# Overall models based on ensemble methods for predicting continuous annealing furnace temperature settings

**A. Sanz-Garcia, F. Antoñanzas-Torres, J. Fernández-Ceniceros and F. J. Martínez-de-Pisón***

The prediction of the set points for continuous annealing furnaces on hot dip galvanising lines is essential if high product quality is to be maintained and energy consumption and related emissions into the atmosphere are to be reduced. Owing to the global and evolving nature of the galvanising industry, plant engineers are currently demanding better overall prediction models that maintain accuracy while working with continual changes in the production cycle. This paper presents three promising prediction models based on ensemble methods (additive regression, bagging and dagging) and compares them with models based on artificial intelligence to highlight how good ensembles are at creating overall models with lower generalisation errors. The models are trained using coil properties, chemical compositions of the steel and historical data from a galvanising process operating in Spain. The results show that the potential benefits from such ensemble models, once configured properly, include high performance in terms of both prediction and generalisation capacity, as well as reliability in prediction and a significant reduction in the difficulty of setting up the model.

**Keywords:** Hot dip galvanising line, Continuous annealing furnace, Data mining, Process modelling, Ensemble methods, Artificial intelligence

## List of symbols

| | |
|---|---|
| Al, Cu, Ni, Cr, Nb | chemical composition of steel, wt-% |
| C, Mn, Si, S, P | chemical composition of steel, wt-% |
| $N$ | total number of simulations or trained models |
| $t$ | computation cost for training a set of models, s |
| THC1 | zone 1 set point temperature (initial heating zone), °C |
| THC3 | zone 3 set point temperature (intermediate heating zone), °C |
| THC5 | zone 5 set point temperature (final heating zone), °C |
| ThickCoil | strip thickness at the annealing furnace inlet, mm |
| TMPP1 | strip temperature at the heating zone inlet, °C |
| TMPP2 | strip temperature at the heating zone outlet, °C |
| TMPP2CNG | strip set point temperature at the heating zone outlet, °C |
| WidthCoil | strip width at the annealing furnace inlet, mm |
| V, Ti, B, N | chemical composition of steel, wt-% |
| VelMed | strip velocity inside the annealing furnace, m $min^{-1}$ |
| $X$ | percentage of training data |

### Subscript

| | |
|---|---|
| TH | maximum/minimum threshold value |
| tr | training dataset |
| tst | testing dataset |
| val | validation dataset |

### Superscript

| | |
|---|---|
| ME | mean |
| SD | standard deviation |

## Introduction

The production of galvanised flat steel products has been increased over the last two decades to meet an increase in global demand, leading the galvanising industry to become a key sector in the fabric of European industry.[1] This growth has increased investments by galvanising companies with a view to increasing not only the production capacity of their continuous hot dip galvanising lines (HDGLs) but also the operational flexibility of their production plants. In fact, the search for greater flexibility is currently

EDMANS Group, Department of Mechanical Engineering, University of La Rioja, La Rioja 26004, Spain

*Corresponding author, email fjmartin@unirioja.es


Published by Maney on behalf of the Institute
Received 19 October 2012; accepted 8 January 2013
DOI 10.1179/1743281213Y.0000000104

considered crucial in meeting the wide variety of customer needs, especially given today's rapidly evolving markets such as vehicle manufacturing.[2] The strategy is clear: the more different products are supplied, the more new markets will be open up. However, this required flexibility may also generate significant problems such as idle times in the continuous mode of operation or reduction in the quality of coatings. Therefore, HDGL plant engineers need to make multiple adjustments when dealing with new products until the right adherence and uniformity in coating are achieved. Reducing the time needed to determine optimal set points is also crucial for saving costs.[3]

This paper focuses on enhancing the accuracy and efficiency with which the operating set points are calculated for a continuous annealing furnace (CAF) on an HDGL. Over the last few decades, many mathematical models based on heat and mass balance have been developed to predict furnace temperature settings.[4] However, those models have some disadvantages such as clear difficulties in tuning model parameters to new product specifications, strong dependence on the experience of the engineer and long computation time to modify models. Recent papers have proved that data driven models based on artificial intelligence (AI) are a good alternative for developing accurate prediction models in the steel industry.[5] To take advantage of these techniques in the modelling of the galvanising process, the main requirements are the availability of historical data, detailed knowledge of the chemical composition of steel and time to train the models.

We propose the use of models based on ensemble methods (EMs) with high overall performance in predicting system set points.[6] The major problem of the aforementioned AI based models is that they do not maintain consistent prediction accuracy with all types of products, especially with products not previously processed. Unlike single AI based models, EMs are built using a set of models and the final output is a combination of the outputs of each individual model.[7] Ensemble methods (EMs) have the same requirements as single models, but two independent studies have shown that combining outputs from multiple predictors increase their generalisation capacity.[8,9] This may make EMs very attractive for working in rapidly changing environments.[10]

In this paper, three EMs, i.e. additive regression[11] (AR), bootstrap aggregating[6] (bagging) and dagging,[12] are developed to adjust CAF temperature settings in an HDGL. Additionally, five data driven models, i.e. least median squared linear regression[13] (LMSQ), linear regression[14] (LR), Quinlan's improved M5 algorithm[15] (M5P), multilayer perceptron neural network[16] (MLP) and support vector machine[17] (SVM), are chosen from literature as basic components of the ensembles. A comparative evaluation procedure is also included to help plant engineers select the best performing model. The results for furnace temperature set point predictions obtained with this method highlight the benefits of using EMs rather than other data driven models for the development of better overall models. Primarily, their higher generalisation capacity reduces the need for continual tuning of model parameters when dealing with new products. The additional ease with which models can be set up is the other crucial advantage for engineers.

In short, EMs have the capacity to provide high performance prediction models for online furnace control systems, leading to very low levels of divergence between the ideal annealing profile and the strip temperature actually measured.

