the single information repository, which is called "The Watson Corpus". For each domain of application, such as for instance law, medicine, engineering or finance, there is a separate unique corpus developed, so that the created information basis contains only relevant to the applied domain information. The collected unstructured and structured information is normalized and refined into the formatted dataset suitable for further analysis.

Further step involves data interpretation, that entails understanding information hidden behind language syntax, grammar, individual terms, and deduction of meaningful information along with focusing on the most relevant lines and paragraphs. Because of such structure, system that has been trained to study chemical engineering is not only able to recognize certain chemical compounds, but also can understand their chemical structures and resolve various synonyms for trivial names and gross formulas due to accumulated information in the corpus. This way, investigation into a specific chemical compound will find all relevant data sources, that contain any information about its structure, trivial name, empirical name, etc. Additionally, likewise, processing of visualized information in form of graphs, drawings, magnetic resonance images, dependency diagrams and other data formats can be implemented to facilitate further information analysis. For adequate natural language processing, system has to be also supplied not only with definitions of main terms and objects, but also with relevant verbs and prepositions in order to understand the relationships between them.

Due to machine learning basis, the cognitive system learns from its preceding iterations and experiences, so that previously studied vocabulary can be applied for interpretation of new

encountered terms based on contextual and syntax clues. In a more detailed way, the cognitive system can for example approximately identify properties of a new developed drug by analyzing context in the information about its side effects and pharmacological effect.

Once, relevant information was collected and normalized, appropriate datasets were gathered into one single corpus and information recognition patterns set up with suitable dictionaries, a set of annotators should be created and applied to the data. The annotators are basically specialized types of computer codes designed to search and extract terms from scientific and technical literature. This way Watson is able to recognize certain nouns and verbs and analyze relations between them, so the object and location of the event can be accurately determined. Development of annotators is a complex process, that is also based upon machine learning and deep natural language processing algorithms which are designed to search for patterns in the analyzed text and learn from them in order to extrapolate relevant information for a certain entity type. The IBM Watson's deep natural language recognition technique is also able to understand different prepositions and articles. For instance, such article as "on" will be recognized as an allocation sign, which will trigger the program code to extract the information about object location. The interface of IBM Watson platform provides an opportunity to visualize annotated datasets for better understanding of the data processing. Moreover, based on these datasets IBM is capable of generating hypothesis about the relationships between the entities by means of predictive analysis, thus, in a way, extracting novel inferences.

After the observation and interpretation of data, the following step is evaluation. Within interface of IBM Watson, evaluation is based upon the exploration purpose. More specifically, the platform develops a holistic visual network depicting all the relevant items and objects and their relationships with each other based on the shreds of evidence found in the processed datasets. In such a manner, at this step novel relationships and dependencies between entities are discovered to develop new hypothesis for further processing and evaluation. Furthermore, the platform is able to conduct a quantitative predictive analysis based on the predefined set of items and objects with the purpose of novel relationships inference, for which there might not be enough evidence. Particularly, the platform utilizes these sets of objects to teach itself to identify similar objects with same text traits and features in other data sources, ranking its findings based on relevance.

As it was mentioned before, human supervisor interaction with Watson can be performed in the natural language manner. Overall logic diagram of IBM Watson architecture for natural language questions processing is represented in the Figure 1.1. First step of question analysis

involves separation of the interrogative sentence into a set of keywords so that relevant answer type is detected by means of grammar parser and semantic analysis. At the second step the initial hypothesis is generated based on the predefined keywords and available data sources. Consequently, the proposed set of hypothesis options is estimated by means of scoring and ranking in reliance on their relevancy. Final step involves merging and ranking of the proposed answers by means of logistic regression empowered machine learning algorithms, which allocate certain probability estimations ("confidence") to each answer (Shariyar Murtaza et al. 2016).

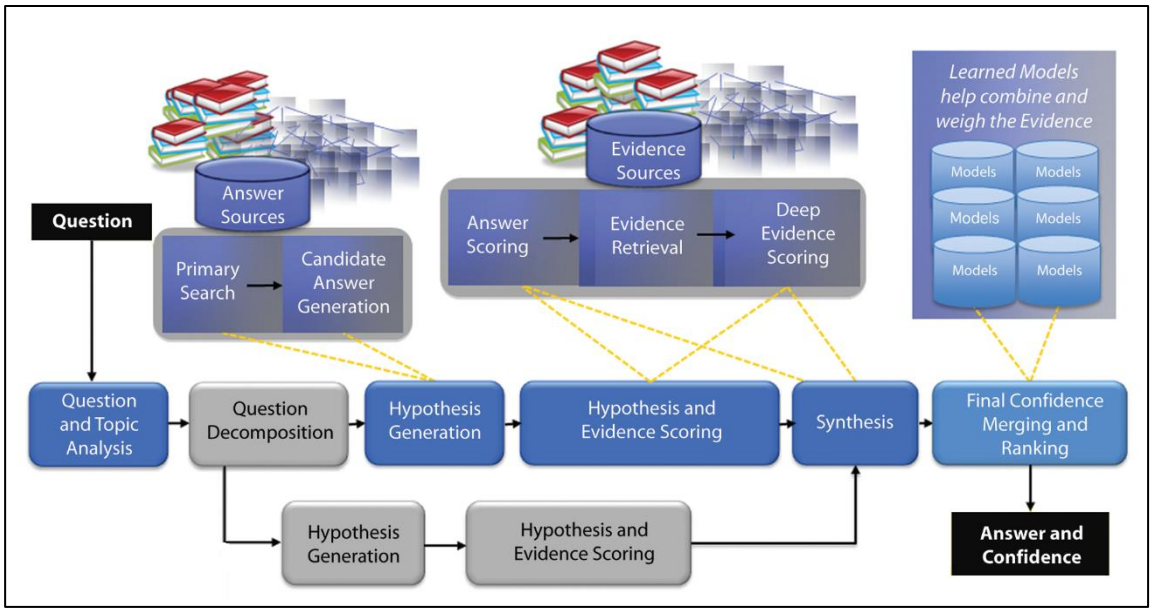


Figure 1.1. IBM Watson natural language processing architecture (Ferrucci 2012)

2.3.2. IBM Watson Analytics platform

IBM Watson Analytics™ is a cognitive based platform that relies on natural language processing for hypothesis generation and predictive analysis (Nagwanshi, Dubey 2018). This platform is able to automatically analyze datasets, estimate data quality, and determine the most suitable statistical approach accordingly. Watson Analytics is divided into four important following sections: refinement, exploration, prediction, and assembly. Refine section enables data initial analysis and manipulation, so that available information can be separated into groups or hierarchies for subgroup analysis. Furthermore, at this step basic calculation operations can be performed. Exploration involves descriptive analysis by means of various data visualization approaches. Additionally, since Watson Analytics supports Watsons's natural language processing algorithms, data can be analyzed by asking questions written in natural language. Initial set of sample questions is automatically generated by the platform once the data was imported. Predictive analysis is performed in reliance on the target attributes selected by a human operator. Assemble feature enables final overall data

combination and visualization, thus creating relevant dashboards, infographics and slide shows. Among the most important statistical tests applied in IBM Watson Analytics™ syntax, the following ones should be mentioned: Analysis of variance (ANOVA), Asymmetry index, Chi-square tests (classification tree and regression tree), Distribution test, Influence test, Model comparison test, Unusually high or low analysis, etc. (Miller 2016, Hoyt et al. 2016).

Among significant limitations of the platform the following aspects should be highlighted:

- For classification purposes with categorical data, Watson Analytics does not provide the user with ability to select the applied statistics algorithms, but rather selects the approach for the user (logistic regression – in most cases);
- Apart from its visualization tools Watson Analytics™ does not provide users with additional assistance in terms of results interpretation, in many cases neglecting the impact of unrelated variables and highly correlated variables. (Hoyt et al. 2016).

Watson Analytics platform has been specifically designed to gain predictive insights from any kind of relatively structured data sources, supporting more detailed analysis with convenient visual interpretation of the obtained results. Moreover, the analytical platform significantly simplifies analysis of new information due to generation of preliminary questions, that can be "asked" about the dataset considering its content, features and properties, thus providing users with relevant starting points for further investigation.

The most significant advantages of Watson Analytics platform can be summarized as follows:

- Computer-aided semi-automatic data discovery with due account for its main contents and properties;
- Guided predictive analysis, assisted with natural language processing algorithms to facilitate and streamline data processing;
- Data exploration enhanced with convenient visualization tools;
- Ability to narrow down the user query, thus directing operator towards more reliable and clear solution highlighting relevant key statistical drivers and hidden interdependencies within the analyzed dataset;
- Substantial reduction of time required for data models' generation due to cloud-based deployment of the platform;
- Data quality analysis and possibility for data refinement through the instrumentality of convenient built-in tools.

3. Metal industry description

In general, technological operations in metal industries can be classified into two basic categories: primary processes and secondary processes. Primary manufacturing operations are mainly related to the extraction of metal from ores and further conversion into metallic materials with controlled amount of impurities. While, secondary processing usually understood as semi-finished or finished parts fabrication from products of primary manufacturing operations (Beddoes, Bibby 2003).

3.1. Iron production and steelmaking

3.1.1. Iron production and steelmaking process description

Iron is the most abundant metal element on our planet – production of steel exceeds approximately 765 million of tons annually. In fact, it is the fourth most abundant element in the earth crust. The most widely used and, therefore, commercially important iron ores are magnetite (Fe_3O_4), haematite (Fe_2O_3), siderite ($FeCO_3$) and awaruite ($FeNi_3$). Additionally, in many cases iron is recovered as a valuable side-stream from other metal production industries (Y. Yang, Raipala & Holappa 2014). Principle schematic representation of iron production and steelmaking technology is provided in the Figure 1.2.

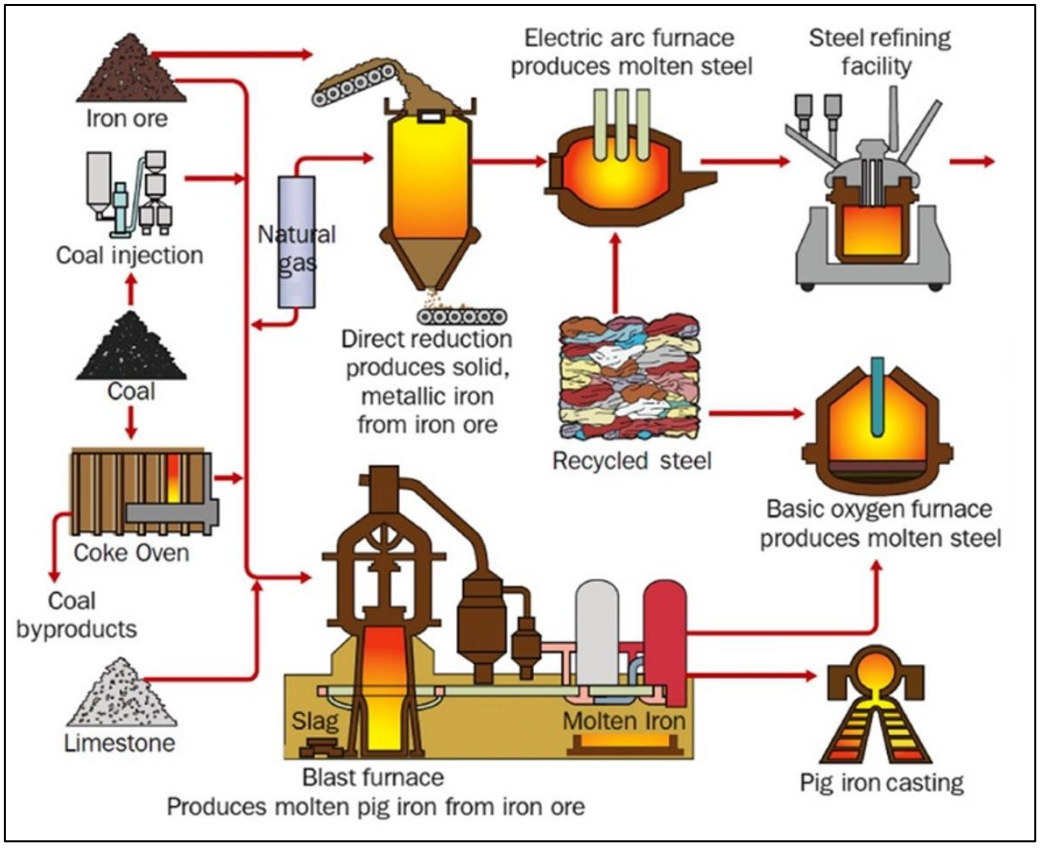


Figure 1.2. Ironmaking and steelmaking principle block diagram (Y. Yang et al. 2014)

In order to prepare iron ores so that they are suitable for smelting and reduction processes they first undergo the process of benefication, which is rather similar to the non-ferrous metal processing industries and typically consists of crushing, screening and concentration either by means of froth flotation or electrostatic/magnetic methods. Agglomeration processes such as sintering, or pelletizing are applied for pretreatment of the extremely fine-grained ore particles to make the materials coarse enough prior to further processing (Y. Yang et al. 2014). Once the beneficiated iron ore is prepared, it is charged into a blast furnace in lump form and as agglomerated sintered fine ore or pellets along with coke and pieces of limestone (as a fluxing agent) for reduction process at 1200-1600^{o}C to produce carbon saturated hot iron. Coke is added for multiple reasons, such as: heat producing fuel, supply of reduction component (mainly CO) and as an agent lowering the metal melting temperature. The obtained molten pig iron in most cases is transferred as liquid hot iron to the converter steelmaking process or can be cast into molds for convenient handling and later used as scrap in electric arc furnaces for melting steel. Converter steel process involves various concepts, such as: basic oxygen process, basic oxygen steelmaking or basic oxygen furnace. All these processes are used to remove adverse impurities such as silicon, carbon and phosphorous as slag or gas phases from the hot iron by means of oxidation with pure oxygen at a temperature of about 2000-2500^{o}C. In the electric steel process graphite electrodes are applied to carry the electrical current of 12 000A at 40V through the furnace roof into the metal charge resulting in temperatures of about 3500^{o}C with supplementary blow of oxygen to accelerate the melting process and enable oxidation process.

Currently, two main commercial alternatives to blast furnace reduction for industrial ironmaking are available: direct reduction process, that involves solid iron production from iron ores and reducing agents (mostly natural gas) in one step and smelting reduction – combination of iron ore reduction (without the use of coke) with smelting process. However, the latter one is still in development and was not extensively applied on industrial scale, while the former one has been widely commercialized for small scale production (Y. Yang et al. 2014). In the direct reduction process, most of the reactions are either solid-solid or solid-gas and proceed at much lower temperature compared to the blast furnace process, which means that without molten slag separation most of the impurities are retained in the final solid product. Natural gas first should be converted into carbon monoxide and hydrogen, which is typically achieved by reforming with carbon dioxide and steam with the presence of nickel-based catalyst at the temperature range of 750-1050^{o}C. Obtained by

means of this process direct reduction iron is widely used in electric arc furnace steelmaking process (Battle et al. 2014).

3.1.2. Iron production and steelmaking side streams identification

Principal side streams of the most common commercialized iron production and steelmaking processes are provided in the Table 1.1.

Table 1.1. Product streams in the iron production and steelmaking manufacturing steps

Processing step	Inlet streams	Side streams
Blast furnace reduction	Iron ore, coke, limestone, hot air (500-900^{o}C, 0.3 MPa)	Slag, Blast furnace top gas, Flue dust
Direct reduction process	Iron ore (fines, lumps, pellets), natural gas or syngas	Exit gases, Flue gases
Converter steelmaking process	Molten hot iron, lime, oxygen (up to 1.5 MPa)	Slag, Waste gases, Flue dust
Electric steel process	Pig iron, scrap materials, direct reduction iron, lime, coal, ore	Slag, Waste gases, Flue dust

Slag as a main side steam of iron blast furnace reduction is a liquid material which floats on the molten metal surface with temperature of about 1550^{o}C. It is produced as a result of addition of limestone, which is combined with impurities contained in the metal ore, thus forming the slag. In general, slags typically contain oxides of various metals such as: aluminium (15%), magnesium (less than 3%), calcium (44%), silicon (up to 35%), iron (less than 1%) and other minor components (sulfur, phosphorous, oxides of titanium, potassium, sodium). Typically, slag at this stage is formed at volumes of 150-300 kg per ton of hot metal (depending on the inlet iron burden composition) with approximate calorific value of 140 kJ/kg. Since, currently, there is no efficient way to use this heat energy, a lot of research and development efforts are being made to find an effective solution for blast furnace slag sensible heat recovery. Roughly 1% of the slag containing more than 50% of iron is recovered by means of magnetic separation and reused in the blast furnace. After cooling, solidifying and crushing slag can be potentially reused in processes of concrete production or can be utilized as a railroad ballast and material in road construction. Furthermore, a granulated slag can be produced by spraying water through the hot liquid slag, resulting in formation of high porous structure, that can be used for cement production or as a fertilizer. (Beddoes, Bibby 2003, Habashi 1999, Y. Yang et al. 2014).

Slag as a side stream of oxygen blowing process is formed from the added lime and contains oxides of silicon and phosphorous. The primary slag components are calcium silicates ($CaSiO_3$ and Ca_2SiO_4), which are formed at the beginning of the process (Jalkanen, Holappa 2014).

Slags produced from electrical steel process contain mainly oxides of calcium, silicon, magnesium and iron (up to 10-15%), while other impurities are aluminium, manganese and phosphorous. These slags are chemically stable and typically utilized in road or dam construction. In terms of electric arc furnace steelmaking process, control of slag formation remains a very important issue for potential development, because it has a significant direct impact on such aspects as: electric arc stability, product dephosphorization, protection of refractory surfaces etc. Energy content of this stream can be approximately estimated as 185 kJ/kg and average rate of production equals 100 kg per ton of raw steel. It can be either utilized in construction area, with preparatory stabilization, or reused directly in the electric furnace (Madias 2014, Habashi 1999).

Blast furnace top gas typically contains nitrogen (up to 60%), carbon monoxide (about 27%), carbon dioxide (no more than 12%), hydrogen and trace amounts of poisonous cyanogen (0.2-2 g/m^3). Because of high carbon monoxide content, blast furnace top gas is reused as a fuel for preheating of the inlet air blast from the room temperature to about 1250^oC. Additionally, because of its relatively high pressure the blast furnace top gas can be used for electricity generation in a recovery turbine. The gas stream leaves the furnace with outlet temperature of approximately 120-370^oC. Calorific value of the blast furnace gas is approximately in the range of 2850-3560 kJ/m^3 (Habashi 1999). Recently, a top gas recycling furnace concept has been developed for carbon dioxide emission reduction – a mixture of carbon monoxide, carbon dioxide and hydrogen from the top gas stream is sent to the separation plant, to remove CO_2, and concentrate CO and H_2 with further reuse as reducing agents in the iron blast furnace (Y. Yang et al. 2014).

Converter steelmaking flue dust consists mainly of iron oxide particles (up to 70%), gangue (about 15%) and carbonaceous materials. This dust contains approximately 40% of solid fraction and is typically collected in special precipitators or Venturi scrubbers as mud that is then sent to the thickening and filtration processes. The obtained filtrate is agglomerated at a sintering plant and after that reused in the blast furnace. Typical amount of dust formed is approximately 10-30 kg per ton or raw steel produced (Jalkanen, Holappa 2014).

Electric arc steelmaking flue dust is formed typically in the amount of 10 kg per ton of steel produced and consists mainly of oxides of iron, zinc and calcium; other impurities also include lead and cadmium. This dust is typically collected in bag filters and sent to treatment in order to stabilize the metals (zinc, lead, cadmium) so that they are transformed to less soluble state and become physically immobilized. Another option, extensively applied in European countries, is recycling of the dust in the electric furnace itself. Additionally, electric furnace dust can be utilized to enrich low zinc containing ores in Waelz process for

zinc production. Currently, similar concepts are being considered for iron recovery from this dust stream (Madias 2014).

Exit gases of the direct reduction process consist mainly of carbon monoxide and hydrogen mixture – up to 70%, carbon dioxide and water. Approximate temperature of the exit gases is about 400^{o}C. Part of this gas stream is burned to provide energy for the gas reforming process, while the remaining part is mixed with fresh gas (Battle et al. 2014).

Waste gases of the oxygen blowing process typically have temperature of approximately 1400-1600^{o}C and calorific value being equal 7000 kJ/m^3. In the case of electric arc steelmaking process, off-gases account for approximately 250 kWh of energy output per ton of produced raw steel. In general, these gases are purified by means of scrubber or electrostatic filter and utilized in heat steam boilers (Madias 2014).

3.2. Aluminium production

3.2.1. Aluminium production processes description

Among metallic materials aluminium is almost the most produced one, being second in tonnage after steelmaking. The most important ore for aluminium production is bauxite ($Al_2O_3 \cdot H_2O$ and $Al_2O_3 \cdot 3H_2O$), which contains 75% of the hydrated metal and is usually extracted by open-pit mining. Major part of refined bauxite is converted into aluminium by means of two processes: the Bayer process for conversion of bauxite to alumina (Al_2O_3) and the Hall-Heroult process for the pure aluminium production (Kvande 2011). The principle process block diagram of alumina production is provided in the Figure 1.3.

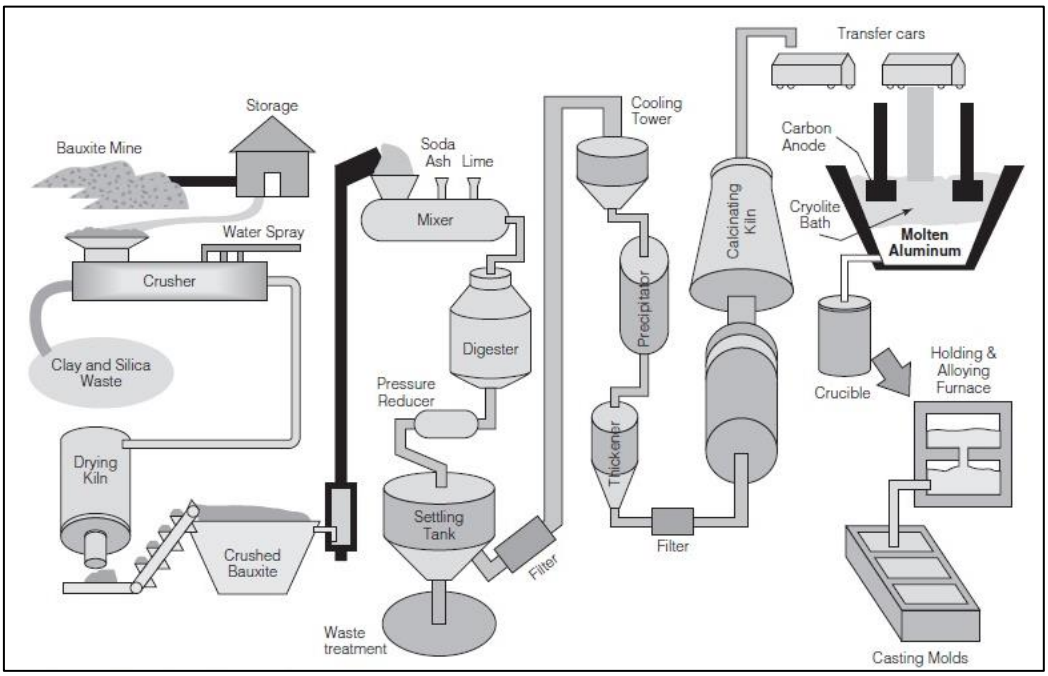


Figure 1.3. Aluminum production principle process block diagram

The Bayer Process currently dominates alumina production – it is a series of chemical reactions implemented on a large scale for digestion of bauxite at high temperature and in the presence of caustic soda to produce aluminium oxide that is pure enough for electrolysis (Tabereaux, Peterson 2014, Metson 2011). First, the ore is ground and pulverized into powder form, which is mixed with process liquor containing caustic soda to form a slurry that is sent to the digestion step in mild steel autoclaves. In digestion process, the slurry is treated by caustic soda solution at 110-270^{o}C under pressure of up to 3.5 MPa depending on the bauxite composition, so that aluminium containing compounds are converted into the sodium aluminate, while silicium, iron and titanium oxides along with other impurities settle out forming the red mud (Totten, MacKenzie 2003, Beddoes, Bibby 2003). The obtained sodium aluminate solution is passed through a number of heat exchangers and blow-off tanks to recover heat and reduce pressure. Afterwards this stream is sent to the clarification processing step where the red mud is separated from the sodium aluminate enriched solution in a settling tank. This settling process is facilitated by addition of flocculants and thickening agents. Once, the red mud is removed, the sodium aluminate is precipitated from the cooled filtered liquor by agitation after addition of solid seed crystals of alumina hydrate in several stages in a series of stirred precipitation tanks at 25-30^{o}C. The obtained coarsely settled agglomerated crystals are filtered, washed and sent to the calcination process in either rotary kiln or fluidized bed calciners at approximately 950-1000^{o}C to remove the remaining chemically bonded water. Heating is typically done by carbonaceous fuel combustion.

Further processing step – the Hall–Heroult process, involves reduction of aluminum by electrolysis in cryolitic bath – molten mixture of cryolite (Na_3AlF_6) and fluoride additives with periodically replaced carbon electrodes in the absence of water, to avoid hydrogen evolution. Anodes are usually covered with crushed bath-alumina particles to reduce significant carbon losses due to excessive oxidation. In general, this process is conducted in special electrolysis cells called potlines, under the influence of high-amperage (350-400 kA at total cell voltage of 4-5 V) direct electric current so that positively charged aluminum ions are electrodeposited on the negatively charged cathode surface where they are converted to liquid metal. A rectangular thermally insulated steel box lined on the outside surface with carbon serves as a cathode (Tabereaux, Peterson 2014, Kvande 2011). Gas generated during electrolysis containing carbon monoxide, carbon dioxide and fluorides is continuously discharged from the cell. Accumulated pure metal is periodically removed by means of tapping or siphoning from the bottom of the cell (Totten, MacKenzie 2003).

Another feasible alternative process for pure aluminum production is the carbothermic reduction of alumina with carbon to produce aluminum carbide, that is further processed at the temperature of 2000^{o}C to acquire pure metal. However, even though this process requires much less energy, it has not yet been applied on the industrial scale and currently is still under research (Tabereaux, Peterson 2014).

3.2.2. Aluminium production side streams identification

Principal side streams of the most common commercialized aluminium production processes are provided in the Table 2.

Table 1.2. Product streams in the aluminium production manufacturing steps

Processing step	Inlet streams	Side streams
Bauxite digestion	Pulverized bauxite powder, Caustic soda	Red mud solid slurry
Alumina precipitation	Alumina slurry, water, alumina hydrate crystals	Mother liquor, waste water bleed
Calcination	Aluminum hydrate	-
Electrolytic reduction	Alumina, cryolite, carbon anodes.	Spent pot lining, waste gases, dross, pot wash wastewater

Spent pot lining is solid waste, generated by the Hall-Heroult process – approximately 2-4 % of aluminium produced. It consists mainly of carbon, sodium aluminium fluoride (6-10%), sodium (up to 19%) and cyanides (about 1%). Heat capacity ranges from 11 up to 18.5 MJ per kg. The waste is considered as hazardous due to high cyanides content – therefore, additional treatment before disposal is necessary. It has been used for cement, mineral wool and asphalt materials production. Carbonaceous content of this lining is well burned at high temperature and can provide significant amount of heat. However, in most cases current disposal technologies include temporary storage with further burial in landfills. Significant research efforts are being made to find opportunity for reuse of this stream. Among potential applications for spent pot lining, reused fluorides extraction and further utilization as fluxing elements are considered (Habashi 1999, Totten, MacKenzie 2003).

Red mud (amount equals approximately half of alumina produced) – is a major side stream consisting mainly of various bauxite impurities and oxides of titanium and iron, desilication products and quartz. Red mud is currently disposed in secured land-fills or under the sea where the alkalinity can be diluted. It can be subjected to the dry slacking process for concentration after which is used to support heavy equipment or light-industrial buildings (Habashi 1999, Beddoes, Bibby 2003, Kvande 2011). Approximate dried red mud

composition can be described as follows: 57% iron oxide, 14-15% aluminum oxide, 7-8% silica, 5-6% titanium oxide, and up to 8% sodium oxide (Totten, MacKenzie 2003).

Mother liquor of alumina precipitation process containing sodium hydroxide and silica is concentrated by evaporation and reused in bauxite leaching processing step (Totten, MacKenzie 2003).

Waste gases – as a side stream of alumina electric smelting consist mostly of carbon monoxide (up to 90%) and carbon dioxide, produced due to anode consumption, sulfur dioxide, gaseous hydrogen fluoride and particulates of fluorine compounds, formed as a result of volatilization of cryolite bath elements and fluorides reactions with air humidity. Total approximate amount of waste gases production of this processing step is estimated as 1.6 kg per kg of pure aluminum produced. This gas is sent to the bag filtration with consequent wet scrubbing process, where fluorine compounds are separated and returned to the electrolytic cell. Typical outlet temperature of these waste gases is approximately 85-120^{o}C (Totten, MacKenzie 2003, Tabereaux, Peterson 2014).

Dross is a side stream of molten aluminum processing – in most cases it consists of free aluminum droplets, aluminum oxides, lead, chromium and cadmium. Two major types of this dross exist: white dross that is acquired as a result of pure aluminum smelting, and black dross containing impurities of salt fluxes and metals oxides.

Pot wash wastewater – a side stream that is produced as a result of waste pot treatment – water is applied to fracture the exhaust bath, metal and lining materials. This stream is considered to be toxic due to high cyanides (up to 200 mg/L) and fluorides (up to 600 mg/L) concentrations, therefore, it cannot be discharged without preliminary treatment. Typically, this water is collected and reused for next spent pots refining (Habashi 1999).

3.3. Copper production

3.3.1. Copper production processes description

Copper containing ores (chalcopyrite $CuFeS_2$, bornite $CusFeS_4$ and chalcocite Cu_2S) are usually extracted either in open pit (0.5% copper) or underground mines (0.5-2% copper). Currently, about 80% of the produced copper is obtained by such sequence as: benefication, followed by smelting and after that – electrolytical refining. The other 20% are produced by means of hydrometallurgical processing (Davenport et al. 2002). Principle process block diagram of copper production is represented in the Figure 1.4.

Beneficiation in copper production is typically achieved by froth flotation of the finely crushed copper containing ore followed by gravity solid-liquid separation. Froth flotation

involves addition of chemicals (which make Cu minerals water repellent; lime for pH control, flocullants) and air blowing through the slurry to acquire copper concentrate. As a result of benefication step, copper concentration is increased to approximately 20-30% with small amount of gangue oxides. Optional processing step such as roasting can be applied to prepare copper sulfide concentrates for further pyrometallurgical treatment in order to decrease the sulfur content and partly oxidize iron at 500-800^{o}C depending on the process purpose (Habashi 1999).

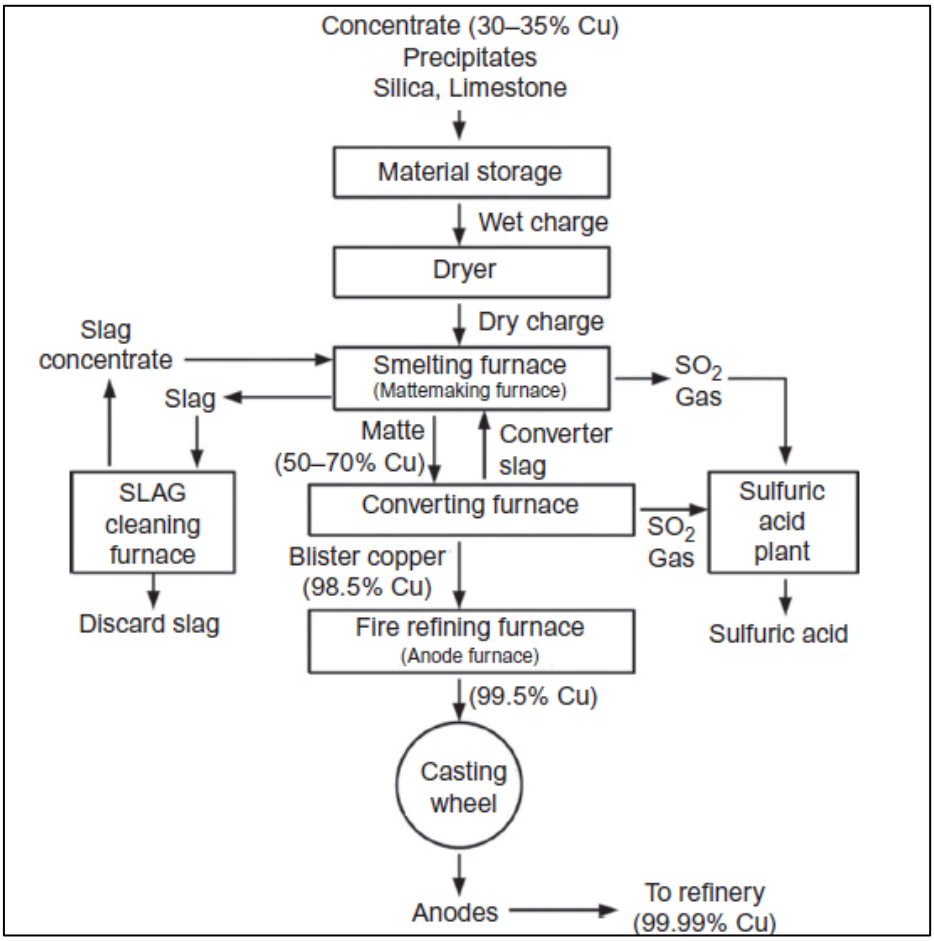


Figure 1.4. Copper production principle process block diagram (Seetharaman 2014).

Further processing step is smelting of acquired copper concentrate that results in production of two immiscible phases: matte – a homogeneous mixture of copper and iron sulfides (45-75% of copper), and smelting slag. Matte smelting can be performed in batch manner (reverberatory and electric furnaces) or continuously (flash smelters and Noranda smelters). Even though the choice of smelting technology depends on the ore concentrate properties, continuous type smelters are being applied in increasing frequency due to the improved overall efficiency and heat recovery (Robert, Leslie & Ingrid 2002). In this process iron and sulfur materials contained in the copper concentrate are oxidized by oxygen enriched air at the temperature of about 1250^{o}C. Fluxing silica agents are added to separate oxidized iron

and other metal impurities as fayalite slag, which is tapped from the furnace (Seetharaman 2014).

Once the molten matte is obtained and separated from the smelting slag, it is charged into a converter, where lime and silica are applied at a temperature of 1200^{o}C to further reduce iron and sulfur content by oxidizing with oxygen-enriched air. Typical process in rotary Pierce-Smith converter is usually performed in batch manner and involves two steps: slag formation stage, when iron sulfide is oxidized and fayalite slag is produced similarly to the smelting process, and copper making, where impure molten copper sulfide, so called "white metal", is further oxidized resulting in blister copper formation (98-98.5% of copper). In case of continuous conversion process (flash converting, downward lance converting or submerged tuyere converting) limestone is applied to produce calcium oxide flux that absorbs iron oxides and forms calcium ferrite slag, that is separated from the molten matte in the electric cleaning furnace at 1250^{o}C (Seetharaman 2014, Robert et al. 2002).

After this, the obtained copper concentrate of "blister copper" is sent to the fire refining step, where air and natural gas are blown through the melt layer in order to remove remaining sulfur and oxygen so that it is suitable for further efficient electrorefining. Typically, this process is performed in rotary refining kilns or hearth furnaces at about 1200^{o}C. During this step optional fluxing agent's addition is possible in order to eliminate impurities such as antimony, arsenic and lead in a form of slag, that is then returned to the smelting furnace (Robert et al. 2002).

The final step is electrolytical treatment in $CuSO_4$-H_2SO_4-H_2O electrolyte in polymer concrete cells. Blister copper is cast into impure copper anodes (4-5 cm thick, 300-400 kg each) and placed in the electrolyte solution, where under direct electrical current (20 000-30 000 A at the voltage of 0.3-0.4 V) copper dissolves and electrodeposits onto stainless steel cathodes (3 mm thick blanks), from which pure copper is afterwards machine stripped (99.99% copper). Utilized thinned anodes are periodically removed, washed and reused for fresh anodes production. Likewise, copper laden cathodes are also periodically removed and replaced by new ones. In order to provide smooth and equal copper deposition on the cathode surface, special levelling agents such as bone glue and thiourea are added to the electrolyte solution. In general, electrolyte temperature is maintained at about 60-65^{o}C by means of steam heating (Habashi 1999, Davenport et al. 2002).

Among hydrometallurgical processes of copper production, the most important on industrial scale are leaching with sulfuric acid or ammonium solutions, solvent extraction and

electrowinning process. Principle process block diagram of the hydrometallurgical production of copper is provided in the Figure 1.5.

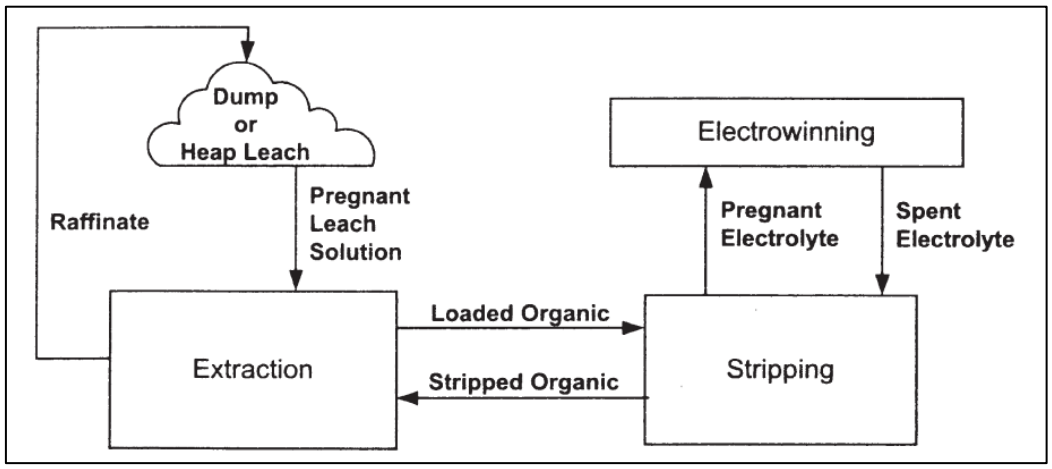


Figure 1.5. Copper hydrometallurgical production principle process block diagram (Kordosky 2002)

Currently, most of the leaching processes are performed in a heap leaching manner applying dilute sulfuric acid through a system of polymer pipes to the top surface of the copper concentrate heap, allowing acid to leak through it under nearly ambient conditions. Additionally, to provide sufficient amount of oxygen, air is supplied by means of perforated polymer pipes at the heap base. As a result, copper dissolves from minerals into an aqueous pregnant solution forming a copper laden liquor which is collected by a sloped surface underneath the leached heap and sent to the further processing by means of solvent extraction (Davenport et al. 2002, Kordosky 2002). In some cases, certain bacteria strains and addition of ferric salts to accelerate the leaching process have been applied. Average copper recovery during this step varies between 30 and 70% (Robert et al. 2002).

Solvent extraction process involves contacting pregnant leach solution with copper-specific liquid organic extractant (aldoximes or ketoximes in petroleum-based distillate) in order to produce relatively pure copper sulfate solution suitable for electrowinning process. Copper forms a complex organic compound with solvent molecule forming a separate dense layer, which is then removed and treated with sulfuric acid to break the complex and dissolve copper sulfate, achieving overall metal extraction of 90% (Robert et al. 2002).

Electrowinning process is substantially different from electrolytical refining in two aspects: insoluble conductive anodes are utilized instead of soluble ones and overall voltage of applied electrical current is much higher as well – 2.4-3 V. Anodes are inert rolled lead-based alloy sheets containing also tin (1.5%) and calcium (0.1%) to improve corrosion resistance and strength, while cathodes are made from stainless steel. Pure metallic copper

is electrodeposited on cathodes producing oxygen at the anode and sulfuric acid in the solution (Davenport et al. 2002, Robert et al. 2002).

