TUCS

Björk | Holopainen | Wikström | Saxen |
Carlsson |Sihvonen

ANALYSIS OF BLAST FURNACE TIME SERIES DATA WITH ANFIS

Turku Centre for Computer Science

TUCS Technical Report
No 1094, Nov 2013

ANALYSIS OF BLAST FURNACE TIME SERIES DATA WITH ANFIS

Kaj-Mikael Björk

Arcada University of Applied Sciences and Åbo Akademi University,
Institute for Advanced Management Systems Research

Markus Holopainen

Åbo Akademi University, Institute for Advanced Management
Systems Research

Robin Wikström

Åbo Akademi University, Institute for Advanced Management
Systems Research

Henrik Saxen

Åbo Akademi University, Department of Chemical Engineering

Christer Carlsson

Åbo Akademi University, Institute for Advanced Management
Systems Research

Miika Sihvonen

Ruukki Metals Oy, Raahe, Finland

TUCS Technical Report
No 1094, Nov 2013

Abstract

Analyzing processes that cannot fully be measured is necessary for better decision making. In this paper we analyze some blast furnace data with the help of ANFIS (Artificial Neural Fuzzy Inference System). The entire process is described, from the application, preprocessing the data and the analysis made. Also the results from the best ANFIS models are compared to results with more traditional ARMA-model theory (as a benchmark). Special features in this industrial case study, the blast furnace process, includes by exhibiting high levels of noise and complexity.

Keywords: Blast Furnace, ANFIS, Decision Making, Identification

TUCS Laboratory
Name of the laboratory

1. Introduction

The escalating business climate of the twenty-first century has forced companies to study parts of their processes that at first glance might seem impossible to investigate. By examine issues which previously were considered to be unthinkable, the companies try to gain an advantage towards their competitors. These studies are facilitated with new, computationally powerful, data mining and soft computing techniques that have emerged during the last decades, making it possible to analyze data in completely new way, opening the possibility to gain better insights and understandings of the process and how it can be improved.

Such data mining or modeling techniques allow for trend discovery or correlation analysis between different factors affecting the process. The goal of the analysis may be to build predictive forecasting models, to find alternate actions to be taken, or simply to gain a deeper understanding of the underlying influencing elements. One of the industries facing fierce competition is the steelmaking industry. According to a review by the Association of Finnish Steel and Metal Producers [2012], the ongoing political and economical crisis in Europe has radically increased price competition. As steelmaking is a global operation, it is usually among the first to be affected by a financial decline. The industry is also faced with increasing costs due to emissions trading- and sulphur directives of the European Union. These are examples of aspects that contribute to a general focus on production as well as a striving towards optimization of processes.

In this paper, the blast furnace for the steel making process is examined. As the process is conducted at extreme conditions, it is impossible to observe what is occurring inside the furnace. Nevertheless, optimizing this part of the process would greatly improve the overall process and quality of the final product. By implementing the ANFIS (Artificial Neural Fuzzy Inference System) model on collected blast furnace data, we aim at optimizing the blast furnace process, or more specifically predicting the performance indicator η_{CO} which describes the gas utilization rate in the furnace.

This article is structured as follows: Section II presents the blast furnace process and introduces relevant concepts and theories associated with ANFIS. Section III introduces the material used as inputs for the modeling phase. Section IV explains how the modeling phase was conducted and presents the results received from the ANFIS modeling. Finally, Section VI summarizes the article and gives some conclusions and possible future research directions.

2. Previous studies in blast furnace modeling

Even if the blast furnace process has not, according to the authors' knowledge, been modeled with an ANFIS approach, many other blast furnace models have certainly been created. The following section presents an introduction to blast furnace process and a short introduction to the Soft Computing field, including ANFIS, is given.

The blast furnace

A crucial part in the steelmaking process involves the blast furnace. The main operating task of the furnace is to melt and reduce oxygen from the iron ore before it is sent to a steel plant for further refinement. The process is run continuously, with iron-bearing materials and coke being charged from the top of the furnace (Geerdes et al., 2009). Air is preheated to temperatures of 1200 C° and other additional fuel is blasted in through pipes in the lower region. The hot air reacts with the coke and the additional fuel, which causes a burning flame. The generated gas ascends and gradually softens and melts the iron ore, which helped by gravity descends. The ore is simultaneously reduced of oxygen and the final products, hot metal and slag, is collected at the bottom of the furnace and also periodically removed through tapholes. Chemical analyses of the hot liquid metal are carried out to determine the quality of the product, after which it is sent to a steel plant for further processing. It has to be mentioned that the furnace process is continuous, meaning that each charging cannot be undone. Usually, the optimal charging distribution is commonly found individually for each furnace by trial and error. The process of reducing oxygen from the ore is crucial and directly related to the quality of the hot metal output of the furnace. There are two main reactions taking place in the furnace:

- The direct reduction reaction, present in the lower part of the furnace. This reduction is dependent on expensive coke.
- The indirect reduction reaction, present in the upper part of the furnace. This reduction is more cost effective, as the gas can itself remove oxygen from the ore at this stage.

The efficiency of the indirect reaction is often expressed as the gas utilization rate, which is considered an important performance indicator of the furnace. As the blast furnace is continuously running, it is impossible to embed sensors inside the furnace, turning the blast furnace into a black-box model. Instead, furnace operation relies heavily on experienced engineers for process monitoring and control. To aid operation of the furnace, external sensor data, other process information and the accumulated expertise of the engineers is used to monitor and adjust the process accordingly. The process is highly complex from a chemical point of view as it involves numerous factors, nonlinear relations and a certain level of randomness.

In an ideal scenario, there is a column consisting entirely of coke at the very centre of the furnace, through which gas flows upward and branches out towards the sides. If the column radius is too small, the gas will not reach the uppermost layers, but will branch out toward the walls at the bottom of the furnace, cool down too fast and slow down reduction reactions. If the column radius is too large, however, a large part of the gas will flow straight through the center, not contributing to the ore reduction and going to waste. In the ideal case, some gas is supposed to ascend at the walls. Optimal gas flow control is based on maintaining the balance between central and wall gas flow through optimized burden distribution (Nath, 2002 and Danloy et al. 2001). Although an optimal burden distribution theoretically could be calculated, it is influenced by a large number of parameters such as size and coarsity of materials and the distribution of the previous charge. Small-scale experiments have been conducted by Yu and Saxén (2010) but real-world applications are not yet viable. Thus burden distribution is normally found individually for each furnace through trial and error. The aforementioned nonlinearity and complexity properties, however, speak for more intelligent soft computing methods, suggesting that the analysis and modeling process requires techniques from both domains to be combined.

