Efficient Approximation of Time Consuming Models for their use in Optimization Frameworks

Marco Vannucci, Giacomo F. Porzio, Barbara Fornai, Valentina Colla, Marco Vannocci
TeCIP institute, PERCRO Laboratory
Scuola Superiore Sant'Anna
Pisa, Italy
mvannucci{g.porzio, b.fornai, colla, m.vannocci}@sssup.it

Abstract — Several widely used model optimization techniques such as, for instance, genetic algorithms, exploit an intelligent test of different input variables configurations. Such variables are fed to an arbitrary model and their effect is evaluated in terms of the output variables, in order to identify their optimal values according to some predetermined criteria. Unfortunately some models concern real world phenomena which involve a high number of input and output variables, whose interactions are complex. Consequently the simulations can be so time consuming that their use within an optimization procedure is unaffordable. In order to overcome this criticality, reducing the simulation time required for running the model within the optimization task, a novel method based on the combination of clustering and interpolation techniques is proposed. This technique is based on the use of a set of pre-run simulations of the original model, which are firstly used to cluster the input space and to assign to each cluster a suitable output value within the output space. Subsequently, in the simulation phase, an ad-hoc interpolation is performed in order to provide the final simulation results. The proposed method has been tested on two complex models related to the steel making industry: the first one concerns the optimization of blast furnace, the other one the operation of a EAF scrap pretreatment plant. The proposed approach has obtained good results in terms of accuracy and time-efficiency.

Keywords – simulation, modeling, clustering, optimization

I. INTRODUCTION

Process optimization is a key issue in many real world applications and, in particular, in the industrial field in order to improve product quality, maximize the production, minimize consumptions and working times [1] [2].

When the relations between the process parameters and the process outputs to be optimized are not completely known, the strategy adopted by many optimization methods consists in the test of different parameters configurations which are fed to a model simulating the interested process. Subsequently, on the basis of the output of the model, new sets of parameters are tested until the optimum is reached. It is worth to cite, among these optimization methods, several well known evolutionary algorithms such as genetic algorithms (GA) [3], which performs the candidate solution search simulating natural selection, the particle swarm optimization (PSO) [4], which explores the search space mimicking the movements of birds swarms and the ant colony optimization (ACO) [5] method, which reproduces the search for resources of a ants population. Among the previously mentioned optimization methods, GAs are the most widely used and known and well represent this class of approaches; for this reason GAs are taken as reference optimization method within this paper without limiting the generality of the discussion.

All these methods are based on the intensive exploitation of a model of the process, which is frequently run in order to assess the performance of the tested candidate solutions, thus, if the time required by this latter model simulation is large, such methods are unusable due to the computational burden. Let us consider an exemplar realistic situation that depicts the problem entity: if GAs are used for the optimization of a process whose simulation requires 10 minutes and 50 individuals are employed through 100 generations, the whole optimization process requires more than 1 month of continuous computation.

This problem has been encountered for the optimization of two flow-sheeting models referring to a steel plant and, in particular, a blast furnace (BF) and a scrap pretreatment plant. The complexity of the models which relate the inputs and the outputs of the two models made impossible any optimization step and raised the issue of the reduction of simulation time.

In order to overcome this criticality a novel method based on the combined use of clustering and interpolation techniques has been developed. This approach aims to the substitution of the time consuming model simulation with one approximation which is as reliable and fast.

The paper is organized as follows: Sec. II presents two industrial problems concerning the BF performance optimization and the operation of the pyrolysis process for zinc removal whose criticality in the use in an optimization framework led to the development of the novel approach. In section III the proposed method is described in detail and in sections IV and V its performances are evaluated on the approximation of a several representative literature functions and on the approximation and subsequent optimization of the
previously introduced industrial problems. Conclusions and final remarks are drawn in section VI.

II. CRITICALITIES IN THE OPTIMIZATION OF COMPLEX TIME CONSUMING MODEL

The BF is the heart of the iron and steelmaking production process. Its energy consumption and top gas emissions, especially those responsible for climate changes, are the most intensive of the whole production chain: the optimization of its operation is therefore of fundamental importance. An effective tool to achieve such an optimization is to carry out a process simulation which takes into account the basic laws of thermodynamics as well as the chemical equilibria and reactions underlying the functioning of the BF. A commercial software which has proven to be useful in the literature [10] for such a modeling is Aspen Plus of the Aspentech Engineering suite. Aspen Plus is a flowsheeting program which enables to carry out and easily visualize energy and material balances, simulate scenarios and solve simple optimization problems.