3.3.2. Copper production side streams identification

Principal side streams of the most common commercialized copper production processes are provided in the Table 1.3.

Table 1.3. Product streams in the copper production manufacturing steps

Processing step	Inlet streams	Side streams
Benefication (froth flotation)	Copper containing ore, water, chemicals, thickeners	Wastewaters, flotation tailings
Matte smelting	Copper concentrate (20-30% Cu), silica flux	Smelting slag, sulfur dioxide bearing off-gas
Conversion	Matte, lime, silica, limestone	Converter slag, sulfur dioxide bearing off-gas
Electrolytic refining	Copper concentrate (98-99% Cu), sulfuric acid	Anode slime
Copper leaching	Copper concentrate (20-30% Cu), sulfuric acid/ammoniacal solutions	Leach waste
Solvent extraction	Pregnant leach solution	Raffinate
Electrowinning	High-copper concentrated electrolyte, sulfuric acid	Oxygen, sulfuric acid

Flotation tailings usually account for approximately 98% of the inlet ore amount. Flotation tailings are stored in large dams from which water is removed and treated. Solids usually consist of common rock materials (in most cases – silicates), heavy metals (up to 15%), pyrites in form of sulfate gangue (Robert et al. 2002).

Smelting slag from matte smelting process step is highly viscous liquid with density of 3.1-3.6 g/cm^3, that usually contains up to 40-50% of iron oxides and approximately the same amount of silica (typically 30-40%). Slags contain 1-2% of copper, usually in forms of dissolved copper sulfide, slagged copper oxide and silicate. Among other compounds alumina, quicklime and magnesia are also commonly found. Slags are usually disposed as waste or sold as products with similar to basalt properties. After cooling (initial temperature 1200^{o}C) this product can be used for river-bank protection or as a railway ballast. Currently two methods are applied for slag cleaning: froth flotation (concentrated slag is returned to the feed and tailings are discarded) and treatment in special electric arc slag furnaces (involves slag reduction with production of additional amount of matte). In some cases, this slag can be used as a trace element in fertilizers due to its copper and non-ferrous metals content (Habashi 1999, Davenport et al. 2002).

Converter slag has similar to smelting slag properties but contains 4-8% of copper – mostly dissolved copper in form of copper oxide or silicate. Approximate batch converter slag composition can be described as follows: 35-50% of total iron (20-25% in form of magnetite), 15-30% of silica, alumina, zinc oxide and calcium oxide each less than 5%, about 1% of magnesium oxide. In general, continuous matte conversion slag contains more copper (about 14-20% on average) and calcium oxide (15-20%). Due to high copper contents these slags are usually reused in the conversion process. Outlet temperature of the conversion process slag is in the range of 1150-1200^{o}C (Davenport et al. 2002).

Anode slime represents less than 1% of anode weight. The slime stream typically contains: gold < 4%, silver < 25%, lead sulfate 5-10%, arsenic 0.5-5%, antimony 0.5-5%, bismuth 0.5-2%, copper 20-50%, selenium 5-20%. Additionally, gypsum, silica and alumina are present due to anode casting process. Anode slimes are generally treated by leaching of copper with sulfuric acid, precious metal separation by means of electrolysis and hydrometallurgical recovery of selenium (Habashi 1999).

Leach waste – heap of total undissolved by leaching ore remnants containing acidic effluents and toxic metals.

Raffinate – copper depleted solution containing approximately 11-12 kg/m^3 of sulfuric acid and 0.3-0.7 kg/m^3 copper, that is sent back to the leaching process to pick more copper.

Sulfuric acid acquired by electrowinning process is recirculated to the solvent extraction process.

Sulfur dioxide bearing off-gas from matte smelting process on average contains 10-60% SO_2 and is usually combined with sulfur dioxide bearing gas from conversion process and captured for sulfuric acid production. Smelter off-gases also usually contain substantial amount of dust (up to 0.3 kg/m^3), nitrogen, small amount of carbon dioxide and water vapor. Dust usually consists of unreacted ore concentrates, particles of fluxing agents and droplets of mate or slag, along with trace amounts of arsenic, antimony, bismuth and lead. Approximate outlet temperature – 1200^{o}C. In terms of batch matte conversion process, sulfur dioxide content is typically less than 18%, while for the continuous process its concentration is approximately 25-40% (Davenport et al. 2002).

3.4. Zinc production

3.4.1. Zinc production processes description

Most of zinc is produced from sulfide materials (Sphalerite ZnS – is the most important zinc containing source) principally by either high-temperature processing (10% of the world zinc production) or by hydrometallurgical processing (86% of the world zinc production). Before

processing zinc ores are always crushed, ground and concentrated by means of froth flotation. Principle pyrometallurgical and hydrometallurgical zinc production flowsheets are provided in the Figures 1.6(a) and 1.6(b) respectively.

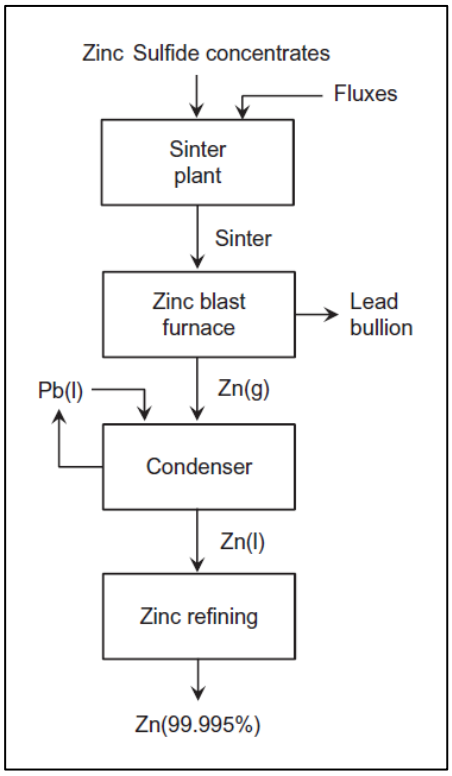


Figure 1.6. (a). Principle flowsheet of pyrometallurgical zinc production (Sohn, Olivas-Martinez 2014)

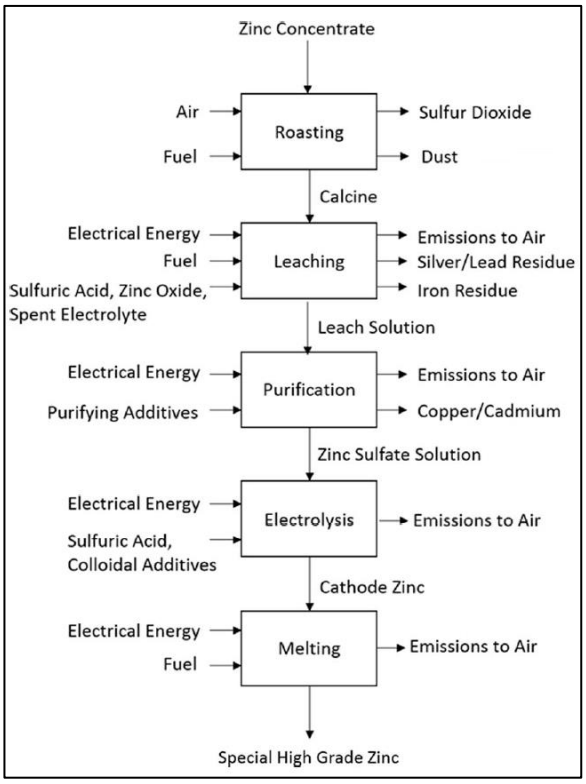


Figure 1.6. (b). Principle flowsheet of hydrometallurgical zinc production (Genderen et al. 2016)

In terms of high-temperature zinc processing, the first step is roasting of ore concentrates to prepare metal for further extraction, which involves removal of excessive sulfur content and conversion of zinc-containing sulfidic materials to the oxide form – calcine. Currently, most of the zinc ore roasting processes are carried out either in a fluidized bed reactor at 900-950^{o}C or at a sintering plant at about 1450^{o}C. Afterwards, the obtained sinter product containing zinc and lead sulfides and oxides is conveyed to a zinc shaft furnace where it is reduced by carbon monoxide at a temperature of approximately 900-1150^{o}C. Effective heat energy utilization improvement is achieved by countercurrent operation manner in the furnace, so that the heat produced by exothermic reaction of coke combustion is applied to heat the downflow of charged material (Habashi 1999, Sohn, Olivas-Martinez 2014).

Metallic zinc leaving the furnace in off-gas (5-7% of zinc) is charged into condensing unit, where it is rapidly quenched and condensed by contacting with sprayed molten lead in a special spray condenser at the temperature of 550^{o}C. Further separation of zinc from molten lead is based on temperature dependency of zinc solubility in lead and typically involves cooling down the formed metals mixture to the temperature of about 440^{o}C with subsequent

separation of crude zinc by liquation (Vignes 2011, Bernasowski, Klimczyk & Stachura 2017).

The last step is refining of acquired crude zinc by two-stage fractional distillation in parallel working columns with silicon carbide trays in two stages. First, the crude zinc stream is feed at temperature of about 580^{o}C at the column where less volatile elements (mostly iron and copper) are removed along with part of zinc at the bottom section, while the vapor from the top column section is charged in the second column for cadmium separation. In the second column the bottom product represents high purity zinc, while the upper product is cadmium enriched vapor stream, that is sent for further cadmium extraction. Bottom product of the first column is charged in a liquation furnace, where hard zinc and impure lead are produced (Sohn, Olivas-Martinez 2014, T. Chen 2012, Vignes 2011).

Regarding hydrometallurgical production of zinc, processing usually involves such steps as roasting (calcining), leaching, purification and electrowinning. For hydrometallurgical processing zinc containing materials still need to be converted to oxides by the same roasting process as described above (Habashi 1999). Next step is leaching, where produced calcine containing on average 90% of zinc oxide is dissolved into aqueous solution by sulfuric acid to maximize zinc recovery and minimize iron impurities in the leach solution. In most cases to improve zinc recovery and to decrease leach solution contamination with iron, the process is carried out at the temperature of approximately 70-80^{o}C in several stages in multiple tanks with counter current movement of liquids and solids between stages (Wood et al. 2015). In order to remove potentially harmful impurities for electrowinning process, leached zinc acid solution is purified by means of cementation process, that involves addition of zinc dust in the presence of alumina for copper and cadmium removal and addition of antimony/arsenic oxides for nickel and cobalt removal. The impurities are separated from zinc solution as a solid residue by means of filtering, so that purified solution contains less than 1 mg per liter of adverse metallic foreign substances. The purification step is of crucial importance since most of the metallic impurities have lower hydrogen overpotential than zinc and, therefore, will promote hydrogen evolution, thus deteriorating the electrowinning process. (Free, Moats 2014). Currently, processes for removal of impurities by means of fluidized bed electrolysis or solid-liquid ion exchange are being developed (Habashi 1999).

The product from purification process is a solution suitable for effective electrowinning called neutral feed, from which zinc is recovered as a solid metal cathode. Neutral feed is mixed with spent cathode and introduced into concrete electrowinning cells, lined with polyethylene, where direct electric current (at 3.4-3.5 V) is applied and as a result oxygen is

produced on the anode (lead-silver alloy with silver content of 0.5-1%) and zinc metal on the cathode (aluminium blanks). Because of the low standard reduction potential of the metal in order to ensure efficient zinc deposition and hinder hydrogen evolution at the cathode high current densities of approximately 400-600 A per m^2 are typically applied. Zinc laden cathode plates are periodically removed every 48-72 hours, so that the electrodeposited high-purity zinc is then mechanically stripped off and casted into special high-grade zinc ingots. To decrease anode lead corrosion trace amounts of manganese are added to the electrolyte solution to form protective manganese oxide layer on the anode surface (Free, Moats 2014, Genderen et al. 2016).

Novel zinc direct smelting process involves pyrometallurgical treatment of zinc concentrates at about 1250-1300^{o}C in a submerged lance furnace, producing volatilized zinc-oxide fume, that is already suitable for leaching and electrolysis. Thus, roasting processing step can be avoided. After precipitation, the acquired zinc oxide stream has approximate concentration of zinc up to 63% and low content of gangue materials (Wood et al. 2015).

3.4.2. Zinc production side streams identification

Principal side streams of the most common commercialized zinc production processes are provided in the Table 1.4.

Table 1.4. Product streams in the zinc production manufacturing steps

Processing step	Inlet streams	Side streams
Ore concentrates roasting	Zinc sulfide concentrate	Sulfur dioxide bearing off-gas
Zinc blast furnace	Sinter, carbon monoxide, metallurgical coke	Lead bullion, slag
Zinc refining (distillation)	Zinc bearing off-gas	Cadmium containing dust, hard zinc, impure lead
Leaching	Calcine (90% of ZnO and 10% of zinc ferrites), sulfuric acid, lime, spent electrolyte	Leach residue, sulfuric acid containing wastewaters
Purification	Leached zinc-acid solution, zinc dust, antimony or arsenic oxide, copper sulfate	Copper cake, cadmium cake
Electrowinning	Natural feed, strontium or barium carbonate	Electrolytic sludge

Hard zinc – contains up to 20% of lead and 2% of iron on average and is usually sent to recycling in the process.

Impure lead – contains approximately 2% of zinc and can be sold to the lead production facilities, where it is typically mixed with crude lead from smelter and further refined.

Leach residue contains iron phases – up to 30% (mostly in forms of jarosite – 40% and goethite – 35%), lead 5-10%, silver 170-1800 ppm, sulfur – up to 12%, silica 5-10%, arsenic, antimony and trace amounts of many other heavy metals. Because of relatively high lead and cadmium content, these residues are often classified as physically unstable and hazardous materials, therefore additional treatment is required. Typical waste management approach involves stabilization by addition of Portland cement and storing on-site in tailings impoundments (T. Chen 2012, Creedy et al. 2013, Wood et al. 2015). However, valuable metals can be recovered from leaching residues by means of combination of froth flotation and pyrometallurgical (such as Outotec's top submerged lance technology with reductant coal) or hydrometallurgical processing (sulfating roasting and leaching) (Habashi 1999, Creedy et al. 2013).

Copper cake – is usually sold to copper smelting facilities.

Cadmium cake – contains more than 80% of cadmium that can be sold.

Electrolytic sludge is formed at the bottom part of the electrowinning cells by flaking of magnesium and lead oxides. It is usually returned to the leaching processing step for recycling, where magnesium oxide is utilized as oxidant (Free, Moats 2014).

Sulfur dioxide bearing off-gas from the roasting processing step typically contains up to 10% of sulfur dioxide (Habashi 1999).

Slag from the sintered zinc reduction processing step usually consists of non-reduced oxides and is separated by means of settling. Typical composition of this stream is as follows: zinc – up to 8%, iron oxide – about 32-34%, silica – up to15-23%, calcium oxide – 17-18%, alumina – approximately 9%, sulfur – 7-8%, trace amounts of lead (1.2-1.4%) and copper (less than 0.2%) (Vignes 2011, Bernasowski et al. 2017).

3.5. Lead Production

3.5.1. Lead production processes description

Currently, primary lead production by refining of lead sulfide containing ores by means of such processes as sintering, blast furnace smelting, and pyrometallurgical refining contributes to approximately 70% of the produced metal. Principle process block diagram of lead production is represented in the Figure 1.7.

Once, lead ores are crushed, ground and concentrated by froth flotation, the obtained lead concentrate is smelted in the sintering plant, where it is subjected to hot air stream in order to burn off the sulfur at about 1200^{o}C, resulting in sulfur dioxide and lead oxide formation. Another objective of a sintering plant is to produce porous and durable material by means of

partial melting so that it is suitable for further handling and processing in a blast furnace. Therefore, typical lead concentrate sintering process involves two steps: sulfur removal and blast furnace reduction. Currently updraft sintering process is applied, where required amount of air is supplied from the below through the ignited layer of sinter material, that consists mostly of lead concentrate, lime, silica, flux and recycled sintering dust (Habashi 1999, Sohn, Olivas-Martinez 2014). As a result, sinter product containing on average 45-52% of lead and 1-1.7% of sulfur, depending on initial concentrate composition, is produced. Output sinter gases typically consist of dusts, fumes, sulfur dioxide and volatile metallic elements and have temperature of about 200-500ºC (Sinclair 2009).

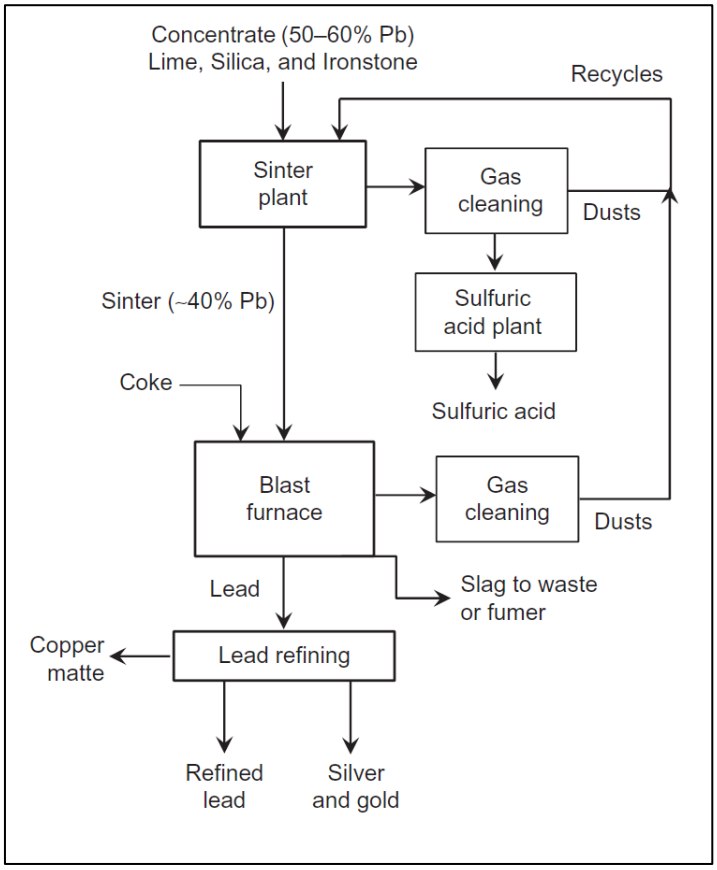


Figure 1.7. Lead production principle process block diagram (Sohn, Olivas-Martinez 2014)

The acquired sinter product is then charged into a blast furnace along with metallurgical coke and fluxing elements (for molten slag formation) for lead oxide reduction. Lead reduction blast furnace represents a counter-current vertical shaft reactor, where downflow of charged materials is heated and reduced by contacting with ascending carbon monoxide containing hot gas. Temperature varies in the range from 200ºC at the top heating zone up to 1500ºC in the lower tuyere zone, where coke is combusted in the presence oxygen enriched air. As a result, two molten layers are formed in the furnace hearth: lead bullion and slag. If the charged material has high sulfur content – a separate sulfidic matte phase is

formed. In case of high arsenic and antimony content in the burden material, an additional side stream – speiss is formed as a solid layer on lead surface (Habashi 1999, Sohn, Olivas-Martinez 2014).

Direct smelting reduction process has been developed to oxidize and then reduce lead sulfide concentrates in a single step at sufficiently high temperature of above 1200^{o}C, thus improving process efficiency by effectively utilizing heat produced via oxidation reactions. However, due to overall complexity, large amount of mechanical equipment units, high capital requirements, environmental issues and occupational health related problems, this concept is considered as not economically viable and impractical in many cases (Sinclair 2009).

Further refining of the obtained lead bullion is typically achieved by pyrometallurgical methods (90% of overall lead production), where various impurities are separated in a series of processing steps, that depends on the present foreign materials. First step is copper removal, which is typically implemented as a two-stage batch process: initial phase involves lead bullion cooling with precipitation of lead rich copper dross that floats to the surface and is then removed by skimming. Afterwards, fine decoppering is accomplished by addition of elemental sulfur to the lead bullion at 320^{o}C to sulfidize the remaining copper. The obtained dross from the fine decoppering process is recirculated back to the smelting step, because of low copper content and high sulfur amount (Habashi 1999).

Subsequent pyrometallurgical refining is softening and it usually involves oxidation of antimony, arsenic and tin by means of air blowing or by molten sodium addition. Softening with oxygen enriched air blast is usually implemented in kettles or reverberatory furnaces at 700-800^{o}C to form lead enriched slag containing oxides of arsenic, tin and antimony, which is continuously tapped out. The other process, which is commonly referred to as Harris process, includes lead bullion treatment with sodium hydroxide or nitrate for the impurities oxidation in an agitated vessel at temperatures above 400^{o}C. Contained impurities form slag, that is removed, granulated and then subjected to further treatment (Sinclair 2009).

Afterwards, precious metal separation is achieved by addition of zinc to the cooled lead bullion. The process is conducted in batch manner in two stages: first step involves addition of zinc crust at the temperature of around 480^{o}C so that solid crust containing silver is formed on the molten lead surface, while in the second step the remaining silver content is further reduced by addition of metallic zinc to the cooled lead bullion at the temperature slightly above its freezing point. Remaining zinc content (0.55-06%) in the lead bullion is commonly removed by means of batch vacuum distillation. This process is typically conducted in a

special agitated kettle under the pressure of 0.05-0.13 bar and operating temperature of approximately 600°C.

Once the molten lead is purified from the remaining zinc, it is subjected to the debismuthizing process, where alkali earth metals (mostly magnesium and calcium) are added to the lead bullion at temperature of about 420°C so that containing bismuth is precipitated in a form of bismuthides and encrusted on the surface. The formed crust is enriched in bismuth and removed by skimming at the end of a batch.

About 10% of the processed lead bullion is carried out by means of electrolytic refining by dissolving lead from impure anodes. Electrolytic refining is possible only after copper removal: lead bullion casted anodes and thin high-purity lead cathodes are used in sulphamate or fluorosilicate acid electrolyte for pure lead electrodeposition on cathode under the influence of electric current with density of 120-150 A/m^2 typically at 0.45-0.6 V. During this process containing impurities form a slime layer (around 1.5-3% of the initial anode weight) (Thornton, Rautiu & Brush 2001, Habashi 1999, Sinclair 2009).

3.5.2. Lead production side streams identification

Principal side streams of the most common commercialized zinc production processes are provided in the Table 1.5.

Table 1.5. Product streams in the lead production manufacturing steps

Processing step		Inlet streams	Side streams
Sintering plant		Lead concentrate	Sulfur dioxide bearing off-gas (4-8% SO_2), Flue dusts
Smelting furnace		Sinter product (45-52% of lead), metallurgical coke, flux	Smelting slag, lead containing particulate materials, matte, exhaust gas
Lead bullion refining	Lead drossing	Lead bullion (98% lead)	Copper containing dross
	Softening	Drossed lead bullion	Softening slag and skimmings or Harris salt slags
	Desilvering	Lead billion	Silver-zinc crust
	Bismuth drossing	Desilverized lead bullion	Bismuth containing dross

Smelting slag can be either acidic (high viscose) or basic (lower viscosity, high corrosivity) depending on the ratio of calcium oxide to silica. Density 3.5-3.8 g/cm^3. Typical weight composition: 25-36% wustite, 2-5% magnetite, 19-25% silica, 12-20% calcium oxide, 1-11% alumina, 1-3% manganese oxide, up to 22% of zinc oxide and 1-4% of lead oxide. Can be reused in production for removal of elements, which oxidize more easily than molten lead (Thornton et al. 2001). This side stream often contains sufficient amount of zinc oxide

to make profitable the recovery by means of reduction in rotary Waelz kilns with other zinc containing intermediates (Habashi 1999). Additionally, processing in special slag fuming furnaces is widely used for high zinc level (16% or more) slags treatment at about 1250^{o}C. Zinc is evaporated and collected by means of bag filters in zinc oxide forms. Thus, zinc content in the smelting slags is reduced to about 1-2% (Sinclair 2009).

Matte – is a sulfidic product of blast smelting process that consists mainly of copper 20-35%, lead 10-25%, iron 20-35%, 5-9% zinc, and 18-25% sulfur. Metals predominantly occur in sulfide forms. Density is in the range of 4.5-5 g/cm^3. In some cases, this side stream can contain silver, gold and other precious metals. In general, produced mattes are sold to smelters for additional copper recovery. Typical treatment involves oxidation with oxygen in special converters with production of blister copper, which can be used for further electrolytic refining.

Flue dust from sintering plant contains cadmium, sulfur (about 10%) and lead (up to 60-70%wt.). Usually, it is collected by wet scrubbing and dry baghouse cleaning. Additionally, cadmium, indium, thallium and selenium tend to concentrate in sintering dusts. Recovered dust is collected and returned for recycling in the sintering mixture (Thornton et al. 2001, Habashi 1999).

Exhaust gas of the blast smelting furnace on average consists mostly of carbon monoxide and carbon dioxide with the ratio varying from 0.4:1 to 1:1 depending on the net heat in the furnace. Approximate sulfur dioxide content is typically less than 0.3% (Sinclair 2009).

Copper containing dross is a complex blend consisting mostly of metallic and sulfidic lead, that is usually melted in the retreatment furnace producing matte phase (lead and copper sulfides), that is suitable for sale to copper smelters, and slag, which mainly contains metals oxides. In case of large arsenic content – formation of speiss occurs, where nickel, arsenic and copper are concentrated. Fumes from the dross retreatment furnace are enriched in such metals as: indium, tin, zinc and arsenic. Additionally, copper dross can be treated hydrometallurgically by oxidation leaching in sulfuric acid-sodium chloride solution followed by solvent extraction and electrowinning process with lead anodes for deposition of copper on stainless steel cathodes (Habashi 1999). In general, composition of copper drosses vary drastically depending on initial impurities, however, in most cases copper content is about 30-45%, sulfur 15-20% and lead 20-50%. Another approach for copper dross processing is the Glover process, which involves addition of sodium metal to upgrade the dross by increasing copper content along with reducing lead concentration, thus making it more suitable for further treatment (Sinclair 2009).

Softening slag and skimmings are a side stream of air blast softening which consists of variable quantities of arsenic, tin and antimony and is usually reduced to hard antimonial lead in a rotary or blast furnace together with soda and char or coal. Side stream of such treatment is arsenic enriched slag, that is typically leached with caustic soda to produce arsenic oxide suitable for sale.

Harris salt slags is side stream of sodium hydroxide/nitrate softening which can be processed to recover caustic soda and produce minor metals in forms of: black antimonate, sodium antimonate, calcium stannate and calcium arsenate. In most cases, Harris slags processing involves leaching with water to remove remaining excessive sodium, separation in cyclone to remove metallic particles of lead and further extraction of tin and arsenic, usually along with selenium and indium.

Silver-zinc crust on average contains up to 65% of entrained lead, 20% of zinc and 10% of silver. It can be refined to produce malleable lead and recover zinc. Refining is done by liquation at a crucible or induction furnace at 650^{o}C where crust is melted into two layers: lead-rich lower, which is charged back to desilvering process and zinc-rich upper which is sent to vacuum distillation, where zinc is recovered in metallic form at a temperature of 600^{o}C. Residue of distillation process contains mostly silver, gold, copper, lead and zinc.

Bismuth containing dross apart from lead and bismuth (3-10%) contains also magnesium and calcium (up to 2% for both), which should be removed prior to further processing. Afterwards, lead can be removed by chlorination with chlorine gas at 700-800^{o}C. Additionally, this dross stream can be refined by the electrolytic method, which involves deposition of lead in pure form at a cathode, while anode is subjected to reduction process where alloy with high (>90%) bismuth content is produced (Habashi 1999).

3.6. Nickel Production

Nickel usually occurs together with cobalt in either sulfide (60% of nickel production) or oxide ores – laterites (40% of nickel production). Laterite ores are typically used for ferronickel production (from saprolite) or nickel matte (from limonite), while sulfide ores are refined to produce high-grade nickel. Principle block diagram of nickel production is provided in the Figure 1.8.

3.6.1. Sulfide ores processing

Sulfide ores refining usually involves concentration, smelting and conversion to metal-rich matte. Concentrate acquired by means of froth flotation of crushed and ground ore, is smelted either by flash or electric furnace smelting at 1300^{o}C and converted into sulfide matte. Electric furnace smelting is achieved by passing a three-phase electric current through

carbon-based electrodes – this process is usually applied when the required temperature is higher. Flash furnace smelting process compared to the electric smelting is not able to attain and accurately control slag temperature. Therefore, although this process requires much less energy and produces sulfur enriched dioxide off-gas, nickel loss due to oxidation is substantial (Moats, Davenport 2014). The smelting process product – matte is then sent to the conversion process to remove iron and iron silicate by oxidation and slag formation caused by oxygen-enriched air blowing through the molten matte with addition of fluxing elements at the temperature of 1200-1300^{o}C (Habashi 1999). Currently, in 95% cases, nickel conversion process is a batch operation performed in the Pierce-Smith converter, that has such serious drawbacks as high sulfur dioxide emissions in the environment and interruption of sulfuric acid production due to charging and skimming procedures. For this reason, extensive research efforts have been made to mitigate these issues and processes of continuous converting and direct flash smelting to the final matte were proposed. However, these concepts have not yet found widespread industrial application (Crundwell et al. 2011).

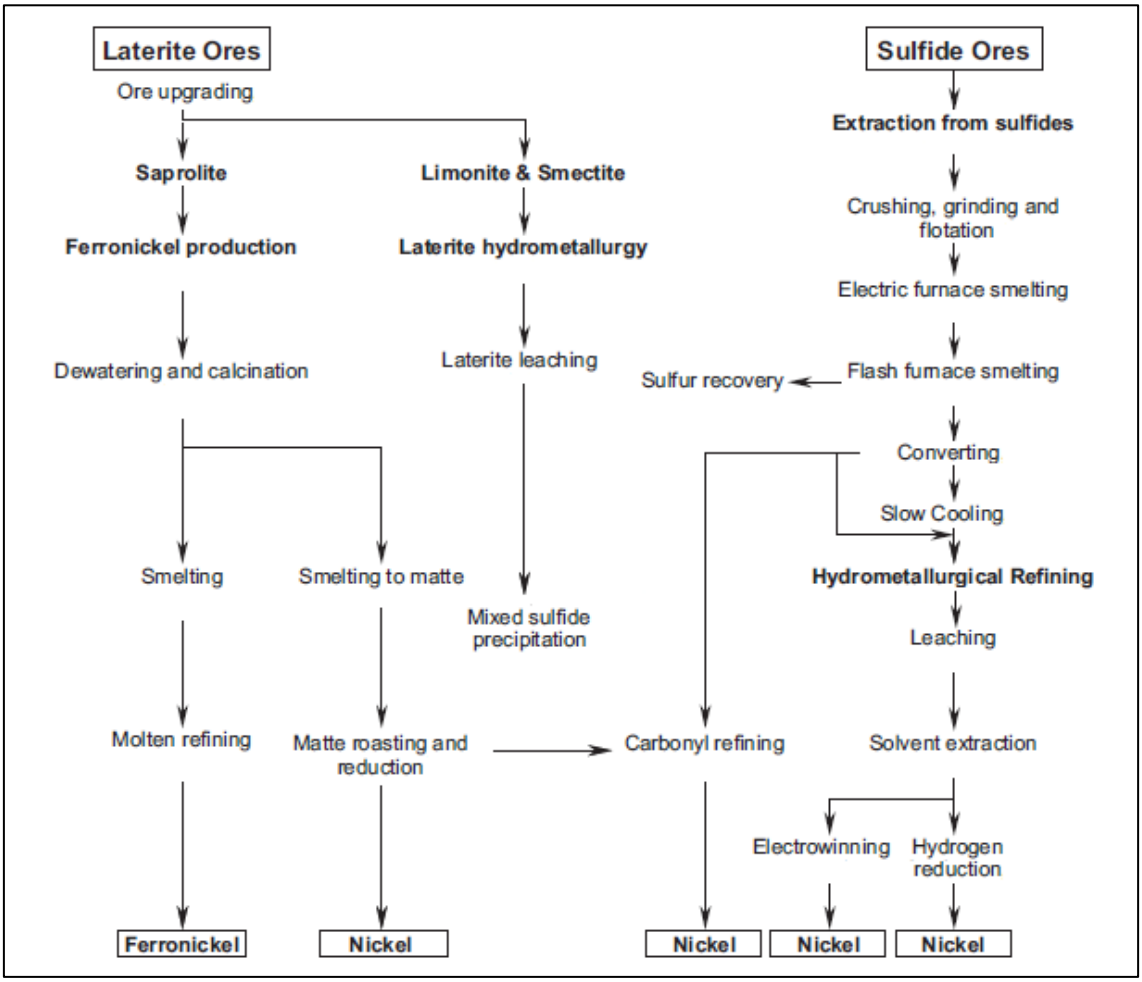


Figure 1.8. Principle process block diagram of nickel production (Crundwell et al. 2011)

The obtained low-iron matte containing 0.5-4% Fe, is then subjected to solidification and cooling with further grinding and magnetic separation to segregate copper sulfide and nickel

sulfide concentrates. Solidification is a slow process, where molten converter matte is cooled down in molds from the 1000°C to 200°C over a period of several days so that its constituents are segregated into separate chemical phases (Habashi 1999).

Acquired nickel concentrates undergo process of leaching by either chlorine gas in chlorine solution or oxygen in ammonia (or sulfuric acid) solution to get pregnant solution, which is purified and sent to solvent extraction to separate nickel and cobalt. Chlorine route involves high-pressure autoclave leaching of granulated matte particles at 150°C with iron or nickel and copper containing chloride solution and chlorine gas to produce nickel enriched pregnant solution and copper sulfide precipitate. In case of ammonia leaching, matte particles are leached in ammonium sulfate solution at 110-120°C and overpressure of approximately 0.85 MPa in horizontal autoclaves to produce nickel enriched solution, which is later boiled to distill ammonia and precipitate copper sulfide (Crundwell et al. 2011, Moats, Davenport 2014).

Obtained by means of leaching sequences nickel enriched solution should be separated from cobalt prior to further treatment. This is typically done by chloride or sulfate solvent extraction processes. Chloride route involves cobalt extraction with tri-isooctyl amine solution producing cobalt laden organic phase, that is later washed from entrained nickel and stripped from the organic solvent, while nickel enriched solution is sent to electrowinning process. In sulfate solvent extraction 2,4,4-trimethylpentyl phosphonic acid solution is applied for cobalt extraction (Moats, Davenport 2014).

Electrowinning process is applied to produce high purity nickel from solvent extraction purified solution by means of direct electrical current of approximately 23 000 A at overall cell voltage of 3 V. Nickel is electrowon from the aqueous chloride electrolyte on the titanium-based blank cathodes which are periodically harvested, and the deposited nickel is stripped by special automated machinery. Conductive but inert lead or ruthenium coated titanium anodes are placed in a permeable polyester diaphragm, that is used to contain chlorine gas, which was generated during the anode reaction (Moats, Davenport 2014, Habashi 1999). Thus, emitted from the anodes chlorine gas is collected along with nickel depleted electrolyte by vacuum. In case of nickel electrowinning from sulfate solutions, higher overall cell voltage of 4 V is applied, and cathodes are placed in polypropylene frames to prevent transfer of the generated at the anode acid to catholyte, so that anodic generation of hydrogen ions, which can adversely affect nickel electrodeposition, is reduced. Generated sulfuric acid is reused in the leaching circuits (Crundwell et al. 2011).

Additionally, currently several other processes such as carbonyl refining and electrolytic refining of molten matte are also used for sulfide ores processing. Carbonyl refining involves formation of gaseous phase of nickel carbonyl followed by further nickel decomposition. Nickel carbonyl is produced by reaction with carbon monoxide at nearly ambient pressure in a rotating kiln at 50^{o}C, while unreacted residue is conveyed to further metal recovery processes. Formed nickel carbonyl is decomposed by contacting with hot air at 240^{o}C so that carbon monoxide is removed, and high-purity nickel pellets are produced (Moats, Davenport 2014).

Electrorefining of nickel concentrates is conducted with cathodes of pure nickel in concrete electrolytic cells filled with nickel sulfide/chloride electrolyte solution and slabs of nickel converted matte which were specifically molded in horizontal matte anodes. Under the influence of electric current at overall cell potential of 3-4 V and density of 250 A/m^2, anodes are dissolved, and nickel is being electrodeposited on the cathode thin sheets. Nickel laden cathodes are washed and sent to the market, while anode scrap is sent for further processing to recover platinum group metals (Habashi 1999, Moats, Davenport 2014).

3.6.2. Laterite ores processing

Laterite ores can be processed either for production of ferronickel, containing approximately 20-40% of nickel and 60-80% of iron, or subjected to hydrometallurgical treatment for production of high-purity nickel. Ferronickel production usually starts with dewatering of ore concentrates, that is followed by calcination at 800-1000^{o}C to remove the remaining water. Both dewatering and calcination processing steps are accomplished in long sloped hot rotating kilns. The main product of the calcination process is hot, dry and partially reduced ore concentrate – calcine at approximately 900^{o}C containing 1.5-3% of nickel and up to 15% of iron, that is suitable for downstream processing.

Further steps include reduction by means of smelting in electric furnace at 1400-1450^{o}C with carbon and refining of molten ferronickel in order to reduce phosphorus and sulfur content. Generally, smelting is done in electrically heated furnace with suspended carbon-based electrodes at approximately 30 000 A. This process involves formation of two immiscible liquid layers: slag and molten ferronickel. These products are separately tapped and ferronickel, enriched with iron and nickel, is conveyed to further refining. Refining involves sulfur and phosphorus removal from the molten ferronickel at temperature of 1550-1600^{o}C. Phosphorous is removed by addition of lime, calcium oxide and blowing of oxygen through special lances resulting in the liquid slag formation on the molten ferronickel surface, which is carefully decanted. Sulfur is removed by adding calcium carbide or carbonate to the

molten ferronickel, separating sulfur as calcium sulfate and forming the slag which is also later decanted. Produced molten ferronickel contains less than 0.02% of phosphorous and 0.03% of sulfur.

Approximately 90% of laterite ores are used for ferronickel production, while the remaining 10% are utilized in the matte smelting process to eventually produce melting-grade nickel with purity of 95-97%. Typically, this route involves such processing steps as dewatering of laterite concentrates, calcining, sulfiding, smelting, converting and refining. Dewatering and calcining processes are done in rotating kilns similarly with the previously described techniques. In the sulfiding step molten sulfur is sprayed into the discharge end of the rotating kiln to produce sulfided calcine, which is further treated in an electric furnace to form molten matte and separate it from slag at 1500^{o}C. The obtained hot molten matte is sent to the downstream processing by means of converting and refining, which were already described above.

3.6.3. Nickel production side streams identification

Principal side streams of the most common commercialized nickel manufacturing processes are provided in the Table 1.6.

Table 1.6. Product streams in the nickel production manufacturing steps

Processing step		Inlet streams	Side streams
Flash furnace / Electric smelting		Nickel concentrate (15% Ni), SiO_2 flux	Molten slag, Sulfur dioxide bearing off-gas.
Conversion		Furnace matte (40% Ni), flux	Molten slag, Sulfur dioxide bearing off-gas.
Solidification and magnetic separation		Converted matte (50-60% Ni)	Copper sulfide concentrate, alloy concentrate
Carbonyl refining		Converted matte (50-60% Ni) / nickel concentrates, 80-95% carbon monoxide	Non-carbonyled residue
Electrorefining of nickel matte		Converted matte (50-60% Ni)	Anode scrap and solid corrosion products
Hydrometallurgical refining of nickel	Leaching	Finely crushed nickel matte (50-60% Ni)	Leach precipitate, iron mud or solid unleached residue
	Solvent extraction	Nickel pregnant chloride solution, tri-isooctyl amine	Cobalt-enriched solution
	Electrowinning	High-purity nickel	Chlorine gas or oxygen gas, depleted nickel solution or sulfuric acid enriched solution
Ferronickel refining	Calcination	Dewatered saprolite concentrate	Dusty off-gases
	Smelting	Calcine	Molten slag, off-gases
	Refining	Molten ferronickel, calcium carbide, calcium oxide	Molten slags

Smelting process molten slag – is silica saturated iron-silicate dross containing on average up to 35-37% of SiO_2, 33-35% of iron, 0.2% of nickel, 0.1% of copper and cobalt. Average outlet temperature is within the range of 1200-1300^oC. Slag from electric smelting furnace typically has lower nickel and copper content – 0.1% and 0.05% respectively, and higher outlet temperature – 1350^oC. In general, molten slag obtained in the electric furnace smelting process is transferred to the disposal area by means of slag pots, while furnace smelting slag can be recycled in the flash furnace for valuable metals recovery. The slag composition in both cases is controlled by fluxing materials addition (Habashi 1999).