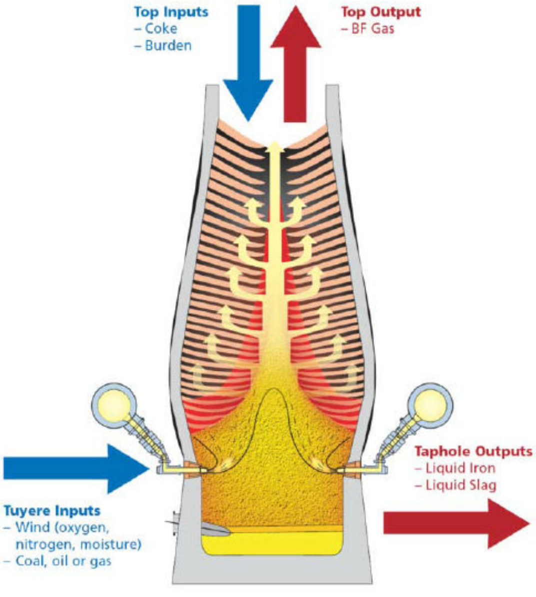


Figure 1: Operating principles of the blast furnace (Geerdes et al. 2009)

There are several attempts to model the blast furnace process, aiming to uncover hidden relationships not immediately visible due to the aforementioned "black-box" characteristics of the furnace. Mathematical models for describing the process have, for instance, been proposed by Nath (2002) and Danloy et al. (2001). Agarwal et al. (2010) states that mathematical models usually prove inadequate due to the high degree of nonlinearity and randomness of the process.

Linear data-driven models have been studied by Saxén and Östermark (1996), Korpi et al. (2003) and Bhattachary (2005). These studies have used linear modeling techniques to study the effects of several explanatory variables of the silicon content in the hot metal part of the output. The results of these studies are promising, but might be outmaneuvered by nonlinear soft computing techniques in accounting for complex nonlinear relations. Nonlinear soft computing techniques have been applied to blast furnace process modeling in, amongst other: Hao et al. (2004), Helle and Saxén (2005) and Pettersson et al. (2007). These studies have used variants of neural networks or neural network-based hybrids to model furnace performance indicators with fairly good results. The outputs have mainly been variables expressing properties of the hot metal output such as the aforementioned silicon content. An interesting approach is presented by Agarwal et al. (2010) where they train a neural network using multi-objective genetic algorithms. The study analyses blast furnace productivity related to two important performance indicators; carbon dioxide content of top gas and silicon content in the hot metal output. The research indicates that a productivity increase implies a compromise in one of the two indicators. Apart from the carbon dioxide content being an explanatory variable in Agarwal et al. (2010), no research could be found which studies the process relationships with carbon dioxide content in furnace top gas as a dependent variable.

To the best of our knowledge, ANFIS has not been applied in this fashion. Agarwal et al. [2010] notes that a model high in complexity is difficult to execute, as it might capture noise in the data treating it as a trend, resulting in an overfit of the data. Contrarily, an overly simple model will fail to find the trends in the data. As ANFIS allows for customization regarding membership functions, inputs and rules, an appropriate degree of complexity is expected to be found. These findings indicate a validity to apply ANFIS to blast furnace process analysis.

ANFIS

Current systems have a lacking in handling imprecise and vague information but still achieving precise and useful results, which is a natural process for a human brain to perform. Due to this, Soft Computing emerged as a sub-area of Computational Intelligence, offering techniques and solutions for computationally deal with imprecise data (Zadeh 1994a-b). The Soft Computing techniques tend to be suitable for combining with other established methods, making it possible to create hybrid systems which are more suitable for problem solving and data analysis. Fuzzy set theory (Zadeh, 1965), has recently attracted more interest, as computers are today more suitable for handling the somewhat computationally intensive calculations imminent in the Soft Computing field.

Another technique affiliated with soft computing is neural networks, inspired from the actual principles of the human brain, creating an artificial network of interconnected neurons (Jang et al., 1997). Due to its complex implementation, a neural network is sometimes regarded as a black-box model. This means that one is only able to see the model's inputs and outputs, not what is going on inside the process. The advantage is the learning abilities, which is why it is often used together with other methods. By including fuzzy sets into the mixture it creates a hybrid approach, called neuro-fuzzy models, which integrates the strengths of both methods.
Jang (1993) and Jang et al. (1997) introduced a class of adaptive networks that perform in the same manners as fuzzy inference systems, called ANFIS. The architecture combines the properties of neural networks and fuzzy logic, creating a dynamic fuzzy inference system similar to the Sugeno fuzzy model (Sugeno and Kang, 1988), built as a network based on the same manner as in neural networks.
To define ANFIS, we assume that the rule base contains two fuzzy if-then rules (Takagi and Sugeno, 1985):

$$Rule\ 1\colon If\ x\ is\ A_1\ and\ y\ is\ B_1, then\ f_1 = p_1x + q_1y + r_1,$$
$$Rule\ 2\colon If\ x\ is\ A_2\ and\ y\ is\ B_2, then\ f_2 = p_2x + q_2y + r_2.$$

The ANFIS architecture thereafter consists of 5 layers:

Layer 1
Each node in the first layer is adaptive and has a function

$$O_{1,i} = \mu_{A_i}(x), \qquad for\ i = 1,2\ or$$
$$O_{1,i} = \mu_{B_{i-2}}(y), \qquad for\ i = 3,4$$

Where x or y is the input to node *i* and A_i or B_{i-2} is a linguistic label, associated with this node. $O_{1,i}$ represents the membership grade of a fuzzy set *A*. Parameters in this layer are referred to as *premise parameters*.