Material components of the entering streams, the chemical reactions that govern it and some process parameters such as flow rate, temperature and pressure are the model inputs. The model outputs useful information such as product composition, temperatures, and flow rates. A scheme of a BF model realized in Aspen Plus is presented in Fig.1.

![Figure 1. Scheme of the Aspen Plus BF model](image)

Such model has been developed to simulate the behaviour of an existing BF and with a focus on top gas generation and its composition, fuel and minerals consumption as well as product generation and electrical power consumption. The model is part of a larger simulation aiming at both the overall optimization - in terms of energy consumption and greenhouse gases generation - of an integral cycle steelmaking plant and the elaboration of possible scenarios where different fuel mix are employed. Therefore a validation phase has been required to assess its accuracy and reliability for a wide range of operating conditions through a comparison of model outputs and real process data.

The flow sheeting nature of the program allows to interconnect different unit operation models in a diagram that represents the desired process. However, since the solution of the problem sequentially proceeds through the different units, when the model becomes more and more comprehensive, the number of components and input variables increase together with the complexity of the simulation, that can be very time consuming. As an example, to run the BF model, more than 35 different input variables need to be specified at each run, generating 23 output variables of interest. Such a high number of parameters, together with the complexity of the flow sheet, resulted in a solution time (depending on the machine characteristics) in the range of tens of seconds. Let us estimate the average simulation time for 1 model run as 20 seconds. Considering GAs as the optimization method and the high number of involved input and output variables, the number of individuals forming the GAs population and the number of GAs generations should be both high, e.g. 100 individuals and 50 generations. This choice corresponds to 28 hours of continuous computation for one single optimization. Such a high computational time is not affordable within the application the model has been developed for, thus a new method for approximating the model characterized by a sensible cut of simulation time has been developed.

A second example which required the use of the proposed approximation method is constituted by another Aspen Plus simulation of a steel-related industrial system. This model is intended to demonstrate the feasibility as well as predicting the key performances of an innovative system for simultaneous preheating and by-products recovery (such as zinc) from steel scrap which has to be subsequently fed into an Electric Arc Furnace (EAF).

The technology on which the system is based is currently in the research and development stage, utilises plastic waste materials as a feedstock (e.g. car tires, automotive shredder residues, waste electrical and electronic equipment) and is based on two alternative process routes: pyrolysis and gasification. In both solutions, the objective is to obtain a synthetic fuel gas rich in hydrogen chloride which is burnt in a furnace containing the scrap. In literature different examples exist in which plastic pyrolysis or gasification processes (as well as coal or biomass pyrolysis and gasification) have been both simulated with Aspen Plus [12] [13] [14]. It was therefore deemed convenient to utilise the software to obtain a reliable process model. The syngas which passes through the furnace shaft heats the burden of steel scrap, simultaneously initialising chemical reactions between hydrogen chloride and zinc oxide on the steel surface and allowing the removal of zinc surface coatings, according to the stoichiometry reported in (1):

\[
\text{ZnO} + 2 \text{HCl} \rightarrow \text{ZnCl}_2 + \text{H}_2\text{O} \quad (1)
\]
The zinc-containing flue gas is extracted from the furnace and purified in a water scrubber unit, zinc is extracted from the gas through further chemical processing, respectively leaching, cementation, and electrolysis steps. The preheated scrap is then charged into the EAF for melting into liquid steel. The process of scrap preheating and zinc removal promises several advantages in terms of both environmental and economical advantages: the EAF operations should require less energy given the heat content of the incoming metal, and the slag formation should be limited; the utilisation of waste plastics reduces the environmental impacts on landfill or disposal of the materials; last but not least the recovery of valuable by-products such as zinc enhances the profitability of the process.

Figure 2. BFD of the preheating and zinc removal system model

In order to evaluate the process performances, a model has been realized which is able to predict the different subprocesses behavior according to variations in the operating parameters; as an example, plastic feed composition and characterization (content of C, H, O, N, S, Cl, percentage of ash and volatile matter), pyrolysis temperature, addition of fuel gas in the preheating furnace, and acid neutralization in the water scrubbing column are fundamental parameters which influence the process results. Once again, the presence of different sub-plants closely interacting with each other (e.g. part of the scrubber top gas is recycled back to the furnace), makes the simulation time consuming and the optimisation procedure unaffordable, especially if more than one objective has to be optimized.

For the purpose of a simplified illustration, simulations were carried out considering the plastic pyrolysis and syngas production route only (thus excluding the gasification module from the simulation) in order to evaluate the trend of two basic performance indicators, zinc production and CO2 emissions, based on the variation of eleven different variables. A single simulation execution requires slightly less than 10 seconds, but when more iterations are required (e.g. to carry out sensitivity analyses) and the CPU resources start becoming more scarce, the computational time increases more scarce, the computational time increases exponentially reaching hours of calculation. It is clear that such performances are not acceptable unless it is required to carry out repeated simulations only once or a few times. The proposed approximation method is therefore an elegant and effective way to overcome the drawbacks of the flowsheeting simulation without giving up on simulation accuracy.

