Genetic Algorithm based variable selection in prediction of hot metal desulfurization kinetics

Tero Vuolio[1)], Ville-Valtteri Visuri[1)], Aki Sorsa[2)], Timo Paananen[3)] and Timo Fabritius[1)]*

1) Process Metallurgy Research Unit, University of Oulu, P.O. Box 4300, FI–90014, University of Oulu, Finland.

2) Control Engineering, University of Oulu, P.O. Box 4300, FI–90014, University of Oulu, Finland.

3) SSAB Europe Oy, Rautaruukintie 155, P.O. Box 93, FI–92101, Raahe, Finland.

*) Corresponding author. E-mail: tero.vuolio@oulu.fi

Keywords: Automated Model Identification, Genetic Algorithm, Optimization, Hot Metal Desulfurization.

ABSTRACT

Sulfur is considered as one of the main impurities in hot metal and hot metal desulfurization is often carried out using injection of fine-grade desulphurization powder. The selection of variables used for predicting the course of hot metal desulphurization requires expert knowledge. However, it is difficult to model the complex interactions in the process and to evaluate a high number of possible variable subsets with manual variable selection techniques. As the amount of data gathered from the process increases, manual variable selection becomes too time-consuming and might lead to a suboptimal prediction model. The objective of this work is to execute an automatic variable selection procedure for prediction of hot metal desulfurization based on an industrial scale data set. The variable selection problem is formulated as a constrained optimization problem, in which the objective function is formulated based on repeated leave-multiple-out cross-validation. The implemented solution strategy is a binary-coded Genetic Algorithm (GA). By making use of the developed model, the effect of the main production variables on the rate and efficiency of primary hot metal desulfurization was quantified. The variables related to properties of the reagent and the injection parameters were found to be of great importance.

1. Introduction

Sulfur is considered as one of the main impurities in hot metal and hot metal desulfurization is often carried out using injection of fine-grade desulfurization powder. Hot metal desulfurization with powder injection consists of two main reactions:

i) Transitory contact reaction (reagent-metal)

ii) Permanent contact reaction (slag-metal)

In the case of hot metal desulfurization with a lime-based reagent, the following reaction is considered:

$$< \mathrm{CaO} > + [\mathrm{S}] \rightarrow < \mathrm{CaS} > + [\mathrm{O}] \qquad [1]$$

In literature, the sulfur content in the hot metal has been observed to follow 1st order kinetics, for which the mass balance of sulfur can be written as follows: [1]

$$\frac{\mathrm{d}[\mathrm{S}]}{\mathrm{d}t} = -k_{\mathrm{tot}}([\mathrm{S}] - [\mathrm{S}]_{\mathrm{eq}}), \qquad [2]$$

The total rate of hot metal desulfurization reaction with lime-based reagents has been suggested to be dependent on the following parameters: [1-14]

- active solid surface area in contact with the metal phase,
- feed rate of the particles,
- mass of the metal bath,
- total flow rate of the gaseous compounds,
- mass transfer coefficient in the metal-reagent diffusion boundary layer,
- mass-transfer coefficient in the metal-slag diffusion boundary layer,
- rate of solid-state diffusion in the product phase,
- average residence time of the reagent particles in the metal bath, and
- sulfide capacity of the slag phase.

Predicting the evolution and end content of sulfur for powder injection based hot metal desulfurization has been under extensive research. Regardless of the wide variety of modeling approaches [1-14], the prediction of the end sulfur content in particular has been suggestive. Often, the mathematical descriptions of the physico-chemical phenomena occurring in the system are

not accurate enough for process control purposes. Furthermore, the complexity of such models often leads to relatively long computational times and difficulties in online implementation. Because they are based on true causalities, phenomena-based mathematical models are well-suited for process development and decision support, but cannot account for all sources of process variation, for example irregular wear of processing equipment. This being so, the mathematical modeling approaches must be tuned by identifying the appropriate model variables and parameters. As was established by in our previous work [14], the reasoning behind the inaccurate predictions could be an inadequate data set applied for predictions and uncertainties related to system identification. As suggested by the authors, the data-driven techniques could provide a feasible alternative for less accurate mechanistic models. [14] To illustrate the division in the literature of the modeling approaches presented, some of the previous modeling approaches for hot metal desulfurization are given in **Table 1**.

Table 1. Previous modeling approaches for hot metal desulfurization with powder injection.

Authors	Modeling approach	Reagent	Variable selection	Ref.
Oeters *et al.*	MR	CaO	Manual	[1]
Ohguchi *et al.*	MR	CaO	Manual	[2]
Oeters	MR	CaO	Manual	[3]
Datta *et al.*	ANN	CaC_2	Manual	[4]
Rastogi *et al.*	PMR	CaC_2	Manual	[5]
Deo *et al.*	ANN	CaC_2	Manual	[6]
Deo and Boom *et al.*	ANN	CaC_2	Manual	[7]
Visser and Boom	MR	CaO + Mg	Manual	[8]
Seshadri *et al.*	MR	CaO	Manual	[9]
Rodriguez *et al.*	MR	CaC_2	Manual	[10]
Ma *et al.*	MR	CaO + Mg	Manual	[11]
Barron *et al.*	MR	CaO, CaC_2 and Mg	Manual	[12]
Vargas-Ramirez *et al.*	MR	CaO, Na_2CO_3	Manual	[13]
Vuolio *et al.*	PMR	CaO + $CaCO_3$	Manual	[14]
This work	PMR	CaO + $CaCO_3$	Evolutionary	–

Notes: (MR = Mechanistic reaction model; PMR = Parameterized mechanistic reaction model; ANN = Artificial neural network model

Appropriate system identification is mandatory for building an effective prediction model. A key step in prediction model identification is the variable selection phase. The variable selection can be carried out either manually or by applying a search algorithm. However, as the amount of data and combinations of possible explanatory variables increase, the manual variable selection techniques do not often lead to an optimal subset, not to mention the adequate prediction accuracy and generalizability of the modeling approach. For this reason, the variable selection often needs automatic and more sophisticated methods.