## Description of problem

A continuous HDGL is a well known industrial process composed of several independent sections, each one involving a specific treatment.[18] The heart of the annealing treatment is the CAF, which is divided up as follows: preheating area, heating and holding area, slow cooling area, jet cooling area and overaging area. Inside the CAF, the strip is recrystallised by heating it up and holding it at temperatures of between 750 and 850°C; then, the strip is cooled down at different rates to the liquid zinc bath temperature of between 450 and 470°C. In practice, full adjustment to the annealing profile prescribed can be only guaranteed by feeding the strip into the CAF at a constant speed.[19] For that reason, the control of CAF temperature settings is crucial for proper heating along the steel strip and consequently, has significant influence on the mechanical properties of the steel and the quality of the coating adherence.[20–23]

In this study, the CAF's online control system includes models for estimating three furnace temperature set points. These predictions represent a complex task, mainly due to continuous variations in product specifications and in the chemical composition of the steel used. The current data based models tend to specialise in specific groups of coils and do not maintain their accuracy with products not previously processed.[24,25] Several different models therefore need to be developed and tested to solve this problem, but this task may take up enormous amounts of time and effort. The challenge still lies in developing better overall prediction models capable of working out CAF temperature settings for new coils and different operating conditions while maintaining accuracy with no additional model training phases.

## Related research

The prediction of CAF temperature settings is an unresolved challenge for enhancing the online control of the annealing process.[1] Classical approaches determine temperatures empirically on the basis of a number of process trials that generate prediction models using a set of tables.[26] This method is not efficient because there are high costs associated with in plant trials. The most widely used alternative is to develop mathematical models that consider both thermodynamic properties and heat transfer mechanics inside the CAF.[17,27–31] However, as Prieto *et al.*[32] claim, some furnace characteristics and material properties may change appreciably with different compositions and heat treatment of steel, and this may influence these models significantly. This happens, for instance, with the specific heat capacity of the strip and heating power of burners. So more than just metallurgical knowledge is required to determine precise CAF settings for each coil.[23] Developing models that are based on data may improve prediction capacity because they take into account not only the inherent non-linearities of the annealing process but also the knowhow of plant operators and historical

data.[33,34] Since 1998, several authors have reported on studies related to regression models that use historical data from steel processes.[20,35] In recent years, interest in such models has grown, especially in those based on artificial neural networks (ANNs),[2] genetic algorithm guided neural network (GA-NN) ensemble models,[36] fuzzy logic models,[37] fuzzy ANNs,[38] Bayesian models[39] and Gaussian mixture models.[40] For instance, several applications have been developed applying fuzzy set theory in HDGL for system control and quality management. Kuru and Kuru proposed a new galvannealing control system based on a fuzzy inference system that contributed to a significant improvement in the uniformity and quality of the coating layer running at lower limit of permissible coating values.[41] Recently, Zhang *et al.* proposed a feedforward control method based on fuzzy adaptive model for the thickness control process of galvanising coating.[42]

Likewise, ANNs have been successfully applied in many parts of the galvanising process. Schiefer *et al.*[43] report a combination of clustering and RBFN to improve the predictor of an online galvannealing process control. Lu and Markward[44] reduce the coating weight transitional footage by integrating ANN with the coating weight control system in an HDGL. A 2005 paper by Pernía-Espinoza *et al.*[45] shows the high performance of robust MLP estimating the velocity set point of coils inside the CAF using only their characteristics and furnace temperatures. Conversely, Martínez-de-Pisón *et al.*[46] develop similar MLPs to predict CAF temperature set points but without varying strip velocity.

Finally, the problem of finding optimal ANNs has become more tractable with the advent of genetic algorithms,[47] since which time several models have been reported with better balances between accuracy and complexity in predicting the optimal settings of the annealing process in an HDGL[48] and estimating the strip temperature during the annealing treatment inside the CAF.[49]

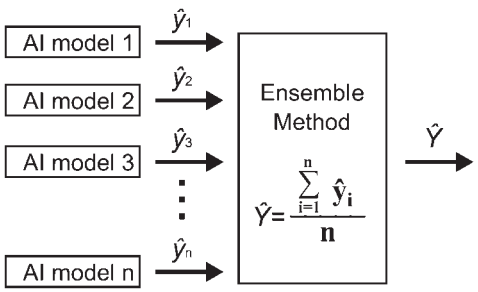
Lastly, Köksal *et al.* and Liao *et al.* presented a complete revision of most relevant data mining techniques developed during the past decade and their applications in steel and manufacturing industry.[50,51]

## Methodology

Many techniques for handling regression tasks have been proposed in literature, but even today, designing automatic methodologies that choose the best performing model for a particular application remains a challenge. One current proposal is based on selecting not only the most accurate model for predicting with stored data but also the model with the highest generalisation capacity. In the following section, the methodology proposed is described on the basis of a comparative evaluation to search for the models with the best balance between accuracy, generalisation capacity and computation cost during the training phase.