III. NOVEL METHOD BASED ON CLUSTERING AND INTERPOLATION TECHNIQUES FOR MODEL APPROXIMATION

As it emerges from the industrial problems described in section II, the main issue concerns the model simulation time which, once embedded into an optimization algorithms requiring iterated runs of the simulation, implies unacceptable computation time.

The approach described in this paper aims at substituting such model simulation with an approximation which should be reliable in terms of discrepancy with respect to the results provided by the original simulation and, mostly, fast enough to allow its use within the optimization framework.

The basic idea is to use the original model in order to perform a reasonable set of predefined simulations and to use the outputs of these simulations for extrapolating the behaviour of the model in the whole input space. Obviously this latter extrapolation phase should have very short computation time in order to fulfill the main method constraints. In a more formal way,

\[ \text{given an input space } I, \text{ an output space } O \text{ and an} \]
\[ \text{original model } M \text{ which maps the elements } x \in I \text{ into an} \]
\[ \text{element } y=M(x) \in O \]
\[ \text{a set of } N \text{ points } \{x_1, \ldots, x_N\} \in I \text{ is chosen and the} \]
\[ \text{corresponding model outputs } \{y_1=M(x_1), \ldots, y_N=M(x_N)\} \in O \]
\[ \text{are calculated} \]
\[ \text{exploiting the couples } (x,y), \text{ an approximate model} \]
\[ M \text{ of } M \text{ is built.} \]

When dealing with problems like the one depicted in Sec. II (i.e. high number of input and output variables linked by very complex relationships), the use of techniques such as neural networks [7] or radial basis function networks [8] is not fruitful due to the high-dimensionality of the problem which requires a high dimension network and, by consequence, an unaffordable number of training samples. The approximation method introduced in this paper addresses to solve complex, high dimensional problems, which can not be faced through standard techniques.

The main characteristic of this method is that it is based on a suitable subdivision of the input space into a set of \( k \) regions, that is performed by means of a clustering algorithm. Subsequently, each region is associated to an input vector \( C_i \) which acts as reference vector for all those observations belonging to that particular cluster. This clustering operation has the advantage of drastically reducing the complexity of the input space at the expense of an
information loss; thus the choice of the clustering algorithm and of the number of clusters is a key issue of the method and should reflect as much as possible the original distribution of data samples.

Once the clusters have been formed and centers determined, a representative output value \( Y_i \) is assigned to each center \( C_i \). Such value is calculated as the mean output of all the observations belonging to each cluster. This assignment operation finalizes the tuning phase (depicted in Fig. 3) of the model approximator.

![Figure 3. Scheme of the tuning phase of the model approximator](image)

Within this work two different clustering methods have been evaluated: the first one employs the k-means algorithm [9], the second one is based on the use of a Self Organising Map (SOM) [6]. K-means is a well known algorithm used to classify or to group objects based on their features into a pre-established number \( K \) of clusters. The clustering is performed by minimizing the sum of squares of distances between data and the corresponding cluster centroid. In the case of k-means clustering, the reference vectors \( C_i \) used within the proposed method coincide with the centroids calculated by the algorithm. The main advantage of this algorithm lies in its speed: in facts, it reaches the convergence in a few iterations. On the other hand, its main disadvantage consists in the quality of the reached solution which, in particular cases, can be far away from the global optimum.

The SOM is a particular kind of artificial neural network which performs unsupervised learning. The SOM is used to map high dimensional input samples into a smaller dimensional space. The main advantage of this method with respects to k-means is that it preserves distribution and topology of the original training samples: the regions of the input space which are more represented in the training set are mapped in deeper detail with respect to other less populated regions. This latter characteristic makes the SOM extremely suitable for the purpose of the designed method. The SOM training algorithm, which mimics human learning, provides a set of neurons, each one representing a particular region of the input space, which are used as reference vectors within the proposed approximation method.