In previous studies concerning the metallurgical field, the variable selection algorithms have been used mainly for identification of the blast furnace (BF) or basic oxygen furnace (BOF) processes. Saxén and Petterson[15] carried out a simultaneous input variable selection and structural optimization of the neural network that was designed for the prediction of silicon content of the hot metal in a blast furnace. By applying a pruning algorithm that was based on the importance of the connective weights in the network model, the authors succeeded in explaining major changes in the silicon content. The relevant variables for the predictive model were the ones associated with a non-zero weight connection.[15] However, using a neural network as the basis for a prediction model needs an extensive amount of data that covers the operational area of the process. Unfortunately, this is rarely available in the case of batch processes like hot metal desulfurization. Similar to Saxén *et al.* [15], Mahanta *et al.* [16] applied an evolutionary neural network and bi-objective genetic algorithm in the optimization of model structure for predicting several operational parameters in a blast furnace. They presented a pareto-optimal set of input variables, i.e. the model structure with a reasonable computational complexity and modeling error. In a study by Wang *et al.* [17], the Random Forest (RF) algorithm was used for prediction of the silicon content in a blast furnace. In their study, the measure of variable importance was related to classification accuracy of the model candidate when applied to an external data set. The algorithm selected the relevant input variables amongst 28 variable candidates. Wang *et al.*[18] presented a variable selection algorithm for a Support Vector Machine (SVM) model designed for the end-point prediction of the of BOF blow. The importance of a predictor variable candidate was measured with mutual information and selected if a threshold value of the quantity was exceeded.

This study presents a method to build a multivariable parametrized prediction model for hot metal desulfurization. The models considered in this study are identified by applying a repeated

leave-multiple-out cross-validation as the objective function, while the proposed objective function is minimized with a binary-coded genetic algorithm. The main objective of the study is to automatically identify a reliable and parsimonious prediction model that is well-suited for process control and optimization purposes.

2. Methodology

2.1. Objective function for variable selection

The objective of the variable selection problem is to find a best possible subset of variables or variables that explains most of the variance in the output vector. [19] Furthermore, this can be formulated as a combinatory optimization problem, in which the prediction error is minimized with respect to the explanatory variables and model parameters. The usual criteria for data-driven problems is either Sum of Squared Errors (SSE) or Mean Squared Error (MSE). In the case of the given rate constant, the linear multivariable model is given as:

$$k_{tot} = b_0 + b_1 x_1 + \cdots b_j x_j + \varepsilon = b_0 + \sum_{j=1}^{k} b_j x_j + \varepsilon \qquad [4]$$

Owing to the non-linear nature of desulphurization kinetics, the interactions between the dependent and independent variables are assumedly non-linear. For this reason, a log-linear form of the prediction equation is considered:

$$\ln k_{tot} = b_0 \ln e^1 + b_1 \ln x_1 + \cdots b_n \ln x_n + \varepsilon = b_0 \ln e^1 + \sum_{j=1}^{k} b_j \ln x_j + \varepsilon \qquad [5]$$

In the light of the reasoning above, the conditioned least-squares objective function can be written as follows:

$$\min \sum_{i=0}^{M} \left(y_i - [b_0 + \sum_{j=1}^{k} g_j b_j \ln x_{i,j}] \right)^2 \qquad [6]$$

$$\text{s.t.} \sum_{j=1}^{k} g_j \leq k_{max};$$

$$b_{min} \leq b_j \leq b_{max} \text{ where}$$

$$\hat{g} = [g_1\, g_2 \ldots g_k]^T ;\ g_j \in \{0,1\};$$

$$\hat{b} = [b_1\, b_2 \ldots b_k];\ b_j \in \mathbb{R}$$

In this case g_j marks whether the corresponding variable *j* is selected for inclusion in the model, such that if g_j = 1, *the variable is selected*. In the case of data set with noise and collinearities, the objective function is multimodal with usually several local minima. For this reason, the model estimate with *n* variable candidates should be the global optimum of the training set with respect to the regression parameters. It can be shown that the global optimum for the parameter vector *b* is obtained with the Moore-Penrose inverse of the data matrix X (*m* × *n*):

$$\hat{b} = (\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T y. \qquad [7]$$

As the parameter identification phase is out of the scope of this study, for the sake of simplicity the constraints set to the model parameters are ignored. In case of a regression model, the minimum of the least squares cost-function approaches zero when the number of predictor variables (*n*) approaches infinity. For this reason, the model with an excessively high number of predictor variables reaches very low values of the objective function for the training set, but fails to predict the changes in the validation set. This phenomenon is referred to as overfitting [19]. As was experimentally proven by Baumann [20], in variable subset selection there often lies a possibility for a chance correlation and overfitting [21]. Consequently, it was suggested that chance correlation and overfitting can be avoided by combining cross-validation with the efficient use of data.[19] With this, it is necessary to employ a cross-validation-based objective function.[19] The proposed methods are LOO (Leave One Out) and LMO (Leave Multiple Out) cross-validation techniques. However, the performance of LMO was superior to the performance of LOO since LOO gave highly over-fitted results compared to LMO when applying a Tabu Search for subset selection. In case of a BCGA, the conditioned cost-function applying a LMO cross-validation can be written as: [20]

$$\min \frac{1}{4N}\sum_{l=1}^{4N}\sum_{j=1}^{n}(y_{i,CV_l} - \left[b_{0,j} + \sum_{j=1}^{k} g b_{i,j} \ln x_{i,j,CV_l}\right])^2, \quad [8]$$

$$\text{s.t.} \sum_{j=1}^{k} g_j \leq k_{max};$$

$$b_{min} \leq b_j \leq b_{max}$$

where

$$\hat{g} = [g_1\, g_2 \ldots g_k]^T \; ; \; g_j \in \{0,1\}$$

$$\hat{b} = [b_1\, b_2 \ldots b_k]; \; b_j \in \mathbb{R}$$

where CV_l stands for internal validation set in cross-validation, n is the number of data points in the testing set and N is the number of data points in the fitting set. The value of the objective function used in the ranking of the model candidate is the mean value of the internally cross-validated modeling result. The split of the data is repeated $4N$ times, where N represents the number of data-points used in the fitting of the model. As shown by Baumann [21] a search algorithm defined in this way converges towards a variable subset with a high prediction performance and generalizability, and sufficiently excludes irrelevant variables from the set. [21]