### Overview of models

Ensemble learning is a paradigm of ML where multiple single models, referred to as base learners (BLs), are fitted and combined to give a better solution to a particular problem. An EM constructs a set of predictors and then combines their outputs to obtain an improved single response (*see* Fig. 1). Specifically, the final output of an EM is the average or weighted average in regression tasks. Two main conditions need to be addressed for improved results to be achieved: high accuracy in all individual models and diversity between their predictions. A brief description of the EMs selected follows:




**1 Basic layout of EMs by averaging a set of numerical inputs**

(i) additive regression (AR) is a metamodel that enhances the performance of a regression base model. Each iteration fits a model to the residuals left by the predictor on the previous iteration. The output is obtained by adding together the predictions of all the models. The main parameter is the shrinkage or learning rate, which helps prevent overfitting and has a smoothing effect but increases the learning time[11]

(ii) bootstrap aggregating (bagging) is a metamodel that averages a number of BLs fitted with a set of different training datasets. These datasets are generated by sampling uniformly and with replacement of the original data (bootstrapping)[6]

(iii) dagging is a technique similar to bagging but in which the training datasets are created using disjoint samples. A number of disjoint, stratified folds out of the data are generated to train a BL with each subset. Dagging is potentially very suitable for BLs that are quadratic (or even worse in terms of time consumption) regarding the lowest number of instances in each training dataset.[12]

The best performing BLs then need to be selected on the basis of the support provided by previous studies.[52] Five algorithms are finally selected here from an initial list of seven, as follows:

(i) least median squared linear regression is a robust linear regression model based on classical least squares regression that minimises the median of the squared residuals. Its major disadvantage is its lack of efficiency because of its slow convergence rate[13]

(ii) linear regression is an improved version of a traditional scheme for linear predictions that includes the Akaike criterion[53] for feature selection and can also deal with weighted instances[14]

(iii) M5P tree (M5P) is an improvement of the original Quinlan's M5 algorithm for regression tasks by generating a decision tree with simple LR models at its leaves[15,54]

(iv) multilayer perceptron neural network (MLP) is a feedforward ANN that uses a supervised learning algorithm called 'back propagation' to adjust the network weights. The criterion selected for measuring the goodness of fit is usually the least mean square (LMS) error.[16] In regression, one hidden layer is usually considered due to the fact that any continuous function can be approximated with only one hidden layer if the number of connection weights is high enough[55]

(v) support vector machine is a technique with the ability to model nonlinearities resulting in complex mathematical equations. Support vector machine separates the input data, which are presented as two sets of vectors in an $n$ dimensional space, by generating a hyperplane in the same space that maximises the margin between the two input sets.[56]

Other prediction models such as simple LR (SLR), $k$ nearest neighbour ($k$-NN) and radial basis forward network (RBFN) were excluded because they previously showed lower accuracy and generalisation capacity regarding the prediction of new cases than the selected models.[46,49,52]

## Evaluation and performance criteria

The most widely employed procedures for evaluating model accuracy and overfitting are hold out validation and $k$ fold cross-validation.[57] In this paper, the former is selected because it deals better with large DBs from industrial processes. The method consists of dividing data into two non-overlapped datasets by stratified random sampling with a division ratio at percentage $\chi_{tr}$ and $\chi_{val}=1-\chi_{tr}$ for training and validation datasets respectively. The division is repeated $N$ times, and the final errors are determined by average, increasing the robustness of the results against skewed data.

The absolute error criteria considered here are root mean squared error (RMSE) and mean absolute error (MAR). The former is expressed as

$$\mathrm{RMSE}=\left\{\frac{\sum_{k=1}^{n}\left[y(k)-\hat{y}(k)\right]^2}{n}\right\}^{1/2} \quad (1)$$

where $y(k)$ is the target output, $\hat{y}(k)$ is the prediction of the model and $n$ is the total number of instances; the latter is defined as

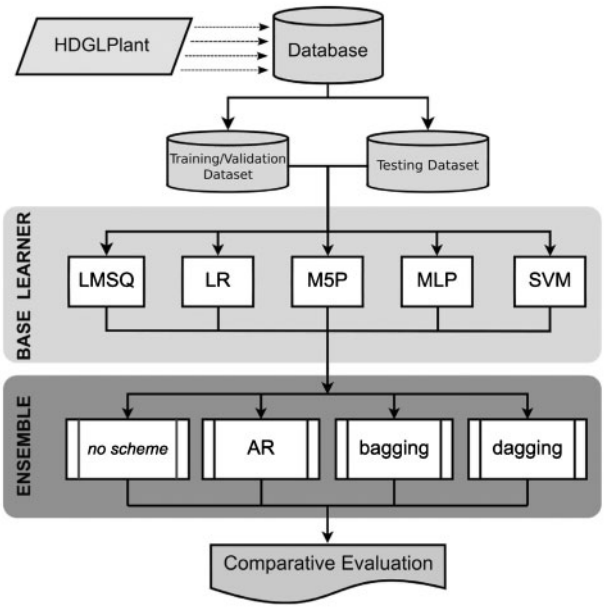
$$\mathrm{MAE}=\frac{\sum_{k=1}^{n}\left|y(k)-\hat{y}(k)\right|}{n} \quad (2)$$

From our knowledge, absolute measures are more important for comparison because they enable the actual application of models to be evaluated directly. The percentage error is also available because the data are normalised. Another performance measure is the coefficient of variation (CV), which compares the amount of variance between sets with different means and is defined as follows

$$\mathrm{CV(RMSE)}=\mathrm{RMSE^{SD}}\left(\mathrm{RMSE^{ME}}\right)^{-1} \quad (3)$$

which represents a comparative measure of the model's stability.

Finally, it is necessary to select those models that show not only good prediction capacity on modelling data but also low generalisation errors. To that end, a testing dataset with coils not included in the modelling dataset is used to check the overall capacity of models.




**2 Conceptual scheme of methodology**

## Description of framework of experiment

The proposed framework involves the complete methodology for finding the best overall model from a set of preselected models. This prior selection is carried out using different setting parameters with each specific model and then selecting the most accurate. Figure 2 summarises the method, which is mainly divided into the following steps:

(i) creation of the DB using the data collected and subsequent division into the modelling and testing datasets by stratified random sampling. This technique homogenises the number of different cases in each dataset

(ii) developing all possible EMs combining every BL and ensemble scheme chosen with a list of preestablished parameters with no prior knowledge of performance. This is an iterative process and the values of the parameters are established according to recommendations by other authors, previous experiences or a series of initial experiments

(iii) checking of models using the testing dataset, which includes new types of steel coils and enables the generalisation error of the model to be assessed

(iv) ranking of models by a comparative evaluation to select the most appropriate for a particular application, in four steps:

*Step 1:* accuracy in predicting known values is evaluated by calculating the $\mathrm{RMSE_{val}^{ME}}$ of $M$ models. Those EMs that show an $\mathrm{RMSE_{val}^{ME}} \le \mathrm{RMSE_{val}^{TH}}$, where $\mathrm{RMSE_{val}^{TH}}$ is a user defined threshold, are the only ones chosen for the next step.