Matte conversion slag represents dense, unreactive and readily disposable nickel-lean waste product containing on average 21-25% SiO_2, 10-20% Fe_3O_4, 20-30 % Fe_2O_3 at outlet temperature of 1200-1300^oC. Slags with low silica content usually have large amounts of magnetite. Specific gravity – 3.0 g/cm^3. The slag in most cases is sent to refining in electric slag-cleaning furnace for additional nickel recovery or returned to the smelting furnace, because it always contains matte droplets.

Sulfur dioxide bearing off-gas as a side stream of electric smelting contains on average 13-25 % SO_2 at temperature of 800^oC, while in case of blast furnace smelting – 20-50% SO_2 at temperature of 1300-1400^oC. As a side stream of conversion process contains usually 2-4% SO_2 and in most cases discarded into the atmosphere (Habashi 1999, Crundwell et al. 2011).

Copper sulfide concentrate consists of 74% of copper, 4% of nickel, silver, selenium and also tellurium. This product is typically sold to copper smelting facilities, where it is converted to blister copper. Silver, tellurium and selenium are recovered from the anode slimes of copper processing (Habashi 1999).

Alloy concentrate is rich in precious metals, such as silver, gold and platinum group metals, contains approximately 65% of nickel and 17% of copper – sent to the carbonyl refining process.

Non-carbonyled residue is a solid side stream of carboxylation process that contains 40-50% of nickel plus other metals – sent to additional refinery for nickel recovery. Average stream outlet temperature – 50^oC (Crundwell et al. 2011).

Leach precipitate from chlorine leaching process contains on average up to 50-55% of copper, 35-40% of sulfur, 5-7% of nickel as unreacted nickel sulfide, 0.5% of cobalt and 2% of iron. Outlet temperature – about 70^oC. Further processing involves transportation to the copper production plant for copper and precious metal recovery and sulfuric acid production (Moats, Davenport 2014).

Iron mud – is a side stream of chlorine leaching process produced as a result of iron precipitation by oxidation with chlorine gas and hydrolysis with nickel carbonate. It consists mostly of iron (up to 43%) and also contains trace amounts of nickel (less than 1.5%).

Solid unleached residue – is a solid by-product of nickel concentrate ammonia leaching that on average consists of 25% of iron (mostly in form of $Fe_2O_3 \cdot H_2O$), 20% of silica, 0.5% of cobalt and up to 4% of nickel at temperature of 105-120^oC. This stream is washed from the process and discarded in a tailings storage (Crundwell et al. 2011).

Cobalt-enriched solution – is a side stream of chloride solvent extraction process that is sent to washing with hydrochloric acid water solution to remove the entrained nickel and obtain high-purity cobalt solution.

Chlorine gas is emitted at the anodes in case of electrowinning from chloride electrolyte and collected, dried and compressed for further reuse in the chloride leaching process. Oxygen gas – a product of electrowinning from sulfate solutions is released to the environment. Depleted nickel solution and sulfuric acid enriched solution from the corresponding electrowinning processes are utilized in the relevant leaching steps (Crundwell et al. 2011).

Dusty off-gases of calcination process are enriched with carbon dioxide, water and nitrogen, have outlet temperature of 250^oC. Dust from this stream is collected by means of electrostatic precipitation and pelletized for recycling in the calcination kiln. Approximate amount of produced dust is 15% of the kiln inlet feed. Off-gases from the calcine smelting process have almost similar composition, except higher carbon monoxide content, and higher temperature – 900^oC. They are collected and dedusted in cyclones and baghouses and generally discarded in the atmosphere or reused in laterite ore dewatering processing step.

Molten slag as a side stream of calcine smelting process consists mainly of silica (30-50%), magnesium oxide (29-33%), calcium oxide (1-7%) and iron oxide (5-15%) at outlet temperature of approximately 1550^oC and density of 3.0 g/cm^3. Nickel content – 0.1-0.2%. In most cases discarded as a waste or can be granulated by means of water spraying and applied as a building material or metallurgical fluxing element.

EXPERIMENTAL PART

4. Focus and aims of the work

Within the limits of this work the application of available IBM Watson analytical tools has been studied for detailed analysis of several metal industries. The general objectives included the following targets: understanding of potential opportunities and drawbacks of Watson Analytics platform in terms of application for technical data analysis; in-depth study of industrial processes related information with the purpose of revealing hidden patterns and interdependencies between key parameters and product properties; comparative analysis of the platform analytical and predictive capabilities. To achieve these targets a series of databases related to metal production industries containing diverse information has been collected and prepared. Further analysis initially involved determination of key drivers for the characteristics of the most important products followed by assessment of statistical significance of these drivers and development of predictive models. Databases of various volumes and properties have been studied for this analysis in order to better understand Watson Analytics capabilities. Additionally, with the purpose of evaluation of the results and main features of the platform, a comparative study with the linear regression data analysis tool has been performed.

Furthermore, natural language processing capabilities of IBM Watson software have been studied through the instrumentality of the company's cognitive search platform – Watson Discovery service. Hence, a data collection containing up-to-date research articles about processing and characterization of side streams and wastes in metal industries has been prepared and analyzed. The general objective of this study is to customize data ingestion and processing algorithms with a view to further analysis and detection of the most relevant scientific articles

5. Data preparation

For successful application of IBM Watson Analytics platform large quantity of data first needs to be collected and appropriately refined. For efficient and accurate analysis and pattern recognition it is advisable that the prepared database contains major part of information in unified numerical form. Within the scope of this work a number of databases related to aluminum and copper production were gathered and transformed into the format suitable for further analysis.

5.1. Aluminum production

As it was mentioned previously in the literature review, extensive emissions of perfluorocarbons and fluorides produced as a result of cryolite bath volatilization during electrolytic reduction is a significant issue of the aluminum production technologies. Therefore, according to the data, obtained from the International Aluminum Institute, a database containing information about world aluminum production rates, utilized energy sources and corresponding emissions of perfluorocarbons and fluorides was collected for the period from 1990 to 2016 (The International Aluminium Institute 2017). The collected database consisted of 15 columns and 16 rows, thus containing 240 information cells. More detailed description of the information collected in this database is provided in the Table 2.1.

Table 2.1 – Aluminum production database description

Column	Description
Alumina production	Primary alumina production in the given year, tons
Coal	Amount of energy for the aluminum production obtained from coal, TJ
Oil	Amount of energy for the aluminum production obtained from oil, TJ
Gas	Amount of energy for the aluminum production obtained from natural gas, TJ
Electricity	Amount of electricity applied for aluminum production, TJ
Prebake	Total fluorides emissions from plants based on pre-bake technology, tons
Prebake intensity	Intensity of fluorides emission from plants based on pre-bake technology, kgF/tAl
Soderberg	Total fluorides emissions from plants based on Soderberg technology, tons
Soderberg intensity	Intensity of fluorides emission from plants based on Soderberg technology, kgF/tAl
CF4	Overall carbon tetrafluoride emissions, Gg
C2F6	Overall Hexafluoroethane emission, Gg
PFC	Overall perfluorocarbon emissions, $ktCO_2e$
PFC intensity	Intensity of perfluorocarbon emissions, $ktCO_2e/tAl$

5.2. Copper concentrate smelting in submerged-tuyere furnace

Information about copper production by means of Vanyukov submerged-tuyere technology in Norilsk Copper plant was collected for further analysis with IBM Watson Analytics platform. Information was available for two slightly different operating units: VF-2 and VF-3[1].

[1] VF-2 – Vanyukov Furnace-2; VF-3 – Vanyukov Furnace-3

5.2.1. Process description

Vanyukov process involves sulfide materials smelting in a vigorously mixed slag-matte bath with the use of heat from oxidative reaction. The crucial differences of this technology are oxygen enriched blow and furnace charge with small amount of matte. The process is continuous and is typically achieved in a shaft type furnace. Oxygen enriched air is fed into the molten mass through tuyeres, which are located symmetrically on the both sides of a furnace. Tuyere belt divides the molten bath into two zones: subtuyere and abovetuyere sections. Such processes as dissociation, water evaporation, melting and oxidation of the furnace charge, smelting products formation, matte droplets coalescence occur in the upper zone, which is efficiently mixed by inlet oxygen blow. Smelting products are separated due to difference in specific gravity into bottom matte phase and upper slag phase in the more stable subtuyere zone. Slag discharge is performed by means of slag siphon pipe installed in the middle part of the subtuyere zone, where non-ferrous metals content is the lowest.

This process can be performed in autogenous manner – by utilizing the heat of exothermic oxidative reactions, or in semi-autogenous manner – when lack of heat is compensated by additional natural gas combustion.

Furnace charge heating starts during its upwards vertical movement towards surface of melt ending at the bubble section of the furnace. Efficient mass and heat transfer processes in the bubbling molten bath result in formation of homogeneous slag-matte emulsion. Apart from iron and copper oxides furnace charge also contains silica, burnt lime, alumina and magnesia, which cause formation of slag during the smelting process. Specific slag composition depends mostly on wustite, silica and burnt lime concentrations. Schematic visual representation of the Vanyukov smelting is provided in the Figure 2.1.

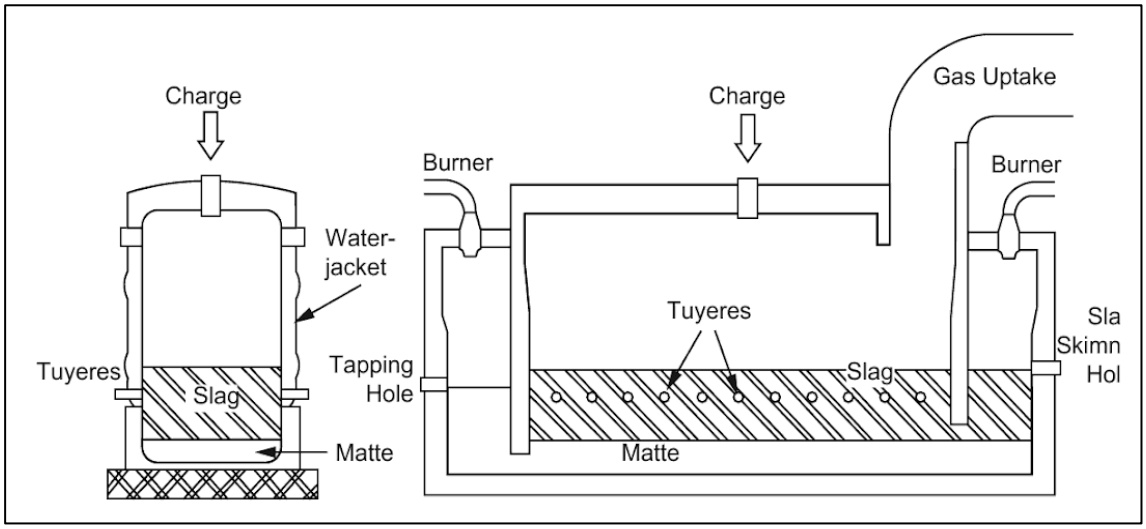
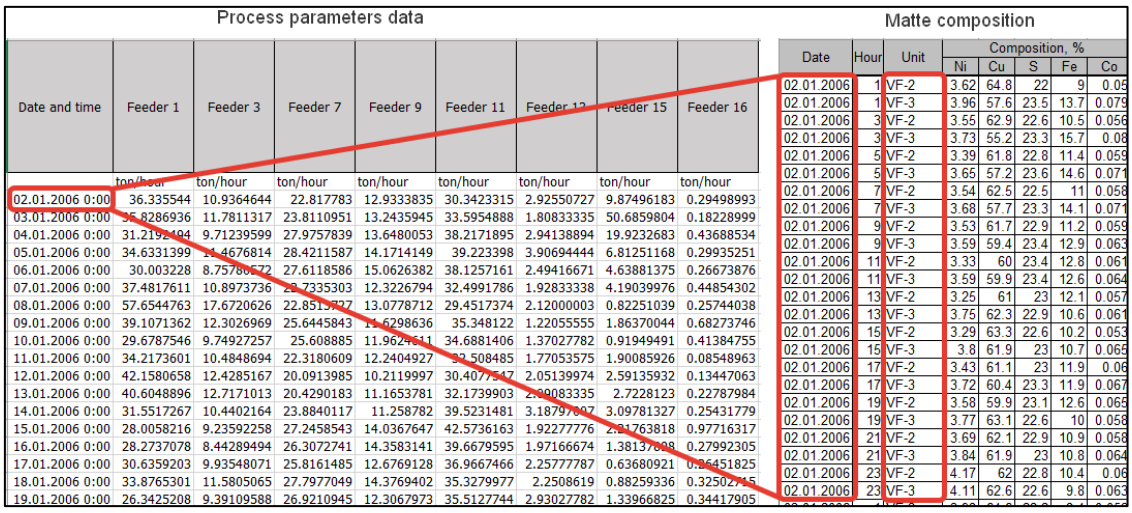


Figure 2.1 Vanyukov submerged-tuyere smelting (Schlesinger et al. 2011)

Copper loss in slag can be either due to mechanical losses, which are associated with insufficient sedimentation time of small sulfide droplets, or due to dissolution losses, which are related to dissolution and oxidation of copper in the slag. Dissolution losses account for approximately 80% of all copper losses in the slag.

5.2.2. Data preparation

The available datasets contained information about process parameters, matte composition and slag composition in separate files with .xls extension. Process parameters values were available for every 24 hours, while slag composition analysis was available for every 4 hours, and matte composition – for every 2 hours as it is represented in the Figure 2.2 Therefore, the obtained databases were not comparable and, thus, not suitable for direct analysis. Moreover, another encountered problem resided in the fact that there were inconsistencies and gaps in products composition analysis.



Process parameters data

Date and time	Feeder 1	Feeder 3	Feeder 7	Feeder 9	Feeder 11	Feeder 13	Feeder 15	Feeder 16
	ton/hour	ton/hour	ton/hour	ton/hour	ton/hour	ton/hour	ton/hour	ton/hour
02.01.2006 0:00	36.335544	10.9364644	22.817783	12.9333835	30.3423315	2.92550727	9.87496183	0.29498993
03.01.2006 0:00	35.8286936	11.7811317	23.8110951	13.2435945	33.5954888	1.80833335	50.6859804	0.18228999
04.01.2006 0:00	31.2192494	9.71239599	27.9757839	13.6480053	38.2171895	2.94138894	19.9232683	0.43688534
05.01.2006 0:00	34.6331399	11.4676814	28.4211587	14.1714149	39.223398	3.90694444	6.81251168	0.29935251
06.01.2006 0:00	30.003228	8.75780672	27.6118586	15.0626382	38.1257161	2.49416671	4.63881375	0.26673876
07.01.2006 0:00	37.4817611	10.8973736	29.7335303	12.3226794	32.4991786	1.92833338	4.19039976	0.44854302
08.01.2006 0:00	57.6544763	17.6720626	22.851[illegible]27	13.0778712	29.4517374	2.12000003	0.82251039	0.25744038
09.01.2006 0:00	39.1071362	12.3026969	25.6445843	11.[illegible]6298636	35.348122	1.22055555	1.86370044	0.68273746
10.01.2006 0:00	29.6787546	9.74927257	25.608885	11.9624611	34.6881406	1.37027782	0.91949491	0.41384755
11.01.2006 0:00	34.2173601	10.4848694	22.3180609	12.2404927	32.508485	1.77053575	1.90085926	0.08548963
12.01.2006 0:00	42.1580658	12.4285167	20.0913985	10.2119997	30.4077547	2.05139974	2.59135932	0.13447063
13.01.2006 0:00	40.6048896	12.7171013	20.4290183	11.1653781	32.1739903	2.[illegible]083335	2.7228123	0.22787984
14.01.2006 0:00	31.5517267	10.4402164	23.8840117	11.258782	39.5231481	3.18797[illegible]07	3.09781327	0.25431779
15.01.2006 0:00	28.0058216	9.23592258	27.2458543	14.0367647	42.5736163	1.92277776	2.[illegible]1763818	0.97716317
16.01.2006 0:00	28.2737078	8.44289494	26.3072741	14.3583141	39.6679595	1.97166674	1.38137[illegible]08	0.27992305
17.01.2006 0:00	30.6359203	9.93548071	25.8161485	12.6769128	36.9667466	2.25777787	0.63680921	0.[illegible]6451825
18.01.2006 0:00	33.8765301	11.5805065	27.7977049	14.3769402	35.3279977	2.2508619	0.88259336	0.32502[illegible]15
19.01.2006 0:00	26.3425208	9.39109588	26.9210945	12.3067973	35.5127744	2.93027782	1.33966825	0.34417905

Matte composition

Date	Hour	Unit	Composition, %				
			Ni	Cu	S	Fe	Co
02.01.2006	1	VF-2	3.62	64.8	22	9	0.05
02.01.2006	1	VF-3	3.96	57.6	23.5	13.7	0.079
02.01.2006	3	VF-2	3.55	62.9	22.6	10.5	0.056
02.01.2006	3	VF-3	3.73	55.2	23.3	15.7	0.08
02.01.2006	5	VF-2	3.39	61.8	22.8	11.4	0.059
02.01.2006	5	VF-3	3.65	57.2	23.6	14.6	0.071
02.01.2006	7	VF-2	3.54	62.5	22.5	11	0.058
02.01.2006	7	VF-3	3.68	57.7	23.3	14.1	0.071
02.01.2006	9	VF-2	3.53	61.7	22.9	11.2	0.059
02.01.2006	9	VF-3	3.59	59.4	23.4	12.9	0.063
02.01.2006	11	VF-2	3.33	60	23.4	12.8	0.061
02.01.2006	11	VF-3	3.59	59.9	23.4	12.6	0.064
02.01.2006	13	VF-2	3.25	61	23	12.1	0.057
02.01.2006	13	VF-3	3.75	62.3	22.9	10.6	0.061
02.01.2006	15	VF-2	3.29	63.3	22.6	10.2	0.053
02.01.2006	15	VF-3	3.8	61.9	23	10.7	0.069
02.01.2006	17	VF-2	3.43	61.1	23	11.9	0.06
02.01.2006	17	VF-3	3.72	60.4	23.3	11.9	0.067
02.01.2006	19	VF-2	3.58	59.9	23.1	12.6	0.069
02.01.2006	19	VF-3	3.77	63.1	22.6	10	0.058
02.01.2006	21	VF-2	3.69	62.1	22.9	10.9	0.058
02.01.2006	21	VF-3	3.84	61.9	23	10.8	0.064
02.01.2006	23	VF-2	4.17	62	22.8	10.4	0.06
02.01.2006	23	VF-3	4.11	62.6	22.6	9.8	0.063

Figure 2.2. Inconsistency between available databases

Considering the immense volume of databases, manual correction was not possible and, therefore, special algorithms were developed by means of MATLAB™ for calculation of mean values of matte and slag compositions for corresponding days, so that information about process parameters would be relevant to the product properties. These algorithms were developed taking into consideration possible mistakes or gaps in the data and different processing units. MATLAB scripts with more detailed explanation are provided in the Appendix 1. As a result, after application of the developed scripts and discarding all unsuitable information cells, the two following databases were obtained:

- Vanyukov Furnace - 2 – containing 41 columns and 239 rows, totally – 9214 information cells;
- Vanyukov Furnace - 3 – containing 34 columns and 320 rows, totally – 10515 information cells;

Part of the Vanyukov Furnace - 2 database is provided in the Appendix 2 as an example for better representation of the obtained results. More detailed explanation of information contained in the databases is provided in the Table 2.2.

Table 2.2 – Vanyukov process database description

Column	Description
Flux feed (sand) & Flux feed (sandstone)	Fluxing agents charge, ton/hour
Feed sulfide materials	Sulfide materials charge, ton/hour
Coal feed	Coal addition rate, ton/hour
Reverts feed	Reverts materials charge, ton/hour
Exhaust gases temperature	Temperature of exhaust gases after the furnace, ^{o}C
Oxygen content	Oxygen concentration in the air-oxygen blast mixture, %
Air-oxygen mixture flowrate	Flowrate of air-oxygen mixture through tuyeres, m^3/hour
Air-oxygen mixture pressure	Inlet pressure of air-oxygen mixture through tuyeres, kgf/cm^2
Natural gas flowrate	Flowrate of natural gas for temperature regulation, m^3/hour
Water pressure (left side/ right side)	Water delivery for the furnace water jacket, kgf/cm^2
Water temperature entrance/exit (left side/right side)	Temperature of cooling water, ^{o}C
Matte Ni/Cu/Co/S/Fe	Nickel/Copper/Cobalt/Sulfur/Iron content in matte, %
Slag Ni/Cu/Co/S/Fe/ $CaO/MgO/SiO_2/Al_2O_3$	Nickel/Copper/Cobalt/Sulfur/Iron/ Quicklime/Magnesia/Silica/Alumina content in slag, %

Additionally, information related to weather conditions at the plant region during the corresponding time span was collected and integrated for the analysis. More specifically, information contained data about such parameters as: precipitation (cm), snow-cover (cm), humidity (%), ambient pressure (mm Hg) and temperature (^{o}C).

5.3. Matte and slag composition database

In order to conduct a separate more detailed analysis of obtained products properties and check the possibility to analyze the interdependencies between various matte and slag constituents' content, an additional database was created. Since, IBM Watson Analytics platform is based upon pattern recognition algorithms, this database was created by collecting as much comparable information as possible. Therefore, similarly with the previous cases, MATLAB™ was applied to extract relevant information, considering any possible mistakes, inconsistencies or gaps in the matte composition and slag composition databases.

Since, slag composition analysis has been performed less often, more specifically every 4 hours, time samples for slag compositions were taken as a basis. This way, the created algorithm firstly extracts information sets corresponding to the suitable time values into separate matrices. Afterwards, the algorithm compares the obtained matrices line by line for the purpose of finding any discrepancy of date, hour and unit. In case if no mismatch has been detected – the "approved" lines from both databases (slag composition and matte composition) are collected into one matrix. In case if a mismatch is detected – the nonrelevant line is removed from the corresponding database. More detailed description of the algorithm is provided in the Appendix 1. The obtained resulting database consists of 17 columns and 2983 rows, totally – 50 622 information cells.

Analysis of this database will highlight how effectively IBM Watson Analytics platform can work with databases of different volumes. Particularly, comparison of the results of matte and slag compositions analysis between this database and databases obtained in the section 5.2.2 is of interest because of significant difference in rows quantities.

6. Data analysis

Once, the necessary information has been collected and structured, the obtained datasets in .xls format were uploaded into IBM Watson Analytics cloud drive. The data quality index values assigned by default are represented in the Table 2.3.

Table 2.3 – Collected datasets quality index

Database name	Database volume	Quality index, %
Aluminum production	240 cells	87
Vanyukov Furnace 2	9 214 cells	84
Vanyukov Furnace 3	10 515 cells	83
Matte and slag composition	50 622 cells	69

As it can be seen from the provided information, prepared datasets were estimated as reasonably qualitative and suitable for further analysis without the need for additional refinement. However, quality index for matte and slag composition dataset was evaluated as the lowest compared with the other databases with a significant difference. This can be explained by the following factors:

- Occurring repetition of certain data values;
- Possible imbalance in data;
- Influential categories;
- Outliers;
- Data skewness.

The above described possible reasons for data quality deterioration in case of matte and slag composition dataset, can be also explained by the significant quantity of rows and considerable overall database volume, which in turn may have enhanced the negative effect of the mentioned factors.

6.1. Aluminum dataset analysis

Analysis of data package related to the aluminum annual production and corresponding emissions rates has been performed with the purpose of determination of energy sources that had the most significant statistical impact on emissions intensity. Within the limits of Watson Analytics platform in terms of key statistical drivers' determination, the available parameters are rated according to the values of their predictive strength, calculated based on the generated model. Predictive strength – is a measure of a certain parameter importance in terms of its ability predict the specific outcome.

It was revealed, that among the key drivers for emission intensity for plants based on the pre-bake aluminum production technology, amount of energy produced by coal was associated with 60% of predictive strength, while electricity and natural gas accounted for 56% each. In case of emissions intensity for Soderberg technology, coal energy was associated with 85% of predictive strength, while for gas and oil the values were 78% and 64% respectively. It should be taken into account that during the analyzed time period aluminum production and energy consumption increased, while overall emissions rate decreased due to improvement of existing production units and implementation of new ones. Therefore, more detailed analysis of the exact key drivers statistical impact is deteriorated by possible confusion because of this discrepancy. Additionally, this issue is aggravated by a low amount of available information, therefore, the obtained graphical representation for the combined effect of key drivers contains missing values and does not provide a clear vision of interdependencies (Figure 2.3.).

Therefore, even though the data quality index and predictive strength value are relatively high for most cases, it can be concluded that more detailed information regarding aluminum production should be collected for accurate analysis of the key driver statistical impact. Moreover, predictive analysis is not possible for this data package, because Watson Analytics platform was not able to develop the predictive models due to low amount of available data.

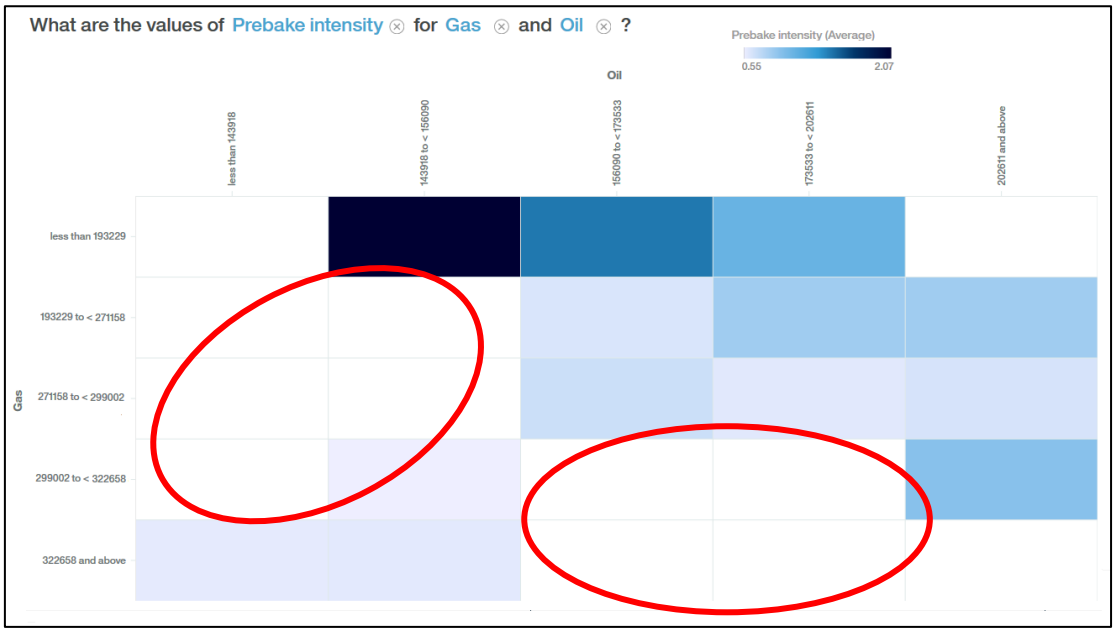


Figure 2.3. Emissions intensity of aluminum production by pre-bake technology influenced by amount of energy produced from natural gas and oil

6.2. Copper concentrate smelting process analysis

6.2.1. Process analysis in terms of matte properties

Although, taking account of all the features and content of the analyzed dataset, Watson Analytics platform automatically generates a default initial set of questions that can be used for further investigation, it was decided to use a novel set of questions in order to conduct a more detailed analysis.

Initially, to get a clear understanding of basic patterns and interdependencies within the dataset, the most significant key drivers have to be defined. Therefore, a combination of drivers that provides increased predictive strength for the most important constituents of matte and slag streams has been determined. This way, content of copper was analyzed for matte product ("Matte Cu"), while content of silica, iron and alumina were analyzed for slag stream ("Slag SiO2", "Slag Fe" and "Slag Al2O3" respectively). The platform interface provides an opportunity to analyze both singular impact of separate drivers as well as the combined impact of two drivers. The graphical representation is provided in a form of so-called "spiral visualization", where the more significant drivers are located closer to the spiral's center. An example of drivers of copper content in matte is represented in the Figure 2.4. More detailed information regarding data analysis procedure is provided in the Appendix 3.

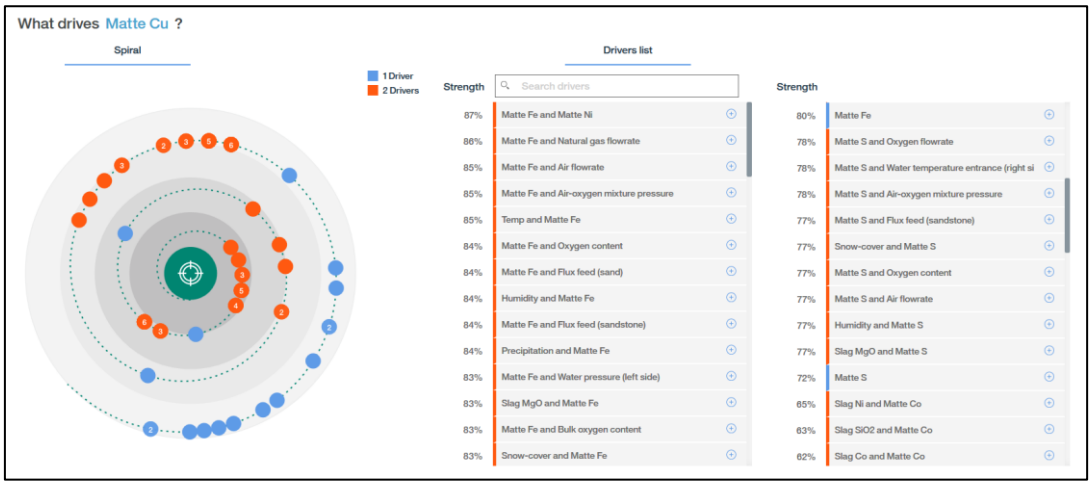


Figure 2.4. Determination of the key drivers for matte copper content by Watson Analytics spiral (Vanyukov Furnace 2 database)

As it can be seen from the provided information, the most significant impact on the copper content in matte stream is associated with iron and nickel content ("Matte Fe" and "Matte Ni" respectively) and also inlet flowrates of air and natural gas. Aside from that, it should be noted that combinations of several drivers typically provide higher predictive strength compared with single parameters. More specifically, combined statistical impact of iron and nickel content in matte resulted in 87% of predictive strength, while iron content by itself accounted for 80%. An additional point to be considered is that, even though iron and sulfur content (80% and 72% respectively) accounted for the highest predictive strength values separately, no relation has been found regarding the combined influence of these two parameters. More detailed information about determination of the most significant drivers of major constituents of matte and slag is provided in the Appendix 4.

After determination of the most significant drivers for matte and slag composition, more detailed information regarding the exact impact of each driver has been extracted. In case of copper content in matte analysis – a linear regression based approach (ANOVA) has been automatically applied by Watson Analytics platform because copper content has been recognized as a continuous target. The analysis approach is always selected by the platform and cannot be changed or modified. Visual representation is provided in the Figure 2.5. Values of iron and nickel content by default were gathered into 5 corresponding groups for both constituents to facilitate further analysis.

Based on the available analysis details, the number of degrees of freedom was equal 4 both for nickel and iron content – consequently resulting in 16 degrees of freedom (product of iron content and nickel content values).

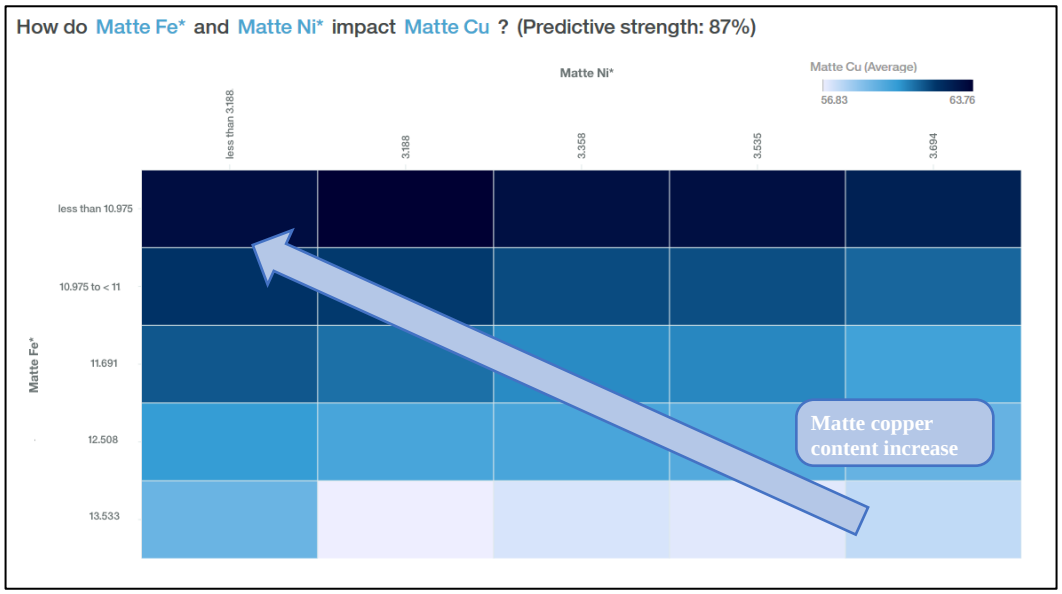


Figure 2.5. Copper content in matte depending on iron and nickel content

Additionally, Watson Analytics provides an opportunity to assess the model's statistical significance presenting calculated *F value*[2] – in this case it was equal 2.28 which is considered by the platform as statistically significant in terms of multivariate analysis. After determination of the key drivers and their statistical significance, an additional tool of Watson Analytics platform was used to conduct a predictive analysis. Predictive analysis typically provides an opportunity to observe the impact of more than two key drivers on the target variable.

In terms of IBM Watson Analytics platform interface, predictive model can be generated automatically in a form of decision tree or a set of decision rules. Example of predictive analysis application for copper content in matte is represented in the Figure 2.6. The generated predictive models are estimated based on their predictive strength. In this case, it's important to understand that the meaning behind predictive strength estimation of a predictive model is slightly different from the same term in the context of key drivers' determination. More specifically, the predictive strength value of a decision tree evaluates the degree to which the generated decisions, represented by each separate branch of the tree, predict the certain outcome of the target variable.

This predictive analysis feature facilitates more detailed investigation by describing which combination of factors is more likely to result in a specific outcome of the target variable. Reading from left to right each branch of the tree is a unique pattern that leads to a likelihood of a certain outcome. It should be noted that variables located more to the left are associated

[2] *F value* is a statistical term which is frequently used to determine whether the generated model is statistically significant. It is calculated as the ratio of the explained variance by the model to the unexplained variance. More detailed description can be found in Appendix 5.

with the increased statistical impact compared with the variables from the right. Conversely, standard deviation decreases from left to right, thus predictive strength is improved in this direction.

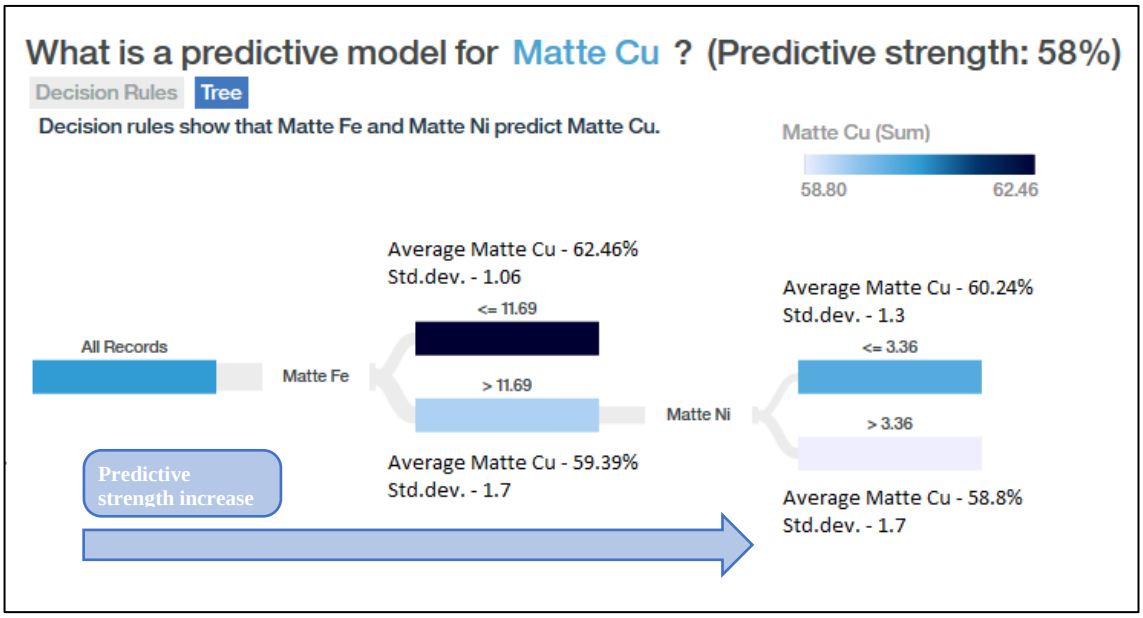


Figure 2.6. Matte copper content decision tree (Vanyukov Furnace 2 database)

In this case, only the most significant key drivers were considered within the decision tree for matte copper content predictive analysis. This shallow result can be explained by the insufficient amount of data, more specifically relatively low number of rows, and possible imbalanced influence of other variables. Therefore, additional more detailed dataset with increased volume has been studied with respect to the same aspects. The database contained information about similar processing unit and was described previously in the section 5.2.2. Key drivers' analysis for copper content in matte applied to this database revealed that, combined impact of cobalt and iron content is even more significant compared with iron and nickel. Apart from this, air and natural gas flowrate values were also associated with increased statistical significance in comparison with the previous case, while combined impact of nickel and iron content was the same – 87%. More detailed information is provided in Appendix 4 and Figure 2.7.

At the same time, it should be noted that apart from the cooling water characteristics, statistical impact of single key drivers remained much about the same as with the previous case.

Figure 2.7. Determination of key drivers of copper content in matte by Watson Analytics spiral (Vanyukov Furnace 3 database)

However, in terms of predictive analysis, only iron content in matte has been recognized as statistically significant predictor, so that nickel content was not mentioned in the generated model. The graphical representation of the decision tree is provided in the Figure 2.8.