The covariance of independent variables, i.e. multicollinearity, can be treated by applying penalty factors. Collinearity between the selected independent variables is a typical problem in selecting a multiple linear regression model and can lead to weak prediction performance and poorly interpretable model structure.[22] To avoid the existence of collinearity in the selected variable subset, the correlation between the independent variables is analyzed and a penalty term is added to the objective function. For this purpose, the concept of a variance inflation factor (*VIF*) matrix is introduced. In the *VIF* matrix, a single element is defined with the correlation matrix of X. The expression for an element in the *VIF* matrix is:

$$VIF_{i,j} = \frac{1}{1 - \text{corr}(X_{i,j})^2}. \quad [9]$$

The value used in the penalty function is:

$$VIF_{\max} = \lambda_i \max\left(\frac{1}{1-\mathrm{corr}(X)^2}\right), \quad [10]$$

where λ_i is a corresponding penalty constant.

2.2. Binary-Coded Genetic Algorithm for Variable Selection

The applicability of genetic algorithms in combinatory optimization problems, of which variable selection is a good example, has been proven in various studies, [23-30] in which a binary-coded genetic algorithm (BCGA) is used for the optimal subset selection for either a principal component regression, partial least-squares regression or multiple linear regression models. As an example, Kepplinger *et al.*[29] and Sorsa *et al.*[24] got realistic and reliable results by applying BCGA with repeated leave-multiple-out cross-validation as the objective function for model selection, [24,29] whereas in the approach of Barycki *et al.*[30] the objective function in variable selection was based on the training set only, and the validation of the models was carried out with leave-one-out cross-validation. [30]

Metaheuristic optimization algorithms, like the Genetic Algorithm, are feasible alternatives in situations where exhaustive search is too time-consuming. This is often true especially when using repeated objective function evaluations for a single iteration round. It can be deduced that the success of the GA in variable selection is based on the coding of the chromosome and on the capability of the algorithm to operate in a wide search area. Unlike in the parameter identification task, where the binary coding is converted into the decimal form, in a variable selection problem each of the binary coded chromosomes represents the variable subset candidate. In addition, the constraints for the objective function can be implemented with simple scaling factors[31], which has been proven useful along the repeated cross-validation for restricting the number of predictors in the final model. [24]

The steps of the GA can be roughly divided into four steps: 1) initialization, 2) ranking, 3) selection and 4) recombination. [31] The initialization of the population is carried out randomly by tossing a biased coin, and to yield more parsimonious models, 20% of the genes are initialized as ones and 80% as zeros. The ranking of each of the variable subset candidates is based on the relative fitness value of each individual in the population. The formulated

minimization problem is converted into a maximization problem by using the inverse of the objective function. Thus, the fitness of the variable subset candidate using repeated cross-validation as the objective function can be given as:

$$\text{fitness} = \frac{1}{\frac{1}{4N}\sum_{l=1}^{4N}\sum_{j=1}^{n}(y_{i,CV_l} - [b_{0,j} + \sum_{i=1}^{k} gb_{i,j} \ln x_{i,j,CV_l}])^2} - \lambda_1 VIF_{\max} \quad [11]$$

In the selection phase, roulette wheel selection is applied. In roulette wheel selection, the probability of an individual to be selected is proportional to its fitness. To enhance the rate of convergence, Kepplinger *et al.* [29] suggested an exponential transformation of the fitness function to increase the selection probability of the best individuals. Similar as in [32], the number of parents selected for the next population is $n_{pop}/2$.

The recombination of the selected individuals is typically carried out with crossover and mutation. In crossover, pre-selected parts of two individual chromosomes are swapped. The crossover point is selected randomly. In this work, a single-point crossover was used. The crossover rate can be regulated with the crossover probability. The crossover between two individuals occurs if the generated random number is below the pre-selected crossover probability. [31] Nowadays due to the increased computational capacity, the crossover probability is usually well above 0.6.

Trapping at a local minimum can be avoided by allowing a random mutation to occur during the recombination. [31] In this work, a single-bit mutation was used, but the main factor for convergence is associated with crossover. The mutation probability (P_M), i.e. the probability at which the mutation occurs, is a crucial computational parameter in a GA. In the literature, there is no strict consensus for a proper value for P_M. In this work, the mutation probability was chosen to evolve deterministically during the iterations, as was presented by Bäck and Schultz[33]. However, the equation is slightly modified as presented in the second term of the product such that $P_M \to 0$ when $k \to T$. Thus, the mutation probability of a chromosome is given as[33]

$$P_M = P_C \left[\left(2 + \frac{l-2}{T-1}k\right)^{-1} - l^{-1}\right], \quad [12]$$

where l is the length of the chromosome, P_C is the crossover probability, P_M is the mutation probability, k is the iteration and T is the maximum number of iterations. The solution presented yields a convergence that preserves the information within the population, and has a high probability of improving a single solution candidate. To preserve the current best solution in the population, the population is treated with the elitism at each of the iterations. The structure of the proposed variable selection algorithm is presented in **Figure 1**. The outer loop of the algorithm consists of selection, ranking and recombination. In the inner loop of the algorithm, the conditioned data-matrix is constructed and the least-squares solution for each of the model candidates is accessed, after which the fitness of each individual in the population is evaluated $4N$ times based on the repeated cross-validation. The fitness value of an individual is defined as the mean of $4N$ split repetitions.