*Step 2:* the computation cost of training $M$ models ($t$) is measured to discard those EMs that take too long, i.e. $t<t^{TH}$, where the value $t^{TH}$ is a threshold directly proportional to the complexity and size of the modelling dataset.

*Step 3:* a high generalisation capacity is also necessary to maintain accuracy when

models face unseen data. This is evaluated comparing the $(\mathrm{RMSE}^{\mathrm{ME}} \pm \mathrm{LSD})_{\mathrm{val}}$ and $(\mathrm{RMSE}^{\mathrm{ME}} \pm \mathrm{LSD})_{\mathrm{tst}}$, where LSD represents the least significant difference for independent values. The first value should be at least equal to the second but never much lower if the behaviour of the model is to be considered as satisfactory in terms of generalisation capacity.

*Step 4:* finally, the stability of the EMs is evaluated by comparing two terms: $\mathrm{RMSE}^{\mathrm{SD}}_{\mathrm{tst}}$ and the ratio $\mathrm{CV}_{\mathrm{tst}}$ for a total number of $M$ models. Stable EMs are able to keep low variability in output errors and show very low CV when the input data are changing constantly.

In this paper, the case under study is the prediction of CAF temperature settings in an HDGL, but the method can be applied in other processes.

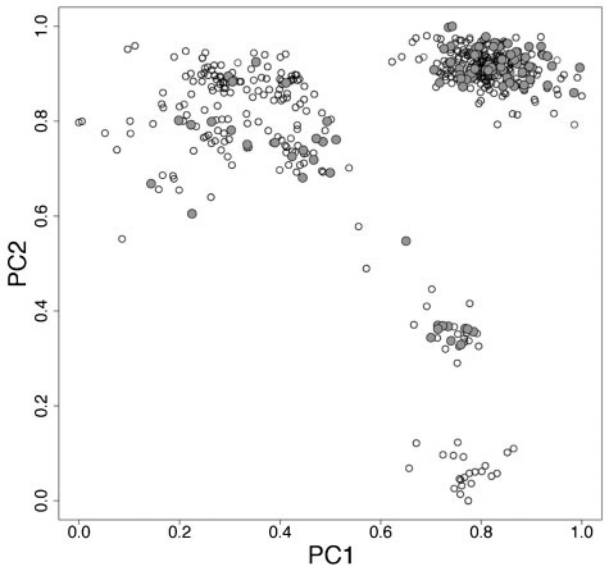
## Hot dip galvanising line database for experiment

The three furnace temperature set points and other attributes were measured on an HDGL operating in Spain. The raw dataset was formed by 56 284 observations of 2436 types of coil, sampled every 100 m along the strip under different processing conditions and chemical compositions. In 2003, Martínez-de-Pisón preprocessed the raw data filtering for wrong records, detecting outliers and removing redundant variables.[26] Later, a principal component analysis[58] (PCA) was carried out to reduce the number of inputs related to the chemical composition of steel (C, Mn, Si, S, P, Al, Cu, Ni, Cr, Nb, V, Ti, B and N) to only the first seven principal axes (from PC1 to PC7), which are independent to one another and cover 87·44% of the original variance.[52] This technique appears as a black box to plant engineers and undermines the physical relationships within the chemical composition and the annealing process. Nevertheless, PCA was mainly motivated by its proved reliability, reducing the amount of data from industrial process datasets since the data used to be redundant and the variables are usually correlated.

The modelling dataset was finally formed by 49 000 instances from sampling original data until the number of coil types was homogenised. The set of inputs comprise 12 attributes (WidthCoil, ThickCoil, VelMed, TMPP1, TMPP2CNG, PC1, PC2, PC3, PC4, PC5, PC6 and PC7), and the outputs are the CAF's three temperature set points (THC1, THC3 and THC5). In the case of the test dataset, this was acquired several weeks after the data for the modelling dataset were gathered. As Table 1 shows, the test dataset is formed by 5000 instances containing 59 different types of coil but only 25 different types of steel. Finally, both datasets were normalised. Using the same datasets in this paper as in previous studies enables us to compare past models with our new proposal.[52]

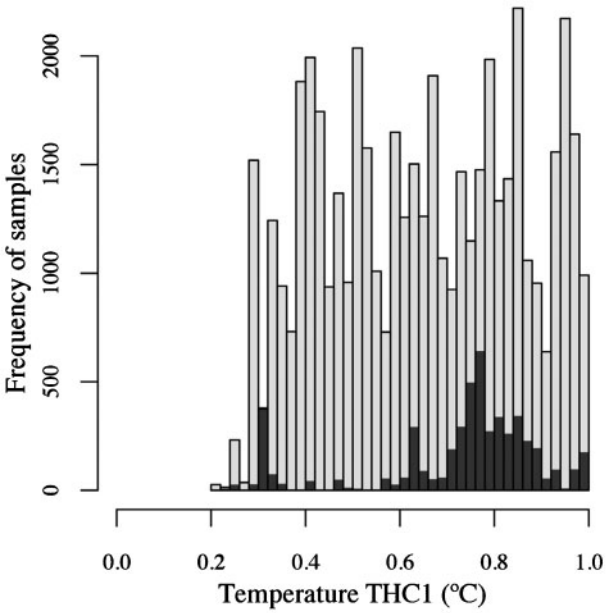
**Table 1 Testing DB used to test generalisation and flexibility of models**

| Description | Values |
|---|---|
| Number of coils in DB | 59 |
| Different types of steels in DB | 25 |
| ThickCoil (range) | 0·601–0·775 mm |
| WidthCoil (range) | 805–1180 mm |