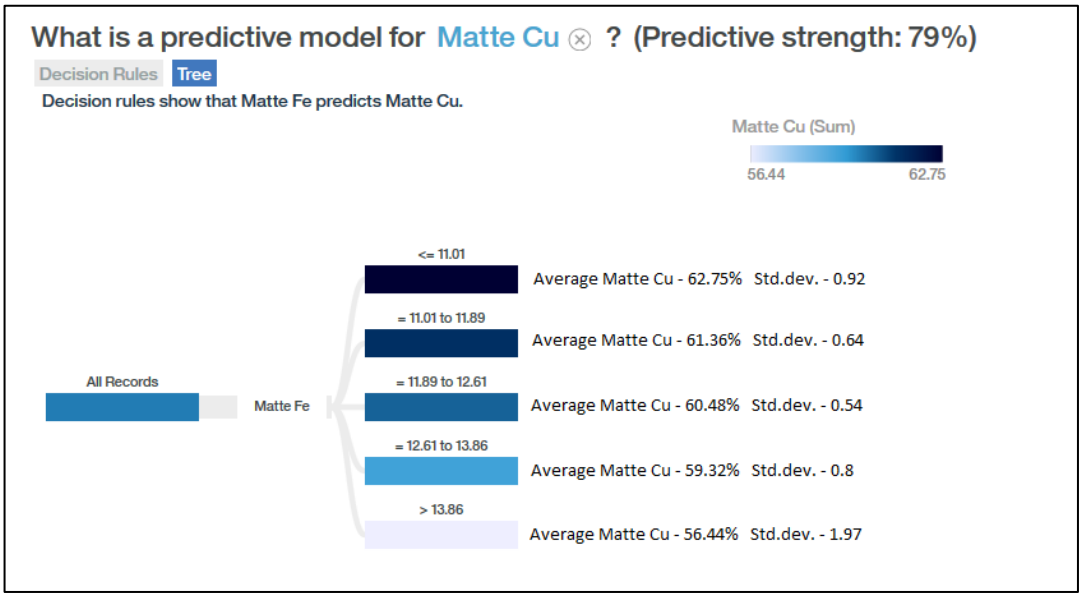


Figure 2.8. Matte copper content decision tree (Vanyukov Furnace 3 database)

The generated model has higher predictive strength – which can be explained by slight simplification since impact of nickel content has been omitted within the scope of this model. On the other hand, this predictive model provides an opportunity to get a more detailed analysis of the matte composition from the perspective of iron content. Additionally, it should be noted that value of standard deviation increases as the iron content range expands.

6.2.2. Process analysis in terms of slag properties

Furthermore, composition of the main side-stream of the process, namely slag, has been analyzed in a similar manner. The most significant constituents of the stream are silica, iron, and alumina; thus, their concentrations have been analyzed with the purpose of revealing relevant interdependencies and patterns. The first step of the analysis also involved determination of the most significant statistical drivers by the same approach that was previously described.

Among the key drivers for silica content, iron and copper concentrations in slag ("Slag Fe" and "Slag Cu" correspondingly) have been associated with the higher statistical strength with values of 73% and 35% respectively. In terms of process parameters slight influence of oxygen (16%) and natural gas (13%) inlet flowrate has been determined. Combined impact of several drivers was also associated with increased statistical strength. More specifically, the most significant impact was associated with combined influence of iron content in slag and oxygen inlet flowrate – 78% (Figure 2.9). More detailed description of other variables is provided in the Appendix 4.

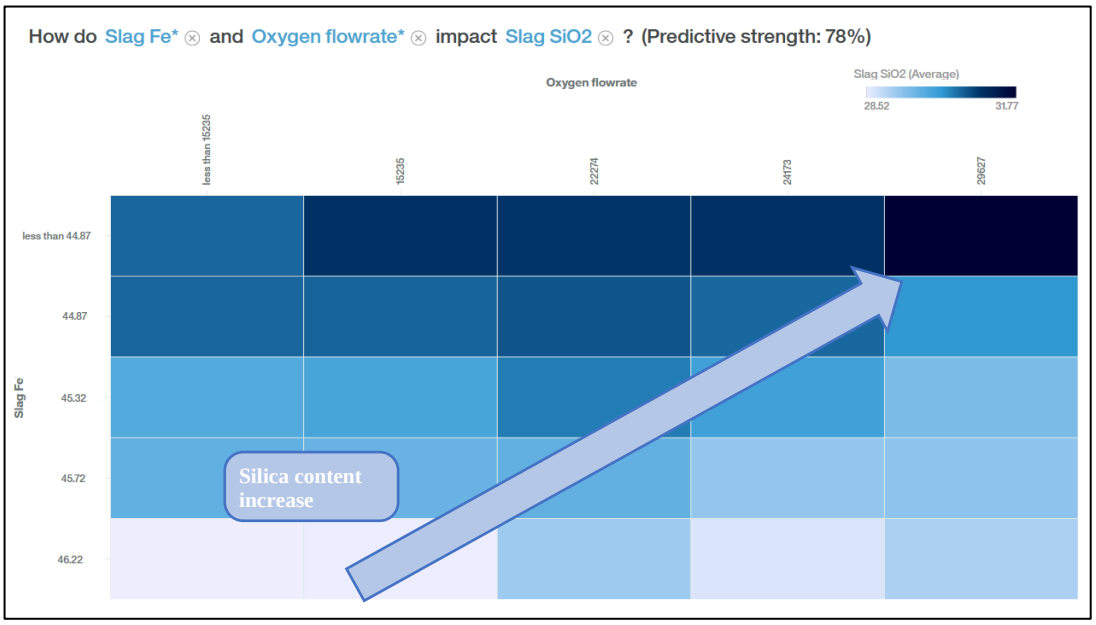


Figure 2.9. Silica content in slag depending on oxygen inlet flowrate and slag iron concentration

Judging by the obtained results, it can be observed that even though both parameters, namely iron concentration and oxygen flowrate, were considered to be statistically important, impact of iron content was more significant.

The conducted predictive analysis revealed also significant statistical impact of cobalt content in slag on the silica concentration, yet no information regarding oxygen or air

flowrates was not recognized in the generated predictive model (Figure 2.10). It can be seen that iron concentration in slag is also associated with increased statistical significance compared with other parameters.

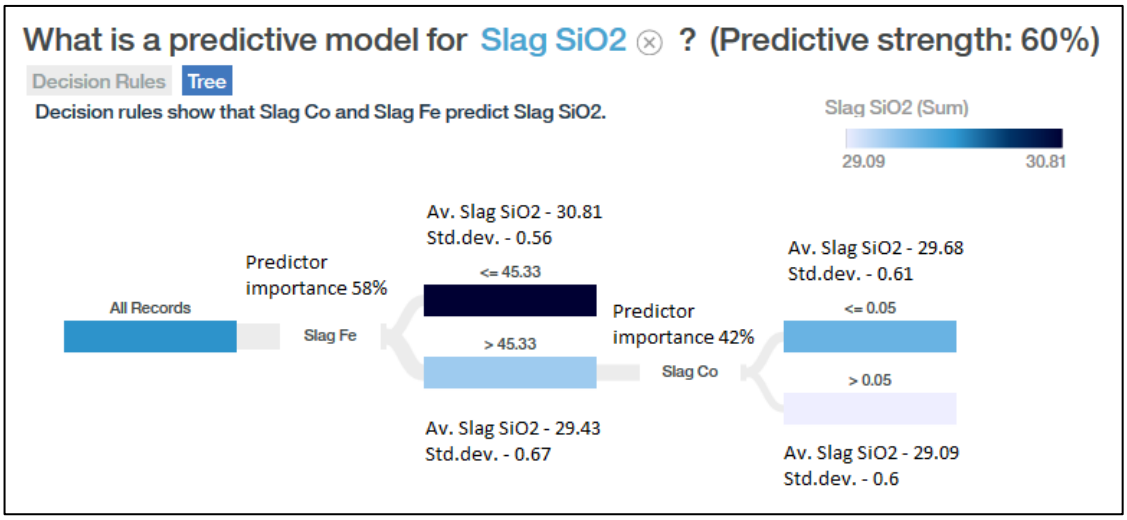


Figure 2.10. Slag silica content decision tree (Vanyukov Furnace 2 database)

Analogous study of silica concentration in slag conducted by analysis of the more detailed database (Vanyukov Furnace 3 dataset) revealed approximately similar results, yet with slight deviations. More specifically, in terms of single key drivers impact, statistical significance of iron and copper concentrations in slag slightly decreased for both parameters, particularly to the values of 68% and 32% respectively, while influence of inlet oxygen flowrate remained the same as with the previous case. For multivariable analysis – combined impact of iron and cobalt concentrations in slag was recognized to be the most significant with statistical strength of 80%. Additionally, increased statistical influence of combined iron concentration in slag with matte constituents' content on silica concentration should be mentioned in terms of analysis of the extended dataset. More detailed information is provided in the Appendix 4. Predictive model for this case is represented in the Figure 2.11.

Further analysis conducted for iron concentration in slag confirmed significant interdependence between iron and silica content. Silica concentration in slag has been recognized as the most significant statistical driver for iron content – 70% and 67% for databases with lower and bigger volumes respectively. Concentration of copper and alumina in slag were also associated with relatively high statistical strength, yet for bigger dataset their importance was lower. More detailed information related to the key drivers of iron content is provided in Appendix 4. Developed predictive model for iron content in case of both datasets was similar to the results of silica concentration predictive analysis and did not provide sufficient detailed description.

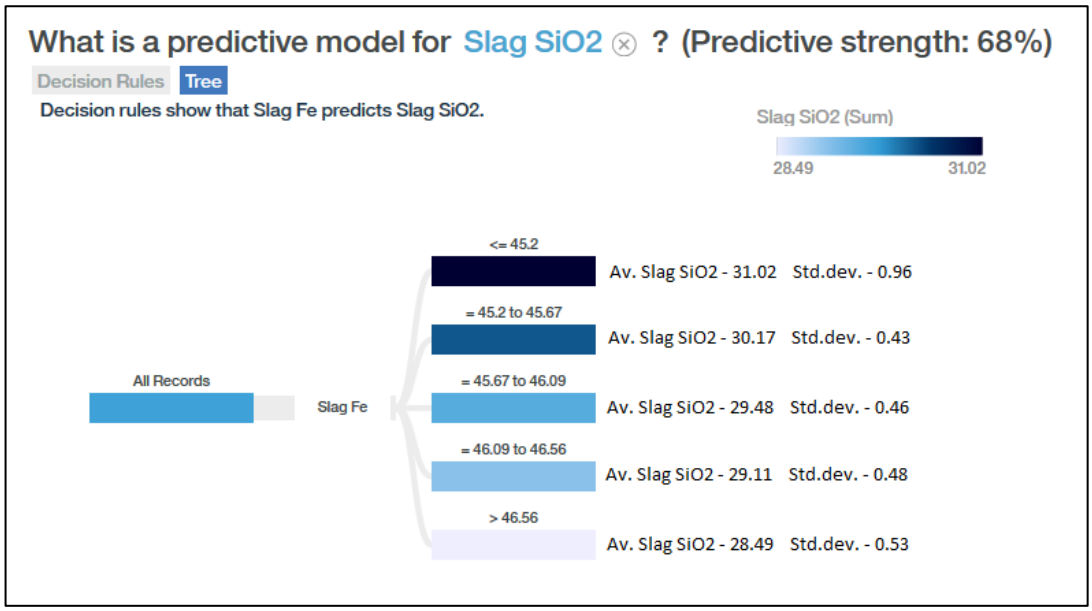


Figure 2.11. Slag silica content decision tree (Vanyukov Furnace 3 database)

Similar analysis of alumina concentration in slag conducted for the both databases did not reveal any single key drivers with significant statistical strength. Iron and quicklime content in slag were associated with relatively significant statistical impact of around 20% for both databases. In terms of multivariable analysis, even though combined impact of some parameters resulted in statistical strength above 50%, more detailed analysis of these parameters did not reveal any clear patterns or interdependencies for both databases.

6.3. Matte and slag composition data analysis

Since, as it was mentioned in the literature review, increased data volume facilitates analysis with Watson Analytics platform due to the fact that with large datasets there are more opportunities to find specific entities and patterns hidden within the data – a more detailed database describing matte and slag composition has been studied with the purpose of revealing interdependencies between various constituents' content. This time, database, described previously in the section 5.3, contained significantly more information, even though the number of columns has been decreased. Likewise, firstly key drivers of copper content in matte have been analyzed prior to further analysis. Concentrations of iron, sulfur and cobalt have also been recognized as the most significant separate drivers with slight deviation of approximately 2-4% compared with the previous analyses. In terms of multivariable analysis – combined impact of iron and sulfur was associated with the highest predictive strength value, while combined impact of iron and nickel was not mentioned. Significant increase in available for analysis data provided an opportunity to generate a more

robust and detailed predictive model. The obtained decision tree is presented in the Figure 2.12.

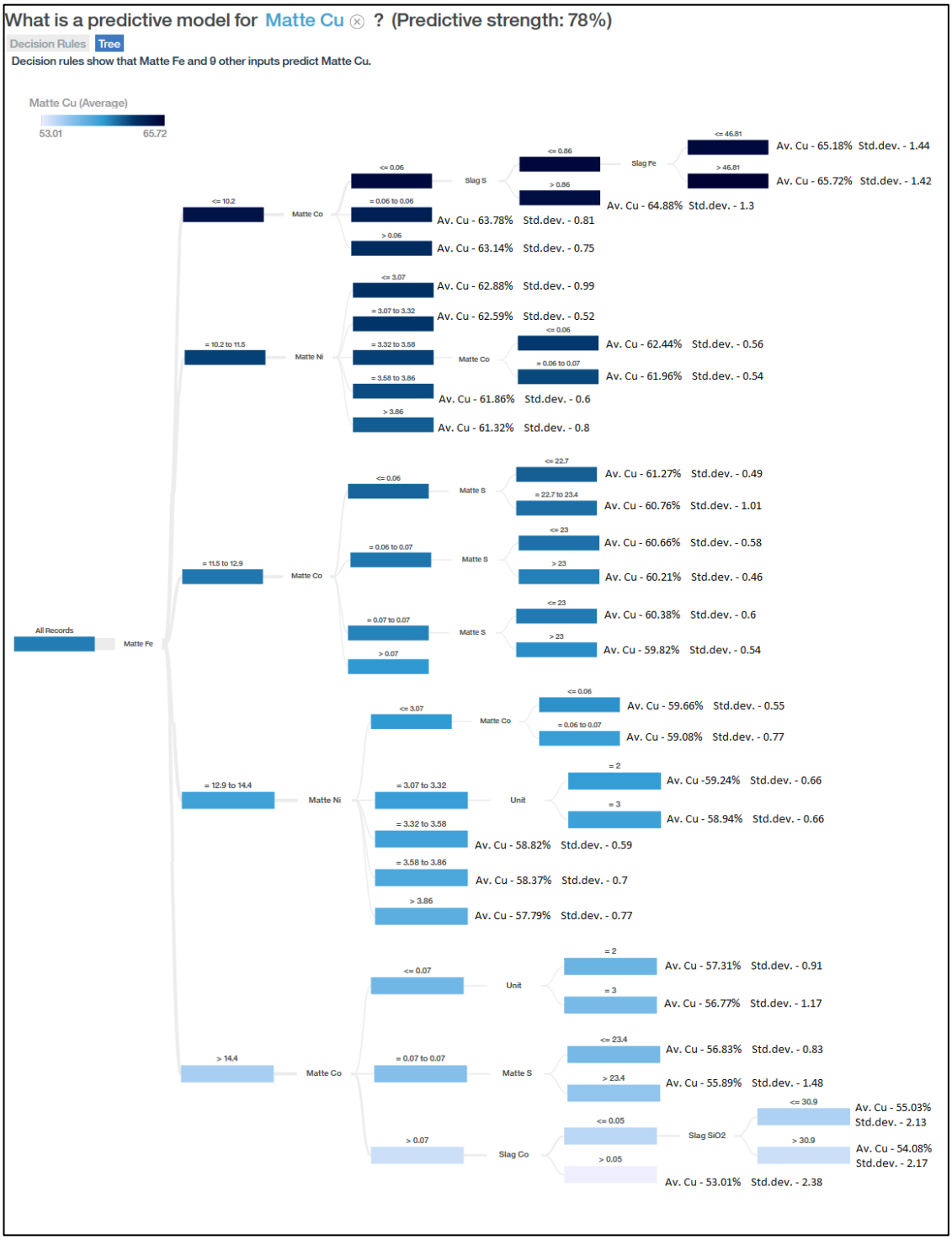


Figure 2.12. Matte copper content decision tree (Matte and Slag Composition database)

In this case, similar to the predictive models, generated based upon the previous datasets, iron content in matte has been associated with the most significant statistical impact and divided into 5 groups. Additionally, for each group a more detailed description is provided considering other constituents in matte and slag composition. However, matte composition characteristics were still associated with more significant statistical impact compared with

the slag constituents content. More specifically, according to the generated model, slag composition has been recognized as a relatively significant predictor with the combination of cobalt content in matte. Judging by the above provided results, it also should be noted that in case when slag compositions was analyzed as a predictive driver – standard deviation was relatively higher. More detailed analysis of the predictors' importance is provided in the Table 2.4.

Table 2.4 – Statistical importance of predictors for matte copper content analysis

Predictor description	Predictor importance
Matte iron content	24 %
Matte cobalt content	17 %
Slag cobalt content	10 %
Matte sulfur content	9 %
Slag silica content	9 %
Slag iron content	7 %
Furnace unit (VF-2 or VF-3)	7 %
Slag sulfur content	6 %
Matte nickel content	4 %

Moreover, even though the generated model for prediction of copper content in matte is much more diverse and complicated, the predictive strength index is quite high – 78%. This can also be explained by opportunity to find complex correlations and patterns to support analytical treatment and identification due to large data volume. Combined impact of sulfur and copper content was associated with 47% of statistical strength. The developed based on this dataset predictive model is characterized by much more detailed description of the process and increased predictive strength.

Furthermore, to confirm the previously determined interdependence between iron and silica content values in slag, a similar analysis has been conducted for the slag stream. In terms of silica content in slag ("Slag SiO2"), iron concentration ("Slag Fe") was also recognized to be one of the most important drivers and associated with 62% of predictive strength, followed by copper (32%) and alumina (19%) concentrations ("Slag Cu" and "Slag Al2O3" respectively).

Graphical representation of the predictive model in a form of decision tree is represented in the Figure 2.13. More detailed analysis of the predictors' importance is provided in the Table 2.5. According to the obtained results it can be concluded that the most important driver for silica concentration in slag stream is iron content. The developed predictive model is far

more detailed and considers all the most important influencing factors. Moreover, predictive strength is also higher due to increased volume of available data.

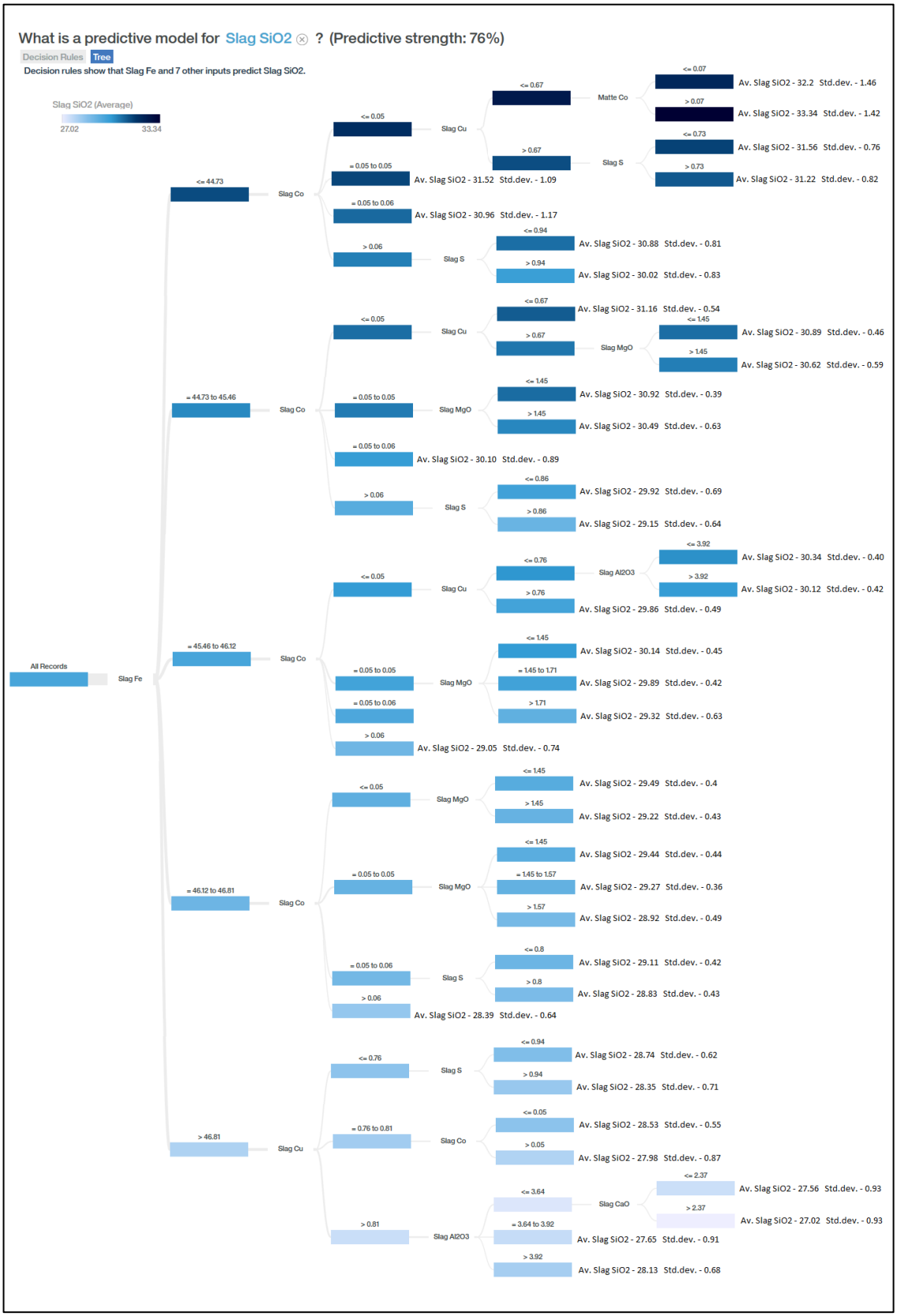


Figure 2.13. Slag silica content decision tree (Matte and Slag Composition database)

Table 2.5 – Statistical importance of predictors for slag silica content analysis

Predictor description	Predictor importance
Slag iron content	30 %
Slag cobalt content	18 %
Slag sulfur content	14 %
Slag alumina content	12 %
Slag copper content	11 %
Slag quicklime content	8 %
Matte cobalt content	4 %
Slag magnesia content	4 %

However, in terms of multivariable analysis for iron concentration combined impact of key drivers was associated with lower statistical significance – combination of silica and alumina concentrations resulted in only 33% of predictive strength. The developed predictive model further confirmed statistical significance of impact of silica and copper content on iron concentration in slag. The generated decision tree is quite similar to the case of silica content in slag. The most significant predictors for iron content in slag streams are provided in the Table 2.6.

Table 2.6 – Statistical importance of predictors for iron content analysis

Predictor description	Predictor importance
Slag silica content	27 %
Slag copper content	19 %
Matte copper content	13 %
Slag alumina content	12 %
Slag quicklime content	11 %
Slag magnesia content	9 %
Slag cobalt content	9%

It is worth highlighting that among the most significant predictors for iron content is slag, copper concentration (primary product) in matte was recognized. Particularly, within the developed predictive model its impact was considered together with copper and alumina concentration in slag.

For alumina concentration in slag, the obtained results were similar to the previous analyses results. More specifically, iron and silica content were recognized as key drivers, yet predictive strength values for these parameters were relatively low – 19% and 15%. Combined impact of several parameters was approximately at the same level – 19% for silica and sulfur concentrations. In terms of predictive analysis, the generated model was characterized by only 40% of predictive strength. Taking into consideration high volume of available data, such a coefficient of predictive strength points out independence of alumina

concentration from content of other constituents in the slag stream. However, increased quantity of predictors and their importance related to matte characteristics should be taken into account. More detailed description of key predictors for alumina concentration in slag is provided in the Table 2.7.

Table 2.7 – Statistical importance of predictors for alumina content analysis

Predictor description	Predictor importance
Slag cobalt content	20 %
Slag iron content	18 %
Slag quicklime content	17 %
Slag sulfur content	12 %
Matte sulfur content	8 %
Matte cobalt content	8 %
Matte copper content	7 %

7. Comparative analysis of the Watson analytical capabilities

In order to conduct a comparative analysis of the Watson Analytics platform features with other available software applications it was decided to examine and evaluate the similar parameters and databases by an additional third-party program with the purpose of revealing possible advantages and drawbacks of the Watson's platform. For this reason, Modde Pro™ software by Sartorius Stedim Biotech® has been selected. This software provides an opportunity to apply a set of linear algorithms for multivariable data analysis to gain insights and analyze complex relations within large datasets.

Modde Pro – is an advanced application initially intended for efficient design of experiments, which is based upon a set of linear regression equations for data analysis. However, broad functionality of the program also facilitates data analysis for determination of main interdependencies and key drivers. In Modde interface, the analyzed dataset is divided into two sets of parameters – factors (influential parameters) and responses (variables influenced by factors). Most of the application's tools are developed for optimization of the response parameters and determination of their key statistical drivers. This way, Modde provides an opportunity to assess statistical impact of certain parameters and direct the "responses" towards the desired level by optimization of the corresponding factors. These factors can be characterized either as qualitative or quantitative. Modde supports limited quantity of factors – up to 32 in total, which significantly limits analysis of large datasets. In order to avoid possible mistakes in data handling, Modde has a built-in set of comprehensive raw data analytical tools. The collected and prepared data is used to develop a mathematical model (multiple linear regression or partial least squares based

models can be generated), which is capable of representing complex relations within the dataset between factors and responses. Similar to the most of data analytics tools, Modde also supports various features and services for efficient visualization to facilitate interpretation and evaluation of the results and to improve decision making. Apart from this, Modde also provides an opportunity to evaluate risks of certain setups and guide the user towards more robust and optimal settings to achieve the desired outcome. Modde Pro has been used to facilitate scientific research for many various projects in different areas: petrochemical industry, pharmaceuticals, medical research, pulp & paper, food and plastics industry (Wensley 2017).

7.1. Application of Modde Pro for analysis of moderate size dataset

7.1.1. Data preparation

Contrary to cloud-based Watson Analytics platform, Modde Pro is a conventional software application that should be operated from a computer and requires excessive processing power for efficient analysis (particularly severe requirements for CPU performance).

For the data preparation, similar approach to that described in section 5 has been applied, with one significant difference – once the dataset is uploaded into Modde Pro, all the empty spaces within the database should be deleted – otherwise further analysis is impossible. In other words, Modde Pro is extremely sensitive to the quality of the analyzed information and is not able to treat data with missing values. Additionally, after import and preliminary preparation of the data, variables in the database need to be sorted – more specifically, "experiment number" column should be selected as a basis (in our case – separate dates or time spans, when the process parameters and corresponding product characteristics have been analyzed) and target value should be assigned as a "response" prior to further analysis. At this stage it is advisable to determine the desirable upper and lower limits of the target variable value. For this reason, optimal composition of the matte and slag streams as stated in the manufacturing specification of melting department of Norilsk Nickel Copper Plant are provided in the Tables 2.8 and 2.9 respectively.

Table 2.8. Matte optimal composition

Compounds	Total	Cu	Ni	Fe	S	Others (mostly Co)
Cu_2S	70.6	56.5	-	-	14.1	-
FeS	19.0	-	-	12.1	6.9	-
Ni_3S_2	6.2	-	4.5	-	1.7	-
Fe_3O_4	2.9	-	-	2.1	-	0.8
Others	1.3	-	-	-	-	1.3
Total	100	**56.5**	**4.5**	**14.2**	**22.7**	**2.1**

Table 2.9. Slag optimal composition

Element / Compound	Lower limit, %	Upper limit, %
Cu	0.55	0.75
Ni	0.13	0.2
Fe	43.7	46.8
S	0.9	1.1
SiO_2	29	32
CaO	1.7	2.4
MgO	1.2	1.6
Al_2O_3	4.2	4.5

7.1.2. Key drivers analysis for single target value

After successful data import, a predictive model has been generated based on the available information and fitted by either MLR (Multiple Linear Regression) or PLS (Partial Least Squares) approach. The applied model type is also determined automatically considering structure and features of the analyzed dataset, but in contrast with Watson Analytics platform, this can be changed – however, it is advised to use the default model type.

Since the purpose of this comparative analysis is assessment of Watson Analytics platform with less complicated Modde Pro software, the similar variables have been analyzed. Therefore, in order to create the required predictive model copper content in matte has been set as a "response variable" specifying the optimal value. MLR model has been automatically generated with the following parameters:

- R2 (fraction response variation explained by the model) is equal to 0.995;
- Q2 (model predictive power) is equal to 0.988;
- RSD (Residual Standard Deviation – variation of the response not explained by the model) is equal to 0.1666.

The difference between R2 and Q2 is smaller than 0.3 and both coefficients are close to 1, thus the generated model can be considered as reliable (Eriksson, Eriksson 2000). In case when PLS model has been applied for fitting – the obtained results were less satisfactory. More specifically, coefficients R2 and Q2 were equal to 0.961 and 0.915 respectively, while RSD index was 0.4499. Therefore, default MLR model has been selected for further analysis. "Observed vs Predicted" plot showing model validity is represented in the Figure 2.14.

The most important statistical drivers for matte copper content in terms of Modde Pro interface were determined by means of coefficients plot, which provides graphical representation of the model terms significance. This representation is provided in the Figure 2.15. In terms of Modde Pro interface, the statistically significant predictive

coefficient is determined as one with a large distance from the zero level (either positive or negative) as well as having an uncertainty level that does not extend across the zero level.

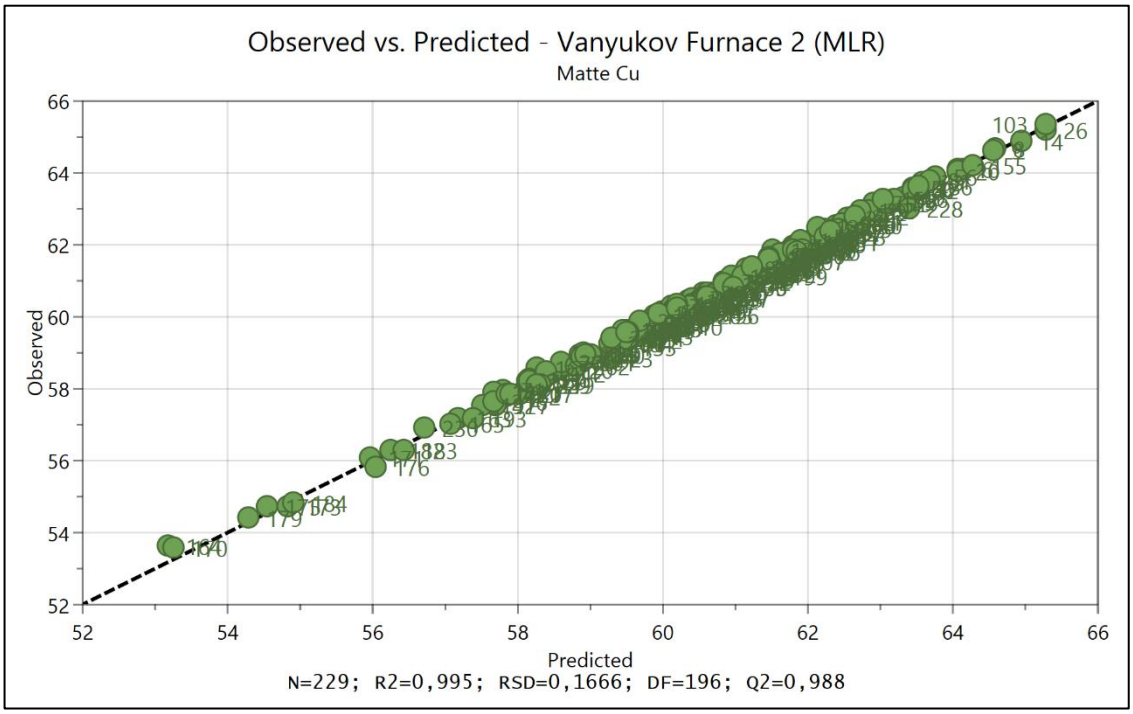


Figure 2.14. Observed vs Predicted plot for matte copper content model

The least significant drivers have been excluded from this analysis, therefore a slight increase in values of R2 and Q2 can be noted. Similar to the results obtained by means of Watson Analytics platform among the most significant predictive drivers the following variables have been recognized:

- concentrations of iron, nickel and sulfur in matte;
- concentrations of silica and iron in slag;
- cooling water temperature (both inlet and outlet) and pressure;
- air-oxygen mixture flowrate.

In order to get more detailed description of the key drivers' impact on the copper content in matte, a series of prediction plots has been generated for the most significant predictors. The results are provided in the Figures 2.16 (a-d). This plot not only visualizes the effect of the predictor on the target variable but also provides an opportunity to assess the confidence of prediction for various cases (dotted lines).

One significant assumption in terms of predictive analysis in Modde Pro is that the plot displays the target values of the selected response when the factor varies over its range, all other factors in the design are kept constant at their averages.

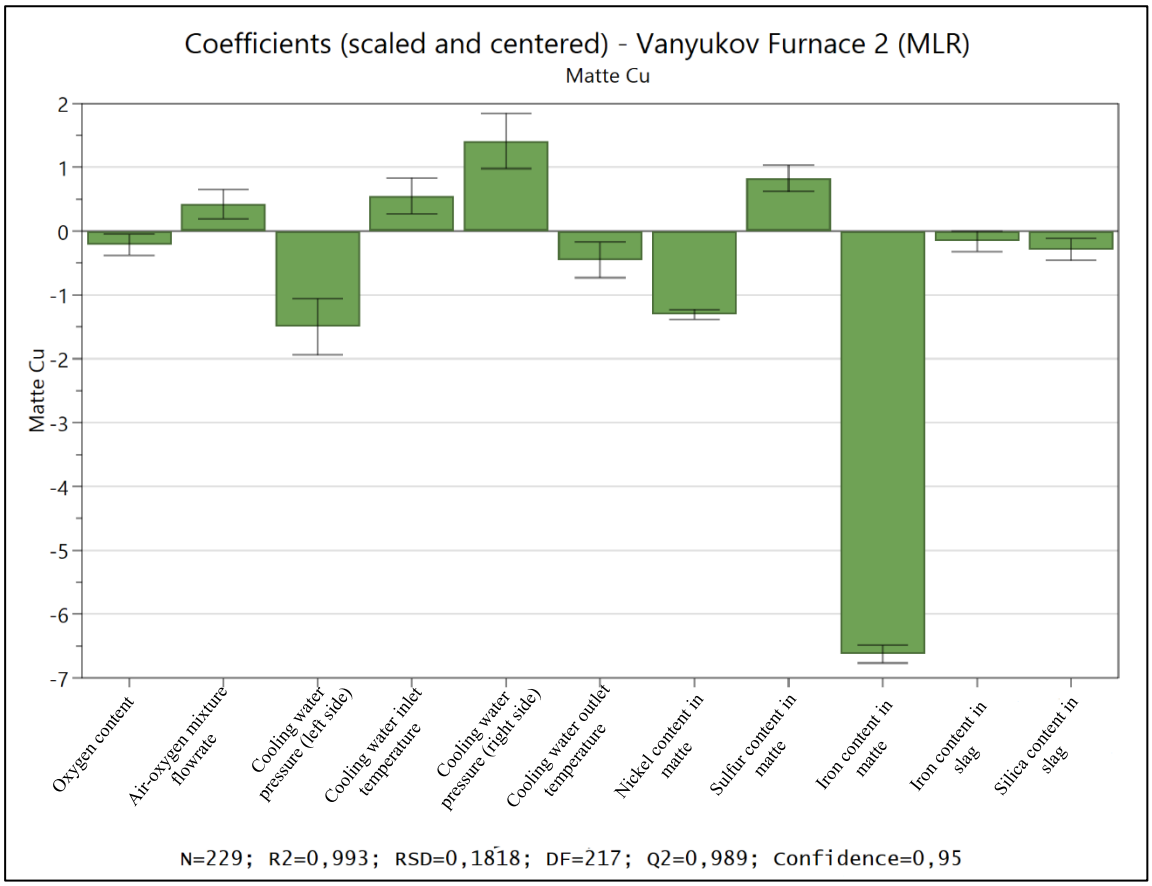


Figure 2.15. Predictive coefficients for matte copper content model

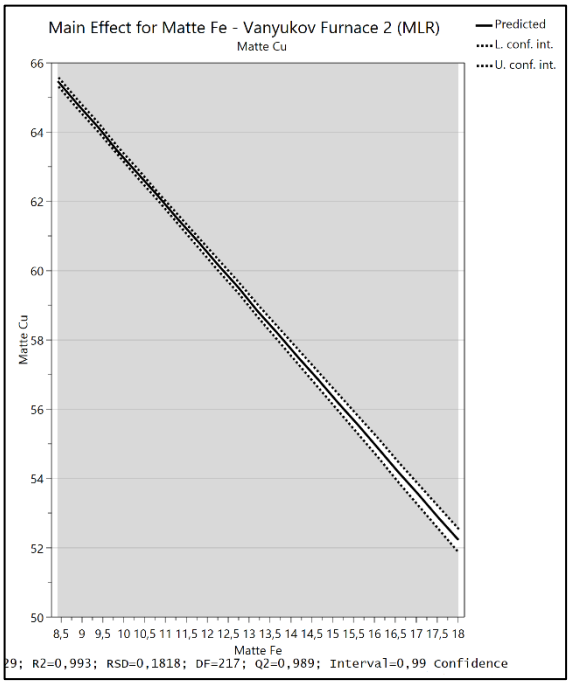


Figure 2.16 (a) Prediction plot for copper content in matte vs iron content in matte

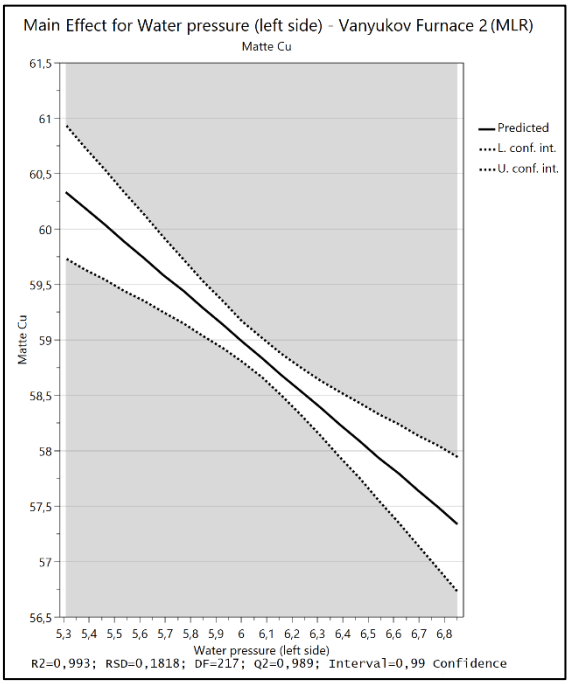


Figure 2.16 (b) Prediction plot for copper content in matte vs cooling water pressure

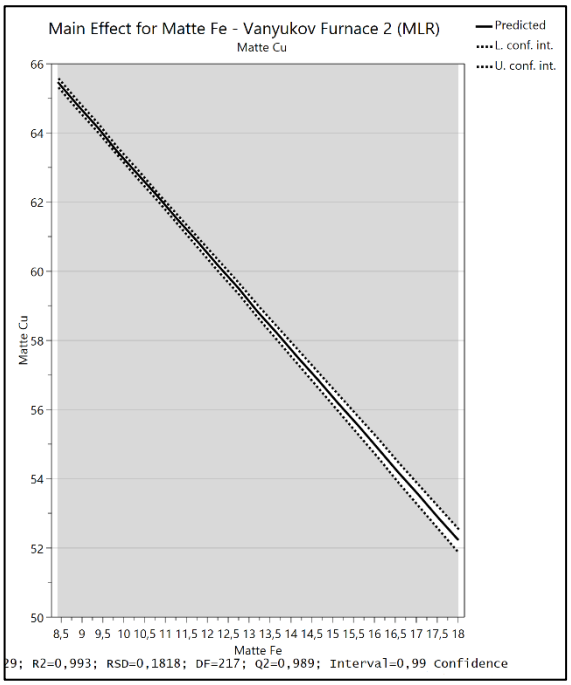


Figure 2.16 (c) Prediction plot for copper content in matte vs nickel content in matte

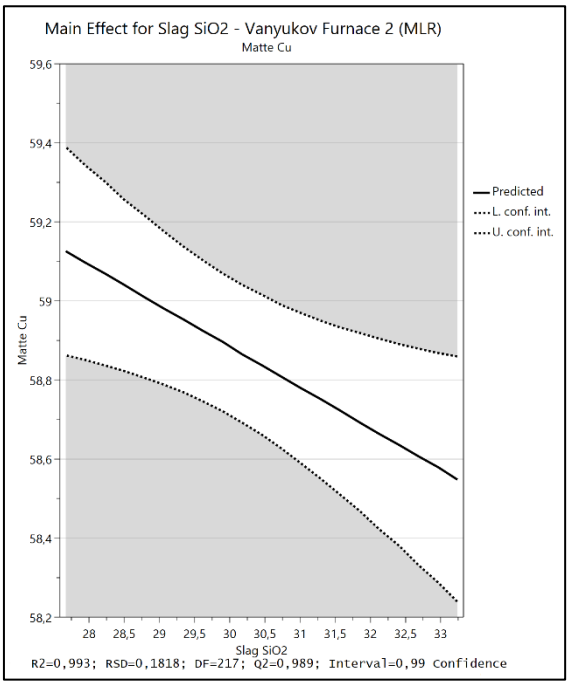


Figure 2.16 (d) Prediction plot for copper content in matte vs silica content in slag

In contrast with Watson Analytics platform, Modde Pro interface is developed to analyze information based on the optimal values of the target variable. More specifically, there is a number of tools available to determine the conditions under which the responses can be kept within the optimal limits with a certain level of probability. This way, copper content in matte has been studied depending on iron content and cooling water temperature (the most significant predictors). The obtained results in a form of graphical representation is provided in the Figure 2.17.