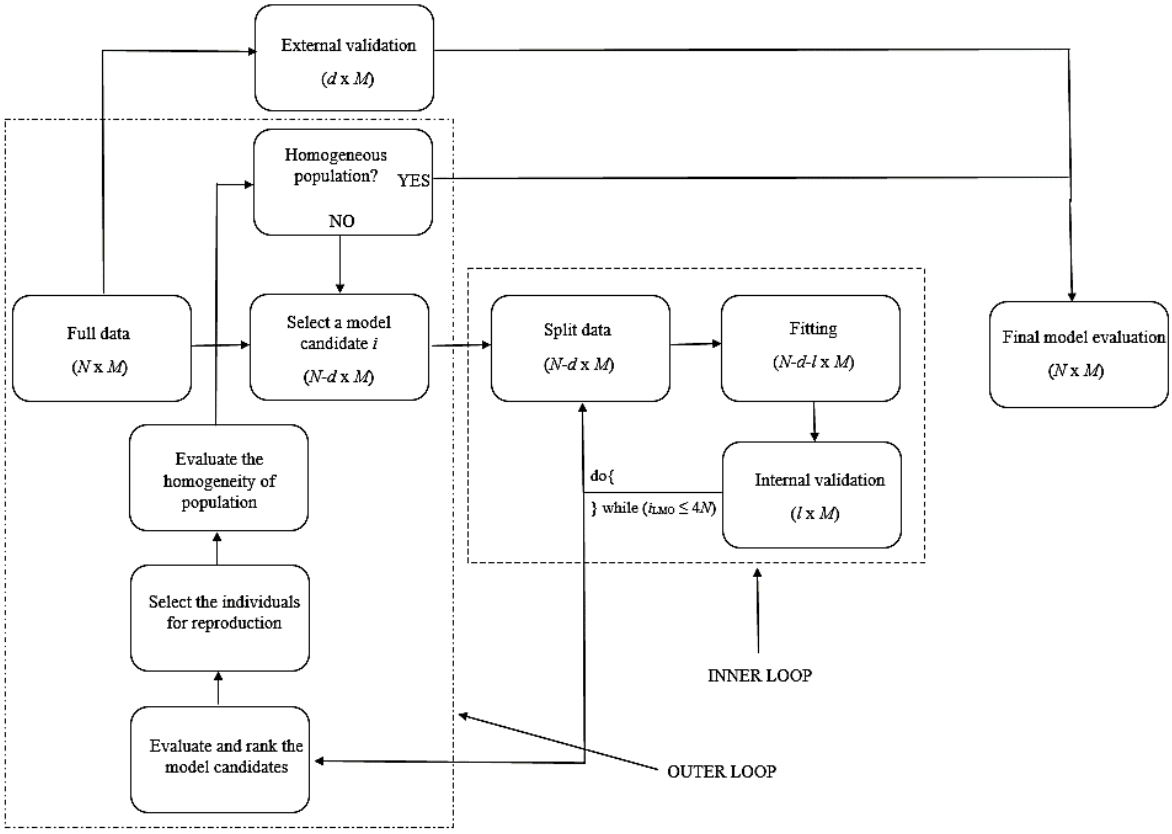


Figure 1. Cross-validation as an objective function when applying Genetic Algorithm for variable subset selection.

3. Experimental data and evaluation of the fit

The experimental data in this study was adopted from Vuolio *et al.*[14] and is collected from the primary hot metal desulfurization process at SSAB Raahe, Finland. The considered data matrix consists of 23 variable candidates, so the number of possible variable subset combinations is $2^n - 1 = 2^{23} - 1 = 8388607$, which practically rules out the exhaustive search as the selection strategy. The most important variables that are included in the data set are the injection parameters, reagent properties, temperature of the hot metal and chemical compositions of slag and metal phases before injection. The selected dependent variable is the rate constant for the desulfurization. Prior to variable selection, the data was pre-treated with outlier removal. After outlier-removal procedure, the full data set consists of 40 rows of data overall.

The hot metal samples were taken instantly before and after desulfurization treatments to obtain a representable set of samples, and to minimize the effect of sulfur resulfurization via the permanent phase contact. During the injections, the carrier-gas flow rate and immersion-depth of the injection lance were held constant, which is why the value of Q_{tot} can be considered as a pure function of temperature and injection rate of limestone at constant pressure. The production data from the hot metal desulfurization plant was expanded by measuring the particle size distributions of the reagent prior to injection. The particle size distributions for the reagents were determined by laser-diffraction analysis prior to injection. The analysis of the hot metal samples was carried out by the C-S-combustion method and by X-Ray Fluorescence (XRF). The analysis of the slag phase was carried out for S with the C-S-combustion method and for the rest of the compounds with X-Ray Diffraction (XRD) and XRF. The existing oxide phases in the slag were qualitatively evaluated with XRD, whereas the oxide phases were calculated based on the chemical composition measured with XRF. The hot metal samples were taken instantly before and after desulfurization treatment to obtain a representative set of samples, and to minimize the effect of sulfur resulfurization via the permanent phase contact. During the injections, the carrier-gas flow rate and immersion-depth of the injection lance were held constant.