Layer 2
Opposite to the nodes in layer 1, the nodes in the second layer are fixed and labeled Π, the output is the product of all the incoming signals:

$$O_{2,i} = w_i = \mu_{A_i}(x)\mu_{B_i}(y), i = 1,2.$$

The outputs from the nodes are equal to the firing strength of a rule.

Layer 3
The nodes in the third layer are fixed and labeled *N.* The *i*th node creates the ratio of the *i*th rule´s firing strength to the sum of all rules´ firing strengths:

$$O_{3,i} = \overline{w_i} = \frac{w_i}{w_1 + w_2}, i = 1,2.$$

The outputs of this layer are referred to as *normalized firing strengths.*

Layer 4
Every node *i* in this layer is an adaptive node with a node function

$O_{4,i} = \overline{w_i} f_i = \overline{w_i}(p_i x + q_i y + r_i)$,

Where w_i is a normalized firing strength based on layer 3 and $\{p_i, q_i, r\}$ the parameter set of this node, referred to as *consequent parameters*.

Layer 5
The fifth and final layer consists of a single fixed node, labeled Σ, calculating the overall output as the sum of all incoming signals:

$$overall\ output = O_{5,i} = \sum_i \overline{w_i} f_i = \frac{\sum_i w_i f_i}{\sum_i w_i}$$

With these 5 layers, an adaptive network is structured, having similar functionality to a type-3 fuzzy inference system. The modifiable parameters of an ANFIS are composed of premise and consequent parameters. The premise parameters will modify the rule membership functions and thus will cause nonlinear changes from the system's inputs to outputs. They are therefore nonlinear parameters. The consequent parameters modify the output functions, which are linear, thus these parameters are linear. Jang et al. (1997) notes that nonlinear optimization methods could be used for the training of an ANFIS, the proposed hybrid method requires much less computation and will be faster.

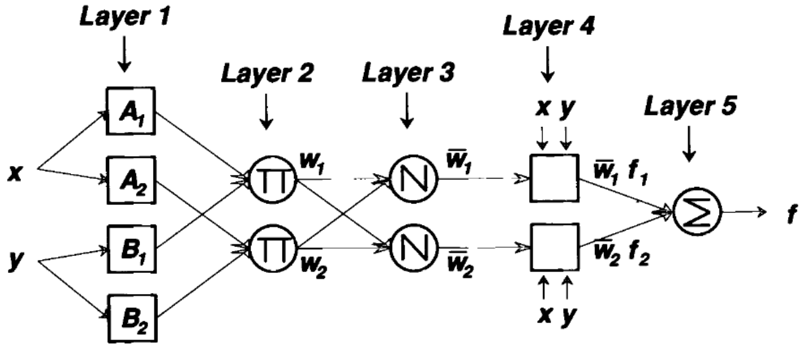


Figure 2: Type-3 ANFIS model (Jang, 1993)

The ANFIS hybrid learning algorithm is composed of two passes, the forward pass and the backward pass. These two constitute a training epoch, which is run a number of times until a specified model fit is obtained, according to different error criteria. In the forward pass, the premise parameters are given an initial estimation based in the input vectors, then the node outputs are calculated layer by layer until the last layer is reached, after which the consequent parameters are estimated by a Least-Squares Estimator LSE. Next, the algorithm calculates the output errors and propagates them backwards towards the input end, accumulating a gradient error vector. When the first layer has been reached, the premise parameters are updated by gradient descent methods.

3. Using ANFIS for Blast Furnace modelingEven if the blast furnace process

The analysis and modeling of the process is a key element in gaining a deeper understanding of the complex relationships and ongoing activities within the blast furnace. The performance indicator gas utilization rate was chosen as the target series for the conducted analysis and modeling. As noted in Chatfield (2000), an important question in multivariate analysis is whether or not to use the target series itself as an input to enhance model performance. In this study the target series was not included as an input to emphasize the focus on causal relationships. The different factors affecting the process with different time lags require that historical time series data of the process for a prolonged period to be obtained.

ANFIS is expected to be able to handle the high complexity of the process, as well as to account for the nonlinear correlation which may be present in the data. Conventional time series techniques are to be used simultaneously in the analysis and pre-processing of the data. A well-known data mining framework, CRISP-DM, is to be used to guide the analysis and modeling process. Expert knowledge of the furnace is utilized at the initial analysis stage, but otherwise the modeling process is entirely data-driven. In this section, the data, the inputs and the time lags are presented as well as the preliminary modeling and model identification.

The Data

The need for high-quality data and comprehensive pre-processing of the data is essential. For the study, three months of detailed operational blast furnace data was used and, additionally, three more months of data was used for the evaluation conducted in the end of this article. The data consisted of three segments:

- Charging data, consisting of precise amounts of each separate material, structured by charge and including a time code.
- Continuous process data, such as temperature data of external sensors and detailed specifics of the hot blast. Structured as averages.
- Hot metal analysis data, describing the quality of the final product

Discussions with a blast furnace expert suggested an aggregation of data to averages per hour, as narrower aggregation would yield noisy results. The charging data was assumed to obtain a total amount charged per hour. The final set of preprocessed data contained 36 columns and 2208 rows, corresponding to 36 inputs including 3 columns of time data and 16 columns of temperature sensor data, and to 2208 hours of total data available. Scattered throughout the data was the short stops in charging, spanning from 1 hour stops to maximal stops of 13 hours.

One important blast furnace performance indicator is η_{CO}, measuring the ratio of carbon monoxide converted to carbon dioxide. It evaluates performance of the data understanding indirect reduction reaction taking place in the upper part of the furnace. As this reaction does not consume coke, it is less costly and has the potential of cost savings if understood properly. It is expected of ANFIS to have adequate modeling capability to capture the complex and nonlinear relationships present in the data.
At this stage, the initial time series were converted to graphs and reviewed in conjunction with the expert. The primary goal of this initial review was to identify portions of the time series data that showed anomalies that could not be explained as normal behavior of the furnace process. These portions of data were considered to be misleading for the modeling process and were therefore deleted. The charging caused empty values in charging data series and where also removed.