Once the tuning phase is completed, the obtained data structure can be used to simulate the model \( M \) by means of the function \( M: x \in I \rightarrow y \in O \). This approximation stage is put into practice by means of interpolation operations which exploit the created clusters and the output values assigned to each of them. This process can be summarized as follows:

- a new input vector \( x \in I \) is given.
- The reference cluster \( C_w \) of vector \( x \) is determined according to the proximity criteria of the clustering algorithm (normally euclidean distance).
- A set of \( n \) clusters \( \{C_{p1}, \ldots, C_{pn}\} \) which are nearest to the winning cluster is determined. The number of neighbours \( n \) is a parameter of the method and the performance of the algorithm according to its values are evaluated.
- The output vector \( Y_w \) associated to \( C_w \) and the output vectors \( \{Y_{p1}, \ldots, Y_{pn}\} \) associated to \( \{C_{p1}, \ldots, C_{pn}\} \) are respectively calculated
- The output \( M(x) \) is calculated as the weighted sum of \( Y_w \) and \( \{Y_{p1}, \ldots, Y_{pn}\} \) as shown in (2):

\[
M(x) = Y_w + \sum_{i=1}^{n} Y_{pi} W_i
\]

where \( W_i \) are weights that are computed by considering the distance of \( C_{pi} \) from the input vector \( x \): the closer \( x \), the higher the weight \( W_i \) associated to \( Y_{pi} \). In fact, \( W_i = (1 / d_i) \cdot F \), where \( d_i \) is the distance between the input vector and the center and \( F \) is a normalization factor which makes the sum of all the weights unitary.

The above described simulation stage is affected by the choice of the neighbours of the winning neuron which contribute to the output calculation and to their number. The choice of the neighbours, given their number \( n \), differs depending on the adopted clustering method:

- if the clustering is performed through the K-means algorithm, the winning center and the closest \( n \) centers are selected according to euclidean distances
- if clustering is performed through the SOM approach, the framework of the SOM itself is used to identify the \( n \) neighbours. More in detail: the structure of the SOM contemplates a neighborhood relation among the neurons: once the winning neuron is identified, such structure is exploited to identify the selected neurons. The SOM topology adopted within this work is the so-called grid topology (see Fig. 4 where neighbours selection is also depicted).
The method described in section III has been tested on a set of problems involving the approximation of pre-existing models. More in detail it is firstly tested on the approximation of literature functions of different degree of complexity which should assess the effective goodness of the proposed approach. This test campaign is described in detail in section IV.A.

The method has also been used for the optimization of the two industrial optimization problems introduced in section II. In these cases the plant models are firstly approximated by means of the proposed approach exploiting real data collected on the steel plant, subsequently the obtained approximated models have been used within an optimization task aiming to the improvement of the process operation.

In Sec. IV.B and IV.C, the proposed algorithm is employed for substituting the models related to the industrial applications handled in this work. For those problems, different configurations of the approximator are tested in terms of both the clustering method and the number of neighbors which are taken into account. Subsequently the approximated versions of the models are used within a GA-based optimization process instead of the original ones. Also in this case different configurations of the approximator are tested. The results obtained on these tests are described and analysed in section V.

A. Approximation of test functions

The proposed method has been tested on a set of well known sample functions in order to assess its performance. The functions included in this set, widely used for the evaluation of optimization algorithms due to their characteristics, are the following: sphere, Rastrigin, Ackley and Rosenbrock. The input space dimension for each function has been set to 50 for increasing their complexity and making them interesting for the scope of this work. The detailed description of the functions, including their domains is shown below where D is the dimension of the domain:

The sphere function:

\[ f(x) = \sum_{i=1}^{D} x_i^2 \] (3)

where \(-5.12 \leq x_i \leq 5.12, \text{ for } i = 1, 2, \ldots, D\)

The Rastrigin function:

\[ f(x) = AD + \sum_{i=1}^{D} [x_i^2 - A \cos(2\pi x_i)] \] (4)

where \(-5.12 \leq x_i \leq 5.12, \text{ for } i = 1, 2, \ldots, D\) and A is a constant whose value is 10.

The Ackley function:

\[ f(x) = -20 e^{-0.2 \sqrt{\frac{1}{D} \sum_{i=1}^{D} x_i^2}} - e^{\frac{1}{D} \sum_{i=1}^{D} \cos(2\pi x_i)} + 20 + e \] (5)

where \(-10 \leq x_i \leq 10, \text{ for } i = 1, 2, \ldots, D\)

The Rosenbrock function:

\[ f(x) = \sum_{i=1}^{D-1} \left( 100(x_i^2 - x_{i+1})^2 + (1 - x_i)^2 \right) \] (6)

where \(-10 \leq x_i \leq 10, \text{ for } i = 1, 2, \ldots, D\)

In the first step of this testing campaign all the functions have been approximated through the proposed approach, by exploiting the SOM clustering and by testing different configurations (with different number of clusters and neighbors considered for the interpolation stage) and training datasets, in order to evaluate the effect of the variation of these parameters.