4. Results and discussion

The variable selection algorithm was employed for different combinations of computational parameters in order to evaluate their significance. During the simulations it was observed that the most reliable results were obtained with population size of n_{pop} = 200 individuals and with a crossover probability of P_C = 0.9. It was also noticed that the selection algorithm was not sensitive to crossover probability. This property was associated with a relatively high mutation probability. For this reason, the implementation strategy for reproduction can be considered sufficiently robust. The maximum number of iterations was not pre-defined as the homogeneity of the population was chosen as the convergence criterion. The mutation probability evolved deterministically during the iterations of the GA phase. It was observed that realistic prediction results were obtained when data was split in the validation loop such that 56% was used for training, 31% for internal validation and 13% for external validation. The interdependence of the computational parameters of the GA and repeated cross-validation were evaluated by analyzing the performance of the prediction models. The quantitative figures of merit for model performance were the coefficient determination R^2, mean absolute error for prediction (MAE) and sum of squared error (SSE). The selection probability of a variable was associated with its importance.

4.1. Model selection and performance evaluation

The convergence of the variable selection algorithm is illustrated in **Figure 2**. It is seen that the algorithm converges towards a low value of the objective function presented in **Eq. 8**, and the average of the population approaches the stationary point after 10 iterations. It should be noted that as at the beginning of the search, the population is very far from the optima, Figure 2 is scaled such that it illustrates the convergence rate of the whole population properly. From the shape of the line with triangular markers it can be observed that the deterministic mutation schedule increases the diversity of the population, as the rate of change in the population average has a decreasing trend. For these reasons it can be said that the algorithm often not only converges to a feasible minimum, but is capable of going through various subset candidates and excluding the irrelevant ones.

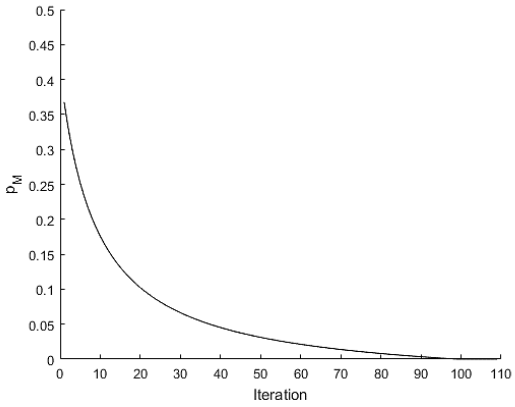


Figure 2. Convergence of the variable selection algorithm and the evolution of the deterministic mutation probability.

Figure 3 shows the prediction result for the logarithmic case. It is seen that the selected subset can explain the changes in the rate constant, even though the model contains only linear interactions between the independent and dependent variables. The quantitative measures of fit can be considered very good, as the coefficient of determination for the full data set is high R^2 = 0.85 and mean absolute error defined for the rate constant is relatively low MAE = 0.0145 (1/min). The corresponding quantities for the end sulfur content by applying the proposed model for the rate constant are R^2 = 0.87 and MAE = 0.0012 (wt-%), which are consistent with the values reported in our previous work [14]. The quantitative measures of fit can be considered highly sufficient, as the prediction model form, as well as the parameters, cannot be considered optimum because of the non-optimal model structure and the linear objective function. However, the modeling results can be considered very promising.

The average performance of the internal validation data set can be considered as good as for the external data set, as the coefficient of determination and the squared error are very sensitive to the deviation, which results in slightly smaller R^2 and SSE values for the internal data set. However, as the variable selection algorithm reduced the dimensionality of the original data set significantly and the suggested model has very high figures of merit, the algorithm can be considered efficient. The best selected subset with the corresponding data splits and quantitative measures of the model for each of the data set splits are given in **Table 2**.

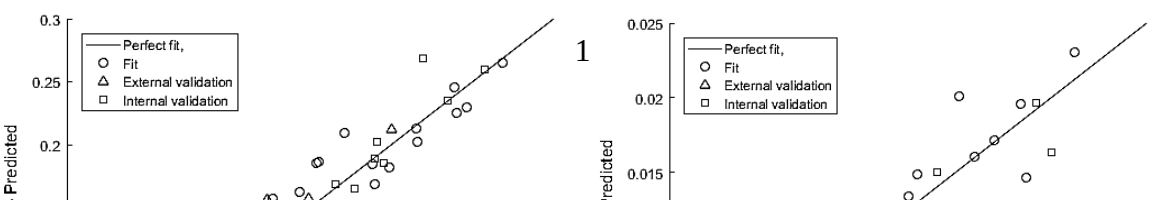


Figure 3. Predicted values for the rate constant for different data sets with the variables given in Table 2.

Table 3. The best subset of variables and the prediction performance of the model for the 1st order rate constant and end sulfur content using different datasets. The selected variables were d_{80}, Q_{tot}, $\dot{m}$, m_{Fe} and P.

y	Data-set	R^2	MAE	SSE	(%) of full data
k_{tot} (1/min)	Training	0.87	0.0145 (1/min)	0.0073	56
	Internal validation	0.81	0.0182 (1/min)	0.0067	31
	External validation	0.95	0.0053 (1/min)	0.0002	13
	Full data	0.85	0.0145 (1/min)	0.0142	100
$[S]_t$ (wt-%)	Training	0.85	0.0014 (wt-%)	$7.21 \cdot 10^{-5}$	56
	Internal validation	0.88	0.0013 (wt-%)	$3.93 \cdot 10^{-5}$	31
	External validation	0.98	0.0004 (wt-%)	$1.23 \cdot 10^{-6}$	13
	Full data	0.87	0.0012 (wt-%)	$1.13 \cdot 10^{-4}$	100

4.2. Analysis of the selected variables

The parameters for the best selected subset are presented in **Table 4**. As the objective function used is different from the one that was used in our previous work[14], the parameter values are only in qualitative agreement.[14] The reasoning behind the selected variables as well as the model parameters are presented in the following.

Table 4. The modeling parameters for the best selected subset.

b_0	b_1	b_2	b_3	b_4	b_5
1.50±0.64	0.15±0.02	0.09±0.01	-0.08±0.01	-0.13±0.05	0.13±0.06

Notes: The parameters are presented in a form of $E(x) \pm \sigma$, where E(x) is the mean of $4N$ repetitions with corresponding standard deviation.