The subsequent data about the stops in charging showed that it would require some hours after charging was resumed before the process would stabilize. The decision was made to remove a number of data after each stop in charging equal to the duration of the stop. For a stop lasting 5 hours, these 5 hours and the following 5 hours of "stabilization data" would be removed. After reviewing all series, the raw input series of blast volume showed a few dips in the data significantly lower than average. These dips were deemed by the expert to be abnormal and all data, for which the blast volume was below a threshold of $1 * 10^5$ was deleted. The above-mentioned steps to ensure process continuity reduced the amount of data to 81.5% of the original, or to 1800 data points.

The Inputs

Charging data entails precise amounts of burden materials and coke charged into the furnace per time unit. The raw data series contain two different types of pellets, quartz, sinter, scrap metal and two different types of coke. A secondary data series obtained from the charging data was burden height, measured by a sensor after each charge. According to experts it would be wise to combine all the series of iron bearing sources, as well as the two different coke series. This would minimize the amount of total available inputs. Secondly, the combined series would be scaled in relative to the total amount of material charged. Thirdly, a few of the raw input series should be combined to support existing theory.

The four series derived from the charging data are:

- Amount of iron-bearing material charged as percentage of total material charged
- Amount of coke charged as percentage of total material charged
- Amount of calculated iron charged divided by calculated coal charged
- Amount of iron-bearing material charged divided by blast volume

The first input, amount of iron-bearing material charged as percentage of total material charged, was calculated by adding the two pellet input series to the sinter input series and dividing by the total amount of pellets, sinter, coke and scrap metal charged per hour.

The second input, amount of coke charged as percentage of total material charged, was determined by adding the two coke input series and dividing by the total amount of material charged as above. The series resembles the first input series if it would be inverted, caused by the amount of charged scrap metal being low in relation to the other charged materials.
The third input, amount of iron divided by coal, was calculated as

$$\frac{Fe}{C} = \frac{0.65 * PE + 1 * SM + 0.61 * SI}{0.87 * CO}$$

where PE = amount of pellets, SM = amount of scrap metal, SI = amount of sinter and CO = amount of charged coke. This series also has a clear visual resemblance to the first series, and it is expected that it thus may be omitted during the ANFIS modeling phase.

The fourth and final input of the charging data, amount of iron-bearing material charged divided by blast volume. It was calculated by adding the two input pellet series to the sinter series and dividing by blast volume obtained from process data.

The final series concern the top gas exiting the furnace. Geerdes et al. (2009) argues that hydrogen competes with carbon monoxide in reducing oxygen, causing a relation between hydrogen content in the top gas and the η_{CO}-content. Thus the hydrogen content in top gas is added as an input series.

The final series derived from process data is the output series η_{CO}, describing the performance of the indirect reduction reaction in the furnace. This is the output series that was used in the ANFIS model.

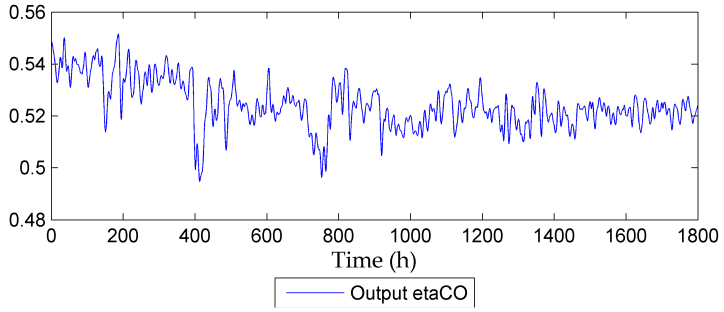


Figure 3: Output ηCO

Time lags

The input series affects the furnace process and the output η_{CO} in different ways. For instance, input series related to the blast affect the process almost instantly, while changes in amounts of iron ore charged have a significantly longer delay, measured in hours, before notable changes in the process can be detected. The ANFIS model is expected to find the correlations within the series, but finds these based strictly on changes occurring within the same hour. Thus a controlled synchronization by introducing delay or lag in the input series is appropriate.

In the first modeling phase, the goal is to utilize ANFIS to find the optimal time lags for each input. The procedure used is to systematically shift each input time series relative to the output η_{CO} , train and evaluate ANFIS and study the results to determine the optimal time lag for each series. It is expected that the blast- and top gas-related series have a low optimal time lag, while charging data inputs will have a higher time lag.

The ANFIS method depends on a training data set and a validation data set to be able to develop a model. The choice of how to split the data into training and validation sets was determined during preliminary testing to be an important criterion to the overall ANFIS performance. As discussed in Kohavi (1995), choosing a portion of the data with different characteristics, for instance a higher average, as training data, may result in a bias. As ANFIS attempts to validate the model against validation data with a lower average, model performance may be affected. Considering the output series η_{CO},the first third of the data has a significantly higher mean, and thus this series is in risk of inducing such a bias.

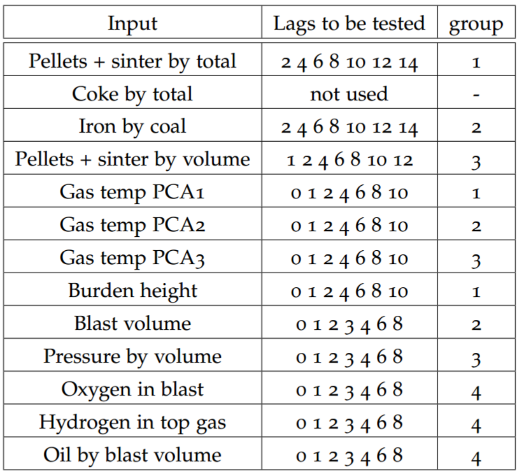
To avoid the training data and validation data selection bias, a method referred to as *k*-fold cross-validation (Kohavi, 1995 and Olson and Delen, 2008) can be utilized. In this method a fold size *k* is chosen, which determine the number of runs required. The data is separated into *k* equal portions. For each run, a different portion of the data is used as validation data, while the remaining portions are used as training data. Using *k*-fold cross-validation (Viaene et al. 2005), with a chosen the fold size of *k* = 5, the resulting *k* output models and their respective error values are averaged to obtain the overall model output and accuracy.