**3 Principal component analysis projection of HDGL dataset using first two principal components: circles indicate modelling data, and grey dots indicate testing data**

It is advisable to take a final overview of the data before the results are explained, because the accuracy of data driven models is directly dependent on the problem that needs to be solved. Figures 3 and 4 show the modelling and testing data distributions using the first two principal axes (PC1 and PC2) and the histogram of THC1 respectively. Figure 3 reveals the existence of a large group of similar coils and several other small groups. The testing data also look evenly distributed enough over the whole range of modelling data but are extremely sparse, like most industrial data (Fig. 4). These data features decrease the prediction capacities of data driven models,




**4 Histogram of attribute temperature *THC1*: light grey bars indicate modelling samples, and dark grey ones indicate testing data**

**Table 2 Parameter specifications for BLs**

| Algorithm | Parameter | Values |
|---|---|---|
| LR | Ridge | 0·1, 0·3, …, 4 |
| | Attribute selection | Greedy |
| LSMQ | Sample size | 4, 8, …, 400 |
| M5P | Min. instances/leaf | 2, 4, …, 80 |
| | Prune | True |
| | Unsmoothed model | False |
| MLP | Learning rate | 0·1, 0·2, …, 0·8 |
| | Momentum | 0·1, 0·2, …, 0·6 |
| | Num. iterations | 50 000–10 000 |
| | Num. hidden neurons | 3, 5, 7, …, 40 |
| | Decay learning rate | True |
| | Autoreset network | True |
| | Validation set size | 20% |
| | Validation threshold | 15, 25, …, 55 |
| SVM | *C* | 1·0 |
| | *E* | 1·0E−12 |
| | $\varepsilon$ parameter | 0·001 |
| | Tolerance | 0·001 |
| | Kernel type | Polynomial |
| | Poly. exponent | 1, 1·02, …, 2·50 |

**Table 3 Parameter specifications for ensemble schemes**

| Scheme | Parameter | Values |
|---|---|---|
| AR | Shrinkage | 0·2, 0·3, …, 1 |
| | Num. iterations | 10, 20, …, 80 |
| Bagging | Subset size | 10, 20, …, 100 |
| | Num. learning rounds | 15, 20, …, 80 |
| Dagging | Num. folds | 2, 3, …, 15 |

but the article shows that our proposed EMs are a better way of dealing with these problems than single models.

## Results and discussion

In this paper, we mainly analyse three types of EM to predict the temperature set points (THC1, THC3 and THC5) for a CAF on an HDGL. However, the results concerning THC1 are only reported to summarise the information, due to the fact that the results from training the THC3 and THC5 prediction models are similar to the first one.

The performance of models predicting THC1 is evaluated following the steps and criteria of the proposed methodology. These evaluations are carried out with the statistical software R-project 2·11[59] running on a dual quadcore Opteron server with Linux SUSE 11·2. Additionally, the models are implemented using the WEKA workbench[60] and R-project packages AMORE[61] and RWeka.[62]

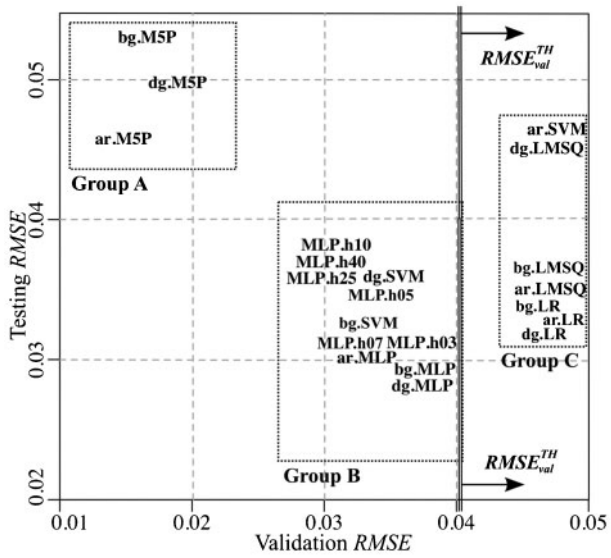
The ranges of the values of the setting parameters used to generate models with different configurations are listed in Table 2 for BLs and Table 3 for EMs. Validation and testing simulations were repeated $N=10$ times for each particular configuration, selecting only the best case for each algorithm. Table 4 presents the mean and standard deviation of both error measures (RMSE and MAE) for the best models and also the total computation time to set up a number of $M=10$ models with the same setting parameters. Likewise, Table 5 shows the results reported by Martínez-de-Pisón *et al.*[52] from earlier research into MLP models. These results allow us to discuss the advantages and disadvantages by comparing the two

**Table 4 Results of modelling process for THC1 prediction models (sorted by $\mathrm{RMSE}_{\mathrm{val}}^{\mathrm{ME}}$)**