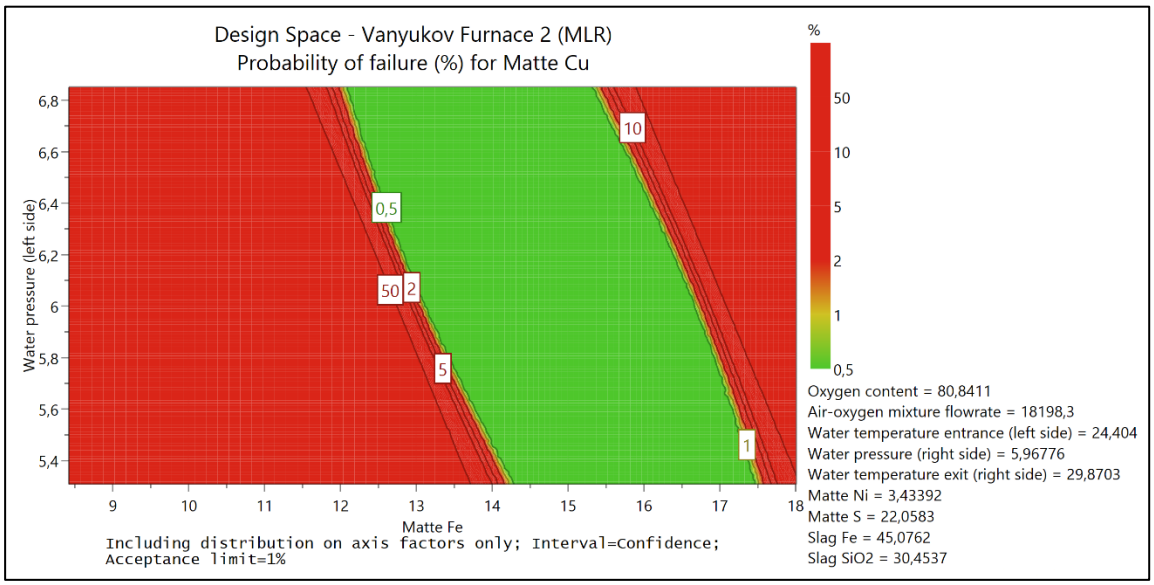


Figure 2.17. Design analysis for the copper content in matte depending on iron content in matte and cooling water pressure

According to the provided graph – required copper content in matte can be maintained within the green section with 0% failure probability, considering that other parameters (apart from iron content in matte and cooling water pressure) are kept constant. Deviation from the green section will result in a corresponding failure probability, as it is highlighted in the Figure 2.17. Results of this design analysis comply with the results of key drivers' evaluation and reiterate that iron content in matte has the most significant statistical impact.

7.1.3. Key drivers analysis for several target values

Modde Pro provides an opportunity to select several target variables at the data preparation stage. Therefore, to assess simultaneous analysis of several target variables copper, sulfur and iron concentrations in matte (the most important constituents) have been selected as "response" variables. Similar to the previous case, prediction model has been fitted with MLR regression method. More detailed description of model characteristics is provided in the Table 2.10.

Table 2.10. Summary of fit for the multivariable prediction model

Target variable name	R2	Q2	RSD
Copper concentration	0.844	0.786	0.892
Iron concentration	0.782	0.674	0.204
Sulfur concentration	0.857	0.807	0.671

As it can be seen from the obtained results, the model accuracy has been significantly decreased, when several target variables were set as responses. More specifically, RSD values increased drastically in comparison with previous case, while Q2 was decreased. However, R2 and Q2 values are negligible different and still close to 1, therefore, the generated model can be used for further analysis. Plots of observed and predicted values plots for each response variable are represented in the Figure 2.18.

In comparison with Figure 2.14 where the similar plot has been represented for single target variable, a significant deviation of the observed parameters from the generated model can be observed in case of several target variable analysis. Graphical representation of the most significant predictive drivers for each matte constituent is provided in the Figure 2.19.

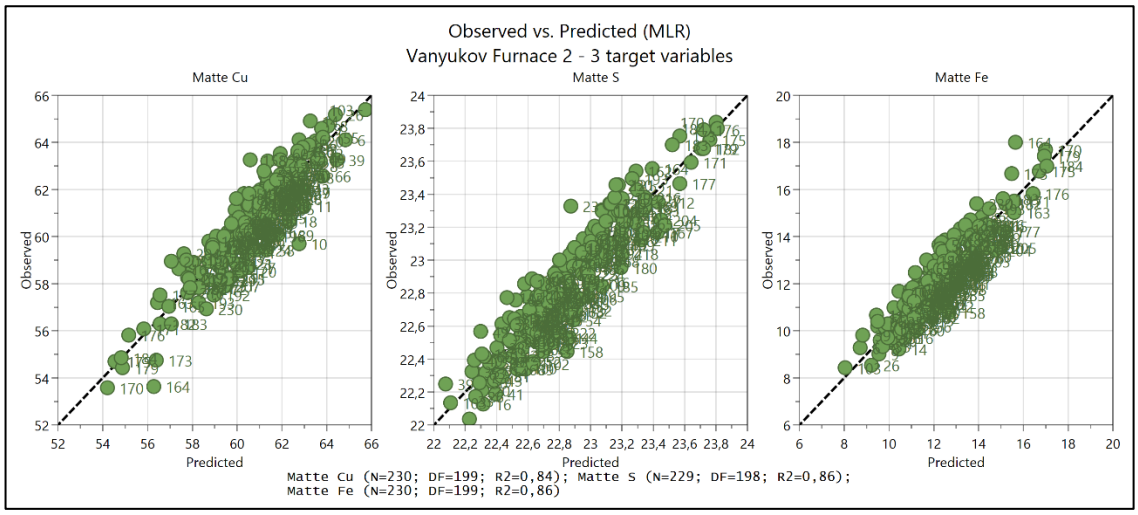


Figure 2.18. Observed vs Predicted plots for several target variables analysis

It is worth noting that variables, which were set as response parameters cannot be included and considered in model generation, thus, in terms of copper content in matte key drivers' analysis – iron concentration has not been recognized as an important statistical driver. Consequently, the same applies for iron and sulfur content values. Therefore, cooling water characteristics and oxidative agents supply rate have been considered as statistically significant predictors.

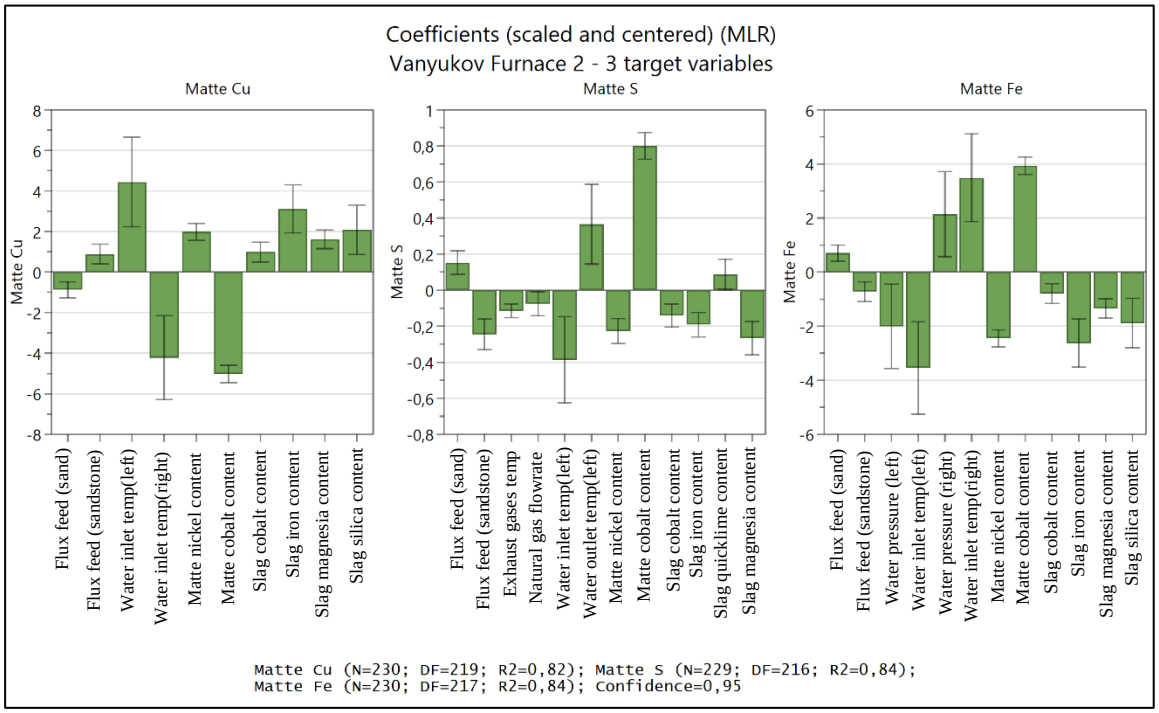


Figure 2.19. Predictive coefficients for matte copper, sulfur and iron content model

Additionally, a significant increase in uncertainty range in comparison with single target key drivers (Figure 2.15) should be noted. Furthermore, to evaluate the statistical importance of the most significant predictors (in this case – cobalt content in matte and cooling water pressure), a sensitivity analysis of their impact on optimal target variables values has been

conducted similarly with the previous case. The graphical representation of the obtained results is provided on the Figure 2.20.

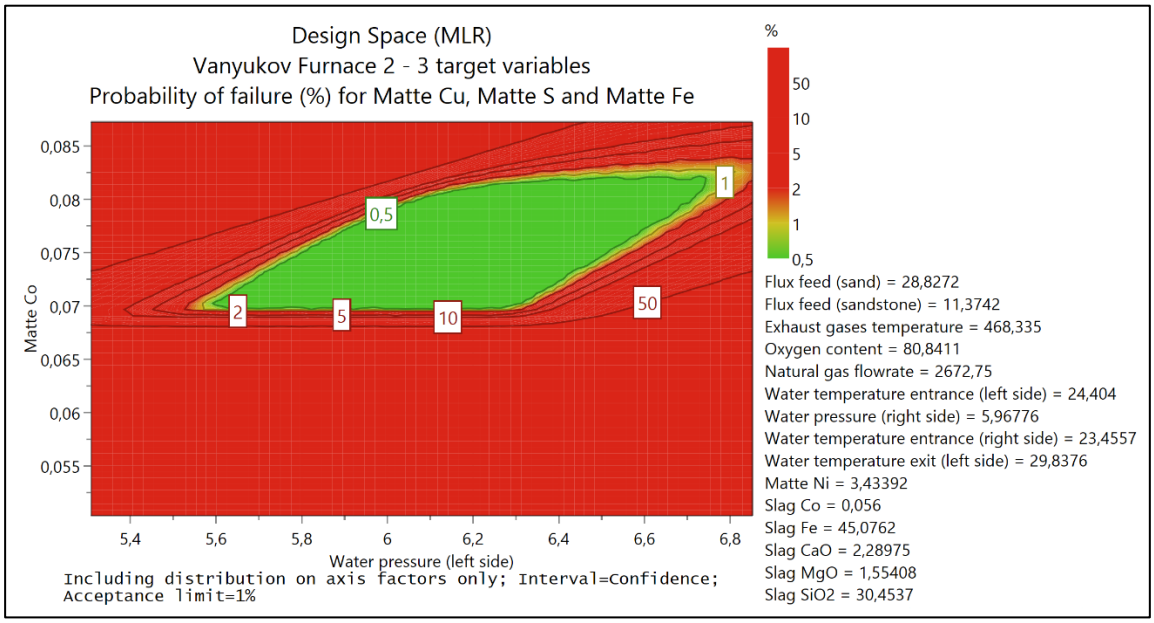


Figure 2.20. Design analysis for the copper, iron and sulfur content in matte depending on matte cobalt concentration and cooling water pressure

Therefore, it can be concluded that even moderate deviations of cobalt concentration in matte or cooling water parameters can cause a significant increase in failure probability.

7.2. Application of Modde Pro for analysis of considerable size dataset

In order to provide insights into Modde Pro ability to work with databases of significant volume, similarly with the section 5.3, database related to matte and slag composition with more than 50 000 information cells has been imported. As it was previously described Modde Pro is able to develop more accurate mathematical model with fewer target variables. Therefore, only copper concentration in matte has been set as a response parameter, while other constituents were selected as quantitative factors for target variable prediction. In a similar manner with the previous analyses, the default model has been generated with MLR regression method. Although the analyzed dataset contained plenty of information related to the most statistically significant driver for copper content in matte, namely iron concentration, the generated model is associated with slightly inferior characteristics:

- R2 coefficient equaled to 0.983;
- Q2 coefficient equaled to 0.981;
- Model validity coefficient was negative (-0.2), which can be explained by extremely small replicate error, due to large quantity of relatively similar values – more specifically, the model error was significantly larger than the replicate error (Bourdon

et al. 2013). Moreover, such a value of this coefficient can indicate extremely high sensitivity in the test (Sartorius Stedim Data Analytics 2017).

- Model reproducibility coefficient equaled to approximately 0.995;
- RSD coefficient was equal to 0.4225.

For better representation of the model accuracy, observed vs predicted plot is provided in the Figure 2.21.

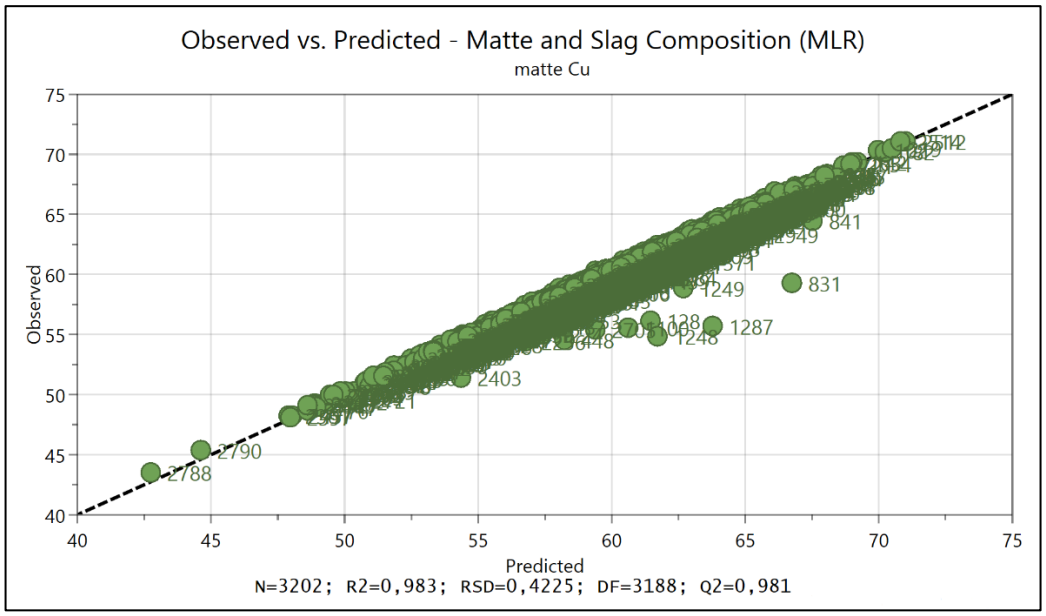


Figure 2.21. Observed vs Predicted plot for matte and slag composition model

Although, overall predicted values comply well with the observed points, a considerable deviation in several cases should also be noted. Main statistical drivers for copper content in matte are represented in the Figure 2.22.

In a similar manner to the preceding analysis of matte copper concentration, iron, nickel and sulfur content have been recognized as the most significant statistical drivers. However, key drivers' significance values assigned by the Modde were considerably higher and uncertainty range decreased as well.

In terms of simultaneous analysis of several target variables, the model accuracy has been severely decreased analogously with the previous case. However, with increased database volume deterioration of model characteristics, namely R2 and Q2, has been even more significant. Therefore, no further analysis was conducted, because of low model accuracy.

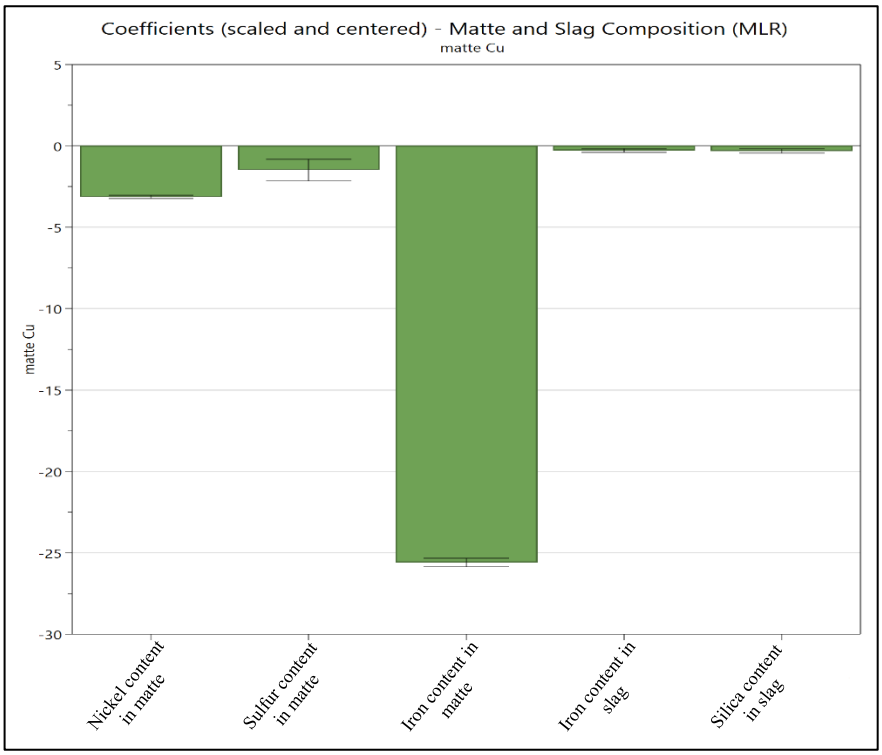


Figure 2.22. Predictive coefficients for matte copper content model with increased dataset volume

7.3. Key findings

Comparative analysis of the Watson Analytics platform with conventional software application for statistical analysis based on linear regression algorithms, namely Modde Pro by Sartorius Stedim Biotech®, revealed and highlighted several important features of the Watson platform.

- Cloud-based deployment of Watson Analytics platform offers the possibility to minimize the requirement for computing power, thus facilitating convenient and efficient data analysis on any device. In contrast, in case of data analysis with Modde Pro, some iterations (particularly design space analyses) were associated with significant time consumption and load on CPU (quad-core Intel Core™ i7-4710HQ CPU @ 2.50 GHz has been utilized).
- Determination of the most significant statistical drivers with Watson Analytics platform provides an opportunity to analyze the simultaneous effect of several factors for any of the variables. In case of data analysis with Modde Pro, linear regression models developed for simultaneous study of several target variables were associated with lower statistical accuracy, namely decreased R2 and Q2 coefficients and significant residual standard deviation.

- Watson Analytics platform is more tolerant to lower data quality. More specifically, databases with constantly repeated or missing values can be subjected to the analysis, yet with decreased accuracy. In contrast, Modde Pro requires perfect data quality (the application is particularly sensitive to the presence of missing values), therefore additional efforts have to be made for data preparation prior to further analysis.
- Watson Analytics facilitates more detailed analysis of key drivers by convenient interface and tools for data visualization, therefore providing an opportunity to evaluate impact of each factor on the analyzed variable.
- The interface of Watson Analytics platform is more user-friendly and does not require extensive background in statistics and data science or preliminary training for efficient data analysis. However, at the same time the platform has a closed system, that does not support any modification or alteration of the applied methods and algorithms. While Modde Pro offers the possibility to apply different regression models and change them at any time although the number of available options is limited.
- Data analysis with Watson platform provides more accurate and robust results when bigger datasets are analyzed. However, in terms of Modde Pro, when larger datasets have been studied – the model validity and accuracy experienced a slight decrease in comparison with moderate size data packages analyses.
- Convenient and multifaceted tools for predictive analysis with Watson Analytics offer a possibility to elaborately evaluate the impact of all available factors on the target variables by graphical representation of a decision tree, with description of corresponding decision rules.
- In comparison with Modde Pro, Watson platform does not support optimization of affecting statistical factors for adjusting the target variable. Data analysis in Modde Pro oriented more towards data optimization with the purpose of achieving required optimal values of the target variable, while Watson Analytics facilitates pattern and interdependencies recognition within the available data.

Overall, it can be concluded that Watson Analytics platform is more suitable for efficient analysis of large datasets with the purpose of revealing hidden interdependencies and patterns, also providing more detailed analysis of certain cases by convenient data visualization. On the other hand, a linear regression based Modde Pro facilitates analysis of moderate size data packages with the purpose of target variables optimization.

8. Discussion of the results

8.1. Aluminum dataset analysis

Analysis of dataset containing information regarding aluminum production and corresponding emission intensities and rates with Watson Analytics platform did not reveal any relevant patterns or interdependencies between utilized energy streams and volumes of emissions. The most significant problem that was encountered during data analysis with the platform is insufficient volume of available information and its low structuredness. More specifically, even though the platform was able to determine the most significant statistical drivers for emissions intensity – more detailed analysis was not possible due to missing values and incapability of the platform to divide the analyzed information set in accordance with the number of degrees of freedom. Furthermore, the analysis was hindered by the low coherence of the information in the dataset, since only general information about production rates and emissions was available. However, Watson Analytics revealed significant statistical relation with predictive strength of 60% between perfluorocarbons emissions intensity and amount of energy produced by coal, electricity and gas, while other sources of energy accounted for 56%. Relatively similar results were obtained for fluorides emission for production facilities based on pre-bake technology – namely, coal was associated with the most significant statistical impact of 60%, while electricity and gas accounted for 56% each. Therefore, it may be concluded, that in terms of aluminum smelting, increased consumption of coal for energy production resulted in higher harmful fluoride emissions in comparison with cases when energy was obtained by electricity or natural gas.

8.2. Matte composition analysis

Analysis of the copper concentrate smelting by submerged-tuyere technology provided more coherent results due to significant amount of available data. Matte product has been analyzed with the focus on copper content in order to determine possible patterns and interdependencies with slag composition and processing parameters. According to the analysis of the key statistical drivers, concentrations of iron and sulfur had the most significant influence on the copper content in matte with predictive strength values of 80% and 72% respectively. First of all, it can be explained by the fact that content of sulfur in matte depends largely on the concentration of non-ferrous metals – due to lower mass fraction of sulfur within oxides of these metals. Furthermore, the possible reason for this lies in various properties of feed compositions – large fraction of initial raw material represents sulfide concentrates. Therefore, variations in concentrations of sulfur and iron in the primary

feed may have a significant impact on concentrations of sulfur and copper in products since copper present mostly in the form of Cu_2S. Moreover, increased concentration of iron under the smelting conditions encourages oxidation of copper, thus resulting in formation of copper oxide (Davenport et al. 2002).

$$FeO + Cu_2S \rightarrow FeS + Cu_2O$$

The produced copper oxide tends to dissolve in slag, thus the overall copper content in matte is decreased. Graphical representation of the obtained result with Watson Analytics Platform in terms of iron concentration impact on copper content in matte is represented in the Figure 2.23.

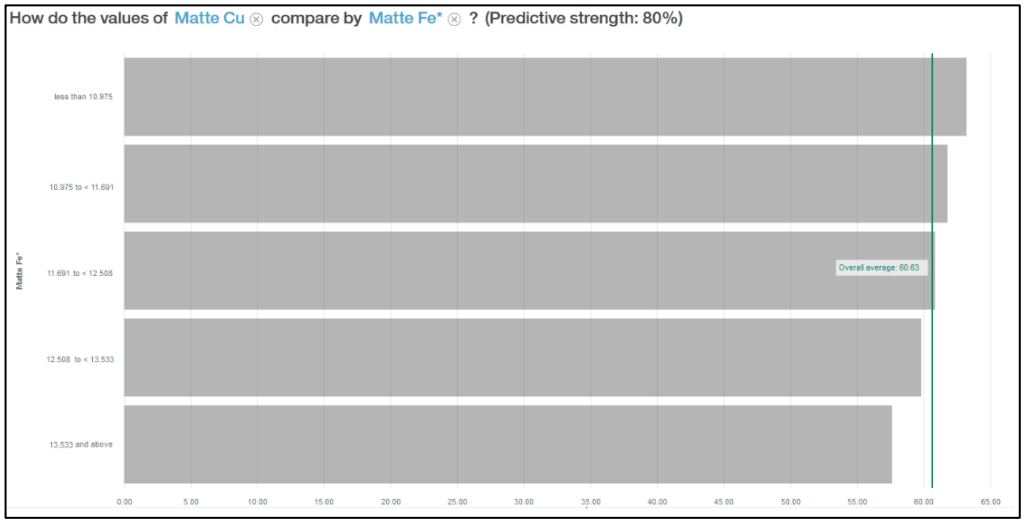


Figure 2.23. Impact of iron content in matte on copper concentration in matte

These results were further affirmed by additional analyses of similar databases with increased volume. However, in case of the analysis of the largest dataset containing information solely about matte and slag compositions the results were slightly different. Although iron and sulfur concentrations were recognized as the most important key drivers, the statistical significance of these parameters was reduced – more specifically to 72% and 56% for iron and sulfur respectively. Taking into consideration that pattern recognition algorithms of Watson Analytics platform work more effectively in terms of larger datasets analysis – it can be concluded that these results should be considered as more reliable.

In terms of multivariable analysis, simultaneous impact of several key drivers was associated with increased statistical significance in most cases. This way, simultaneous impact of iron and sulfur concentration was associated with 87% of statistical strength in all databases, which can be explained by the previously described aspects of the process. However, combined effect of iron concentration with such processing parameters as: natural gas and oxygen inlet flowrates, cooling water characteristics and oxygen concentration were

associated with increased impact – above 80%, while single effect of these parameters did not exceed 35%. It should be also mentioned that simultaneous impact of iron and nickel concentration was also associated with high statistical significance – 87%. Significant impact of oxygen concentration and inlet flowrate can be explained by its influence of oxidizing conditions, which strongly affects product composition. More specifically, oxygen flowrate accounted for the most significant impact on copper concentration in matte in case of high or average iron content (Figure 2.24). Similar results were obtained for the other parameters related to the oxidation process.

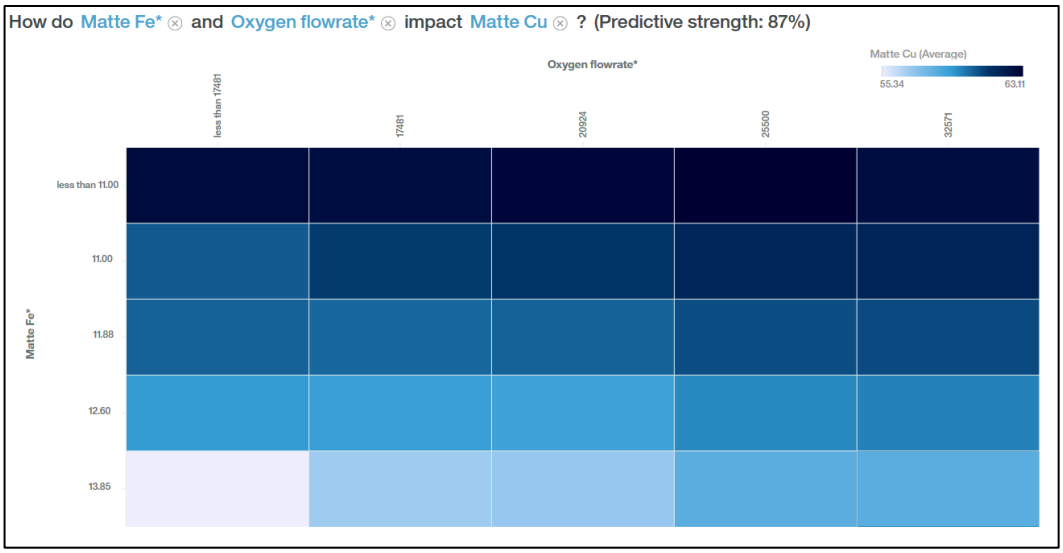


Figure 2.24. Copper content in matte depending on inlet oxygen flowrate and iron concentration in matte

Additionally, cooling water inlet and outlet temperature values were also associated with relatively high impact on the copper concentration. More specifically, increased inlet and outlet temperature were associated with decreased concentration of copper in matte. It can be concluded that increased heat withdrawal from the furnace resulted in slight decrement of matte grade (Figure 2.25).

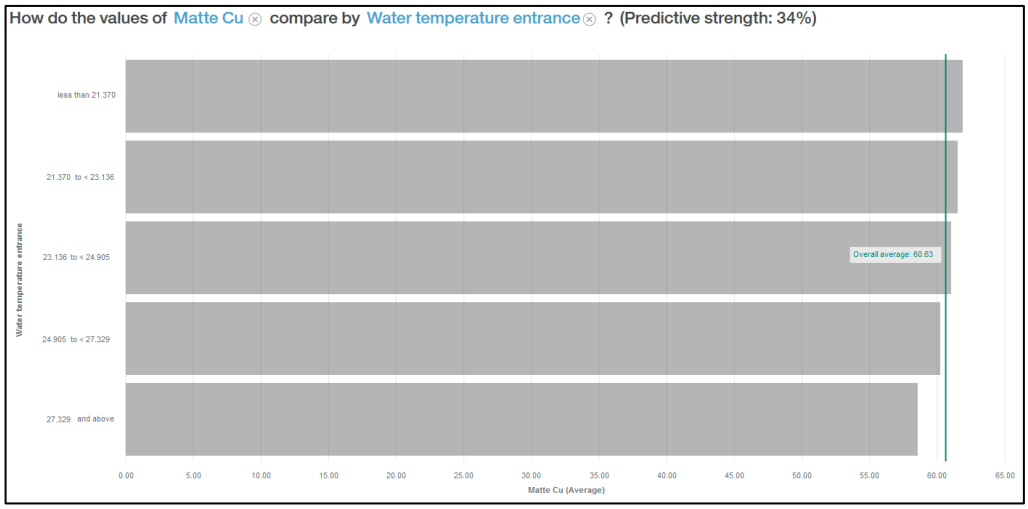


Figure 2.25. Copper concentration in matte depending on cooling water characteristics

Additionally, slight increase in matte copper concentration has been revealed with increase of slag copper concentration. Moreover, intensified deviation of slag copper concentration in case when matte copper content was beyond the optimal range should be noted (Figure 2.26). Similar to the previous case, it can be explained by deviation of the primary feed properties or processing parameters fluctuation.

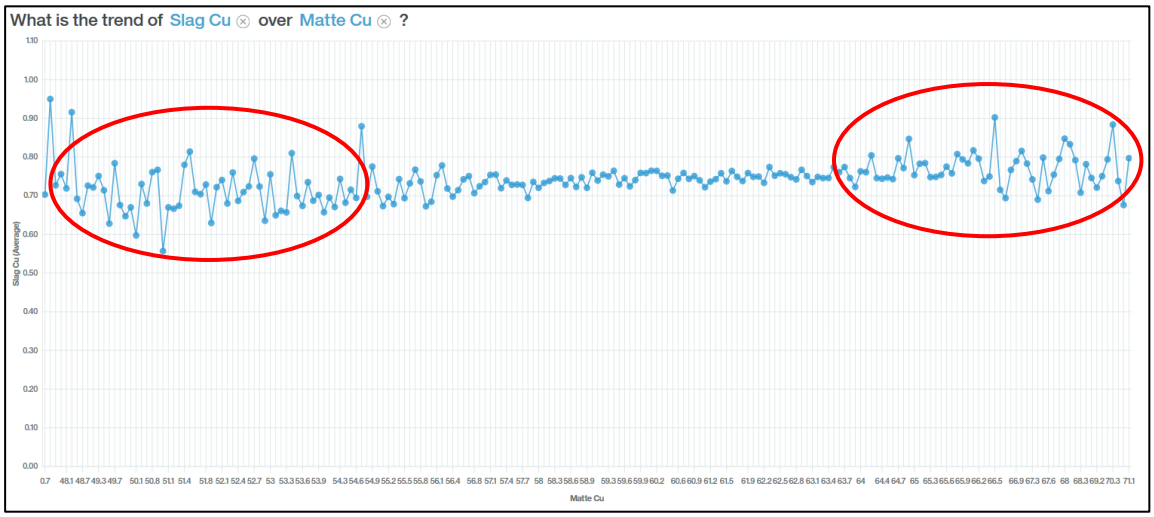


Figure 2.26. Copper concentration in matte vs copper concentration in slag

8.3. Slag composition analysis

Analysis of the slag composition was conducted with the focus on concentration of the most significant constituents – namely, iron, silica and alumina. Additionally, concentration of copper in slag was also studied. Among the most significant statistical drivers for silica content, concentrations of iron and copper in slag were recognized – with approximately 70% and 35% of statistical significance for the both constituents respectively. However, similar to the matte composition case – results obtained from the analysis of the largest dataset indicated lower parameters' significance – 62% and 33% for iron and copper respectively. More specifically, increased content of iron resulted in significant decrease of silica concentration (Figure 2.27).

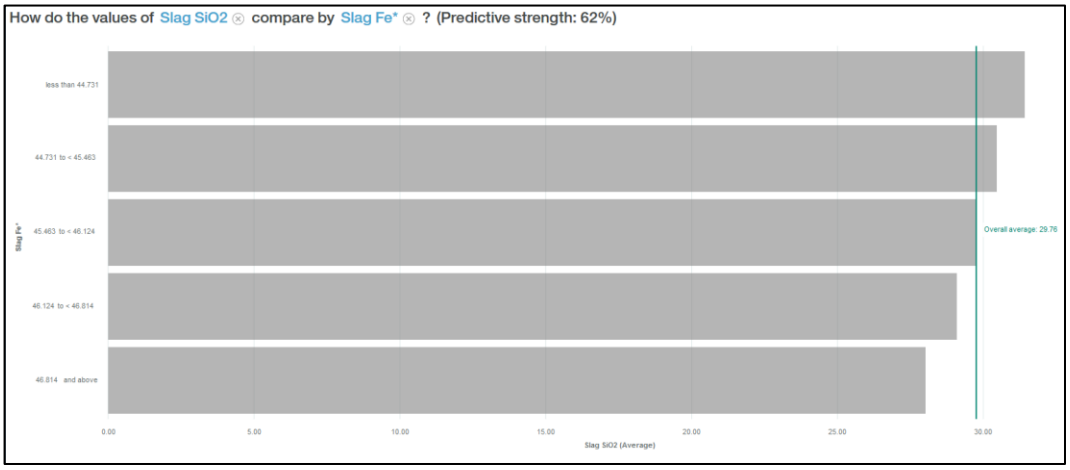


Figure 2.27. Average silica concentration in slag depending on iron content in slag

Other constituents were associated with negligibly small statistical impact on silica content. However, it is worth mentioning that even though the statistical impact of solely sulfur was associated with only 11% of the model strength, in terms of multivariable analysis combined impact of sulfur and iron or copper resulted in significant increase of statistical strength. More precisely, sulfur had a significant impact on silica content in cases of high and average iron concentrations (Figure 2.28).

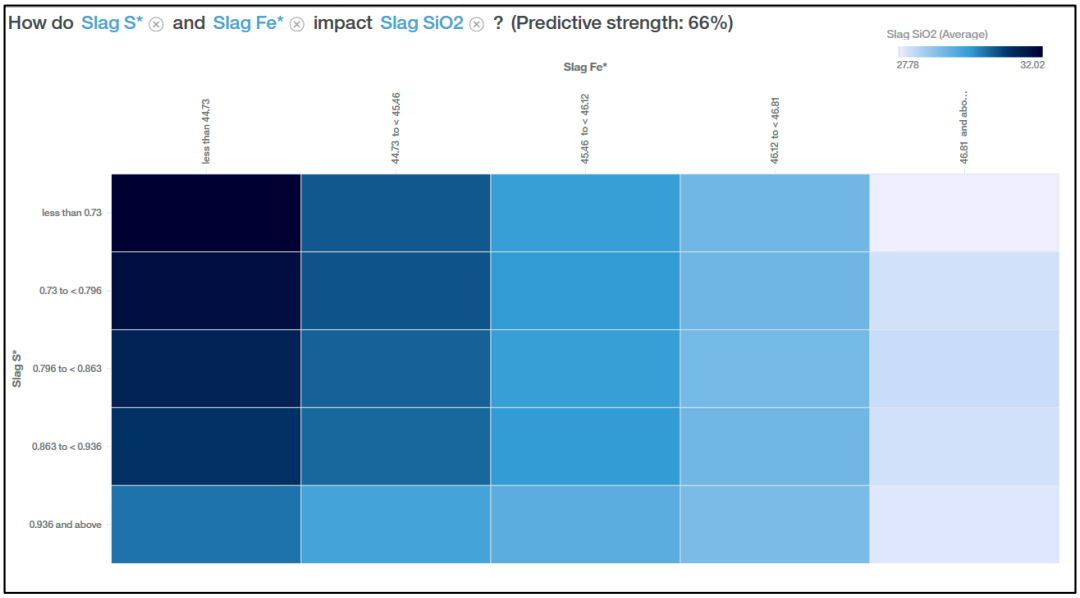


Figure 2.28. Silica content depending on slag iron and sulfur concentrations

Judging by the obtained results, represented on the above provided heatmap, it can be concluded that silica concentration in slag significantly increases with simultaneous decrease of sulfur and iron. As it was already described – increased oxygen flowrates and concentration improved oxidizing conditions, thus increasing silica content. In terms of processing parameters, as it was mentioned in the previous section – the most significant impact was associated with oxygen concentration and inlet flowrate. Cooling water characteristics (pressure and temperature) were recognized as relatively important drivers only in combination with iron content. According to the obtained results, increased heat withdrawal from the furnace resulted in decreased silica formation in slag in case of low and medium iron content (Figure 2.29). It is worth highlighting that single driver effect of cooling water temperature, without considering iron concentration, was not recognized as statistically significant by the platform.

Results of the predictive analysis conducted by Watson Analytics platform further affirmed statistical significance of the mentioned parameters. However, it should be mentioned that within the generated predictive model only slag constituents' concentrations were used for predictive analysis. More specifically, copper content was associated with higher predictor

importance in case of high iron content, while concentrations of cobalt and sulfur were considered to be more important in instances of low and average iron content.