4. Preliminary modeling phase

It is well known that the complexity of the model increases exponentially with every added input parameter. This fact impacts the ANFIS modeling procedure in two ways; in limitations to model complexity due to available amount of data, and in exponentially rising total durations of ANFIS runs. By trial and error, different complexity settings were tested, the number of inputs in the ANFIS model was fixed to 5, as increasing the complexity mostly increased the runtime but barely affected the performance. Of course, the duration of the runs is an important factor to consider. Duration times exceeding several hours or even days are not viable. Considering the case of optimal input lags for the "Pellets and sinter" input series, the optimal time lags are expected to be among the values [2, 4, 6, 8, 10, 12, 14]. With 7 similar lag intervals for each of the 12 input series, the total sum rises to 84. In cases where it can be established that there is no interdependence between some or all of the inputs to the output, the search of the model parameters becomes simple. In this case the selection of optimal time lag could then be done separately for each input and the search space would be radically reduced. In the case of the complex blast furnace process however, interdependence between inputs cannot be excluded. If all the possible combinations of inputs and 7 respective time lags would be considered, the resulting number of combinations is very large and very time consuming to analyze as selecting 5 from 84 yields close to 31 million combinations, causing an infeasible duration time if this would be done in a single ANFIS simulation. For the blast furnace model, a compromise between search space allowing for interdependence and limiting it to reduce computation time was implemented.

The modeling procedure is split into two phases:The first phase, denoted the "Preliminary phase", reduces the search space for the optimal time lags for each input series relative to the output series. The goal of the phase is to find the two time lags for each input series yielding the best correlation to the output. These two optimal time lags for each input are to be used in the next phase. The second phase, named "In-depth phase", combines all the inputs with the two optimal time lags with the goal to find the top 5 inputs yielding the best model performance. The training goal of ANFIS was to minimize the average of training and validation RMSE in the ANFIS output versus η_{CO}.

The goal of the preliminary modeling phase is to find the optimal time lags for each input series. The input series related to the blast are expected to have a low optimal time lag, while series related to charging data are expected to have a slightly higher optimal time lag. The compromise between search space reduction and interdependence between inputs was implemented and this phase was divided into four groups. The four groups consisted of four separate simulations based on different groupings of input series. In this manner, ANFIS was given opportunity to find interdependence between these inputs without causing duration times to expand to excessive proportions.



Input	Lags to be tested	group
Pellets + sinter by total	2 4 6 8 10 12 14	1
Coke by total	not used	-
Iron by coal	2 4 6 8 10 12 14	2
Pellets + sinter by volume	1 2 4 6 8 10 12	3
Gas temp PCA1	0 1 2 4 6 8 10	1
Gas temp PCA2	0 1 2 4 6 8 10	2
Gas temp PCA3	0 1 2 4 6 8 10	3
Burden height	0 1 2 4 6 8 10	1
Blast volume	0 1 2 3 4 6 8	2
Pressure by volume	0 1 2 3 4 6 8	3
Oxygen in blast	0 1 2 3 4 6 8	4
Hydrogen in top gas	0 1 2 3 4 6 8	4
Oil by blast volume	0 1 2 3 4 6 8	4

Figure 4: Inputs and lags tested

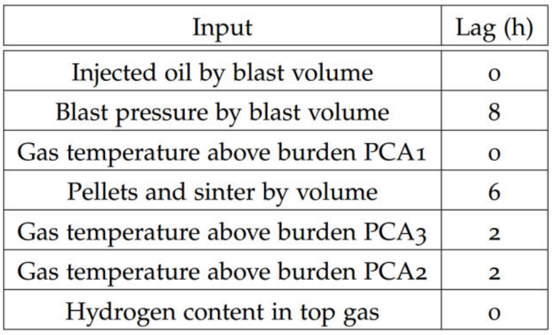
Figure 3 illustrates which inputs and which respective lag search range were used in which group. Combining the inputs in this manner allowed for some interdependency to be found between the inputs. In each group, ANFIS models with 4 inputs were simulated for each unique combination. Choosing uniquely 4 out of 3 inputs * 7 lags = 21 possible results in roughly 6000 combinations. Each simulation yielded RMSE values for the training and the validation data set. These were averaged to produce an "Average RMSE", which was used as the sorting criteria. Thus the inputs and lags could be ranked according to performance.

The problem on how to identify the top inputs and lags from the large tables of data was solved by introducing density plots for visualization and aiding the decision making process. The density plot shows the distribution of the data, similar to a histogram. From the obtained data table, 21 new tables were derived, one for each input series, one for each lag. From these tables, all rows not containing the specific input series and lag, were removed. Now the distribution of the data in relation to "Average RMSE" could be visualized for each lag with density plots. Drawing the density plots for each input with all time lags could then be used to determine which time lags were optimal.

ANFIS model identification

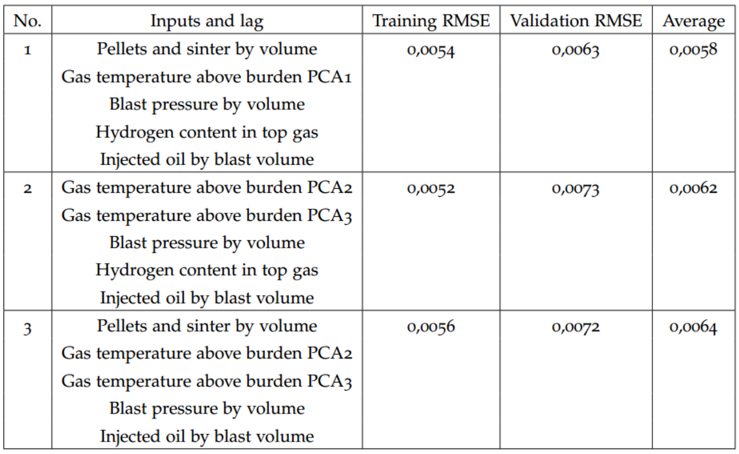
Using the obtained data from the preliminary modeling phase, the goal of the in-depth modeling phase is to determine the top inputs yielding the best ANFIS performance. The phase is divided into two parts. In the first part all possible combinations of choosing 4 out of 12 inputs with best 2 lags would be studied, enabling more interaction than in the previous phase where the inputs were divided into groups. Choosing uniquely 4 out of 24 yields roughly 10600 combinations. This is already a large number and thus the number of inputs was chosen as 4, increasing the number of inputs would create unfeasible running times. The result of the first part, 7 top inputs with one optimal time lag will be utilized in the second part, where the top overall models are to be determined. No other changes were made to the ANFIS simulations, and 5-fold cross validation still was implemented. The same density plot procedure was used to determine the best inputs and lags for ANFIS performance.