The best prediction results for the rate constant, and thus for the end content of sulfur, were acquired with the model that includes the following variables:

1) the particle diameter corresponding a value of 80% in the cumulative volume-based particle size distribution,
2) total gas flowrate,
3) mass flowrate of the reagent
4) the mass of the hot metal.

In addition to the aforementioned variables, three other variables, namely the Chromium (Cr), Phosphorus (P) or Carbon (C) contents in the hot metal are suggested to have a minor effect on the rate constant, but to a significantly lesser extent than the aforementioned variables. The most important variables selected by the algorithm are consistent with implications concerning the rate of transitory contact reaction. In fact, the four main variables chosen by the algorithm correspond to those suggested in an earlier study by the authors[14]. As the interfacial area between the hot metal and the transitory phase is mostly defined by the solid surface area, the particle size distribution of the reagent particles can be considered as a significant predictor. The effect of the mass flowrate can be considered to affect the same attribute as does the size of the reagent particles. The flowrate of the gaseous compounds formed in the decomposition of calcium carbonate contained by the reagent employed in the experimental data can be associated with either scattering of the reagent particles into smaller swarms of particles or to increased stirring of the metal bath. However, as was experimentally proven by Irons [34], the stirring effect related to the decomposition reaction is negligible compared with the bath mixing induced by the carrier gas. [34] In addition, the results given by Lindström *et al*. [35] support the postulated scattering effect. [35]

The effect of composition of the metal bath could be associated with thermodynamic driving force; as the C dissolved in the metal bath is near the saturation limit, and thus acts as a high de-oxidizer, the high C content is beneficial for hot metal desulfurization. However, this

particular case is questionable for two main reasons: Firstly, the initial sulfur content is highly dependent on the operation of the blast furnace, and thus on the carbon content. Secondly, as the hot metal desulfurization operates very far from the thermodynamic equilibrium state, the effect of the activity of oxygen could be meaningful in the case of single particles, but non-observable when concerning the net rate of desulfurization. The reasoning is similar in the case of P and Cr. Consequently, it can be said that the algorithm can be applied in the selection of the most relevant predictor variables amongst a high number of variable candidates.

These results also highlight the dominance of the variables that are related to the transitory phase contact: the transitory contact reaction determines the rate of desulfurization, while the desulfurization via the permanent phase contact is of secondary importance. This can be associated with the magnitude of the interfacial area between the extracting phases, as the rate controlling step is the mass-transfer and not the thermodynamic driving force when the sulfur content in the hot metal operates very far from the thermodynamic equilibrium state, which is often true in the case of industrial hot metal desulfurization. However, it is possible that the effect of slag composition on the overall rate of reaction would be observable from the data with lower concentration areas, namely when sulfur content is well below 0.01 wt-%.

4.3. Evaluation of the robustness of the algorithm

To evaluate the robustness of the variable selection algorithm, the variable selection procedure was repeated 100 times. The significance of each of the predictor variables was evaluated with the rate of selection, i.e. the hit-rate, which is determined as:

$$H\ (\%)\ = \frac{n_{select}}{n_{rep}} \cdot 100, \qquad [13]$$

where H is the hit rate, n_{select} is the number of times that variable i is selected and n_{rep} is the number of repetitions of the variable selection algorithm. From **Figure 4** it is observable that the hit rate for the most significant injection parameters, i.e. particle size distribution, mass flowrate and total gas flowrate vary from 90% to 100%. When the variable selection is carried out with GA, the hit-rate of selecting mass of the hot metal phase as an explanatory variable is only 62.5%, which can be explained by the small variance of the mass of the hot metal in the input data under the conditions of this study. In this case, the small variance in the input data

makes the random split of the data-matrix inefficient. This being so, the relatively low selection probability for the mass of the hot metal relates to experimental conditions rather than to its low explanatory power.