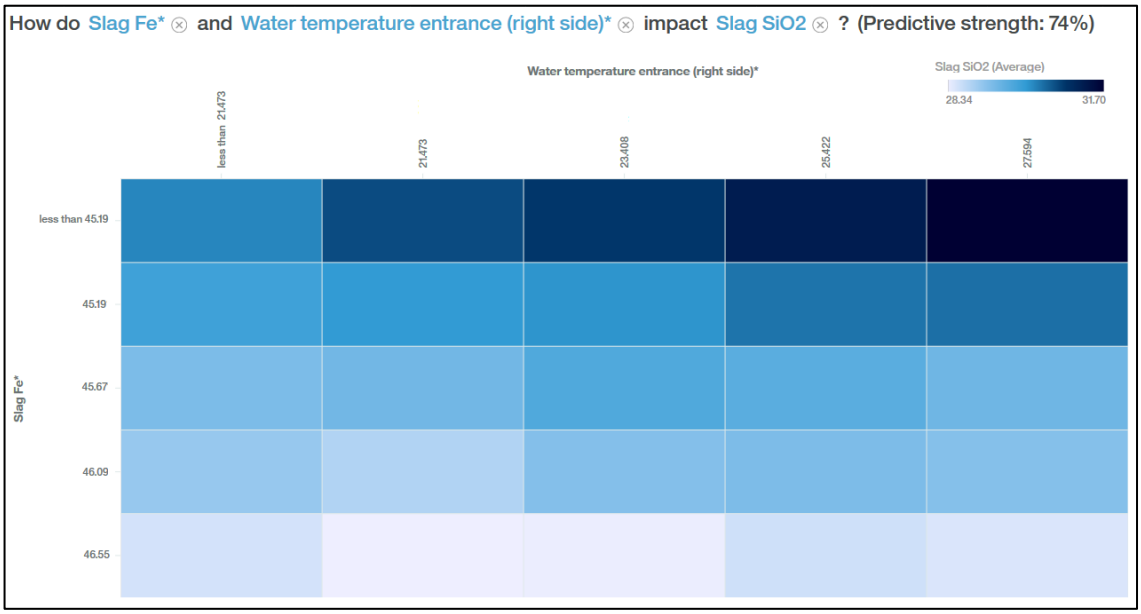


Figure 2.29. Silica content depending on slag iron concentration and cooling water temperature

Analysis of iron content in slag revealed similar results and further highlighted a strongly pronounced interdependence between iron and silica concentrations. However, in this case no relation with cooling water characteristics has been found and statistical impact of the other processing parameters varied greatly in different databases. Therefore, no coherent results regarding process parameters influence on iron content in slag were obtained. In terms of predictive analysis, concentrations of silica, alumina and copper in slag were associated with the most significant predictor importance. More specifically, alumina content was recognized as the most important predictor in case of lower silica content. Copper content had the most significant impact on iron concentration in slag in instances in which silica concentration was high, while cobalt content was associated with increased predictive importance in case of average silica concentration.

This clear correlation between iron and silica content in slag can be explained by the interaction between *FeO*, *FeS* and SiO_2 at temperature above 1200^{o}C and the nature of matte smelting process (Figure 2.30). As it can be seen from the partial phase diagram, silica and iron oxides form a system of two immiscible liquids in a certain region of concentrations during copper smelting. Increased silica addition results in widening of this immiscibility gap and thus iron concentration is changed (Davenport et al. 2002, Seetharaman 2014). However, possibility of impact of different raw material properties and composition also

should be taken into consideration, but within the limits of this work analyzed datasets contained information only about products composition.

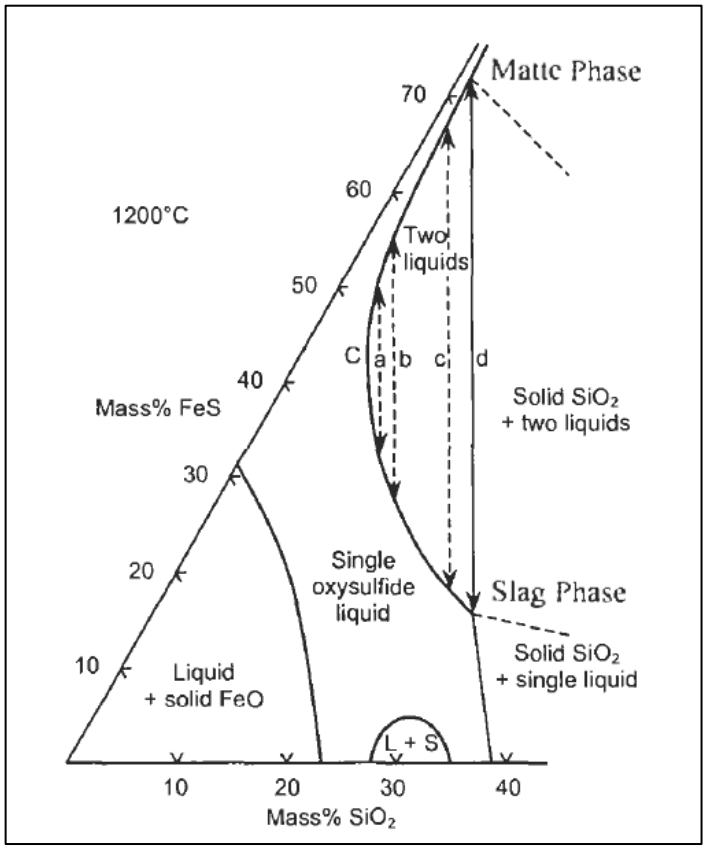


Figure 2.30. Matte and slag phase diagram (Davenport et al. 2002)

In terms of copper content in slag, among the most significant drivers fluxing agents' addition rate, iron, silica and nickel content were determined. Increased concentration of a nickel in slag resulted in drastic increase in copper content in slag.

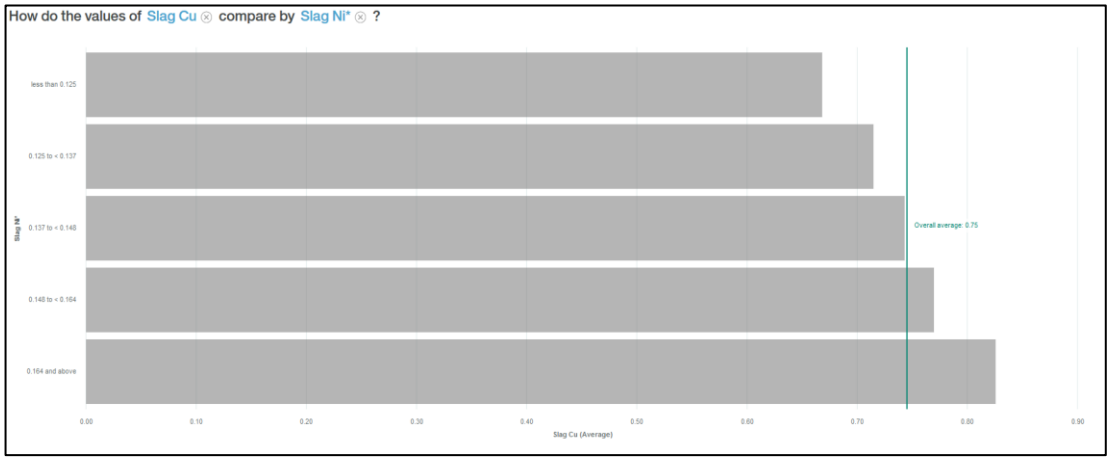


Figure 2.31. Copper content in slag depending on nickel concentration in slag

In a similar manner, increased iron concentration results in high copper content in matte (Figure 2.32). Decreased concentration of Fe_3O_4 improves slag fluidity thus facilitating matte droplets settling – therefore, copper content in slag is reduced (Davenport et al. 2002).

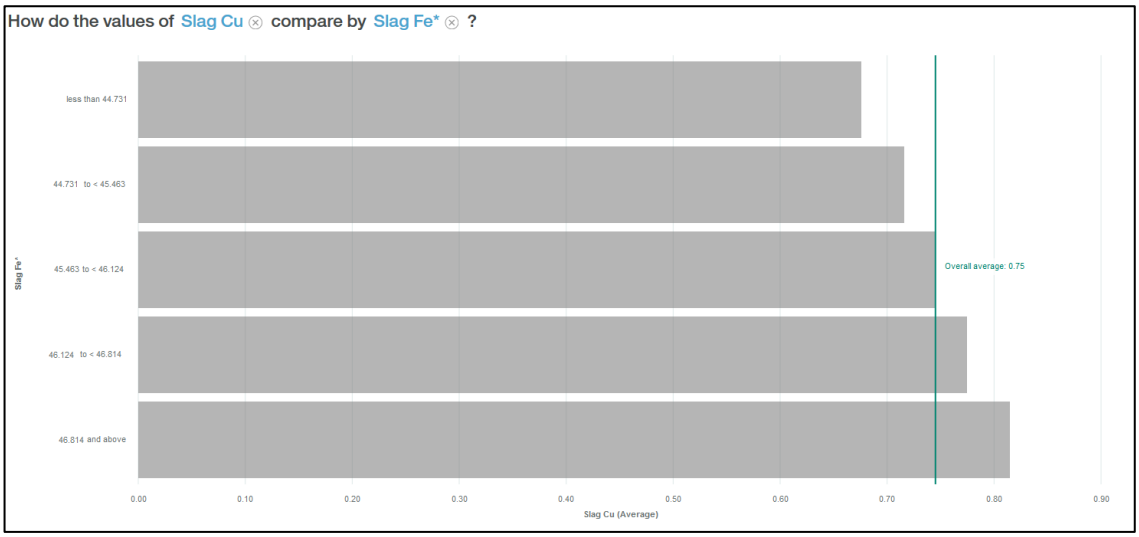


Figure 2.32. Copper content in slag depending on iron concentration in slag

Apart from this, multivariable analysis also revealed statistical significance on the combined impact of sulfur and iron concentrations, even though by itself sulfur was not recognized as an important driver. More specifically, decrease of sulfur concentration in slag in cases of high iron content resulted in significant increase of copper slag concentration.

Additionally, it should be mentioned that predictive strength values of some generated models for determination of copper content in slag were relatively low. However, according to the models' description provided by the Watson Analytics platform the *F value* is statistically significant. Therefore, the developed models can be considered as reliable.

In terms of processing parameters analysis, results obtained from the different databases varied drastically in a similar manner with previously described iron content analysis. It can be explained by slight difference among variables of the analyzed databases, presence of non-uniformities in data and insufficient amount of available information. Moreover, the information about processing parameters was limited and thus some of the crucial characteristics (volumes of produced matte and slags, temperature profiles and raw materials composition) were not available for analysis.

9. Unstructured data analysis with Watson exploration application

Furthermore, in order to investigate Watson's natural language understanding capabilities for classification and processing of information, an additional study has been conducted by application of Watson Discovery service for analysis of a collection of research articles related to specification and determination of side streams and by-products in metal industries. More detailed description of the collected data sources is provided in the Appendix 6. Based on the information gathered during the literature review, research papers relevant to novel technologies for treatment, reprocessing and potential reuse of waste and

by-products of nickel, lead, aluminum and zinc production industries, were collected with the purpose of further analysis.

Watson Discovery service is developed for automated unstructured information processing by applying integrated natural language understanding algorithms in a cloud-based platform. Apart from these features, the platform is also able to discover and determine the most relevant entities, detect anomalies hidden within the data and uncover relations and connections between the most important objects. Additionally, to facilitate data analysis Watson Discovery also provides an opportunity to use the public data – a constantly updated database currently containing more than 16 million pre-enriched documents from various industries and areas. This information is included into the platform's knowledge corpus in order to support querying of the insights – so that relevant news, trends or events can be integrated to facilitate the investigation.

Watson Discovery platform provides an opportunity to process data sources with the following extensions: .doc, .pdf or .html. However, .doc and .pdf files should be firstly subjected to the initial conversion prior to further analysis in the JavaScript Object Notation (JSON) format, which is achieved by built-in configuration service. Although, the platform supports analysis using default set of settings, it is recommended to customize the content processing configuration considering features and properties of the analyzed datasets and desired results.

9.1. Service configuration

9.1.1. Default configuration

Prior to content import, the service should be configured according to the necessary settings, so that information analysis is performed according to the objectives of the investigation. First step of the configuration involves specification of settings related to text processing and determination of entities and objects that should be extracted from the data. This can be achieved by uploading a series of sample files in order to define and adjust data conversion and enrichment algorithms based upon this information. It's crucial that configuration customization is performed prior to data import, because uploaded information is converted and pre-enriched according to the defined configuration. In case if the configuration setup will be changed after data import – these changes will not be applied for the already uploaded documents.

The default setup involves simple normalization and conversion of the data sources based on font sizes and text styles. Data enrichment, in this case, is based solely upon determination

of the most important entities, analysis of the sentiment and classification of the contained information in corresponding categories.

9.1.2. Configuration customization

Configuration customization is performed after initial private data collection generation. The default data ingestion configuration can be edited in three different sections:

- conversion – a set settings that determines how the documents are transformed;
- enrichment – a set of settings that determines which enrichments will be applied and what entities will be extracted;
- normalization – adjustment of the obtained results.

Discovery configuration architecture is represented in the Figure 2.33.

```
{
  "name": "Configuration Name",
  "description": "Descriptive text about the configuration",
  "conversions": {
    "word": {},
    "pdf": {},
    "html": {},
    "segment": {},
    "json_normalizations": []
  },
  "enrichments": [],
  "normalizations": []
}
```

Figure 2.33. Structure of the data injection configuration in Watson Discovery service

Basically, conversion is the first step of the data ingestion and in, our case, it involves conversion of data sources in .pdf files first to the HTML and afterwards to JSON format, as it is presented in the Figure 2.33. Each of these conversion steps is determined by a separate step of settings. Initial conversion from PDF to HTML is based upon a set of tags, that determines headings or subheadings of the documents. This is particularly useful to split the analyzed document in the way the author intended to separate the containing information.

Considering low structuredness of the collected research papers and irregular design of the documents within the limits of this analysis 2 tags were generated to improve PDF to HTML conversion:

- H1 – for specification the document title – bold font with size from 12 to 14;
- H2 – for specification the subheadings – bold font with size from 8 to 10.

Description of the applied configuration for the initial data ingestion process is presented in the Figure 2.34 as exemplified by the uploaded sample document.

```
{
  "extracted_metadata": {
    "title": "Aluminium salt slag characterization
and utilization – A review",
    "author": "P.E. Tsakiridis",
    "publicationdate": "2012-04-04"
  },
  "text": "Aluminium salt slag characterization and
utilization – A review\n\nJournal of Hazardous
Materials 217– 218 (2012) 1– 10\n\nContents lists
available at SciVerse ScienceDirect\n\nJournal of
Hazardous Materials\n\njou rn al h om epage:
www.elsev ier .com/ loc ate /
jhazmat\n\nReview\n\nAluminium salt slag
characterization and utilization – A review\n\nP.E.
Tsakiridis *\n\nDepartment of Mining and Metallurgical
Engineering, National Technical University of Athens
9, Iroon Polytechniou Street, 157 80 Zografou,
Athens, Greece\n\na r t i c l e i n f o\n\nArticle
history: Received 3 February 2012 Received in revised
form 15 March 2012 Accepted 16 March 2012 Available
online 27 March 2012\n\nKeywords: Aluminium salt slag
Properties, Utilization Waste processing\n\na b s t r
a c t\n\nAluminium salt slag (also known as aluminium
salt cake), which is produced by the secondary
aluminium industry, is formed during aluminium
scrap/dross melting and contains 15–30% aluminium
oxide, 30–55% sodium chloride, 15–30% potassium
chloride, 5–7% metallic a..."
}
```


Figure 2.34. Example of PDF to HTML conversion within the data ingestion process

Therefore, it can be seen that even at the initial data ingestion step, the platform accurately recognized and determined title, author and date of the publication. Taking into consideration, that apart from the mentioned entities the system did not recognize any other fields, additional customization of HTML to JSON conversion step is not productive.

Further step involves customization of the data enrichment process. Within the interface of the utilized platform, data enrichment involves incorporation of cognitive metadata which is based upon natural language processing and elements classification.

Since, the analyzed data collection contains mostly specific research related information, the default set of enrichments has been modified according to the aims of the research and dataset properties. For the text field sentiment analysis was discarded, while semantic roles determination, elements classification, keywords and relations extraction were added to the configuration. For other fields, no enrichment criteria were specified. Representation of the enriched sample document is provided in the Figure 2.35.

Figure 2.35. Example of enriched data within the data ingestion process

More specifically, applied data ingestion configuration can be described as follows:

- Semantic role – identification of the most important sentence objects or actions and determination of the relation between them;
- Entities – extraction of information related to people, objects, actions or places mentioned within the analyzed information set – facilitates understanding of the main subjects and the context of the analyzed dataset. This feature is based on a combination of natural language processing and statistical algorithms;
- Concepts – determination of the most significant concepts of the analyzed text, taking also in consideration other present concepts and entities – facilitates more detailed text analysis by identifying also relations and connections between recognized concepts, even if they are not directly mentioned within the analyzed text;
- Keywords – extraction of the most significant and relevant to the document's topic words within the analyzed text.

Normalization section was not edited within the limits of this study because of the low quantity of the recognized text fields.

9.2. Textual data processing

After customization of data ingestion configuration, a collection containing 75 scientific papers related to characterization and processing of by-products and side-streams of nickel, lead, aluminum and zinc production industries was uploaded into cloud drive of Watson Discovery service. Only documents containing less than 50 000 symbols can be processed. In case if larger data source is uploaded – only first 50 000 symbols can be processed and enriched by the platform.

After import of the documents, the platform provides an opportunity to evaluate the preliminary results of data enrichment. More specifically, among all the uploaded research documents the following related concepts were determined: “hydrometallurgy”, “zinc”, “iron”, “oxygen”, “acidity” and “sulfuric acid”. This indicates that most of the analyzed research articles related to by-products treatment are devoted to leaching technologies. It also should be noted that for every document the Discovery service accurately determined title, author and data of the publication. More detailed analysis of the most significant concepts performed by querying in a JSON form (“*term(enriched_text.concepts.text) .top_hits(10)”*) additionally highlighted such terms as: “aluminum”, “hydrochloric acid”, “lead”, “acid” and “PH”, thus confirming assumption regarding high quantity of articles describing leaching and extraction procedures.

Similar analysis, for the most common keywords of the analyzed dataset, performed in a JSON form (“*term(enriched_text.keywords.text,count:10)”*) provided the following results (from the most relevant down to the least relevant):

- Zinc;
- Sulfuric acid;
- Zinc leach residue;
- Blast furnace slag;
- Red mud;
- Zinc recovery;
- Sulfuric acid concentration;
- Solvent extraction;
- Temperature;
- Heavy metal recovery.

Results indicate a large number of research articles devoted to zinc processing and recovery. Additionally, it may be concluded that in case of zinc production waste processing, the applied terminology is more unified and clear in comparison with other research articles, which facilitates data enrichment by natural language processing algorithms. In a similar manner, these results highlight that a number of articles within the analyzed collection describe recovery of heavy metals from the metal industries residues.

9.2.1. Natural language processing improvement

Watson Discovery service also supports data analysis by asking queries in natural language form. However, relevancy of the obtained results, in this case, can be deteriorated, especially if the created collection contains highly specific information. In order to improve natural language processing algorithms and adjust the platform according to the gathered information, a special training should be conducted in order to increase the relevancy of the potential results. More specifically, this training involves generation of a set of example queries with corresponding answers, rated from most relevant to the least one. As a result,

due to machine learning algorithms of the service, the platform teaches itself from these specific cases and applies the “learned” patterns to all new queries. For successful improvement of the natural language capabilities as many queries as possible should be added and rated. Moreover, the platform can request additional queries if the information within the data is confusing or too specific. Furthermore, for efficient training it is crucial that training queries and desired answers have some term overlaps between them – this approach facilitates natural language recognition and machine learning algorithms, thus resulting in improved results relevancy.

Initial queries asked in natural language form without preliminary training provided unsatisfactory results. More specifically, the problem resided in confusion between analyzed metal industries so that, for example, when asked about aluminum production, the service provided results related to research papers about nickel and zinc. Additionally, lack of clarity and accuracy of results has been revealed during initial analyses – especially in cases when such terms as “side stream”, “waste”, “processing” and “utilization” were involved. Therefore, in order to overcome these problems and improve the relevancy of the results, the "learning process of the service" has been focused mainly on the clear separation of analyzed metal industries and “understanding” of the mentioned terms. This way, a set of input training queries was generated so that these aspects are taken into account. More detailed information about the training process is provide in the Table 2.11.

Table 2.11. Description of queries for the results relevancy training

№	Query	Results rating relevant / not relevant
1	*“Aluminum production side stream”*	8 / 15
2	*“Aluminum production waste”*	17 / 9
3	*“Aluminum production side stream processing”*	6 / 11
4	*“Aluminum production side stream characterization”*	4 / 6
5	*“Aluminum production waste processing”*	9 / 5
6	*“Aluminum production waste characterization”*	4 / 2
7	*“Nickel production side stream”*	6 / 16
8	*“Nickel production waste”*	6 / 5
9	*“Nickel production side stream processing”*	3 / 17
10	*“Nickel production side stream characterization”*	1 / 15
11	*“Nickel production waste processing”*	5 / 5
12	*“Nickel production waste characterization”*	4 / 5
13	*“Lead production side stream”*	3 / 11
14	*“Lead production waste”*	7 / 3
15	*“Lead production side stream processing”*	3 / 11
16	*“Lead production side stream characterization”*	4 / 8
17	*“Lead production waste processing”*	7 / 3

Continuation of table 2.11

18	*"Lead production waste characterization"*	3 / 3
19	*"Zinc production side stream"*	5 / 6
20	*"Zinc production waste"*	12 / 1
21	*"Zinc production side stream processing"*	4 / 5
22	*"Zinc production side stream characterization"*	4 / 6
23	*"Zinc production waste processing"*	11 / 1
24	*"Zinc production waste characterization"*	7 / 7
25	*"Aluminum production waste utilization"*	5 / 3
26	*"Nickel production waste utilization"*	4 / 3
27	*"Lead production waste utilization"*	4 / 5
28	*"Zinc production waste utilization"*	6 / 3
29	*"Slag processing"*	8 / 1
30	*"Slag characterization"*	7 / 1
31	*"Leach residue processing"*	7 / 2
32	*"Leach residue characterization"*	5 / 2
33	*"Heavy metals recovery"*	7 / 6
34	*"Aluminum recovery"*	4 / 2
35	*"Nickel recovery"*	2 / 6
36	*"Lead recovery"*	4 / 1
37	*"Zinc recovery"*	5 / 4
38	*"Iron recovery"*	5 / 5
39	*"Application of aluminum production side stream"*	6 / 4
40	*"Application of nickel production side stream"*	1 / 11
41	*"Application of lead production side stream"*	3 / 5
42	*"Application of zinc production side stream"*	2 / 4
43	*"Copper recovery"*	3 / 3
44	*"Cobalt recovery"*	5 / 2
45	*"Solvent extraction"*	7 / 0
46	*"Side stream characterization"*	6 / 4
47	*"Side stream processing"*	9 / 1
48	*"Waste characterization"*	4 / 1
49	*"Waste Processing"*	7 / 2
50	*"Waste application"*	5 / 1

For the analyzed data collection, the minimum required number of input queries determined by Watson Discovery service was 49. Therefore, after all the above described queries and corresponding results ratings were specified, the service began the process of relevancy training. Cloud based deployment of the service negates any load on the computer's CPU, therefore, even though the training process required a substantial amount of time – it was performed autonomously without any processing power consumption.

9.2.2. Data collection analysis

Once the platform finished the learning process, the data analysis section became again available. Therefore, the available information can be analyzed by means of queries in natural language form. Watson Discovery service facilitates data analysis in three different ways: documents search, results analysis and documents filter. Document search feature is based on cognitive search among the data collection with the purpose of finding the most relevant to the input query articles and passages, taking into account also related concepts and entities. Results analysis section provides an opportunity to evaluate the enriched data fields (such as concepts, keywords, semantic roles etc.) in a more detailed way by gathering top values, calculating average values or representing results in various forms (histograms, time slices). Document filter section facilitates filter of the results based on properties and contents of enriched data fields. This section also provides an opportunity to apply different conditions on the obtained results and classify and group the relevant passages.

Aluminum production industry

As it was previously described in the literature review, among the most significant side streams of aluminum production aluminum dross and red mud should be highlighted. During initial analysis it was revealed that some articles in the data collection contain information about processing and reuse of these waste streams. Therefore, after results relevancy training, the available dataset has been analyzed by asking the following question *"Efficient way to process wastes of aluminum production"*. Based on this query the service determined and rated the most relevant research articles, analyzed keywords and concepts for each document and also indicated the paragraphs, which contained the desired information. Visual representation on the obtained results exemplified by the 2 most relevant documents is represented in the Figure 2.36.

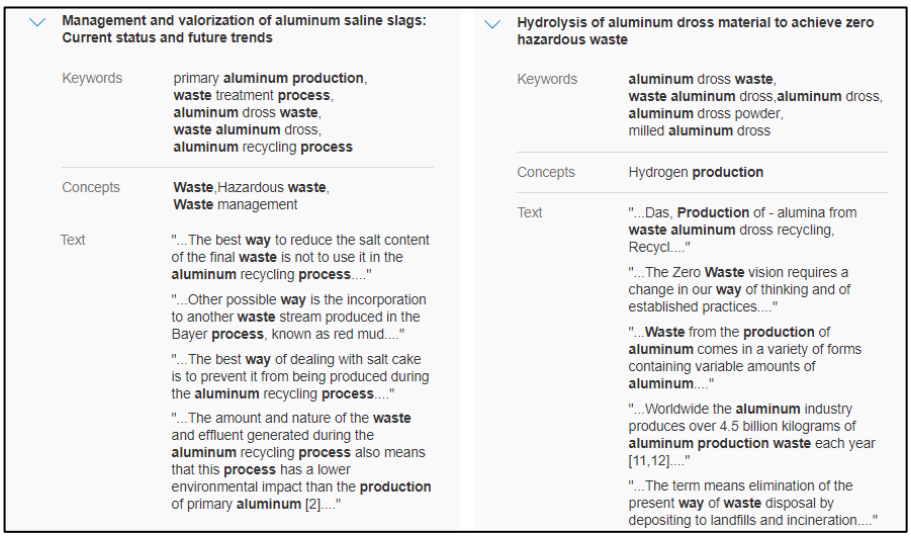


Figure 2.36. Aluminum industry wastes analysis results visualization

More detailed description of the obtained results is provided in the Table 2.12.

Table 2.12 Description of aluminum industry wastes analysis results

№	Research article title	Keywords	Key concepts
1	Management and valorization of aluminum saline slags: Current status and future trends	Primary aluminum Production waste treatment; Aluminum dross waste; Aluminum recycling process;	Waste; Hazardous waste; Waste management
2	Hydrolysis of aluminum dross material to achieve zero hazardous waste	Waste aluminum dross; Aluminum dross powder; Milled aluminum dross	Hydrogen production
3	A promising green process for synthesis of high purity activated-alumina nanopowder from secondary aluminum dross	Extraction process; Developed extraction process; Alumina extraction	-
4	Aluminum recovery as a product with high added value using aluminum hazardous waste	Aluminum dross processing; Waste aluminum dross; Dross recycling process	Waste; Hazardous waste; Aluminum Dross Tailings
5	Molten salt-enhanced production of hydrogen by using skimmed hot dross from aluminum remelting at high temperature	Hydrogen production process; Aluminum remelting process; Dross processing technology	Hydrogen production
6	Tailoring of magnesium aluminum titanate based ceramics from aluminum dross	Aluminum dross waste; Dross product	Aluminum Industry; Aluminum Nitride; Waste
7	Utilization of aluminum plant's waste for production of insulation bricks	Cleaner Production	Waste management
8	The solidification of aluminum production waste in geopolymer matrix	Large-scale aluminum production; Aluminum waste solidification; Industrial waste products	Waste management; Industrial waste; Hazardous waste

The articles in Table 2.12 are arranged by the platform in accordance with their relevance to the asked query. Therefore, judging by the provided results, it may be concluded that most of research technologies related to aluminum dross utilization are associated with hydrogen production, while also a lot of work is currently devoted to treatment of hazardous wastes with the purpose of toxicity reduction.

Nickel production industry

In a similar manner, analysis of nickel industry side streams has been conducted by asking the following query: "*Efficient way to process wastes of nickel production industry"*. Detailed description of the obtained results is provided in the Table 2.13.

Table 2.13 Description of nickel industry wastes analysis results

№	Research article title	Keywords	Key concepts
1	New Slag for Nickel Matte Smelting Process and Subsequent Fe Extraction	Nickel smelting process; Process Slag; Nickel matte	Nickel Smelting; Smelting reduction process
2	Geopolymer prepared with high-magnesium nickel slag: Characterization of properties and microstructure	Geopolymer Nickel slag; Nickel slag substitution; High-magnesium nickel slag;	Fly ash-nickel slag

Continuation of table 2.13

3	Iron, aluminium and chromium co-removal from atmospheric nickel laterite leach solutions	nickel losses; precipitation process; nickel laterite ores;	Nickel; impurity removal efficiency
4	Phase chemical composition of slag from a direct nickel flash furnace and associated slag cleaning furnace	Nickel oxide; Nickel flash furnace	Nickel smelting
5	Utilization of nickel slag using selective reduction followed by magnetic separation	Waste smelter slag; Reduction process; Magnetic separation process;	Nickel slag; Nickel grade

As reflected by the obtained results, most of nickel production waste processing techniques are intended for additional recovery of pure metals by means of various processes. However, it should be mentioned that some of the results were not included in the above provided table, since they were related to the other metal industries. It can be explained by relatively low quantity of relevant responses during results relevancy training in cases when nickel industry when mentioned (section 9.2.1.).

Lead production industry

Likewise, wastes of lead production industry were analyzed in a similar manner with the following query: *"Efficient way to process wastes of lead production industry"*. Results are gathered in the Table 2.14.

Table 2.14 Description of lead industry wastes analysis results

№	Research article title	Keywords	Key concepts
1	Application of granulated lead–zinc slag in concrete as an opportunity to save natural resources	Waste materials; Lead-zinc slag	Waste; Waste management
2	Iron extraction from lead slag by bath smelting	Reduction smelting process; Lead slag	Solid waste; Lead; Step smelting process
3	Reduction in toxicity and generation of slag in secondary lead process	Secondary lead process; Cleaner production; Waste minimization	Lead waste; Hazardous waste
4	Utilization of granulated lead slag as a structural material in roads constructions	Waste management; Zinc slag wastes; Lead slag	Lead; Lead ores; Lead slag
5	Leaching of lead slag component by sodium chloride and diluted nitric acid and synthesis of ultrafine lead oxide powders	Lead slag; Lead oxalate; Lead oxides	Lead-acid battery; Battery manufacturing process
6	Leaching of lead from zinc leach residue in acidic calcium chloride aqueous solution	Rapid leaching process; Oxide lead ores	-

Similar to the previous case, not all the results were included in the above provided table, because of the connection to the other metal industries. However, it can be stated, that most of the novel methods for lead industry waste processing techniques are targeted at minimization of wastes toxicity and poisonousness. Additionally, potential for application of lead processing wastes in construction and roadmaking should be highlighted.

Zinc production industry

Finally, according to the previously described methodology zinc production wastes and side streams were analyzed. The obtained results as provided in the Table 2.15.

Table 2.15 Description of zinc industry wastes analysis results

№	Research article title	Keywords	Key concepts
1	Use of fly ash, phosphogypsum and red mud as a liner material for the disposal of hazardous zinc leach residue waste	Zinc production industry; Red mud waste; Leach residue waste	Zinc; Waste; Hazardous waste; Waste management
2	Recovery of iron from zinc leaching residue by selective reduction roasting with carbon	Zinc; Zinc ferrite; Zinc oxide; Roasted zinc	Zinc; Iron; Roasted metal products
3	Innovative methodology for comprehensive utilization of high iron bearing zinc calcine	Mass zinc production; Zinc calcine; Zinc ferrite	Zinc oxide; Zinc sulfate; Zinc sulfide
4	Leaching and selective zinc recovery from acidic leachates of zinc metallurgical leach residues	Zinc metal production; Zinc	Zinc Study Group
5	Reductive leaching of cobalt from zinc plant purification residues	Zinc plant purification; Total cobalt production	-
6	Selective leaching of zinc from hazardous As-bearing zinc plant purification filter cake	Zinc; Zinc plant; Zinc powder; Metallic zinc; Zinc plant residue	Cobaltic waste products; Alkaline electrowinning
7	Acid leaching kinetics of zinc plant purification residue	Leaching process; Reaction kinetics	Zinc reaction; hydrometallurgical recovery

Taking into consideration the achieved results, it can be concluded that in most cases processing techniques of zinc production wastes are targeted at additional recovery of such metal as iron, arsenic, and zinc. Also, it should be highlighted that most of zinc production wastes are processed involving leaching.

9.3. Discussion of the results

The conducted analysis of the data collection, consisting of research articles related to characterization and processing of wastes and side streams of nickel, lead, aluminum and zinc production industries, affirmed feasibility of application of cognitive search technologies for efficient scientific data analysis. During data ingestion step Watson Discovery service was able to accurately determine articles titles and subtitles, key words, concepts and sematic relations between the most significant objects within the analyzed text. However, textual data analysis without preliminary relevancy training resulted in confusion of metal industries and in some cases even mix up of basic terms. Built-in results relevancy training feature based on machine learning algorithms provided an opportunity to effectively negate this issue and conduct further investigation by queries in natural language form. Overall, after all the preparations and configuration customization, the cognitive search platform was able to effectively find relevant documents and specific paragraphs, which contained the necessary information.

As it was described hereinbefore, based on the collected research articles, in most cases proposed processing techniques of the metal industries production wastes are targeted primarily at the reduction of toxicity and hazardous properties of the wastes. In terms of aluminum production, according to the results provided by the platform, the most relevant waste processing methods were associated with hydrogen production. Regarding nickel production industry – most of the research articles were devoted to additional recovery of various metals (mostly iron). In case of lead production, the most relevant results were associated reduction of hazardous by-products and application of wastes in construction industry. Analysis of zinc production industry revealed that most of the research articles are related to novel leaching techniques with the purpose of additional recovery of other metals.

Although, Watson Discovery service is designed for processing of unstructured information, analysis of the collected dataset has been deteriorated due to non-uniformity of research articles layout, presence of cross-references and literature references, large quantity of tables and figures. Moreover, analysis has also been deteriorated by great number of various specific terms and expression, incidental to metal industries. Additionally, it can be concluded that in instances when more documents related to one topic were collected – the obtained results and the relevancy training process were more accurate and logical.

10. Conclusions

Applicability of cognitive computing technologies for analysis of data related to metal production industries has been studied within the scope of this work by the example of IBM Watson application suite. Detailed numerical data describing process conditions and corresponding products properties in terms of aluminum and copper production industries has been studied by means of IBM Watson Analytics platform. Data collection consisting of up-to-date research articles devoted to processing of metal industries by-products and waste streams has been analyzed by the instrumentality of IBM Watson Discovery cognitive search service.

In terms of numerical data analysis with Watson Analytical platform, a series of algorithms has been developed for preliminary data preparation to facilitate further investigation. The obtained results highlight the crucial importance of substantial volume and quality of the available for analysis information. This way, in some cases no coherent and consistent results were obtained due to low amount and poor quality of the available data. However, in case of copper production related data analysis, the analytical platform proved to be very efficient in terms of determination of key statistical drivers and revealing hidden interdependencies between main parameters. More specifically, the obtained results highlight significant correlation between main matte and slag constituents' concentrations. As for the operational parameters – inlet oxygen flowrate and concentration along with cooling water characteristics during copper smelting process were associated with the most significant impact on product properties. Additionally, in order to assess Watson Analytics possibility to group and classify information, a comparative analysis of the analytical platform with the conventional software based on linear regression algorithms has been conducted. Results indicate, that compared with the conventional data analysis application, Watson analytical platform proved to be more efficient and accurate in cases when bigger datasets were analyzed. Moreover, Watson Analytics is more tolerant to lower data quality and facilitates data analysis by simultaneous determination of key statistical drivers, taking all the parameters into consideration at once, while linear regression based software application is less flexible in this regard. In general, the obtained results from both application were quite similar, even though in case of conventional software – the developed models were characterized with lower statistical accuracy.

In terms of unstructured data processing, data collection containing research articles devoted to processing and reuse of wastes and side streams of nickel, lead, aluminum and zinc production industries, has been prepared and analyzed by means of Watson Discovery

service. The information enrichment and conversion algorithms applied during data ingestion step have been customized according to the aims of the research, considering main properties and features of the analyzed dataset. Furthermore, in order to facilitate data analysis by means of natural language queries, the results relevancy training based on machine learning algorithms has been performed in order to improve understanding of specific terms and text context. As a result, Watson Discovery service efficiently processed unstructured data and determined the most significant keywords, objects and concepts and also recognized semantic relations between them. Through the instrumentality of the Watson Discovery service's cognitive search algorithms the most relevant scientific articles and paragraphs related to efficient way of wastes processing in the metal industries were determined among other documents in the collection. The obtained results indicate that most of the modern research projects are primarily devoted to the reduction of toxicity and hazardous properties of the discarded side streams and by-products.

Although, the studied data analysis platforms require some sort of preliminary data preparation depending on datasets' properties and in some cases processing of industry-specific unstructured information is associated with confused results and decreased accuracy, application of these technologies is extremely relevant in terms of modern scientific research. Furthermore, taking into consideration an incredible amount of available scientific knowledge which is constantly increasing at an exponential rate, the crucial importance of the studied technologies cannot be overrated. Constant improvement of these services aimed at rendering of the mentioned issues also increases importance and significance of novel cognitive data processing software.

References

A. Oord, T. Walters, T. Strohman 2017, , *WaveNet launches in the Google Assistant*. Available: https://deepmind.com/blog/wavenet-launches-google-assistant/ [2018, 01/14].

Arbot Solutions Inc 2017, , *Superior Intelligence with Cognitive Automation*. Available: https://coseer.com/content/white-paper-superior-intelligence-with-cognitive-automation/ [2018, 01/16].

Atomian 2016, , *Atomian technology*. Available: http://www.atomian.com/pdf/atomian-technology-EN.pdf [2018, 01/16].

Battle, T., Srivastava, U., Kopfle, J., Hunter, R. & McClelland, J. 2014, "Chapter 1.2 - The Direct Reduction of Iron" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 89-176.

Beddoes, J. & Bibby, M. 2003, *Principles of Metal Manufacturing Processes*, Elsevier ButtenNorth-Heinemann, Linacre House, Jordan Hill, Oxford.

Bernasowski, M., Klimczyk, A. & Stachura, A. 2017, "Overview of Zinc Production in Imperial Smelting Process", *Iron and Steelmaking*.

Bindi, T. 2017, , *BP Invests $20M into AI Startup Beyond Limits to Transform Oil and Gas Industry*. Available: https://www.beyond.ai/news/bpinvestment [2018, 01/16].

Blue Hill Research 2017, *ROSS Intelligence Artificial Intelligence in Legal Research*, ROSS Intelligence & Artificial Intelligence in Legal Research.

Bourdon, F., Lecoeur, M., Duhaut, M., Odou, P., Vaccher, C. & Foulon, C. 2013, "A validated micellar electrokinetic chromatography method for the quantitation of dexamethasone, ondansetron and aprepitant, antiemetic drugs, in organogel", *Journal of Pharmaceutical and Biomedical Analysis*, vol. 86, pp. 40-48.

Chen, T. 2012, *Honorary Symposium of Hydrometallegy, Electrometallurgy and Materials Characterization*, John Wiley & Sons, Inc, Hoboken, New Jersey, United States.

Chen, Y., Argentinis, E. & Weber, G. 2016, "IBM Watson: How Cognitive Computing Can Be Applied to BigData Challenges in Life Sciences Research", *Clinical Therapeutics*, vol. 38, no. 4, pp. 688--702.

CognitiveScale Inc. 2018, , *Cortex Augmented Intellegence Platform*. Available: https://www.cognitivescale.com/products/#platform [2018, 01/14].