The second part consisted of finding the top overall models based on the top 7 inputs. The ANFIS model was in practice limited to 5 inputs, choosing uniquely 5 out of 7 results in 21 combinations, out of which the top three models, sorted by averaging training RMSE and validation RMSE, are summarized in figure 5.



Input	Lag (h)
Injected oil by blast volume	0
Blast pressure by blast volume	8
Gas temperature above burden PCA1	0
Pellets and sinter by volume	6
Gas temperature above burden PCA3	2
Gas temperature above burden PCA2	2
Hydrogen content in top gas	0

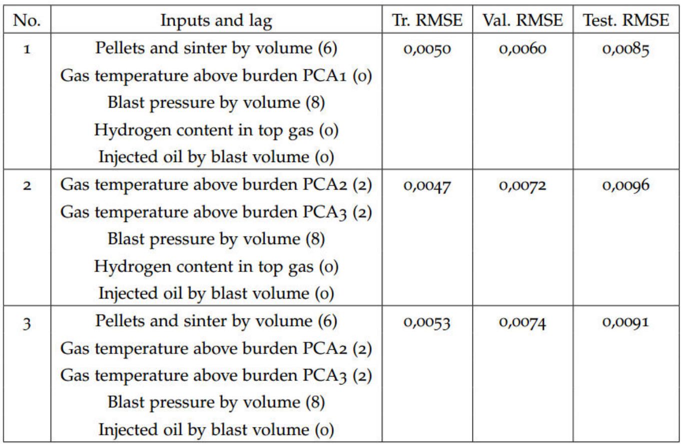
Figure 5: The top 7 inputs



No.	Inputs and lag	Training RMSE	Validation RMSE	Average
1	Pellets and sinter by volume Gas temperature above burden PCA1 Blast pressure by volume Hydrogen content in top gas Injected oil by blast volume	0,0054	0,0063	0,0058
2	Gas temperature above burden PCA2 Gas temperature above burden PCA3 Blast pressure by volume Hydrogen content in top gas Injected oil by blast volume	0,0052	0,0073	0,0062
3	Pellets and sinter by volume Gas temperature above burden PCA2 Gas temperature above burden PCA3 Blast pressure by volume Injected oil by blast volume	0,0056	0,0072	0,0064

Figure 6: Top three models of in-depth modelling phase

During the analyze stage another batch of furnace data was supplied. The data spanned three months and represents another realization of the process. No considerable alteration to the process was reported by the supplier. The data could thus be used to evaluate the top ANFIS models obtained in the modeling phase. The goal of the evaluation procedure is to measure the generalization performance of the ANFIS models when faced with new unseen data. The top 3 models obtained in the modeling phase will be evaluated by simulation of the new data without the new data affecting the model. The new data contained, after the same pre-processing procedure was conducted, 1770 data points.



No.	Inputs and lag	Tr. RMSE	Val. RMSE	Test. RMSE
1	Pellets and sinter by volume (6) Gas temperature above burden PCA1 (0) Blast pressure by volume (8) Hydrogen content in top gas (0) Injected oil by blast volume (0)	0,0050	0,0060	0,0085
2	Gas temperature above burden PCA2 (2) Gas temperature above burden PCA3 (2) Blast pressure by volume (8) Hydrogen content in top gas (0) Injected oil by blast volume (0)	0,0047	0,0072	0,0096
3	Pellets and sinter by volume (6) Gas temperature above burden PCA2 (2) Gas temperature above burden PCA3 (2) Blast pressure by volume (8) Injected oil by blast volume (0)	0,0053	0,0074	0,0091

Figure 7: Top three models of in-depth modeling phase

Figure 6 shows the top 3 models found earlier evaluated with the obtained testing data. As previously mentioned, minor changes to the ANFIS model and the smoothing of the series were required. Thus the residual analysis of the training (and validation) data does not generate equal RMSE-values to the simulations done without these modifications. Residual analysis was applied to the entire testing data portion. The five-fold cross-validation procedure was only used on the training and validation data, while the testing data portion was kept the same during the simulations. The ANFIS output of the testing portion is, however, an average of the five simulations, as is the case with the training and validation output.

Figures 7, 8 and 9 presents the top 3 models with both training and testing data. Original η_{CO} output is in blue, ANFIS output of the training and validation data portion is in red, while testing data output is in green. Overall, the testing data output performance is slightly worse than the output of the training data set.