The approximated models have been trained exploiting different datasets (one for each function) formed by 5000 observations generated by applying the selected functions to random points within the problem domains. Further 3000 observations have been generated with the same procedure for the test of the models.

Each of the obtained approximated models have been compared with the analogous neural networks (NN) models obtained by training a feed forward neural network with the same dataset used for the tuning of the approximated model. More in detail, a different NN model has been used for each function and has been trained in parallel with the presented approximator and exploiting the same data. The employed NN is a two layers feed forward network with one hidden layer: the activation function for the first layer of neurons is the hyperbolic tangent sigmoid transfer function while the linear function is used in the second layer. The NNs are...
trained by means of a variation of the back propagation algorithm, based on Bayesian regularization, which improves the generalization capabilities of the obtained network. Different numbers of neurons have been tested in the hidden layer (form 5 to 50): results refer to the best performing configuration of the network.

The results obtained by the approximation of the described functions are discussed in section V.

B. Approximation and optimization of the BF model

The optimization problem introduced in section II led to the development of an approximation of the Aspen Plus model of the BF. The original model contemplates 37 input variables and 23 outputs, which mainly includes BF process parameters and material compositions and quantities. For confidentiality reasons, further information on the inputs and outputs of the model cannot be provided.

For the development of the clustering and interpolation approximation a dataset has been created. The dataset includes 1000 input configurations and their respective simulations outputs obtained by the Aspen Plus model. The simulation has required 5 hours of computation which is acceptable and is to be performed just once and outside the optimization process (the obtained data will be used also in the future). 75% of these data have been used for the tuning of the approximator, while the remaining 25% for its validation.

Both clustering approaches and different numbers of clusters have been tested on this problem, in order to evaluate the best performing one. Within these tests, the number of active neighbors is varied according to the number of the clusters.

The best performing approximated model obtained by using this approach has been used within an optimization task aiming at improving the BF operation. Due to its speed the obtained approximated model can be exploited within a GA based optimization framework in order to find a suitable configuration of the network. The results achieved by the different tested approximators, the best performing one has been selected to be embedded into a GA-based optimization process, thus the proposed method has been used to create its fast approximation.

As previously introduced the time necessary for performing an Aspen model simulation does not allow its incorporation into a GA-based optimization process, thus the proposed method has been used to create its fast approximation. For this purpose a dataset was formed by exploiting the Aspen model. The so obtained dataset contains 2000 observations and a total of 14 variables (11 input and 3 output) and its composition required 5.5 hours of continuous computation.

75% of these data have been used for the tuning of the approximator, while the remaining 25% for its validation.

Different tests have been performed for the creation of the pyrolysis approximated model by varying the clustering approach, the number of clusters and neighbors.

The results achieved by the different tested configurations of the approximated model are reported in section V. Among the tested approximators, the best performing one has been selected to be embedded into a GA-based optimization aiming at the maximization of the rate zinc removed/CO2 emitted which is the fitness function used by GA.

As for the BF process optimization, similar tests have been pursued by using different configurations of real coded GA by changing main parameters. The same initialization, crossover and mutation functions of the previous application have been employed.

The results obtained on the scrap pretreatment plant optimization obtained exploiting the approximated model are reported in section V.

V. RESULTS AND DISCUSSION

In this section the results obtained by the proposed method on the problems described in section IV are reported. More in detail, firstly the results concerning the approximation of literature functions will be shown in comparison with the NNs based models. In this case the approximate output of the models are compared to the actual values. Subsequently the performance obtained by the designed approximators will be presented for the two handled industrial problems. For these latter applications also
the results concerning the optimization stage based on the approximated model are presented and discussed.

As far as the model approximations concerns, several performance indexes are employed in order to allow a complete assessment of the method. These evaluations are reported in the tables incorporated in this section and include: the absolute mean error, the mean percentage error $\varepsilon_\%$ and the Normalised Square Root Mean Square Error (NSRMSE) $\eta$ [11], which are respectively defined as:

$$\varepsilon = \frac{1}{K} \sum_{i=1}^{K} \left| \frac{y_i - \bar{M}(x_i)}{y_i} \right| \times 100$$  \hspace{1cm} (7)

$$\eta = \frac{1}{\sigma} \sqrt{\frac{1}{K} \sum_{i=1}^{K} \left( y_i - \bar{M}(x_i) \right)^2}$$  \hspace{1cm} (8)

where $K$ is the number of data samples that have been used for the validation stage and $\sigma$ the standard deviation of the measured variable.

The comparison, for all the tests reported in this section, is performed on validation data sets not used in any step of the tuning phase of the approximators.