Contractor, D. & Telang, A. 2017, *Applications of Cognitive Computing Systems and IBM Watson*, 1st edn, Springer Nature Singapore Pte Ltd.

Coseer Inc. 2016, , *Artificial Intellegence and Natural Language Processing*. Available: https://coseer.com/content/white-paper-superior-intelligence-with-cognitive-automation/ [2018, 01/16].

Creedy, S., Glinin, A., Matusewicz, R., Hughes, S. & Reuter, M. 2013, "Ausmelt Technology for Treating Zinc Residues", *World of Metallurgy*, vol. 66, no. 4, pp. 230-235.

Crundwell, F.K., Moats, M.S., Ramachandran, V., Robinson, T.G. & Davenport, W.G. 2011, *Extractive Metallurgy of Nickel, Cobalt and Platinum Group Metals*, Elsevier, Oxford.

Davenport, W., King, M., Schlesinger, M. & Biswas, A. 2002, *Excractive Metallurgy of Copper*, 4th edn, Pergamon.

Digital Reasoning Systems, I. 2017, , *Digital Reasoning Proven Solutions*. Available: http://www.digitalreasoning.com/proven-solutions-content/financial-services-tab#content [2018, 01/16].

Draper, N. & Smith, H. (eds) 1998, *Applied Regression Analysis*, 3rd edn, John Wiley & Sons, Inc, Canada.

Eriksson, L. & Eriksson, L. 2000, *Design of experiments : principles and applications*, Umetrics, Umeå.

Esposito Vinzi, V., Chin, W.W., Henseler, J. & Wang, H. 2010, *Handbook of Partial Least Squares : Concepts, Methods and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Expert System Enterprise 2018, , *Software that reads the wat people do. Company profile*. Available: https://cdn2.hubspot.net/hubfs/2037777/Company_Profile/ENG/Company_Profile.pdf [2018, 01/14].

Expert System Enterprise 2017a, , *Cogito: Human Intelligence for Cognitive Computing*. Available: http://www.expertsystem.com/products/cogito-cognitive-technology/ [2018, 01/14].

Expert System Enterprise 2017b, , *Smart data in smarter oil & gas industry*. Available: http://cdn2.hubspot.net/hubfs/2037777/Brochure/Oil_Gas_brochure_Cogito.pdf?__hssc=79855176.7.1515936030055&__hstc=79855176.74be865b786503f8a419bf47fe55abd2.1515936030055.1515936030055.1515936030055.1&__hsfp=1364048905&hsCtaTracking=4784f352-a81c-4e64-9c2a-a2d7c299b308%7Ce4dcf9d1-a46e-4830-9196-1c94e54f98ca [2018, 01/14].

Fabozzi, F., Focardi, S., Rachev, S., Arshanapalli, B. & Hochstotter, M. 2014, *The Basic of Financial Econometrics: Tools, Concepts, and Asset Management Applications*, John Wiley & Sons, Inc, Hoboken, New Jersey.

Ferrucci, D. 2012, "Introduction to "This is Watson"", *IBM Journal of Research and Development*, vol. 56, pp. 1:1-1:15.

Free, M.L. & Moats, M. 2014, "Chapter 2.7 - Hydrometallurgical Processing" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 949-982.

Genderen, E., Wildnauer, M., Santero, N. & Sidi, N. 2016, "A Global Life Cycle Assessment for Primary Zinc Production", *The International Journal of Life Cycle Assessment*, vol. 21, no. 11, pp. 1580-1594.

Guidi, G., Miniati, R., Mazzola, M. & Iadanza, E. 2016, "Case Study: IBN Watson Analytic Cloud Platform as Analytics-as-a-Service System for Heart Failure Early Detection", *Future Interner*, vol. 8, no. 32.

Habashi, F. (ed) 1999, *Handbook of Extractive Metallurgy*, A Wiley Company, Toronto.

Haenlein, M. & Kaplan, A. 2004, "A Beginner's Guide to Partial Least Squares Analysis", *Understanding Statistics*, vol. 3, no. 4, pp. 283-297.

Headai Ltd 2017, , *Creating Artificial Labor*. Available: http://www.headai.com/ [2018, 01/16].

Hewlett Packard Enterprise Development LP 2017, . Available: https://dev.havenondemand.com/apis [2018, 01/14].

Hoyt, R., Snider, D., Thompson, C. & Mantravadi, S. 2016, "IBM Watson Analytics: Automating Visualization, Descriptive, and Predictive Statistics", *JMIR Public Health and Surveillance*, vol. 2, no. 3.

J. Powles, H.H. 2017, "Google DeepMind and healthcare in an age of algorithms", *Health and Technology*, vol. 7, no. 4, pp. 351-367.

Jalkanen, H. & Holappa, L. 2014, "Chapter 1.4 - Converter Steelmaking" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 223-270.

Kordosky, G.A. 2002, "Copper recovery using leach/solvent extraction/electrowinning technology : Forty years of innovation, 2.2 million tonnes of copper annually", *Journal of the Southern African Institute of Mining and Metallurgy*, vol. 102, no. 8, pp. 445-450.

Kvande, H. 2011, "3 - Production of primary aluminium" in *Fundamentals of Aluminium Metallurgy*, ed. R. Lumley, Woodhead Publishing, , pp. 49-69.

Law, V. 2017, , *Beyond Limits is Spreading NASA's AI to the World*. Available: https://www.beyond.ai/news/2017/10/9/beyond-limits-is-spreading-nasas-ai-to-the-world [2018, 01/16].

Madias, J. 2014, "Chapter 1.5 - Electric Furnace Steelmaking" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 271-300.

Metson, J. 2011, "2 - Production of alumina" in *Fundamentals of Aluminium Metallurgy*, ed. R. Lumley, Woodhead Publishing, , pp. 23-48.

Miller, J. 2016, *Learning IBM Watson Analytics*, 1st edn, Packt Publishing Ltd., Birmingham, UK.

Moats, M.S. & Davenport, W.G. 2014, "Chapter 2.2 - Nickel and Cobalt Production" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 625-669.

N. Pathak, A.B. 2017, *Artificial Intelligence for .NET: Speech, Language, and Search*, Springer Science+, New York.

Nagwanshi, K.K. & Dubey, S. 2018, "Statistical Feature Analysis of Human Footprint for Personal Identification Using BigML and IBM Watson Analytics", *Arabian Journal for Science and Engineering*, vol. 43, no. 6, pp. 2703-2712.

Newstex 2015, , *Security Startup SparkCognition Uses AI to Protect Connected Devices*. Available: https://search-proquest-com.ezproxy.cc.lut.fi/docview/1691023272/fulltext/F53CF8813CA0478DPQ/1?accountid=27292 [2018, 01/13].

Numenta Inc. 2018, , *Leading the New Era of Machine Intelligence*. Available: https://numenta.com/numenta-anomaly-benchmark/ [2018, 01/14].

PerpetuuitiTechnoSoft Inc 2017, , *Automation Redefined with Av3ar*. Available: http://ptechnosoft.com/assets/images/datasheet/av3ar.pdf [2081, 01/16].

PR Newswire Association LLC 2015a, , *Cognitive Computing continues rapid rise in Asia*. Available: https://search-proquest-com.ezproxy.cc.lut.fi/docview/1697558629/citation/516BF1387CB04C87PQ/1?accountid=27292 [2018, 01/14].

PR Newswire Association LLC 2015b, , *CustomerMatrix Unveils First-Ever Cognitive Intelligence Engine for CRM*. Available: https://search-proquest-com.ezproxy.cc.lut.fi/docview/1661983841/ [2018, 01/14].

PR Newswire Association LLC 2014, , *CustomerMatrix Brings the Power of Cognitive Computing Intelligence to CRM: Delivering on the Promise of Data-Driven Revenue & Customer Satisfaction Impact*. Available: https://search-proquest-com.ezproxy.cc.lut.fi/docview/1622045597 [2018, 01/14].

R. Evans, J.G. 2016, , *DeepMind AI Reduces Google Data Centre Cooling Bill by 40%*. Available: https://deepmind.com/blog/deepmind-ai-reduces-google-data-centre-cooling-bill-40/ [2018, 01/14].

Rice, J. 1995, *Mathematical Statistics and Data Analysis*, 2nd edn, Wadsworth Publishing Company, Belmont, California.

Robert, A., Leslie, A. & Ingrid, R. 2002, *The Life Cycle of Copper, its Co-Products and By-Products*, 1st edn, Springer Netherlands.

S.Yegulalp 2016, , *Machine learning for enterprise app makers*. Available: https://search-proquest-com.ezproxy.cc.lut.fi/docview/1772121559/fulltext/A32F175ACC564731PQ/1?accountid=27292 [2018, 01/14].

Sartorius Stedim Data Analytics 2017, *User Guide to MODDE*, , Sweden, Umeå.

Sathi, A. 2016, *Cognitive (Internet of) Things*, 1st edn, Palgrave Macmillan US, New York, United States.

Schlesinger, M.E., King, M.J., Sole, K.C. & Davenport, W.G. 2011, "Chapter 7 - Submerged Tuyere Smelting: Noranda, Teniente, and Vanyukov" in *Extractive Metallurgy of Copper (Fifth Edition)*, eds. M.E. Schlesinger, M.J. King, K.C. Sole & W.G. Davenport, Elsevier, Oxford, pp. 111-125.

Seetharaman, S. 2014, "Chapter 2.1 - Copper Production" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 534-624.

Shariyar Murtaza, S., Lak, P., Bener, A. & Pischotchian, A. 2016, "How to Effectively Train IBM Watson: Classroom Experience", *49th Hawaii International Conference in System Sciences*.

Sinclair, R. 2009, *The Extractive Metallurgy of Lead*, 1st edn, The Australian Institute of Mining and Metallurgy, Victoria, Australia.

Sohn, H.Y. & Olivas-Martinez, M. 2014, "Chapter 2.3 - Lead and Zinc Production" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 671-700.

SparkCognition Inc. 2018, , *A cognitive approach to industrial safety and efficiency*. Available: https://sparkcognition.com/sparkpredict/ [2018, 01/13].

T. Zerucha 2016, , *Digital Reasoning launches Synthesys*. Available: https://www.banklesstimes.com/2016/06/07/digital-reasoning-launches-synthesys-4-june-7/ [2018, 01/16].

Tabereaux, A.T. & Peterson, R.D. 2014, "Chapter 2.5 - Aluminum Production" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 839-917.

Tarafdar, M., Beath, C. & Ross, J. 2017, "Enterprise Cognitive Computing Applications Opportunities and Challenges", *Cognitive Computing*, .

The International Aluminium Institute 2017, , *Perfluorocarbon (PFC) Emissions*. Available: http://www.world-aluminium.org/statistics/perfluorocarbon-pfc-emissions/ [2018, 03/02].

Thornton, I., Rautiu, R. & Brush, S. 2001, *Lead: the Facts*, Ian Allan Publishing, Ltd., London, United Kingdom.

Totten, G. & MacKenzie, D. (eds) 2003, *Hanbook of aluminum. Alloy Production and Mineral Manufacturing*, Marcel Dekker, Inc, New York.

Vignes, A. 2011, *Extractive Metallurgy: Processing Operations and Routes*, John Wiley & Sons, Inc., London.

Wensley, M. 2017, , *MKS Announces Release of MODDE® 12*. Available: https://globenewswire.com/news-release/2017/02/22/926574/0/en/MKS-Announces-Release-of-MODDE-12.html [2018, 05/02].

Wood, J., Coveney, J., Helin, G., Xu, L. & Xincheng, S. 2015, "The Outotec ® Direct Zinc Smelting Process", *Proceedings of EMC 2015*.

Yang, E.M. 2017, , *Cognitive AI Leader Beyond Limits Expands into Healthcare Space*. Available: https://www.beyond.ai/press-releases/2017/10/12/cognitive-ai-leader-beyond-limits-expands-into-healthcare-space-with-new-head-of-health-tech [2018, 01/16].

Yang, Y., Raipala, K. & Holappa, L. 2014, "Chapter 1.1 - Ironmaking" in *Treatise on Process Metallurgy*, ed. S. Seetharaman, Elsevier, Boston, pp. 2-88.

Appendix 1. Applied MATLAB scripts

Extraction of corresponding mean matte composition values for each operating day:

```
clc; clear all; close all
d_in=xlsread('matte_composition'); %initial data import

%initial data refining into suitable format
data=[zeros(length(d_in),1),d_in(:,8),d_in(:,1),d_in(:,3:7)];
day=[1:1:365]; k=1;
for i=1:length(data)-1          %determination of the corresponding day
    if data(i,3)-data(i+1,3)<21 %based on the difference between hours
        k=k;
        data(i,1)=day(k);
    else
        data(i,1)=day(k);
        k=k+1;
        data(i+1,1)=day(k);
    end
end
data(length(data),1)=k;
%extraction and separation of data according to the unit (VF-2 or VF-3)
vf2=[]; vf3=[];
for i=1:length(data)-1
    if data(i,2)==2
        vf2=[vf2;data(i,1:8)]; %extraction of data related to VF-2
    else
        vf3=[vf3;data(i,1:8)]; %extraction of data related to VF-3
    end
end
%calculation of the corresponding mean composition values for VF-2
%considering also gaps and mistakes in the initial data
vf2_r=[]; vf3_r=[]; n2=[]; l2=[0]; n3=[]; l3=[0];
for a=2:k+1
for i=1:length(vf2)
    if vf2(i,1)==a-1
       n2=[n2,vf2(i,1)];
    end
end
l2=[l2,length(n2)];

vf2_r=[vf2_r;a-1,sum(vf2(l2(a-1)+1:l2(a),4))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),5))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),6))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),7))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),8))/(l2(a)-l2(a-1))];
end
%calculation of the corresponding mean composition values for VF-3
%considering also gaps and mistakes in the initial data
for a=2:k+1
for i=1:length(vf3)
    if vf3(i,1)==a-1
       n3=[n3,vf3(i,1)];
    end
end
l3=[l3,length(n3)];

vf3_r=[vf3_r;a-1,...
    sum(vf3(l3(a-1)+1:l3(a),4))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),5))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),6))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),7))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),8))/(l3(a)-l3(a-1))];
end
%data export into convenient xls format
xlswrite('matte2_ready',vf2_r); xlswrite('matte3_ready',vf3_r);
```

Extraction of corresponding mean slag composition values for each operating day:

```
clc; clear all; close all
d_in=xlsread('slag_composition'); %initial data import
%initial data refining into suitable format
data=[zeros(length(d_in),1),d_in(:,14),d_in(:,1),d_in(:,4),d_in(:,5:13)];
day=[1:1:365]; k=1;
for i=1:length(data)-1           %determination of the corresponding day
    if data(i,3)-data(i+1,3)<19  %based on the difference between hours
        k=k;
        data(i,1)=day(k);
    else
        data(i,1)=day(k);
        k=k+1;
        data(i+1,1)=day(k);
    end
end
data(length(data),1)=k;
%extraction and separation of data according to the unit (VF-2 or VF-3)
vf2=[]; vf3=[];
for i=1:length(data)-1
    if data(i,2)==2
        vf2=[vf2;data(i,1:13)]; %extraction of data related to VF-2
    else
        vf3=[vf3;data(i,1:13)]; %extraction of data related to VF-3
    end
end
%calculation of the corresponding mean composition values for VF-2
%considering also gaps and mistakes in the initial data
vf2_r=[]; vf3_r=[]; n2=[]; l2=[0]; n3=[]; l3=[0];
for a=2:k+1
for i=1:length(vf2)
    if vf2(i,1)==a-1
       n2=[n2,vf2(i,1)];
    end
end
l2=[l2,length(n2)];
vf2_r=[vf2_r;a-1,round(sum(vf2(l2(a-1)+1:l2(a),4))/(l2(a)-l2(a-1)),0),...
    sum(vf2(l2(a-1)+1:l2(a),5))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),6))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),7))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),8))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),9))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),10))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),11))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),12))/(l2(a)-l2(a-1)),...
    sum(vf2(l2(a-1)+1:l2(a),13))/(l2(a)-l2(a-1))];
end
%calculation of the corresponding mean composition values for VF-3
%considering also gaps and mistakes in the initial data
for a=2:k+1
for i=1:length(vf3)
    if vf3(i,1)==a-1
       n3=[n3,vf3(i,1)];
    end
end
l3=[l3,length(n3)];
vf3_r=[vf3_r;a-1,round(sum(vf3(l3(a-1)+1:l3(a),4))/(l3(a)-l3(a-1)),0),...
    sum(vf3(l3(a-1)+1:l3(a),5))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),6))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),7))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),8))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),9))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),10))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),11))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),12))/(l3(a)-l3(a-1)),...
    sum(vf3(l3(a-1)+1:l3(a),13))/(l3(a)-l3(a-1))];
end
%data export into convenient xls format
xlswrite('slag2_ready',vf2_r); xlswrite('slag3_ready',vf3_r);
```

Extraction of corresponding matte and slag compositions

```
clc; clear all; close all
datain_slag=xlsread('slag_composition'); %initial data import
datain_matte=xlsread('matte_composition');
%initial data refining into suitable format
data_slag=[zeros(length(datain_slag),1),datain_slag(:,1),...
    datain_slag(:,14),datain_slag(:,5:13)];
data_matte=[zeros(length(datain_matte),1),datain_matte(:,1),...
    datain_matte(:,8),datain_matte(:,3:7)];
%determination of the corresponding day for slag composition
day=[1:1:365]; k=1;
for i=1:length(data_slag)-1
    if data_slag(i,2)-data_slag(i+1,2)<19
        k=k;
        data_slag(i,1)=day(k);
    else
        data_slag(i,1)=day(k);
        k=k+1;
        data_slag(i+1,1)=day(k);
    end
end
data_slag(length(data_slag),1)=k;
%determination of the corresponding day for matte composition
k=1;
for i=1:length(data_matte)-1
    if data_matte(i,2)-data_matte(i+1,2)<21
        k=k;
        data_matte(i,1)=day(k);
    else
        data_matte(i,1)=day(k);
        k=k+1;
        data_matte(i+1,1)=day(k);
    end
end
data_matte(length(data_matte),1)=k;


datam=[]; %extraction of relevant time spans of analysis
for i=1:k

datam=[datam;data_matte(find(data_matte(:,1)==i&(data_matte(:,3)==2|...
       data_matte(:,3)==3)& ...
    (data_matte(:,2)==3|data_matte(:,2)==7|data_matte(:,2)==11 ...
    |data_matte(:,2)==15|data_matte(:,2)==19|data_matte(:,2)==23)),:)];
end

datas=[]; %extraction of relevant time spans of analysis
for i=1:k
   datas=[datas;data_slag(find(data_slag(:,1)==i&(data_slag(:,3)==2|...
       data_slag(:,3)==3)& ...
    (data_slag(:,2)==3|data_slag(:,2)==7|data_slag(:,2)==11 ...
    |data_slag(:,2)==15|data_slag(:,2)==19|data_slag(:,2)==23)),:)];
end

%determination of suitable for further "while" loop number of iterations
if length(datam)<length(datas)
    l=length(datas);
else
    l=length(datam);
end
```

```
%extraction of matte and slag composition results for the same days,
% processing units and time samples
i=1; data=[];
while i~=l
    i=i+1;
  if
datam(i,1)==datas(i,1)&&datam(i,2)==datas(i,2)&&datam(i,3)==datas(i,3)
     data=[data;datam(i,1:8),datas(i,4:12)];
   else if (datam(i,1)~=datas(i,1)||datam(i,2)~=datas(i,2)||...
               datam(i,3)~=datas(i,3))&&(datam(i,1)>datas(i,1)||...
               datam(i,2)>datas(i,2))
          datas=[datas(1:i-1,:);datas(i+1:end,:)];
          l=length(datas);
          if datam(i,1)==datas(i,1)&&datam(i,2)==datas(i,2)&&...
                  datam(i,3)==datas(i,3)
             data=[data;datam(i,1:8),datas(i,4:12)];
          end
   else if (datam(i,1)~=datas(i,1)||datam(i,2)~=datas(i,2)||...
               datam(i,3)~=datas(i,3))&&(datam(i,1)<datas(i,1)||...
               datam(i,2)<datas(i,2))
          datam=[datam(1:i-1,:);datam(i+1:end,:)];
          l=length(datam);
          if datam(i,1)==datas(i,1)&&datam(i,2)==datas(i,2)&&...
                  datam(i,3)==datas(i,3)
             data=[data;datam(i,1:8),datas(i,4:12)];
          end
        end
        end
  end
end
%data export into suitable format
xlswrite('matte_and_slag_composition',data);
```

Appendix 2. Example of the analyzed database – Furnace parameters and products compositions

Date and time	02.01.2006 0:00	03.01.2006 0:00	04.01.2006 0:00
Flux feed (sand)	36.336	35.829	31.219
Flux feed (sand stone)	10.936	11.781	9.712
Feed sulfide materials	69.019	72.459	82.782
Feed revert materials	0.295	0.182	0.437
Exhaust gases temperature	134.728	130.204	115.855
Oxygen content	77.545	78.156	82.188
Air-oxygen mixture flowrate	28665.33	28172.837	27354.132
Air-oxygen mixture pressure	0.971	0.974	0.945
Natural gas flowrate	2670.909	2394.418	2025.801
Air flowrate	12034.058	10803.083	8102.18
Bulk oxygen content	92.024	91.849	90.537
Oxygen flowrate	223050.863	230512.848	251805.867
Water pressure (left side)	5.845	5.807	5.74
Water temperature entrance (left side)	18.653	21.304	23.622
Water pressure (right side)	5.721	5.686	5.619
Water temperature entrance (right side)	18.484	21.039	23.417
Water temperature exit (left side)	25.632	28.58	30.135
Water temperature exit (right side)	24.974	27.578	29.416
matte Ni	3.531	3.886	3.566
matte Cu	61.925	62.917	62.025
matte S	22.8	22.583	22.783
matte Fe	11.167	9.9	11
matte Co	0.058	0.06	0.064
slag Ni	0.154	0.186	0.178
slag Cu	0.74	0.852	0.817
slag Co	0.048	0.053	0.061
slag Fe	45.125	46.795	46.935
slag S	0.722	0.754	0.78
slag CaO	2.146	2.114	2.096
slag MgO	1.623	1.538	1.539
slag Al2O3	3.949	3.722	3.631
slag SiO2	30.633	28.867	28.347

Appendix 3. Description of data analysis procedure with Watson Analytics platform

Analysis with Watson Analytics starts with import of the required dataset. The platform supports import from plenty of various cloud storage sources or from local computer.

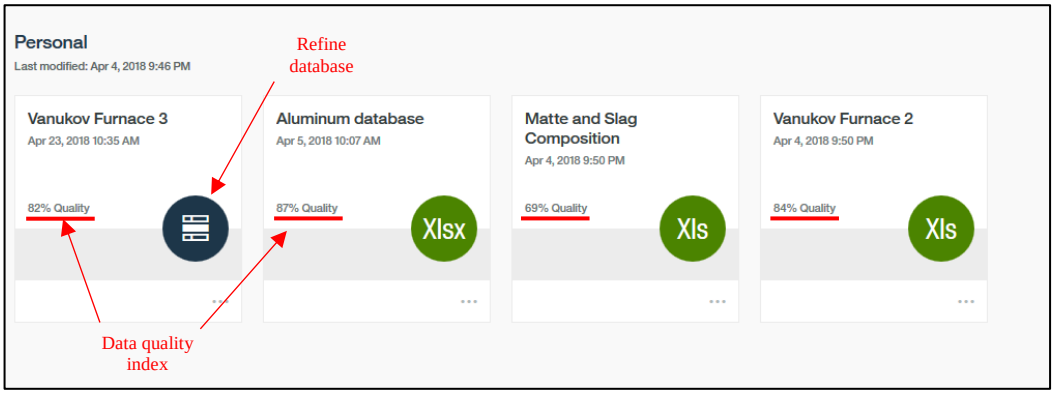


Figure 1. Imported databases in Watson Analytics interface

Once database has been imported, it can be further refined in the platform interface. Data refinement in Watson Analytics provides an opportunity to delete certain rows or columns, do simple calculations, add condition statements, group certain columns into hierarchies and change basic properties of the analyzed parameters.

After required preparation, analysis can be started by clicking at the dataset of interest. Watson Analytics also provides an opportunity to conduct data analysis by means of a question in natural form among all available databases – however, this option can be confusing in case of several similar datasets. Once a certain dataset is accessed, Watson Analytics generates an initial set of questions based on the contained information.

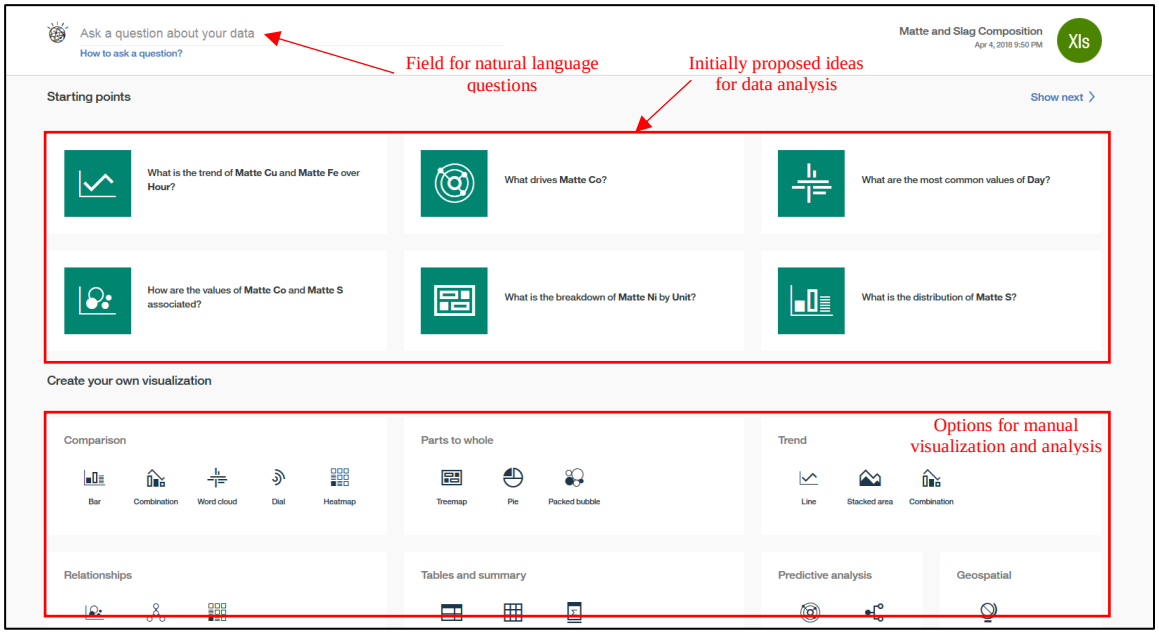


Figure 2. Initial analysis page

In case of new dataset analysis, it's a good practice to firstly determine the most significant statistical drivers for the target parameter. This can be done by means of predictive analysis tab, or by asking a question "What drives [target parameter name]?". The result is represented by means of Watson spiral, where parameters with increased statistical significant are located closer to the center.

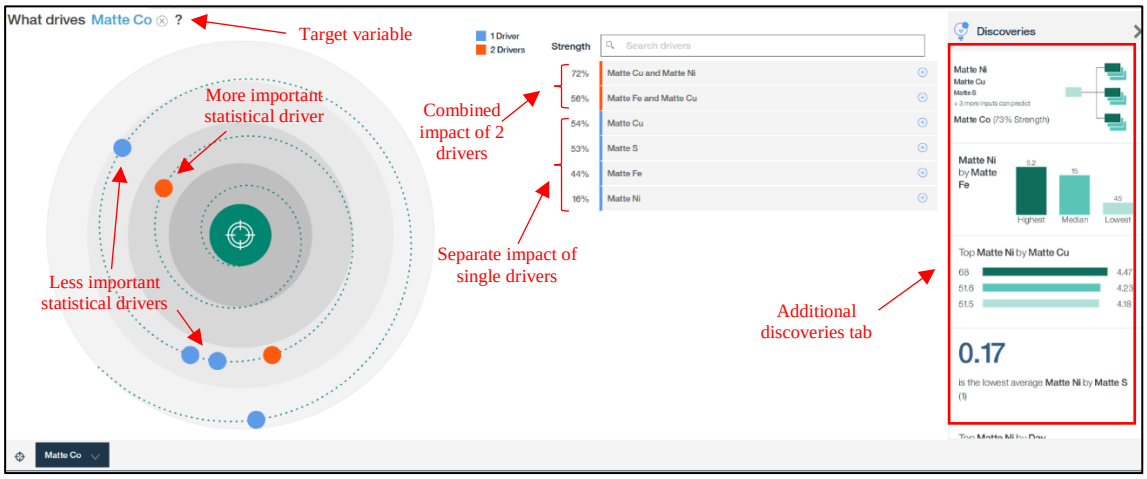


Figure 3. Key statistical drivers determination

Moreover, during key drivers' analysis, Watson Analytics provides a brief description of the most interesting characteristics of the related variables. More detailed analysis of a certain key driver can be accessed by clicking on it. By default, detailed analysis of a simultaneous impact of 2 key drivers is represented in a form of a heat map.

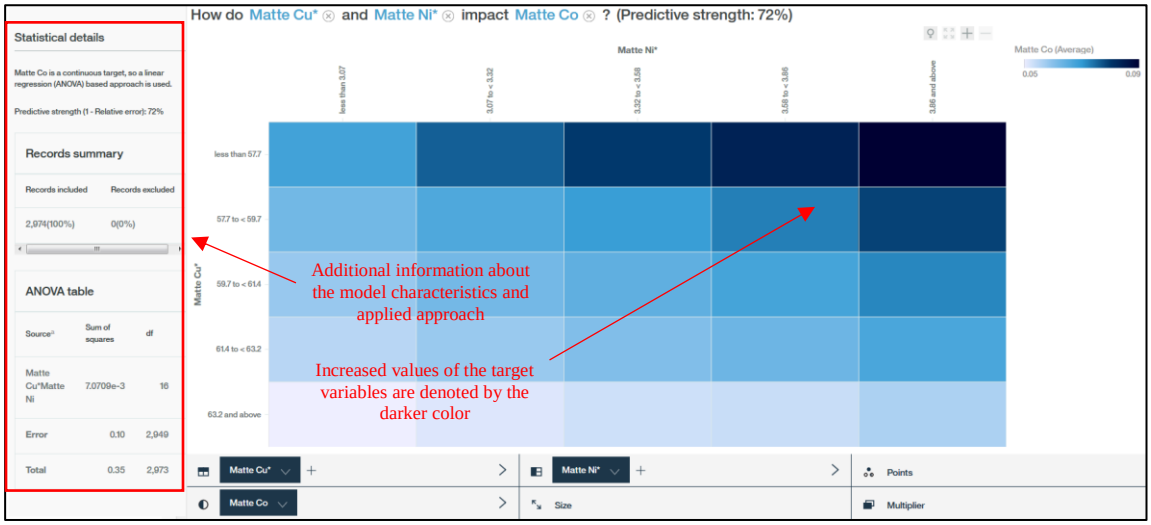


Figure 4. Detailed analysis of a key driver

Further information about simultaneous impact of more than 2 parameters can be obtained by predictive analysis. This can be done by asking a question "What is a predictive model of [target variable name]?" or in a predictive analysis tab. The generated model can be interpreted in a form of decision tree (Figure 2.6 for example) or a set of decision rules. It's

important to note that predictive strength in case of predictive model can be different from the results obtained by a key drivers' analysis.

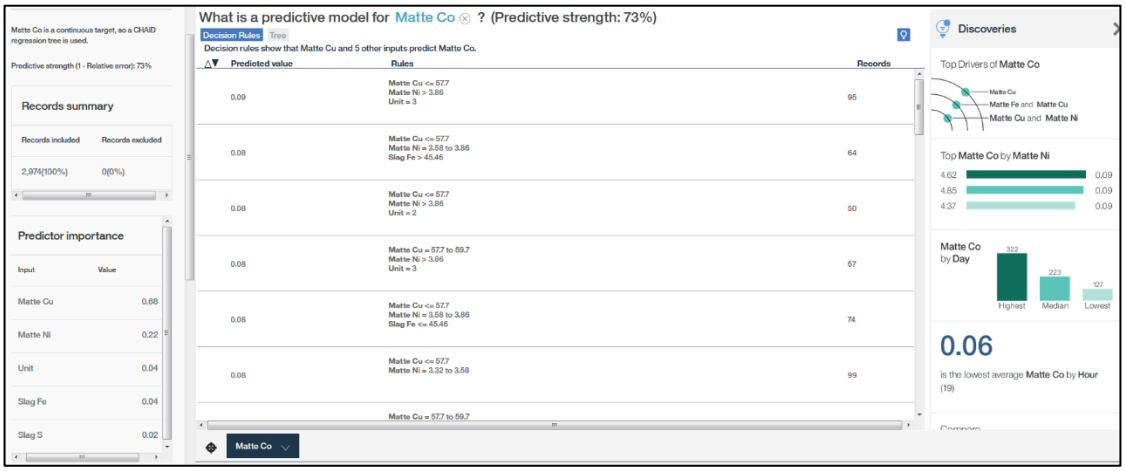


Figure 5. Predictive model in a form of a set of decision rules

Other frequently asked questions for data analysis with Watson Analytics usually involve the following options:

- What is the trend of [target variable name] over [target variable name]?
- What is the contribution of [target variable name] over [target variable name]?
- What is the [target variable name] by [target variable name] by [target variable name]?
- How do the values of [target variable name] compare by [target variable name]?

Watson Analytics algorithms recognize the following keywords for a certain visualization type: compare, trend, contribution, correlation, relationship, breakdown, grouping, where/when/how etc. These keywords should be placed near the beginning of a question. In every visualization format additional conditions can be applied by means of the "Multiplier" tab.

It is worth mentioning that during database preparation it is advisable to select simple and distinctive name for the parameters in order to avoid any problems with natural language processing of Watson Analytics platform.

Appendix 4. Key statistical drivers analyses results

Table 1. Matte copper content drivers (Vanyukov Furnace 2 database)

Predictive strength	Drivers description
Drivers combination impact	
87 %	Matte iron and nickel content
86 %	Matte iron content and natural gas flowrate
85 %	Matte iron content and inlet air flowrate
85 %	Matte iron content and ambient temperature
84 %	Matte iron content and fluxing agent addition rate (sand)
84 %	Matte iron content and ambient humidity
83 %	Slag magnesia content and matte iron content
78 %	Matte sulfur content and air-oxygen pressure
77 %	Matte sulfur content and fluxing agent addition rate (sandstone)
77 %	Matte sulfur content and slag magnesia content
65 %	Slag nickel content and matte cobalt content
63 %	Slag silica content and matte cobalt content
45 %	Cooling water inlet temperature and exhaust gases temperature
42 %	Oxygen and air flowrates
38 %	Sulfide materials feed and fluxing agent addition rate (sand)
Single drivers impact	
80 %	Matte iron content
72 %	Matte sulfur content
53 %	Matte cobalt content
34 %	Cooling water inlet temperature
29 %	Oxygen flowrate
28 %	Cooling water outlet temperature
24 %	Sulfide materials feed rate

Table 2. Matte copper content drivers (Vanyukov Furnace 3 database)

Predictive strength	Drivers description
Drivers combination impact	
90%	Matte cobalt and iron content
88%	Matte iron content and inlet air flowrate
87%	Matte iron content and oxygen inlet flowrate
87%	Matte iron content and oxygen concentration
87%	Matte iron and nickel content
84%	Matte iron content and exhaust gases underpressure
83%	Matte iron content and cooling water pressure
82%	Matte iron content and exhaust gases temperature
78%	Matte sulfur content and inlet air flowrate
76%	Matte sulfur content and oxygen concentration
75%	Matte sulfur and cobalt content
73%	Matte sulfur and cooling water inlet temperature
64%	Matte cobalt content and air inlet flowrate
64%	Matte cobalt content and oxygen concentration
61%	Matte cobalt content and natural gas flowrate
61%	Sulfide materials feed rate and matte cobalt content
Single drivers impact	
79%	Matte iron content
70%	Matte sulfur content
50%	Matte cobalt content
28%	Oxygen concentration
25%	Oxygen inlet flowrate
20%	Air-oxygen mixture pressure

Table 3. Slag silica content drivers (Vanyukov Furnace 2 database)

Predictive strength	Drivers description
Drivers combination impact	
80%	Slag iron content and sulfide materials feed rate
78%	Slag iron content and oxygen inlet flowrate
78%	Slag quicklime and iron content
77%	Slag iron content and exhaust gases temperature
76%	Slag iron content and fluxing agent addition rate (sandstone)
47%	Slag copper content and air inlet flowrate
46%	Slag copper content and cooling water outlet temperature
44%	Slag alumina content and sulfide materials feed rate
43%	Slag alumina and quicklime content
42%	Slag alumina content and fluxing agent addition rate (sand)
42%	Slag alumina content and oxygen inlet flowrate
39%	Slag alumina content and oxygen concentration
38%	Slag alumina content and natural gas flowrate
34%	Slag alumina and sulfur content
33%	Slag sulfur content and oxygen flowrate
Single drivers impact	
73%	Slag iron content
35%	Slag copper content
22%	Slag alumina content
19%	Slag nickel content
16%	Oxygen inlet flowrate
16%	Slag cobalt content
13%	Natural gas flowrate

Table 4. Slag silica content drivers (Vanyukov Furnace 3 database)

Predictive strength	Drivers description
Drivers combination impact	
80%	Slag iron and cobalt content
76%	Slag iron and cooling water outlet temperature
75%	Slag iron content and matte cobalt content
75%	Slag iron content and matte sulfur content
74%	Slag iron content and matte copper content
74%	Slag iron content and matte iron content
73%	Slag iron and oxygen concentration
72%	Slag iron content and inlet air flowrate
71%	Slag sulfur and iron content
40%	Slag copper content and matte iron content
39%	Slag cobalt content and oxygen concentration
37%	Slag cobalt content and oxygen inlet flowrate
35%	Slag magnesia content and oxygen concentration
34%	Sulfide materials addition rate and matte sulfur content
33%	Slag alumina content and oxygen inlet flowrate
32%	Exhaust gases temperature and oxygen inlet flowrate
Single drivers impact	
68%	Slag iron content
32%	Slag copper content
22%	Matte sulfur content
19%	Oxygen concentration
19%	Matte cobalt content
16%	Oxygen inlet flowrate

Table 5. Slag iron content drivers (Vanyukov Furnace 2 database)

Predictive strength	Drivers description
Drivers combination impact	
74%	Slag silica content and matte iron content
55%	Slag alumina and quicklime content
49%	Slag alumina content and oxygen concentration
46%	Slag alumina content and Oxygen inlet content
45%	Slag copper and cooling water outlet temperature
42%	Slag alumina content and sulfide materials addition rate
42%	Slag alumina and air inlet flowrate
42%	Slag alumina and fluxing agent addition rate (sand)
40%	Slag alumina content and oxygen concentration
39%	Slag alumina and sulfur content
38%	Slag alumina and cobalt content
32%	Slag magnesia content and natural gas flowrate
31%	Slag nickel content and natural gas flowrate
29%	Natural gas flowrate and exhaust gases temperatures
29%	Oxygen inlet flowrate and natural gas flowrate
28%	Oxygen inlet flowrate and fluxing agent addition rate (sandstone)
Single drivers impact	
70%	Slag silica content
36%	Slag copper content
27%	Slag alumina content
14%	Oxygen inlet flowrate
13%	Slag nickel content
13%	Natural gas flowrate

Table 6. Slag iron content drivers (Vanyukov Furnace 3 database)

Predictive strength	Drivers description
Drivers combination impact	
75%	Slag silica and alumina content
72%	Slag silica content and air inlet flowrate
72%	Slag silica content and exhaust gases temperature
72%	Slag silica content and natural gas flowrate
71%	Slag silica and nickel content
37%	Slag alumina and oxygen flowrate
34%	Slag alumina and sulfur content
33%	Sulfide materials addition rate and slag alumina content
32%	Slag alumina content and natural gas flowrate
32%	Slag alumina content and cooling water outlet temperature
30%	Exhaust gases temperature and oxygen inlet flowrate
29%	Slag magnesia content and cooling water outlet temperature
29%	Slag sulfur content and cooling water pressure
29%	Slag magnesia content and oxygen concentration
28%	Slag magnesia content and air inlet flowrate
28%	Slag alumina and magnesia content
Single drivers impact	
67%	Slag silica content
27%	Slag copper content
18%	Slag alumina content
17%	Inlet air flowrate
17%	Oxygen concentration
15%	Matte sulfur content

Appendix 5. Brief description of the linear regression methods

Currently, Modde Pro is based upon two regression models for fitting of the mathematical model for data analysis, namely PLS – Partial Least Squares regressions and MLR – Multiple Linear Regression. These regression models are both suitable for analysis of a single dependent variable *Y (n×m)* or simultaneous analysis of several dependent variables – which are also called "responses" within the Modde Pro syntax. The created regression models are used for detailed analysis of the dataset and prediction based on a set of independent variables *X (n×p)* – "factors" (Draper, Smith 1998, Eriksson, Eriksson 2000). The basic structure of the regression model is described by the following equation:

$$\hat{Y} = X \cdot B + E \tag{1}$$

where:
$\hat{Y}$ – predicted dependent variable (or a set of variables);
X – independent variable (or a set of independent variables);
B – a vector *(p×m)* of regression coefficients[3];
E – a vector of residuals.