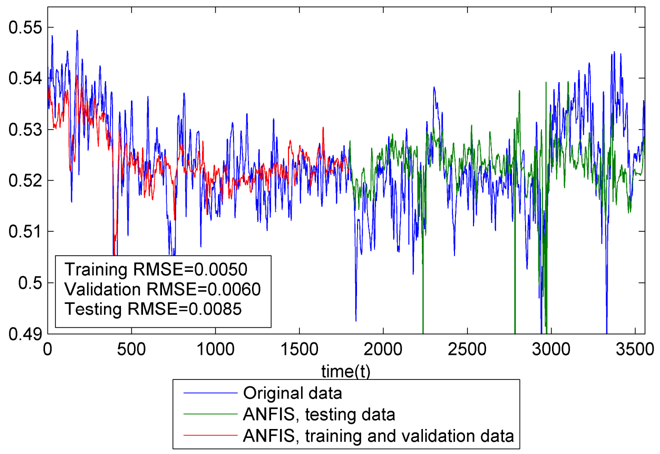


Figure 8: ANFIS output, model 1

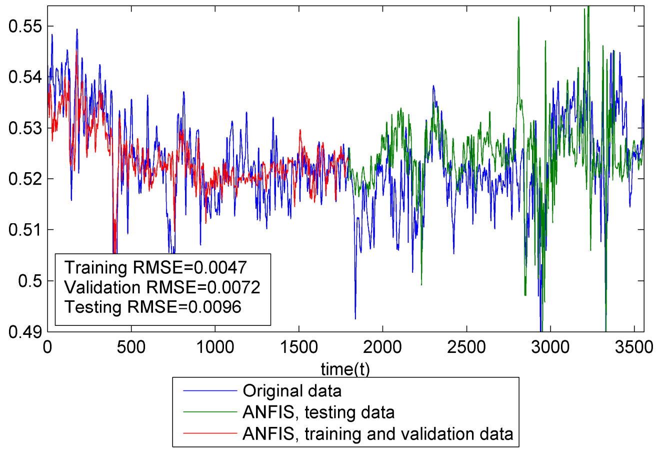


Figure 9: ANFIS output, model 2

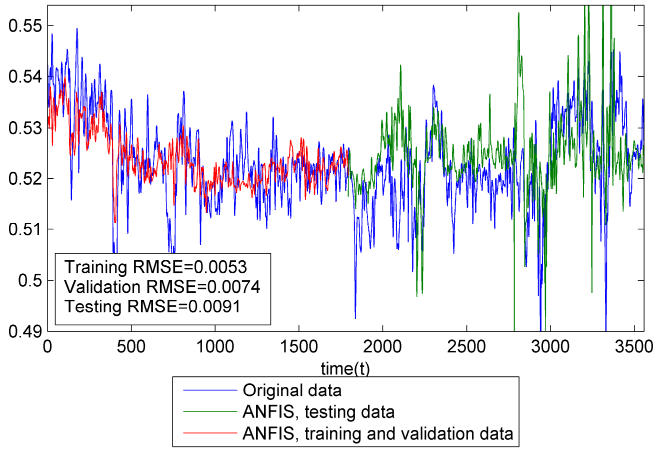


Figure 10: ANFIS output, model 3

Although the high complexity of the blast furnace process was expected to render a linear model unfeasible, it was determined as an important step to implement a comparative linear model to assess the performance of the ANFIS model. As a final phase, the ANFIS model was compared with the multivariate autoregressive (Vector Autoregressive) model (Hatemi and Hacker, 2009). The linear model was implemented directly based on the top three models. The same pre-processed inputs were used, along with minor smoothing and the K-fold cross-validation procedure.

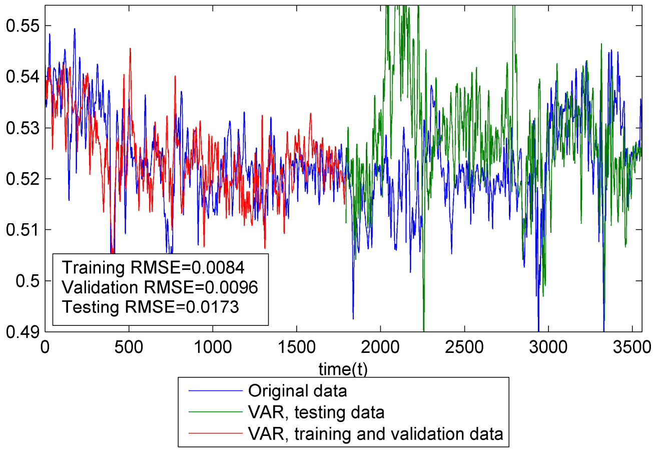


Figure 11: Linear model output, model 1

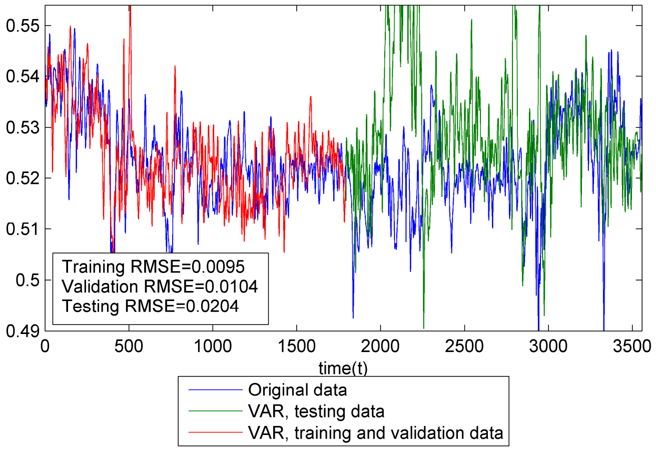


Figure 12: Linear model output, model 2

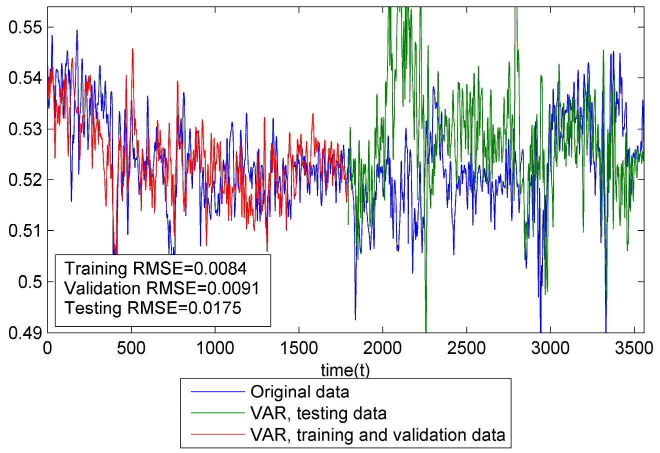


Figure 13: Linear model output, model 3

The data used for training, validation and testing was identical to the ANFIS scenario. The figures 10, 11 and 12 show some promise for the training and validation portions, but poor generalization performance on the testing data. At t = 2000 2200, for instance, all linear models show a radical increase in the output, which is not present in the original testing data during the interval. One possible explanation for this behavior is that there is correlation present, which the linear models pick up, but that the relation is nonlinear. Thus the linear models exhibit poor extrapolation when exposed to previously unseen testing data. As the linear model was not limited to five inputs, the model was also tested with the full seven inputs. Overall model performance, however, showed merely a fractional improvement over the five-input models depicted above and was determined to be negligible.