A. Approximation results for the test functions

Table I reports the results obtained by the proposed approximation system and NNs when coping with the approximation of the mentioned test functions. Figure 5 qualitatively shows the performance on the 2D sphere function and puts into evidence the accuracy of the approximation. As it stands out from the table, the overall results achieved by the approximation system are very satisfactory since the percent error as well as the NSRMSE are extremely low for all the tested functions (also considering the high dimensionality of the problem).

<table>
<thead>
<tr>
<th>Function</th>
<th>Method</th>
<th>Abs. Err.</th>
<th>$\varepsilon_%$</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>Approx</td>
<td>43,9</td>
<td>3,1</td>
<td>0,25</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>45,2</td>
<td>3,4</td>
<td>0,28</td>
</tr>
<tr>
<td>Rastrigin</td>
<td>Approx</td>
<td>60,5</td>
<td>6,1</td>
<td>0,41</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>86</td>
<td>9,4</td>
<td>0,55</td>
</tr>
<tr>
<td>Ackley</td>
<td>Approx</td>
<td>2,1</td>
<td>1,2</td>
<td>0,21</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>2,3</td>
<td>1,2</td>
<td>0,23</td>
</tr>
<tr>
<td>Rosenbrock</td>
<td>Approx</td>
<td>2,1E+5</td>
<td>4,7</td>
<td>0,3</td>
</tr>
<tr>
<td></td>
<td>NN</td>
<td>2,1E+5</td>
<td>4,9</td>
<td>0,36</td>
</tr>
</tbody>
</table>

Comparing the performance of the approximator and of NNs it emerges that the proposed approach overcomes in accuracy this latter models. This behavior is due to the relatively low number of samples available for the tuning of the approximation engine with respect to the dimension of the input and output spaces: the proposed method in facts, thanks to its architecture, requires less data than neural networks in order to provide reliable results. The clustering phase and the partitioning of the input space is done independently from the output space and subsequently the obtained clusters are related to output vectors: this kind of training therefore results more robust when, as in the cases presented in this work and in many industrial applications, little amount of data for the tuning of the models is available.

It also emerges that – as expected - the performance of the approximator is better when facing “simpler” problems, characterized by regular functions. When more complicate input/output relations are handled the performance slowly degrades (it's the case of the extremely irregular Rastrigin function) as the clustered space has to be put into relation with very different output values. Nevertheless the performance of the proposed method – for the previously discussed reason - overcomes more markedly the NNs based approximation of the Rastrigin function in terms of percent error (the performance in this case is 50% more accurate) and NSRMSE.

B. Results on the approximation of the Blast Furnace model and its optimization

The results obtained on the BF approximation problem are shown in table II, where the average values $\varepsilon_\%$ and $\eta$ of the above defined indexes over all the available 23 outputs is considered, as both the performance indexes are normalised, i.e. their values do not depend on the absolute value of the considered output variable. Table II also shows a comparison in terms of results of the use of different clusterization approaches: the k-means and the SOM based approaches: the k-means and the SOM based ones in order to evaluate which one performs better.
Table II. Results on the approximation of the Blast Furnace model through SOM and k-means based clustering

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Clusters</th>
<th>n</th>
<th>(\varepsilon)%</th>
<th>(\eta)</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means 25</td>
<td>2</td>
<td>9.9</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>k-means 50</td>
<td>2</td>
<td>9.8</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>k-means 25</td>
<td>3</td>
<td>10</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>k-means 50</td>
<td>3</td>
<td>9.8</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>k-means 100</td>
<td>4</td>
<td>12.1</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>k-means 250</td>
<td>5</td>
<td>14.8</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>k-means 500</td>
<td>6</td>
<td>18.4</td>
<td>0.9</td>
<td></td>
</tr>
<tr>
<td>SOM 25</td>
<td>2</td>
<td>10</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>SOM 50</td>
<td>2</td>
<td>9.9</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>SOM 25</td>
<td>3</td>
<td>10</td>
<td>0.84</td>
<td></td>
</tr>
<tr>
<td>SOM 50</td>
<td>3</td>
<td>9.7</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>SOM 100</td>
<td>4</td>
<td>9.1</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>SOM 250</td>
<td>5</td>
<td>8.8</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>SOM 500</td>
<td>6</td>
<td>8.4</td>
<td>0.7</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6 depicts in a qualitative way the approximation by comparing the actual and predicted values for one of the considered outputs.

Table II and Fig. 6 show that the results obtained on this complex problem by the proposed method are extremely encouraging. The average percent error is satisfactory and even simply structured approximators, composed by a low number (25-50) of clusters, achieve good performance, as \(\varepsilon \approx 10\%\) represents an acceptable error level for the considered application. The performance of SOM based and k-means based systems are comparable, in terms of percent error and NSRMSE. It is worth to mention that the maximum achieved percentage error is below 13%.