Multiple Linear Regression (MLR)

In terms of multiple linear regression, equation (1) can be rewritten in a more detailed form in the following way:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k + \epsilon \tag{2}$$

where:
β_0 – constant intercept;
$\beta_1 \ldots \beta_k$ – regression coefficients of k independent variables.

Model error – β is a column vector of *(k+1)* regression coefficients including intercept:

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} \tag{3}$$

Therefore, regression coefficient for each independent variable (hereafter referred to as – "factor") describes the average variation in the dependent variable (hereafter referred to as – "response") per unit change in the factor, considering that values of other factors are held constant. Data analysis with MLR typically involves certain assumptions, more specifically:

- Distribution of regression errors is normal with zero mean;
- Regression errors variance is constant;
- Magnitudes or error at different time samples are independent;
- The residuals are assumed to be independent from factors (Fabozzi et al. 2014).

[3] The vector dimension parameter – *p* of the dataset matrix denotes the quantity of terms of the model.

The model fitting and further estimation steps involve construction of functional linear relationship between factors and corresponding parameters. This way, a set of correlation coefficients for the regression model should be determined. In terms of MLR analysis – this step also involves calculation of the correlation coefficients for the factors – to test the possible interaction between them. However, the most important results of this estimation are so-called "*point estimates*" of responses for given values of factors (Fabozzi et al. 2014, Sartorius Stedim Data Analytics 2017).

In distinction with common least squares regression method, which can be simply described by minimization of $\sum(y - \hat{y})^2$ with respect to corresponding regression coefficients, MLR approach involves more complicated principles. More specifically, a set of independent factors x_1, x_2 … x_k and the response with *y-values* form a space of dimension *k+1*. The generated regression model represents a k multidimensional regression hyperplane (with each of the k coefficients determining the slope in the direction of a corresponding factor) that describes the functional relationship between responses and factors. For simplified explanation and demonstration, the described principle is briefly represented in the Figure 1.

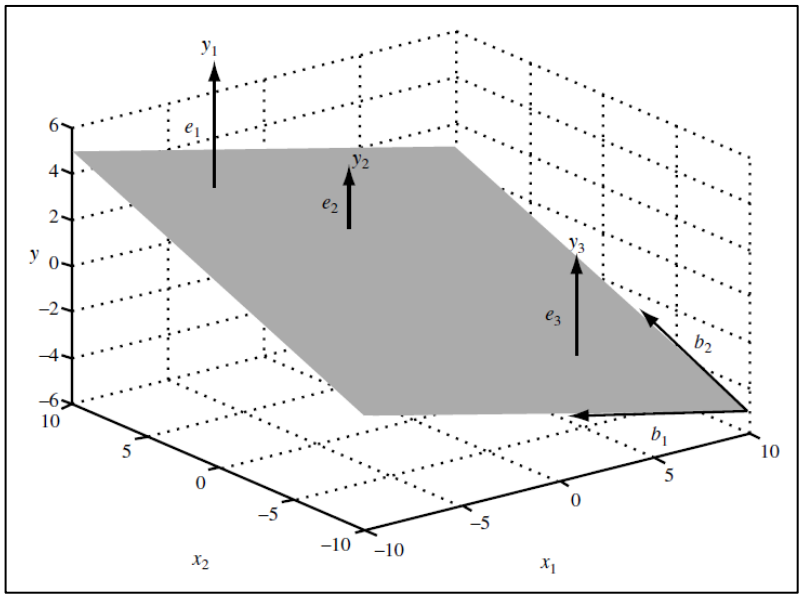


Figure 1. Graphical representation of regression hyperplane for 1 response and 2 factors (Fabozzi et al. 2014)

In this case 2 factors and one response form a 3-dimensional space plot, the regression hyperplane a 2-dimensional plane, that shows functional relation between responses and factor. Observed response values are y_1, y_2 and y_3 – vertical lines between these observation and actual surface of the regression hyperplane represent the error of approximation. In terms of Modde Pro, data analysis based on MLR approach is considered as an optimization

problem. More specifically, the objective is to find the vector β, that minimizes the squared error:

$$\sum_{i=1}^{n}(y-\hat{y})^2 = (y - X \cdot \beta)^T \cdot (y - X \cdot \beta) \quad (4)$$

Therefore, the optimal regression coefficient estimates and estimated residuals can be determined as follows by the equation 5 and 6 respectively (Fabozzi et al. 2014):

$$B = (X^T \cdot X)^{-1} \cdot X^T \cdot y \quad (5)$$

$$E = y - X^T \cdot B \quad (6)$$

These computations are implemented for analysis of the inserted dataset for production of estimations.

Partial Least Squares regression (PLS)

Contrary to the multiple linear regression case, where hyperplanes are generated for the determination of functional relations, the basic principle of partial least squares approach involves a projection of responses and factors to a new space with the purpose of finding a linear regression model. In terms of Modde Pro, partial least squares regression is applied in case if the problem is poorly conditioned and, therefore, the model needs to be additionally regulated by a latent variable method to avoid overfitting (Sartorius Stedim Data Analytics 2017). The general model of multivariable analysis in partial least squares regression can be described by the following equations (Esposito Vinzi et al. 2010, Haenlein, Kaplan 2004):

$$X = T \cdot P^T + E \quad (7)$$

where:
X – factor variable (or a set of factor variables) – $(n \times p)$ matrix;
T – a projected matrix of X – $(n \times l)$ matrix;
P – orthogonal loading matrix – $(p \times l)$ matrix;
E – error term.

$$Y = U \cdot Q^T + F \quad (8)$$

where:
Y – response variable (or a set of response variables) – $(n \times m)$ matrix;
U – a projected matrix of Y – $(n \times l)$ matrix;
P – orthogonal loading matrix – $(m \times l)$ matrix;
F – error term.

Within the PLS regression, error terms are assumed to be independent from other variables and equally distributed. Within Modde Pro, PLS regression can be solved differently depending on the quantity of dependent variables. In case of a single response variable – the

problem is referred to as "PLS1" and basically represents similar to MLR loss function. The optimization in this instance can be described by the following expression:

$$maximise_{\beta \in K_A(X^T X, X^T y)} \frac{1}{2} \|y - X\beta\|_2^2 \quad (9)$$

where:
$K_A(X^T X, X^T y)$ – represents Krylov subspace, correspondingly generated by $X^T X$ and $X^T y$;
A – quantity of components within the regression model;
β – regression coefficient;
$\|y - X\beta\|_2^2$ – denotes matrix norm – $\sqrt{\sum_{i=1;j=1}^{n} (y - X\beta)_{i,j}{}^2}$.

In case of more than one response variable – the optimization problem is generally referred to as "PLS2" – the partial least squares regression algorithm generates a new set of variables by linear combination of original ones for modelling the responses. These created variables, called "*X scores*", are orthogonal and can be described by the following equation (Esposito Vinzi et al. 2010):

$$t_a = X w_a \quad (10)$$

where:
t_a – a new generated variable – X score;
w_a – a vector of weights

The observed response variables are combined in a set of new Y scores in a similar manner, such that:

$$u_a = Y c_a \quad (11)$$

where:
u_a – a new generated variable – Y score;
c_a– a vector of weights

Therefore, within Modde Pro syntax, the optimization problem can be described in the following way (Sartorius Stedim Data Analytics 2017):

$$maximise\ Cov(X_a w_a, Y_a c_a) \quad (12)$$
with the condition of $\|w_a\| = \|c_a\| = 1$

where:
a = 1, 2 … A, A – quantity of components within the regression model;
$Cov(X_a w_a, Y_a c_a) = \frac{1}{n} \sum_{i=1}^{n} \big(X - E(X)\big)\big(Y - E(Y)\big)$ – covariance – the measure of the joint variability (Rice 1995).
$E(X)$, $E(Y)$ – expected values of *X* and *Y* respectively.

The generates sets of scores are collected so that: $T = [t_1, t_2 \dots t_A]$ and $U = [u_1, u_2 \dots u_A]$, which can be applied in equations (7) and (8) respectively.

Graphical representation of the partial least squares regression is provided in the Figure 2. The basic principle, as it was mentioned previously, is that the developed model projects responses and factor on lower-dimensional hyperplanes for each component of the model.

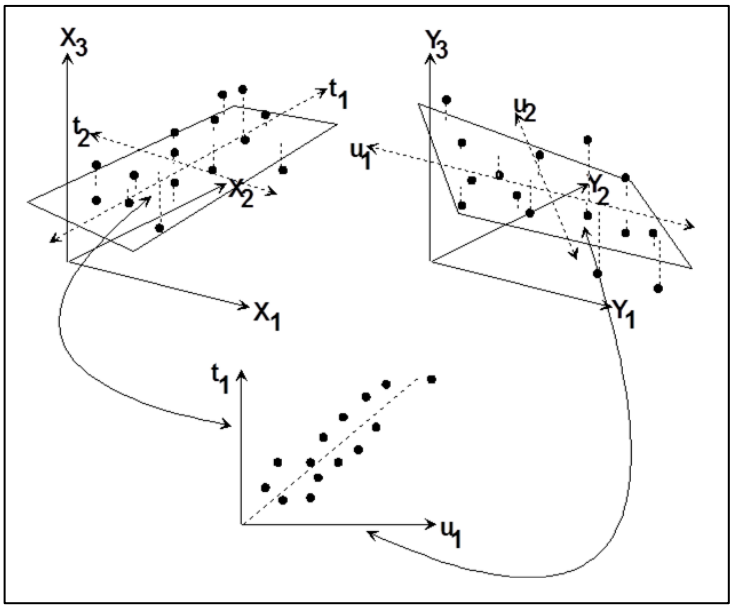


Figure 2. Graphical representation of PLS regression (Sartorius Stedim Data Analytics 2017)

Model characteristics

The goodness-of-fit measure, coefficient of determination – R2 coefficient is one of the most important parameters used in Modde for model characterization – describes fraction of variation that can be explained by the generated regression model.

$$R2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{13}$$

Estimation of the regression model's predictive availability is performed by means of Q2 coefficient – "model predictive power", which can be determined in the following way:

$$Q2 = 1 - \frac{PRESS}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{14}$$

where:
$PRESS$ – prediction residual sum of squares.

Statistical significance of the regression model is determined by means of the *F-test*. F-statistic parameter can be defined as follows (Esposito Vinzi et al. 2010):

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE} \tag{15}$$

where:
MSR – Mean squares of regression;

MSE – Mean squared errors;
SSR – Sum of squares explained by the regression model;
SSE – Unexplained sum of squares;
k – number of independent variables;
n – number of observations.

Statistical significant of the independent variables is determined by means of the *t-test*, which can be described by the following expression:

$$t = \frac{b_j}{s_{b_j}} \tag{16}$$

where:
b_j – sample estimate of a certain regression coefficient;
s_{b_j} – standard error of the coefficient estimate, which can be defined in the following way:

$$s_{b_j} = \frac{SSE}{n-k-1}(X^T X)^{-1}$$

Appendix 6. Textual data collection

Abdulkadir, A., Ajayi, A., Hassan, M. 2015. Evaluating the Chemical Composition and the Molar Heat Capacities of a white Aluminum Dross.

Abyzov, V. 2016. Lightweight Refractory Concrete Based on Aluminum-Magnesium- Phosphate Binder

Abyzov, V. 2017. Refractory Cellular Concrete Based on Phosphate Binder from Waste of Production and Recycling of Aluminum.

Alwaeli, M. 2013. Application of granulated lead–zinc slag in concrete as an opportunity to save natural resources.

Aparajith, B., Mohanty, D., Gupta, M. 2010. Recovery of enriched lead–silver residue from silver-rich concentrate of hydrometallurgical zinc smelter.

Behnajady, B., Moghaddam, J. 2017. Selective leaching of zinc from hazardous As-bearing zinc plant purification filter cake.

Bruckard, W., Woodcock, J. 2009. Recovery of valuable materials from aluminium salt cakes.

Buzatu, T., Gabriel Ghica, V., Iacob, G., Buzatu, M. 2015. Utilization of granulated lead slag as a structural material in roads constructions.

Chen, L., Yang, L., Bin, S., Liu, W. 2014. An Efficient Reactor for High-Lead Slag Reduction Process: Oxygen-Rich Side Blow Furnace.

Coruh, S., Nuri Ergun, O. 2010. Use of fly ash, phosphogypsum and red mud as a liner material for the disposal of hazardous zinc leach residue waste.

David, E., Kopac, J. 2012. Hydrolysis of aluminum dross material to achieve zero hazardous waste.

Davida, E., Kopacb, J. 2013. Aluminum recovery as a product with high added value using aluminum hazardous waste.

Dodoo-Arhin, D., Nuamah, R., Agyei-Tuffour, B. 2017. Awaso bauxite red mud-cement based composites: Characterization for pavement applications.

Dyer, L., Richmond, W., Fawell, P. 2012. Simulation of iron oxide/silica precipitation in the paragoethite process for the removal of iron from acidic zinc leach solutions.

Ewais, E., Besisa, N. 2018. Tailoring of magnesium aluminum titanate based ceramics from aluminum dross.

Fattahi, A., Rashchi, F., Abkhoshk, E. 2015. Reductive leaching of zinc, cobalt and manganese from zinc plant residue.

Gil, A., Albeniz, S., Korili, S. 2014. Valorization of the saline slags generated during secondary aluminium melting processes as adsorbents for the removal of heavy metal ions from aqueous solutions.

Gil, A., Korili, S. 2016. Management and valorization of aluminum saline slags: Current status and future trends.

Haisheng, H., Wei, S., Yuehua, H., Baoliang, J. 2014. Anglesite and silver recovery from jarosite residues through roasting and sulfidization-flotation in zinc hydrometallurgy.

Han, J., Liu, W., Qin, W., Zhang, T. 2017. Effects of sodium salts on the sulfidation of lead smelting slag.

Han, J., Liu, W., Yang, K., Wang, D. 2015. Innovative methodology for comprehensive utilization of high iron bearing zinc calcine.

Huang, X., Arambewela, M., Adkins, R., Tolaymat, T. 2015. Mineral phases and metals in baghouse dust from secondary aluminum production.

Huang, X., Arambewela, M., Ford, R., Barlaz, M. 2013. Characterization of salt cake from secondary aluminum production.

Jankovic, B., Stopic, S., Güven, A., Friedrich, F. 2012. The application of the formalism of dispersive kinetics for investigation of the isothermal decomposition of zinc leach residue in an inert atmosphere.

Jiang, G., Peng, B., Liang, Y., Chai, L. 2017. Recovery of valuable metals from zinc leaching residue by sulfate roasting and water leaching.

Kim, E., Hockmans, L., Spooren, J., Vrancken, K. 2017. Selective leaching of Pb, Cu, Ni and Zn from secondary lead smelting residues.

Koleini, S., Mehrpouya, H., Saberyan, K., Abdolahi, M. 2010. Extraction of indium from zinc plant residues.

Kumar Mandal, A., Ranjan Verma, H., Sinha, O. 2017. Utilization of aluminum plant's waste for production of insulation bricks.

Lee, H. 2013. Separation and Recovery of Nickel from Spent Electroless Nickel-Plating Solutions with Hydrometallurgical Processes.

Li, M., Peng, B., Chai, L., Peng, N. 2012. Recovery of iron from zinc leaching residue by selective reduction roasting with carbon.

Li, P., Wang, J., Zhang, X., Hou, X. 2017. Molten salt-enhanced production of hydrogen by using skimmed hot dross from aluminum remelting at high temperature.

Li, Q., Zhang, B., Min, X., Shen, W. 2013. Acid leaching kinetics of zinc plant purification residue.

Li, X., Wei, C., Deng, Z., Li, C. 2015. Extraction and separation of indium and copper from zinc residue leach liquor by solvent extraction.

Li, Y., Liu, H., Peng, B., Min, X. 2015. Study on separating of zinc and iron from zinc leaching residues by roasting with ammonium sulphate.

Li, Y., Yuan, Y., Liu, H., Peng, B. 2017. Iron extraction from lead slag by bath smelting.

Liu, F., Liu, Z., Li, Y. 2017. Recovery and separation of gallium(III) and germanium(IV) from zinc refinery residues: Part II: Solvent extraction.

Liu, F., Liu, Z., Li, Y. 2017. Recovery and separation of gallium(III) and germanium(IV) from zinc refinery residues: Part I: Leaching and iron(III) removal.

Liu, F., Liu, Z., Li, Y., Wilsom, B. 2017. Extraction of Ga and Ge from zinc refinery residues in $H_2C_2O_4$ solutions containing H_2O_2.

Liu, T., Li, F., Jin, Z., Yang, Y. 2018. Acidic leaching of potentially toxic metals cadmium, cobalt, chromium, copper, nickel, lead, and zinc from two Zn smelting slag materials incubated in an acidic soil.

Mahinroosta, M., Allahverdi, A., 2017. A promising green process for synthesis of high purity activated-alumina nanopowder from secondary aluminum dross.

Meneghetti Faé Gomes, M., Furlanetto Mendes, T., Wada, K. 2011. Reduction in toxicity and generation of slag in secondary lead process.

Min, X., Xie, X., Chai, L., Liang, Y. 2013. Environmental availability and ecological risk assessment of heavy metals in zinc leaching residue.

Mymrin, V., Alekseev, K., Fortini, O., Aibuldinov, Y. 2017. Environmentally clean materials from hazardous red mud, ground cooled ferrous slag and lime production waste.

Nusen, S., Zhu, Z., Chairuangsri, T., Yong Cheng, C. 2015. Recovery of germanium from synthetic leach solution of zinc refinery residues by synergistic solvent extraction using LIX 63 and Ionquest 801.

Nusen, S., Zhu, Z., Chairuangsri, T., Yong Cheng, C. 2016. Recovery of indium and gallium from synthetic leach solution of zinc refinery residues using synergistic solvent extraction with LIX 63 and Versatic 10 acid.

Onisei, S., Pontikes, Y., Van Gerven, T., Angelopoulos, G. 2012. Synthesis of inorganic polymers using fly ash and primary lead slag.

Pan, J., Zheng, G., Zhu, D., Zhou, X. 2013. Utilization of nickel slag using selective reduction followed by magnetic separation.

Penpolcharoen, M. 2005. Utilization of secondary lead slag as construction material

Perná, I., Hanzlícek, T. 2014. The solidification of aluminum production waste in geopolymer matrix.

Qu, Y., Li, H., Tian, W., Wang, X. 2014. Leaching of valuable metals from red mud via batch and continuous processes by using fungi.

Qu, Y., Lian, B., Mo, B. 2013. Bioleaching of heavy metals from red mud using Aspergillus niger.

Raghavan, R., Mohanan, P., Patnaik, S. Innovative processing technique to produce zinc concentrate from zinc leach residue with simultaneous recovery of lead and silver.

Roy, A., Stegemann, J. 2018. Nickel speciation in cement-stabilized/solidified metal treatment filter cakes.

Sadegh Safarzadeh, M., Dhawan, N., Birinci, M., Moradkhani, D. 2011. Reductive leaching of cobalt from zinc plant purification residues.

Sethurajan, M., Huguenot, D., Lens, P., Horn, H. 2016. Leaching and selective copper recovery from acidic leachates of zinc plant metallurgical purification residues.

Shu, Y., Ma, C., Zhu, L., Chen, H. 2015. Leaching of lead slag component by sodium chloride and diluted nitric acid and synthesis of ultrafine lead oxide powders.

Tsakiridis, P. 2012. Aluminium salt slag characterization and utilization – A review.

Tsakiridis, P., Oustadakis, P., Agatzini-Leonardou, S. 2013. Aluminium recovery during black dross hydrothermal treatment.

Urik, M., Bujdoš, M., Milová-Žiaková, B., Mikušová, P. B. 2015. Aluminium leaching from red mud by filamentous fungi.

Vahidi, E., Rashchi, F., Moradkhani, D. 2009. Recovery of zinc from an industrial zinc leach residue by solvent extraction using D2EHPA.

Vakilchap, F., Mousavi, S., Shojaosadati, S. 2016. Role of Aspergillus niger in recovery enhancement of valuable metals from produced red mud in Bayer process.

Vojtech, E., Zdenek, J. 2014. 12 years of leaching of contaminants from Pb smelter slags: Geochemical/mineralogical controls and slag recycling potential.

Waanders, F., Nell, J. 2012. Phase chemical composition of slag from a direct nickel flash furnace and associated slag cleaning furnace.

Wang, K., Li, J., McDonald, R., Browner, R. 2018. Iron, aluminium and chromium co-removal from atmospheric nickel laterite leach solutions.

Wang, L., Mu, W., Shen, H., Liu, S., 2015. Leaching of lead from zinc leach residue in acidic calcium chloride aqueous solution.

Wang, Z., Ni, W., Li, K., Huang, X. 2012. Crystallization characteristics of iron-rich glass ceramics prepared from nickel slag and blast furnace slag.

Xingbin, L., Zhigan, D., Cunxiong, L., Chang, W. 2015. Direct solvent extraction of indium from a zinc residue reductive leach solution by D2EHPA.

Yan, H., Chai, L., Peng, B. Li, M. 2014. A novel method to recover zinc and iron from zinc leaching residue.

Yang, T., Yao, X., Zhang, Z. 2014. Geopolymer prepared with high-magnesium nickel slag: Characterization of properties and microstructure.

Yu, G., Peng, N., Zhou, L., Liang, Y. 2015. Selective reduction process of zinc ferrite and its application in treatment of zinc leaching residues.

Yu, G., Zhang, Y., Zheng, S., Zou, X. 2014. Extraction of arsenic from arsenic-containing cobalt and nickel slag and preparation of arsenic-bearing compounds.

Zhang, C., Min, X., Zhang, J., Wang, M. 2015. Reductive acid leaching of cadmium from zinc neutral leaching residue using hydrazine sulfate.

Zhang, W., Yang, J., Wu, X., Hu, Y. 2016. A critical review on secondary lead recycling technology and its prospect.

Zhao, J., Zhao, Z., Cui, Y., Shi, R. 2018. New Slag for Nickel Matte Smelting Process and Subsequent Fe Extraction.

Appendix 7. Description of the data analysis procedure with Watson Discovery service

Prior to data collection import, the configuration for data ingestion and enrichment has to be customized according to main properties and features of the analyzed data sources. The built-in tool for this configuration customization is separated into three sections: conversion enrichment and normalization. Conversion section provides an opportunity to customize the process of data conversion from PDF or Word files into HTML and after that into JSON format. Interface of the configuration customization section for PDF to HTML conversion step is represented in the Figure 1.

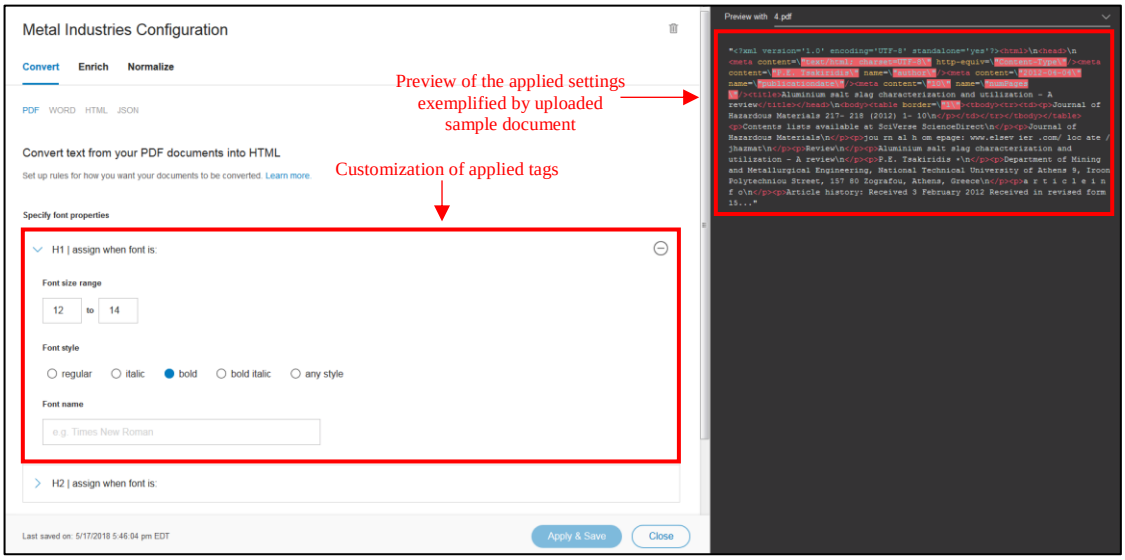


Figure 1. Graphical representation of the data conversion (PDF to HTML) interface within Watson Discovery service

On the left-hand side, a set of settings for characterization of applied tags is located, while representation of the configured settings within the service syntax is presented on the right side in JSON format.

Further step involves conversion of HTML to JSON format, where determined tags can be included or excluded from the further analysis. Interface of the configuration customization section for HTML to JSON conversion step is represented in the Figure 2. Similar to the previous case, the determination of the applied settings is located on the left side, while enriched sample document showing results of applied customization is represented on the right. It is recommended to exclude any scripting statements in order to simplify the further analysis. On the other hand, to facilitate recognition of tags, specified for PDF to HTML conversion, information about font sizes and other text properties should be kept, while corresponding tags can be excluded.

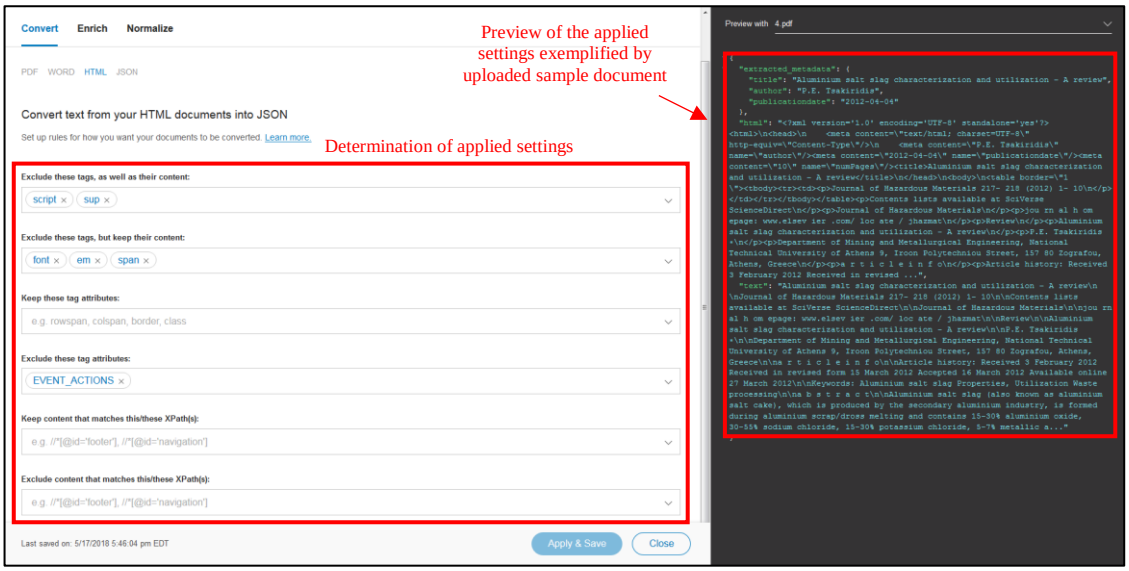


Figure 2. Graphical representation of the data conversion (HTML to JSON) interface within Watson Discovery service

Further step involves determination of data enrichment algorithms. Data enrichment can be either applied to certain recognized fields of the analyzed text or in case if no fields were detected – to the whole dataset. Additionally, at this step desired enrichments should be specified. Figure 3 provides visual representation of the data enrichment section interface.

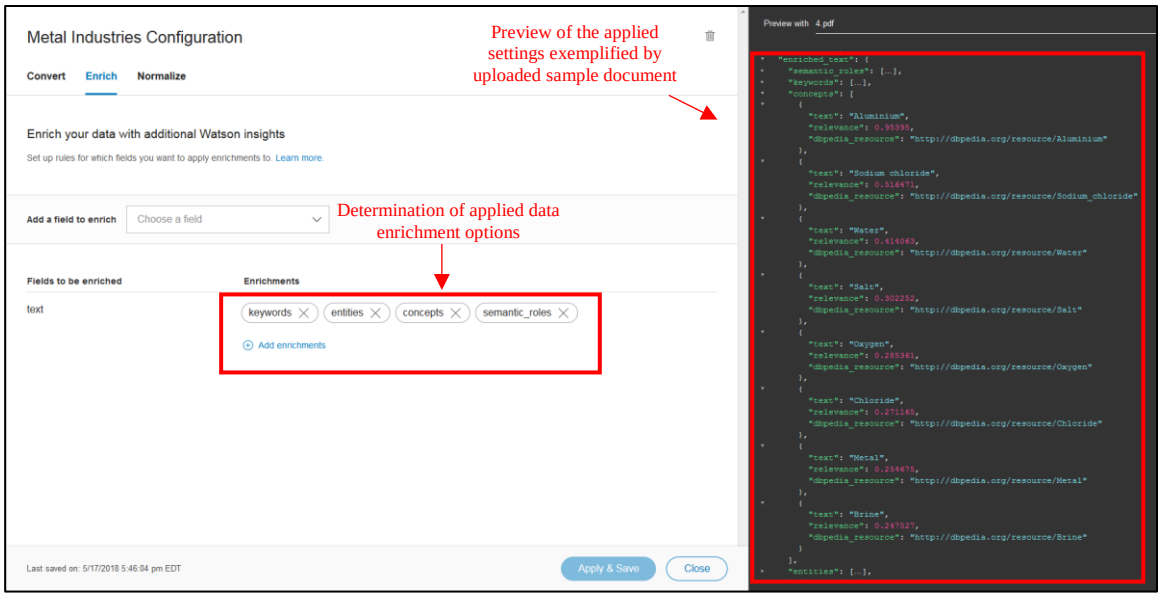


Figure 3. Graphical representation of the data enrichment interface

Similar to the previous cases, the applied settings are specified on the left size of the interface, while preliminary results are displayed on the right. Complete list of available enrichments is represented in the Figure 4.

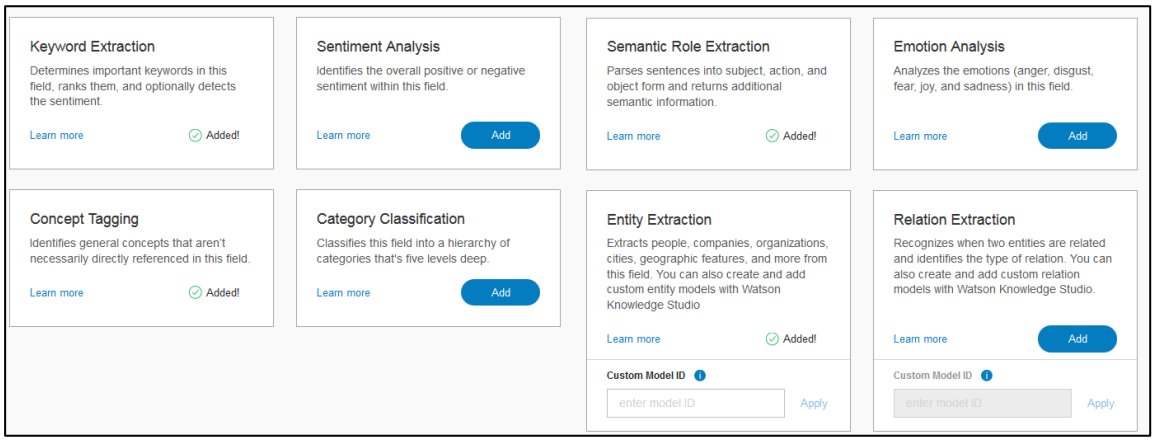


Figure 4. Available data enrichment options

The final step of data ingestion process configuration is normalization section, that provides an opportunity to move, merge, copy, discard or highlight certain enriched metadata fields, based on the previously determined data enrichments.

Once configuration customization has been finished, necessary data files can be uploaded into desired collection to form the analyzed information set. Watson Discovery service provides an opportunity to analyze information in a collection in so-called "query-able" form – so that the most relevant information within the data set is found based on the input query. Apart from this, the service also extracts the most significant results of data enrichment process and represents them in an understandable and clear way (Figure 5).

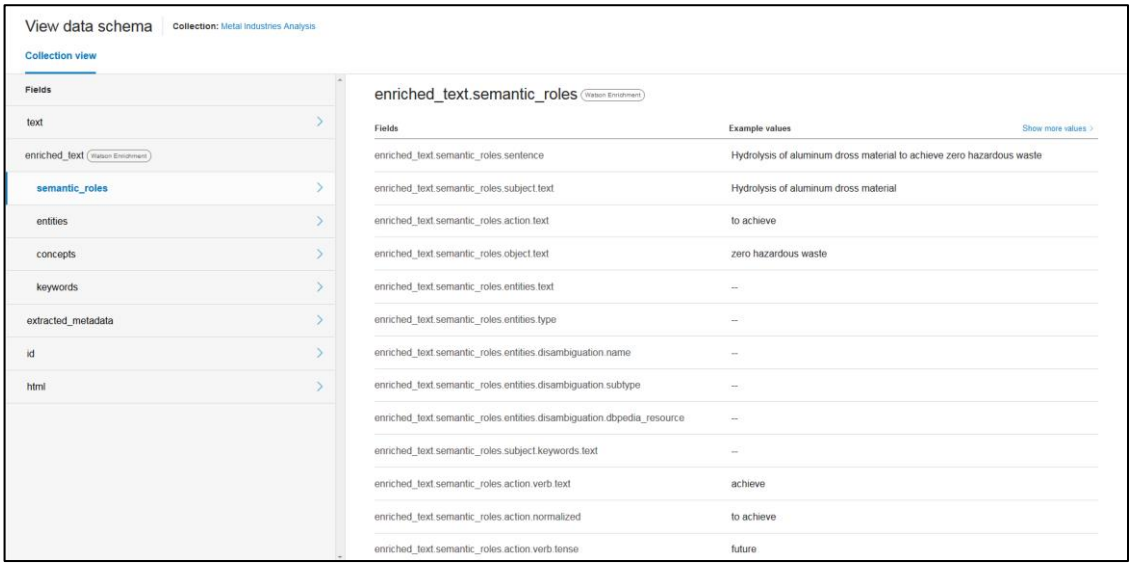


Figure 5. Data enrichment interface

More detailed analysis is performed by directly asking certain queries either in JSON format or in a form of natural language questions. In order to improve accuracy of natural language processing algorithms a special built-in tool based on machine learning algorithms for results relevancy training can be used.

This training process involves asking an example queries and rating of the proposed by the service results according to their relevancy. Graphical representation of the results relevancy training section is provided in the Figure 6.

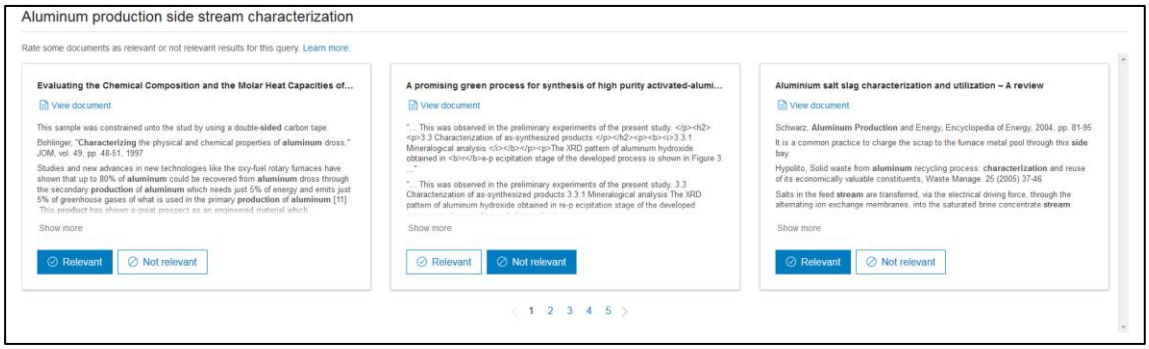


Figure 6. Results relevancy training interface

This way, once enough queries are generated and results are rated by the user, the service begins the process of self-learning based on the specified information. After this step, it is possible to use natural language queries without significant confusion or decrease in accuracy (Figure 7).

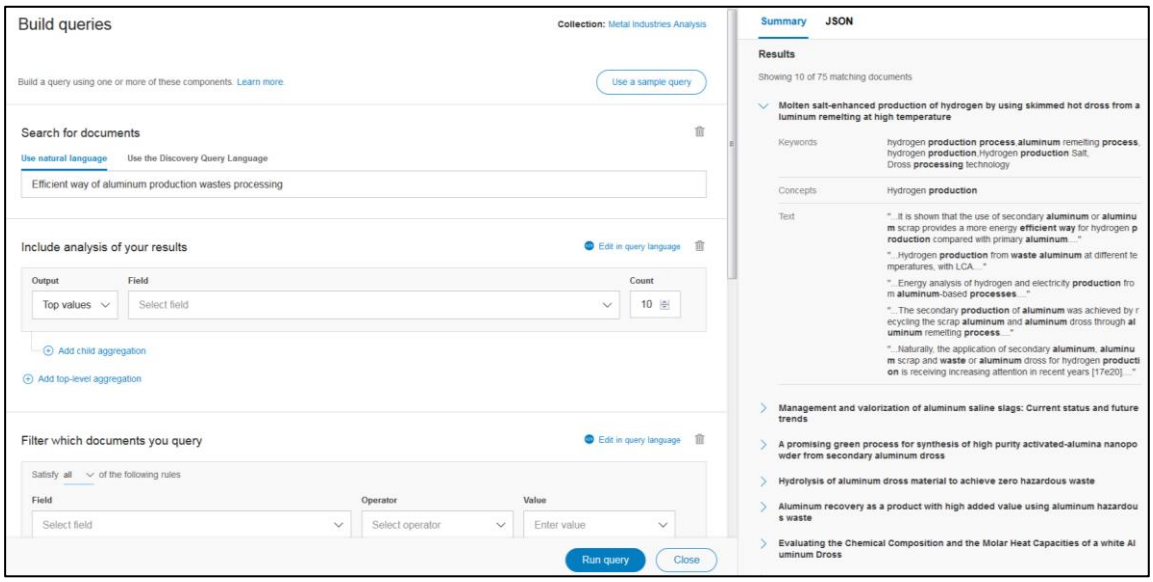


Figure 7. Data collection analysis by means of natural language queries

As it is represented in the Figure 7, Watson Discovery service applies cognitive search algorithms to find the most relevant documents within the analyzed data collection and also determines the most significant paragraphs within these documents. Additionally, further analysis of the results can be achieved by application of various aggregation tools and calculation of minimum/maximum/average values of certain variables recognized within the analyzed data.