5. Conclusions and Future Research

The aim with the research conducted was to produce an ANFIS model which, based a number of explanatory series, could predict the performance indicator η_{CO} describing the gas utilization rate in the furnace. The modeling procedure was allowed to be completely data-driven, except for some expert knowledge that was utilized in the pre-processing stages and in the construction of the final input series, where several series consisted of mathematical combinations of two or more raw input series.

The adequate performance of ANFIS combined with the proposed pre-processing approach resulted in a system which is feasible for real-world industrial application. Discussions with the furnace experts revealed that it would be possible to use the ANFIS model as a support system alongside the existing systems which measure the performance of the furnace. Over time, the model can be tuned in conjunction by experts, for instance by more proficient removal of outliers and anomalies in the training data. This is enabled by the experts having first-hand access and knowledge of changes in the process, which may not show clearly in the data. As both the amount and quality of data rises, there may be a possibility to gradually remove the k-fold cross-validation procedure, finally resulting in a single ANFIS model which allows for deeper understanding of the complex process. At that stage the analytics level of "forecasting and extrapolation" will have been attained and steps towards the final highest levels of analytics may be taken. If this stage is reached, there would be a lot of potential in creating a sophisticated decision support system with predictive capabilities which may be used for further process optimization.

Except involving the tuning of expert into the process other future research directions may involve alternate neuro-fuzzy techniques to be applied to the problem. At the time of writing, other complex models are being implemented based on the pre-processed data. Alternative pre-processing choices are also viable, and may produce better results depending on the modeling technique.

References

A. Agarwal, U. Tewary, F. Petterson, S. Das, H. Saxén, and N. Chakraborti. Analysing blast furnace data using evolutionary neural network and multiobjective genetic algorithms. Ironmaking and Steelmaking, 37(5):353–359, 2010. doi: 10.1179/030192310X12683075004672.

The Association of Finnish Steel and Metal Producers. Structure Review of Metal Refinement,2012.URL
http://www.teknologiateollisuus.fi/file/14190/MJ_rakennekatsaus212.pdf.html.

T. Bhattacharya. Prediction of silicon content in blast furnace hot metal using partial least squares (pls). ISIJ International, 45(12):1943–1945, 2005.

C. Chatfield. Time-Series Forecasting. Statistics (Chapman and Hall/CRC). Chapman and Hall, 2000. ISBN 9781584880639.

J. Danloy, R. Mignon, G. Munnix, L. Dauwels, and L. Bonte. A blast furnace model to optimize the burden distribution. pages 37–48. 60th Ironmaking Conference Proceedings, Baltimore, 2001.

M. Geerdes, H. L. Toxopeus, C. Van der Vliet, R. Chaigneau, and T. Vander. Modern Blast Furnace Ironmaking: An Introduction. Ios PressInc, 2009. ISBN 9781607500407.

X. Hao, P. Zheng, Z. Xie, G. Du, and F. Shen. A predictive modeling for blast furnace by integrating neural network with partial least squares regression. In Industrial Technology, 106-107. 2004. IEEE ICIT '04. 2004 IEEE International Conference on, volume 3, pages 1162–1167 Vol. 3, 2004. doi: 10.1109/ICIT.2004.1490724.

A-S. Hatemi-J, and R, Hacker (2009). Can the LR test be helpful in choosing the optimal lag order in the VAR model when information criteria suggest different lag orders?. Applied Economics, 41(9), 1121-1125.

M. Helle and H. Saxén. A method for detecting cause-effects in data from complex processes. pages 104–107, 2005. doi: 10.1007/3- 211- 27389- 1_25. URL http://dx.doi.org/10.1007/3-211-27389-1_25.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, volume 2, pages 1137–1143. Morgan Kaufmann, 1995.

M. Korpi, H. Toivonen, and B. Saxén (2003) Modelling and identification of the feed preparation process of a copper flash smelter. Computer Aided Chemical Engineering, Elsevier, 2003,(14):731-736, ISSN 1570-7946, ISBN 9780444513687.

N. K. Nath. Simulation of gas flow in blast furnace for different burden distribution and cohesive zone shape. Materials and Manufacturing Processes, 17(5):671–681, 2002. doi: 10.1081/AMP-120016090.

H. Saxén and R. Östermark. Varmax-modelling of blast furnace process variables. European Journal of Operational Research, 90(1):85–101, 1996. doi: 10.1016/0377-2217(94) 00304- 1.

M. Sugeno and G. T. Kang. Structure identification of fuzzy model. Fuzzy Sets Syst., 28(1):15–33, 1988.

T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modeling and control. Systems, Man and Cybernetics, IEEE Transactions on, SMC-15(1):116–132, 1985.

F. Pettersson, N. Chakraborti, and H. Saxén. A genetic algorithms based multi-objective neural net applied to noisy blast furnace data. Applied Soft Computing, 7(1):387–397, 2007. doi: 10.1016/j.asoc.2005.09.001.

Y. Yu and H. Saxén. Experimental and dem study of segregation of ternary size particles in a blast furnace top bunker model. Chemical Engineering Science, 65(18):5237–5250, 2010. doi: 10.1016/j.ces.2010.06.025.

L. A. Zadeh. Fuzzy sets. Information and Control, 3(8):338–353, 1965.

L. A. Zadeh. Outline of a new approach to the analysis of complex systems and decision processes. Systems, Man and Cybernetics, IEEE Transactions on, SMC-3(1):28–44, 1973.

L. A. Zadeh. Fuzzy Logic, Neural Networks, and Soft Computing. Commun. ACM, 1994a, (37):77-84.

L. A. Zadeh. Soft Computing and Fuzzy Logic. IEEE Software, 1994b, (11): 48-56.

Turku Centre *for* Computer Science

Joukahaisenkatu 3-5 B, 20520 Turku, Finland | www.tucs.fi

University of Turku

- Department of Information Technology
- Department of Mathematics

Åbo Akademi University

- Department of Information Technologies

Turku School of Economics

- Institute of Information Systems Sciences

ISBN 978-952-12-2986-2
ISSN 1239-1891