When more clusters are employed, in order to properly face a problem, the SOM based approaches sensibly improve their performance, by achieving \(\varepsilon \approx 8.4\%\) and \(\eta \approx 0.7\).

This latter result, obtained by employing 500 clusters, which is absolutely appropriate for the dimension of the input space, is particularly remarkable. Good results (\(\varepsilon \approx 9\%)\) from this point of view are yet obtained by using 100 clusters. On the contrary, the performance of the k-means based approximators progressively decrease. The SOM acts better as it is able to profitably exploit the big number of available clusters with respect to k-means; the SOM learning algorithm locates the neurons in the input space according to the original distribution of the training data in a very efficient way so, more neurons cover the more populated regions.

The performance of the optimization of the BF operation are shown in table III which reports the results of the most salient tests pursued and highlights the population cardinality, the number of generations through which the GA goes through the search for the optimal solution, the evaluation – also employed as fitness function – in terms of the rate tonnes of produced pig iron / kWh of energy consumption and the time required for the optimization.

Table III. Results of the optimization of the BF model by means of GAS. The evaluation is expressed as the rate between iron ore mass flow expressed in tonne and the energy consumption in kWh.

<table>
<thead>
<tr>
<th>Population</th>
<th>Generations</th>
<th>Evaluation</th>
<th>Time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>0.026</td>
<td>81</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>0.029</td>
<td>148</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>0.029</td>
<td>149</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>0.029</td>
<td>306</td>
</tr>
</tbody>
</table>

The results depicted in table III put into evidence that the solution reached by the four optimization attempted are comparable in terms of evaluation. The first attempt, characterized by the use of 50 individuals through 50 GA generations leads to a different solution with respect to that one – the same – obtained by the three others runs of the GA optimization. This behavior is likely due to the complexity of the explored surface which, if performed by using a low number of individuals through a low number of generations, can leave some zones of the space unexplored or the can stuck algorithm in a local minimum.

On the other hand, when optimization is pursued with a wider number of individuals and through more generations, the reached solution is the same. In these cases (and in others for assessing the validity of the result) the fitness of the solution corresponds to the production of 1.35 tonne of hot metal which corresponds to a consumption of 45.7 kWh of electric energy. It is worth to mention that the obtained solution and its relative evaluation are in line with theoretical knowledge and expectations. Furthermore the output of the model corresponding to the proposed solution has been
verified to be in line with the one provided by the original model. These latter results confirm the goodness and the usability of the proposed method for the exploitation of complex model in an optimization task: on one hand the accuracy of the approximation achieved by the proposed approach ensures the reliability of the solution reached by the optimization process, on the other hand the speed of the approximated model allows its use within the optimization task: since from table III it emerges that the required for one optimization is only few minutes which is definitely satisfactory and grants the usability of the approach.

C. Results on the approximation of the scrap preheating process and its optimization

Table IV shows the results achieved on the approximation of the scrap preheating process model by the proposed approach. These results, expressed in terms of mean percent error ($\varepsilon$%) and NSRMSE ($\eta$), refer globally to the three outputs of the model. Table IV shows the results for a different number of clusters employed and different number of neighbors and compares the k-means and SOM based clusterization approaches.

### TABLE IV. RESULTS ON THE APPROXIMATION OF THE PYROLYSIS MODEL THROUGH SOM AND K-MEANS BASED CLUSTERING

<table>
<thead>
<tr>
<th>Clustering</th>
<th>Clusters</th>
<th>n</th>
<th>$\varepsilon$%</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>k-means</td>
<td>25</td>
<td>2</td>
<td>3.4</td>
<td>0.25</td>
</tr>
<tr>
<td>k-means</td>
<td>50</td>
<td>2</td>
<td>2.8</td>
<td>0.29</td>
</tr>
<tr>
<td>k-means</td>
<td>100</td>
<td>2</td>
<td>2.1</td>
<td>0.28</td>
</tr>
<tr>
<td>k-means</td>
<td>25</td>
<td>3</td>
<td>3.4</td>
<td>0.23</td>
</tr>
<tr>
<td>k-means</td>
<td>50</td>
<td>3</td>
<td>2.8</td>
<td>0.27</td>
</tr>
<tr>
<td>k-means</td>
<td>100</td>
<td>3</td>
<td>2.5</td>
<td>0.26</td>
</tr>
<tr>
<td>SOM</td>
<td>25</td>
<td>2</td>
<td>3.1</td>
<td>0.23</td>
</tr>
<tr>
<td>SOM</td>
<td>50</td>
<td>2</td>
<td>2.1</td>
<td>0.15</td>
</tr>
<tr>
<td>SOM</td>
<td>100</td>
<td>2</td>
<td>1.5</td>
<td>0.11</td>
</tr>
<tr>
<td>SOM</td>
<td>25</td>
<td>3</td>
<td>3.5</td>
<td>0.26</td>
</tr>
<tr>
<td>SOM</td>
<td>50</td>
<td>3</td>
<td>2.4</td>
<td>0.17</td>
</tr>
<tr>
<td>SOM</td>
<td>100</td>
<td>3</td>
<td>1.5</td>
<td>0.11</td>
</tr>
</tbody>
</table>