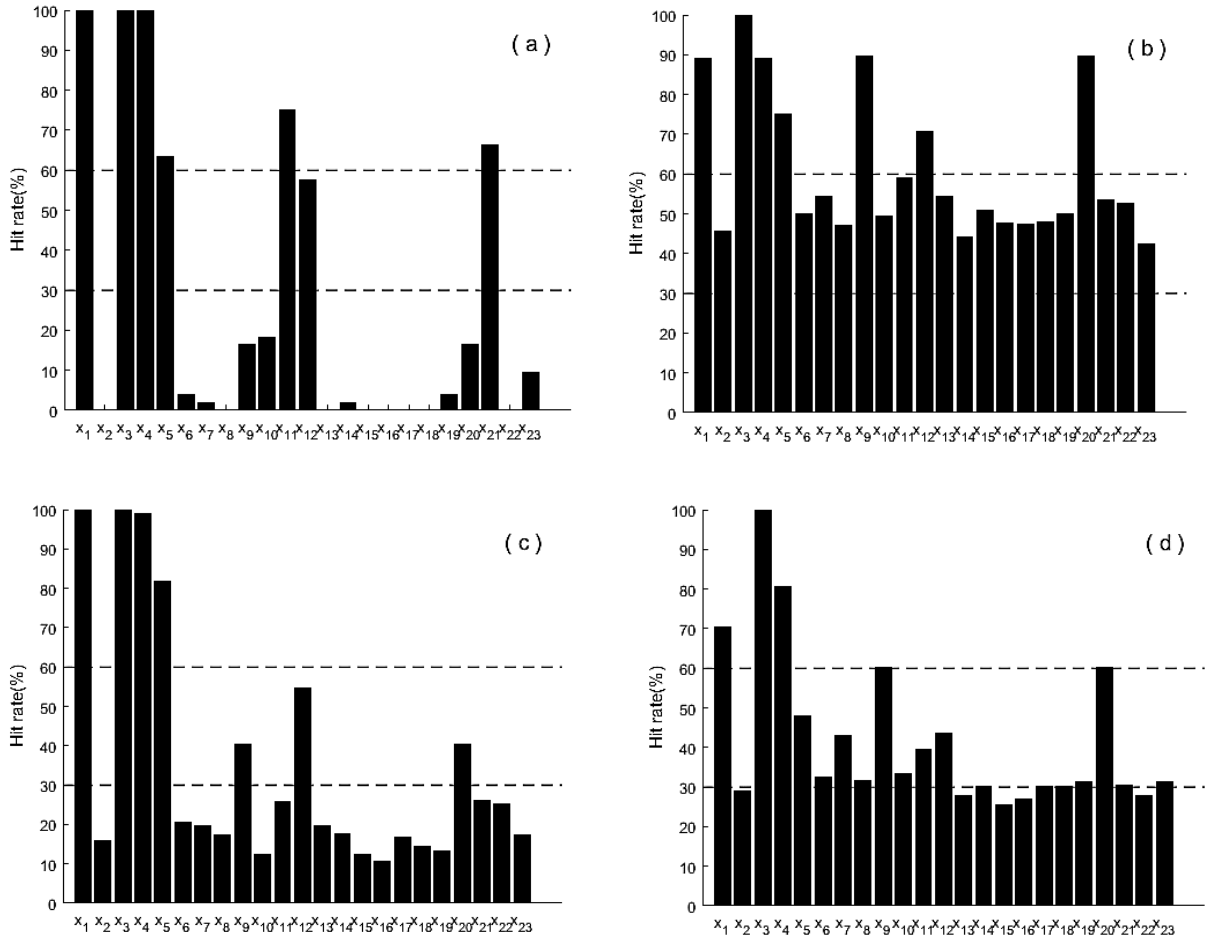
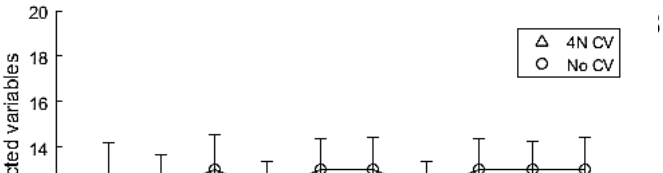


Figure 4. *Selection probabilities of predictor variables for different cross-validation scenarios with different* population *sizes. The dotted lines represent the chosen levels of significance. a)* n_{pop}= 100, *4N LMO-cross-validation; b)* n_{pop} *= 20, no cross-validation; c)* n_{pop} *= 20, 4N LMO cross-validation; d)* n_{pop} *= 100, no cross-validation.*

In **Figure 5,** the most frequent number of selected predictor variables between repetitions with the corresponding standard deviations is presented. It is clearly seen in figure that a Genetic Algorithm that uses repeated cross-validation as the objective function tends to select parsimonious models with high repeatability, which is consistent with the results of the previous studies using the same kind of objective function for variable selection problems[19-21,24,29]. The bar graph on the left hand side supports this interpretation, as the selection frequency distribution between the repetitions is highly skewed to left in the case of the cross-validated objective function.

The increase in the population size used would intuitively result in an increased selection probability of the important variables, as the probability that a good variable subset candidate is included in the population increases with the population size. However, under the conditions of this study, the effect of population size on the convergence of the algorithm seems to be dependent on the objective function. In the case of a noisy data set there are several local minima to which the algorithm can converge. As the number of local minima is large, there is no significant improvement achieved by increasing the number of initial guesses, but actually the performance of the algorithm is worse. In the case of repeated cross-validation, the increase in the population size increases the probability of selecting a parsimonious model that contains only the relevant variables, which is mainly because the predictive power of the explanatory variables is evaluated for multiple data splits and not for only the full training set. However, the increase in the population size increases the computational load per iteration. As an example, for a population size of 20 individuals, it takes around 20 seconds for the algorithm to converge. With a relatively large population (~200 individuals), the computational time is around 7 minutes. For this reason, the proper population size is a compromise between the computational load and the desired accuracy and reliability. In the light of the reasoning above, it is evident that a multi-objective function based GA outperforms the single-objective function in variable selection problems. This being so, the results of this study not only support the benefits of using search algorithms in the model selection for complex experimental data sets, but underline the significance of cross-validation in building prediction models. However, the dependent variable needs to be carefully selected in order for the effect of explanatory variables to be observable from the data.



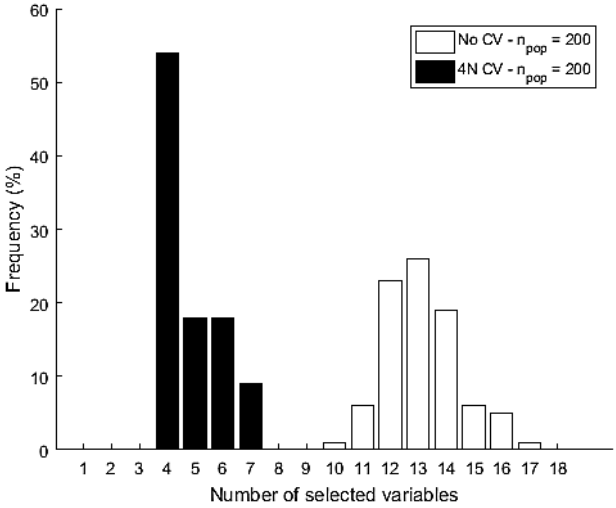


Figure 5. A comparative illustration of the effect of the objective function on the most frequent number of selected explanatory variables between the repetitions as a function of the population. size.

5. Conclusions

A Genetic Algorithm-based variable selection was applied to the prediction of hot metal desulfurization kinetics. Based on the results of this study, the proposed variable selection algorithm can reveal the most significant variables for a multivariable regression model, and can also extract collinear or otherwise irrelevant variables from a sufficiently large industrial scale data set. It was observed that the most relevant variables were:

1) the particle diameter corresponding a value of 80% in the cumulative volume-based particle size distribution,
2) total gas flowrate,
3) mass flowrate of the reagent
4) the mass of the hot metal.