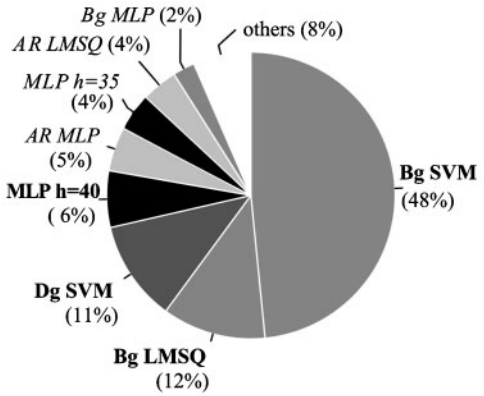
| | Training error | | | | Validation error | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | $\mathrm{RMSE}_{\mathrm{tr}}^{\mathrm{ME}}$ | $\mathrm{RMSE}_{\mathrm{tr}}^{\mathrm{SD}}$ | $\mathrm{MAE}_{\mathrm{tr}}^{\mathrm{ME}}$ | $\mathrm{MAE}_{\mathrm{tr}}^{\mathrm{SD}}$ | $\mathrm{RMSE}_{\mathrm{val}}^{\mathrm{ME}}$ | $\mathrm{RMSE}_{\mathrm{val}}^{\mathrm{SD}}$ | $\mathrm{MAE}_{\mathrm{val}}^{\mathrm{ME}}$ | $\mathrm{MAE}_{\mathrm{val}}^{\mathrm{SD}}$ | *t*/s |
| AR M5P | 0·0124 | 0·0003 | 0·0065 | 0·0001 | 0·0148 | 0·0009 | 0·0071 | 0·0002 | 1799 |
| Bg M5P | 0·0148 | 0·0001 | 0·0079 | 0·0001 | 0·0166 | 0·0008 | 0·0084 | 0·0002 | 3492 |
| Dg M5P | 0·0176 | 0·0003 | 0·0096 | 0·0002 | 0·0189 | 0·0014 | 0·0100 | 0·0002 | 117 |
| Bg SVM | 0·0315 | 0·0001 | 0·0194 | 0·0001 | 0·0314 | 0·0003 | 0·0195 | 0·0001 | 396 000 |
| AR MLP | 0·0339 | 0·0003 | 0·0237 | 0·0003 | 0·0336 | 0·0004 | 0·0236 | 0·0002 | 4000 |
| Dg SVM | 0·0336 | 0·0002 | 0·0218 | 0·0001 | 0·0337 | 0·0004 | 0·0218 | 0·0002 | 93 000 |
| Bg MLP | 0·0376 | 0·0003 | 0·0263 | 0·0002 | 0·0378 | 0·0008 | 0·0263 | 0·0004 | 20 000 |
| Dg MLP | 0·0382 | 0·0003 | 0·0268 | 0·0002 | 0·0380 | 0·0004 | 0·0267 | 0·0002 | 6243 |
| AR LR | 0·0477 | 0·0003 | 0·0351 | 0·0001 | 0·0473 | 0·0007 | 0·0349 | 0·0003 | 194 |
| AR LMSQ | 0·0481 | 0·0001 | 0·0350 | 0·0001 | 0·0475 | 0·0001 | 0·0348 | 0·0001 | 33 156 |
| Bg LR | 0·0475 | 0·0002 | 0·0350 | 0·0001 | 0·0476 | 0·0005 | 0·0350 | 0·0002 | 151 |
| Bg LMSQ | 0·0484 | 0·0002 | 0·0349 | 0·0001 | 0·0476 | 0·0005 | 0·0342 | 0·0002 | 94 949 |
| Dg LR | 0·0475 | 0·0002 | 0·0350 | 0·0001 | 0·0478 | 0·0005 | 0·0352 | 0·0003 | 17 |
| Dg LMSQ | 0·0482 | 0·0003 | 0·0351 | 0·0001 | 0·0484 | 0·0005 | 0·0351 | 0·0002 | 1357 |
| AR SVM | 0·0510 | 0·0003 | 0·0413 | 0·0001 | 0·0511 | 0·0006 | 0·0419 | 0·0002 | 8608 |

**Table 5 Results of modelling process of MLP models predicting for *THC1* (sorted by $\mathrm{RMSE}_{\mathrm{val}}^{\mathrm{ME}}$) (source: Martínez-de-Pisón *et al.*[52])**

| | Training errors | | | | Validation errors | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Algorithm | $\mathrm{RMSE}_{\mathrm{tr}}^{\mathrm{ME}}$ | $\mathrm{RMSE}_{\mathrm{tr}}^{\mathrm{SD}}$ | $\mathrm{MAE}_{\mathrm{tr}}^{\mathrm{ME}}$ | $\mathrm{MAE}_{\mathrm{tr}}^{\mathrm{SD}}$ | $\mathrm{RMSE}_{\mathrm{val}}^{\mathrm{ME}}$ | $\mathrm{RMSE}_{\mathrm{val}}^{\mathrm{SD}}$ | $\mathrm{MAE}_{\mathrm{val}}^{\mathrm{SD}}$ | $\mathrm{MAE}_{\mathrm{val}}^{\mathrm{ME}}$ | *t*/s |
| MLP *h*=40 | 0·0303 | 0·0020 | 0·0211 | 0·0021 | 0·0305 | 0·0020 | 0·0212 | 0·0022 | 52 167 |
| MLP *h*=10 | 0·0308 | 0·0008 | 0·0209 | 0·0008 | 0·0309 | 0·0009 | 0·0210 | 0·0008 | 9958 |
| MLP *h*=35 | 0·0306 | 0·0029 | 0·0213 | 0·0031 | 0·0311 | 0·0033 | 0·0214 | 0·0032 | 34 582 |
| MLP *h*=07 | 0·0313 | 0·0004 | 0·0216 | 0·0006 | 0·0317 | 0·0007 | 0·0217 | 0·0006 | 9932 |
| MLP *h*=05 | 0·0325 | 0·0004 | 0·0225 | 0·0003 | 0·0328 | 0·0007 | 0·0226 | 0·0003 | 7401 |
| MLP *h*=03 | 0·0355 | 0·0005 | 0·0252 | 0·0003 | 0·0356 | 0·0005 | 0·0252 | 0·0004 | 3643 |




**5 $RMSE_{val}^{ME}$ versus $RMSE_{tst}^{ME}$ for models evaluated**




**6 Comparative percentages of computation costs ($t$) for modelling 10 models: models discarded due to $t>50\,000$ s are shown in boldface; label 'others' includes all other models with $t<10\,000$ s**

approaches. This paper also provides support for adjusting MLPs using LMS instead of least mean log squares (LMLS) as the error criterion because LMS costs three times less in terms of computation than LMLS for similar results.