As it emerges from table IV the estimations obtained through the approach presented in this paper are extremely satisfactory: the approximation error is very low ($\varepsilon$% < 4% for all the performed tests) independently on the configuration of the method parameters. The SOM based clusterization approaches overcome those based on the k-means algorithm in almost all the performed tests, especially when a high number of clusters is employed: this latter consideration confirms that locating the centroids of the clusters according to original data distribution is a fruitful strategy and should be employed in the future.

The best performing test is the one based on the use of 100 clusters, which is in line with the dimension and the complexity of the problem. In this case the average percent error is 1.5% and the NSRMSE is 0.11, the time necessary for one simulation is in the order of 1/10 seconds: these figures allow the use of the approximated model within the GA optimization framework.

The most noticeable results of the optimization are shown in table V which reports the GA configuration and the value of the fitness functions for some of the performed tests. The performance of the optimization is expressed as the rate between kg/hr of zinc removed and kg/hr of CO$_2$ produced by the process.

All the different runs of the optimization process lead to the same result which corresponds to the removal of 25.5 kg of zing per hour with a hourly production of 16 tonne of CO$_2$.

The achievement of the same optimal solution by different configurations of the GA optimizer confirms the goodness of the approach. Moreover the solution of the problem obtained through GA is in line with expectations.

The time required for the optimization is extremely low and allows its use in the industrial context where a prompt solution to this kind of problems is desirable.

The results obtained on the optimization and approximation of the pyrolysis model related to the zinc removal process are analogous to those provided by the blast furnace operation optimization and confirm the usability in terms of reliability and speed of the approach proposed in this work.

### TABLE V. RESULTS OF THE OPTIMIZATION OF THE PYROLYSIS MODEL BY MEANS OF GAS. THE EVALUATION IS EXPRESSED AS THE RATE BETWEEN REMOVED ZINC AND EMITTED CO$_2$.

<table>
<thead>
<tr>
<th>Population</th>
<th>Generations</th>
<th>Evaluation</th>
<th>Time (sec.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>1.6E-3</td>
<td>16</td>
</tr>
<tr>
<td>50</td>
<td>100</td>
<td>1.6E-3</td>
<td>35</td>
</tr>
<tr>
<td>100</td>
<td>50</td>
<td>1.6E-3</td>
<td>26</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>1.6E-3</td>
<td>55</td>
</tr>
</tbody>
</table>

### VI. CONCLUSIONS AND FUTURE WORK

In this paper a novel method based on clustering and interpolation has been introduced. This method has been developed in order to approximate time consuming models within optimization tasks, where the iterated runs of the model leads to impractical computation times. Firstly the input space is clustered and a representative output vector is assigned to each cluster; afterwards, in the simulation phase, on the basis of the similarity of input vectors to the formed clusters, an approximation of the model output is provided. The efficiency of the proposed method has been proven on literature functions and on two industrial problems related to
the steel industry and involving the models of a blast furnace and a scrap preheating process. The mentioned models have been approximated and the obtained approximated versions of the models have been used within an optimization process based on genetic algorithms.

The pursued tests of proposed method achieved encouraging results and demonstrated the feasibility of the approximation of the original model within the optimization task.

Moreover the analysis of the results shows that between the considered clustering methods, SOM and K-means, the first one obtains better results and is therefore preferable. In the future the proposed approach will be used for the optimization of more plants within the steel works in order to assess its performance on the optimization of a complete steel production chain.

ACKNOWLEDGMENT

The work described in the present paper was developed within the projects entitled “Development of tools for reduction of energy demand and CO2 emissions within the iron and steel industry based on energy register, CO2 monitoring and waste heat power generation” (Contract No. RFSR-CT-2009-00032) and “Processes and technologies for environmentally friendly recovery and treatment of scrap” (Contract No. RFSR-CT-2010-00004) that have received funding from the Research Fund for Coal and Steel of the European Community. The sole responsibility of the issues treated in the present paper lies with the authors; the Commission is not responsible for any use that may be made of the information contained therein.

REFERENCES