The selected variables as well as the identified model parameters are consistent with the findings of previous studies. It was also found that the selection probability of the most relevant variables can be increased by applying a relatively large population for the Genetic Algorithm, which makes use a repeated leave-multiple-out cross-validation as the objective function. However in the case of the process under study, for optimized prediction performance the identification of the model parameters needs to be conducted by making use of a non-linear objective function. All and all, the proposed approach can be considered as a robust alternative for model selection based on similar data-sets.

Acknowledgements

This work was conducted within the Symbiosis of Metal Production and Nature (SYMMET) research program funded by Business Finland. The financial support of Technology Industries of the Finland Centennial Foundation, the Tauno Tönning foundation, the Finnish Foundation of Technology Promotion as well as the Finnish Cultural Foundation are also acknowledged.

NOMENCLATURE
SYMBOLS AND ABBREVIATIONS

A	Area	m^2
b_i	Regression coefficient for a variable i	–
k_{tot}	1st order rate constant for hot metal desulfurization	1/s
m	Mass of a phase	kg
R^2	Coefficient of determination	–
y	Output variable	–
$\hat{y}$	Predicted output variable	–
w	Mass fraction	–
X	Data–matrix	–
[ ]	Species dissolved in hot metal	–
( )	Species in slag phase	–
{ }	Species in gas phase	–
< >	Solid species	–
ANN	Artificial neural network	–
CV	Cross-validation	–
MLR	Multivariable linear regression	–
MAE	Mean absolute error of prediction	–
SVM	Support vector machine	–
SSE	Sum of squared errors	–

References

1. Oeters F, Strohmenger P, Pluschkell, Arch. Eisenhüttenwes. 1973, 44, 727.
2. Ohguchi S, Robertson DGC, Deo B, Grieveson P, Jeffes JHE, Ironmaking Steelmaking 1984, 11, 202.
3. Oeters F., Steel Res. 1985, 56, 69.
4. Datta A, Hareesh M, Kaira PK, Deo B, Boom R., Steel Res. 1994, 65, 466.
5. Rastogi R, Deb K, Deo B, Boom R., Steel Res. 1994, 65, 472.
6. Deo B, Datta A, Kukreja B, Rastogi R, Deb K., Steel Res. 1994, 65, 528.
7. Deo B, Boom R. Fundamentals of Steelmaking Metallurgy, Prentice Hall International, Hertfordshire, United Kingdom, 1993.
8. Visser H-J, Boom R., ISIJ Int. 2006, 46, 1771.
9. Seshadri V, da Silva CA, da Silva TA, von Krüger P., ISIJ Int. 1997, 37, 21.
10. Rodríguez YC, Múzquiz GG, Torres JRP, Vidaurri LER.,Adv. Mater. Sci. Eng. 2012; 2012: 1.
11. Ma W, Li H, Cui Y, Chen B, Liu G, Ji J, ISIJ Int. 2017, 57, 214.
12. Barron MA, Hilerio I, Medina DY., OJAppS 2015, 5, 295.
13. Vargas-Ramirez M, Romero-Serrano A, Morales R, Angeles-Hernandez M, Chavez-Alcala F, Castro-Arellano J., Steel Res. 2001, 72, 173.
14. T. Vuolio, V. -. V. Visuri, S. Tuomikoski, T. Paananen,T. Fabritius, Metall. Mater. Trans. B 2018, 49, 2692.
15. H. Saxén, F. Petterson, ISIJ Int. 2007, 47, 1732.
16. B. Mahanta, N. Chakraborti, Steel Res. Int. 2018, 89, Article number 1800121.
17. W. Wang, J. Liu, X. Liu, The Open Automation and Control Systems Journal 2015, 7, 966.
18. X. Wang, M. Han, J. Wang, Engineering Applications of Artificial Intelligence 2010, 23, 1012.
19. K. Baumann, N. Stiefl, *J. Comput. Aided Mol. Des.* 2004, 18, 549.
20. K. Baumann, QSAR Comb. Sci. 2005, 24, 1033.
21. K. Baumann, Trends in Analytical Chemistry 2003, 22, 395.
22. F. Harrell: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis, Springer, New York, NY, USA, 2001.
23. A. Barros, D. Ruthledge, Chemometrics and Intelligent Laboratory Systems 1998, 40, 65.

24. A. Sorsa, *Prediction of Material Properties Based on Non-destructive Barkhausen Noise Measurement*, Doctoral Thesis, University of Oulu, 2013, Oulu, Finland.
25. D. Broadhurst, R. Goodacre, A. Jones, J. Rowland, D. Kell: *Analytica Chimica Acta* 1997, 348, 71.
26. U. Depczynski, V. Frost, K. Molt, *Analytica Chimica Acta* 2000, 420, 217.
27. C. Fan, F. Xiao, S. Wang, *Applied Energy* 2014, 127, 1.
28. M. Ohenoja, A. Sorsa, K. Leiviskä, Computers 2018, 7, Article number 60.
29. D. Kepplinger, P. Filzmoser, K. Varmuza, Unpublished work, https://arxiv.org/pdf/1711.06695.pdf.
30. M. Barycki, A. Sosnowska, K. Jagiello, T. Puzyn, J. Chem. Inf. Model 2018, 2467.
31. D. Goldberg: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, Boston, MA, USA, 1989.
32. A. Gharahbagh, V. Abolghasemi: *WASJ*, 2008, 5, pp. 137–142.
33. T. Bäck, M. Schütz: Proceedings of the 9th International Symposium on Methodologies for Intelligent Systems, 1996, 158.
34. G. Irons, Ironmak. Steelmak., 1989, vol. 16, 28.
35. D. Lindström, P. Nortier, D. Sichen: Steel Res. Int. 2014, 8676.