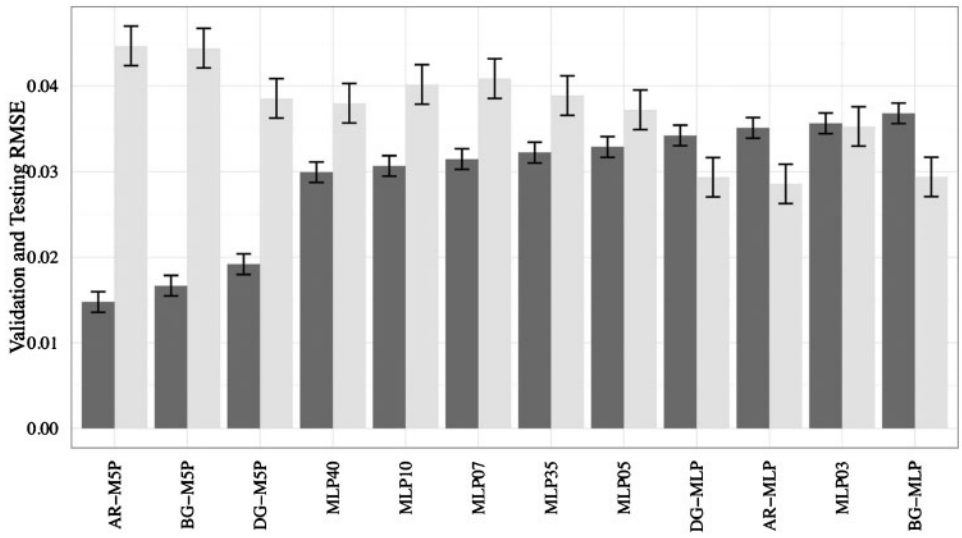
First, Table 4 shows that $RMSE_{val}^{ME}$ was <1·9% for every EM using M5P as its BL. Indeed, the $RMSE_{val}^{ME}$ values of M5P models, marked as group A in Fig. 5, are significantly lower than the rest. This suggests that the application of any ensemble scheme to a tree based regressor generates large models composed of several specialised trees with hundreds of leaves. These results contrast with those obtained from EMs that include MLP or SVM as their BL (except AR-SVM), marked as group B in Fig. 5, with $RMSE_{val}^{ME}$ ~3·5%. Table 4 also shows that the application of any ensemble scheme has a great influence on $RMSE_{val}^{SD}$ and $MAE_{val}^{SD}$, in terms of reducing them in comparison to the corresponding BL. AR-SVM and all the EMs that include LSMQ or LR as their BL (*see* group C in Fig. 5) present a $RMSE_{val}^{ME}$ higher than the predefined threshold $RMSE_{val}^{TH}=4{\cdot}0\%$. Thus, they fail to assure enough accuracy in predicting temperature THC1 and are therefore discarded.

In the next step (step 2), the time for training models is evaluated to check that they are generated in a time that is reasonable in proportion to the size of the modelling task. The time limit for training $M=10$ models ($t^{TH}$) is set at 50 000 s. Both Tables 4 and 5 show that the computation costs associated with SVM based EMs and those MLP based EMs with high numbers of neurons in the hidden layer ($h=40$) are higher than $t^{TH}$. For that reason, bagging SVM, dagging SVM and MLP with $h=40$ are automatically discarded. Note that the computation times for the training phase summarised in Table 5 differ from the results previously published because the simulations are repeated to gather new times in equal conditions for all models. The comparative percentages are shown in Fig. 6 to highlight significant differences between algorithms. However, the BLs that make up EMs can be independently trained in parallel, drastically decreasing the training time of EMs.

The model's reliability in predicting THC1 from unseen data is analysed in step 3. Table 6 shows the testing errors of a number of $M=10$ models with the setting parameters that generate the lowest $RMSE_{val}^{ME}$. It is observed that the best performing of all the models in Step 1 provides the worst results in step 3, i.e. $RMSE_{tst}^{ME}$ of M5P based EM is higher than 4·5% when $RMSE_{val}^{ME}$ was <1·9% in step 1. We illustrate $RMSE_{val}^{ME}$ and $RMSE_{tst}$ together in Fig. 7 to support the selection of the best overall models. The error bars represent least significant difference values to visually identify whether an error is significantly lower than others. The figure leaves no doubts that M5P based EM shows the lowest generalisation capacity when working with new coils and, conversely, MLP with only three neurons in the hidden layer (MLP03) and MLP based EM (AR MLP, Bg MLP and Dg MLP) are the best overall models. The main reason is that these last four models show high capacities to achieve a good balance between validation and testing errors because their structure is less complex

**Table 6 Testing errors of THC1 prediction models (sorted by $RMSE_{tst}^{ME}$), including MLP models from previous article (source: Martínez-de-Pisón *et al.*[52])**

| Algorithm | $RMSE_{tst}^{ME}$ | $RMSE_{tst}^{SD}$ | $MAE_{tst}^{ME}$ | $MAE_{tst}^{SD}$ | CV |
|---|---|---|---|---|---|
| Dg MLP | 0·0284 | 0·0003 | 0·0218 | 0·0004 | 0·0106 |
| Bg MLP | 0·0294 | 0·0003 | 0·0233 | 0·0002 | 0·0102 |
| AR MLP | 0·0298 | 0·0001 | 0·0228 | 0·0002 | 0·0034 |
| MLP $h=07$ | 0·0322 | 0·0019 | 0·0241 | 0·0016 | 0·0590 |
| MLP $h=03$ | 0·0322 | 0·0019 | 0·0249 | 0·0018 | 0·0590 |
| Bg SVM | 0·0329 | 0·0001 | 0·0241 | 0·0002 | 0·0030 |
| Dg LR | 0·0332 | 0·0001 | 0·0258 | 0·0001 | 0·0030 |
| AR LR | 0·0332 | 0·0003 | 0·0257 | 0·0002 | 0·0090 |
| Bg LR | 0·0333 | 0·0001 | 0·0258 | 0·0001 | 0·0030 |
| Bg LMSQ | 0·0334 | 0·0001 | 0·0262 | 0·0001 | 0·0030 |
| ARLMSQ | 0·0348 | 0·0009 | 0·0278 | 0·0008 | 0·0259 |
| MLP $h=05$ | 0·0361 | 0·0044 | 0·0274 | 0·0041 | 0·1219 |
| Dg SVM | 0·0364 | 0·0007 | 0·0273 | 0·0006 | 0·0192 |
| MLP $h=40$ | 0·0368 | 0·0017 | 0·0274 | 0·0013 | 0·0462 |
| MLP $h=35$ | 0·0378 | 0·0019 | 0·0281 | 0·0014 | 0·0503 |
| MLP $h=10$ | 0·0390 | 0·0026 | 0·0289 | 0·0018 | 0·0667 |
| Dg LMSQ | 0·0452 | 0·0052 | 0·0289 | 0·0013 | 0·1150 |
| AR M5P | 0·0458 | 0·0059 | 0·0303 | 0·0029 | 0·1288 |
| AR SVM | 0·0465 | 0·0002 | 0·0377 | 0·0002 | 0·0043 |
| Dg M5P | 0·0497 | 0·0136 | 0·0293 | 0·0034 | 0·2736 |
| Bg M5P | 0·0610 | 0·0009 | 0·0299 | 0·0001 | 0·0106 |



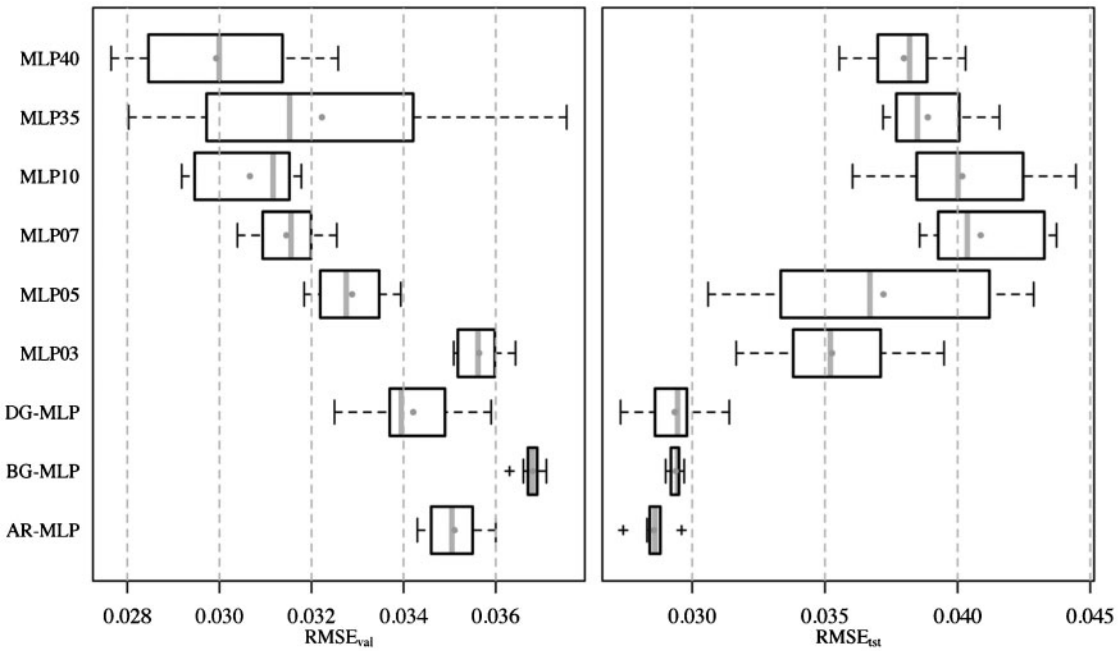

**7 Bar plot with error bars of $RMSE_{val}^{ME}$ and $RMSE_{tst}^{ME}$: dark and light grey bars correspond to validation and testing errors respectively; error bars represent least significant difference between means; thus, two overlapping bars indicate non-significant difference and non-overlapping bars indicate opposite**

and not so specialised in coils already treated. Moreover, some models sometimes present lower $RMSE_{tst}^{ME}$ than $RMSE_{val}^{ME}$ because the validation dataset may be more complex than testing dataset.

Finally, the stability of models is measured and compared by determining the $RMSE_{tst}^{SD}$ and $CV_{tst}$. Stability is formally defined as the degree to which a technique creates repeatable results, given different datasets sampled from the same data. The aim is to find those models that have the lowest $RMSE_{tst}^{SD}$ and $CV_{tst}$ at the same time. Models are not very reliable if they predict different values each time there are slight changes in the input data or even in some of the setting parameters of the model itself. Table 6 shows that EMs that use MLP as BL are the more stable because their $CV_{tst}$ are at least four times lower than each single MLP. Additionally, Fig. 8 lends support to previous observations by illustrating $RMSE_{tst}$ as a box plot of the testing error distribution. It can be claimed that EMs using MLP as BL have a narrower $RMSE_{tst}$ distribution than any single data driven model.

## Conclusions

This paper mainly presents three ensemble schemes that combine the outputs of several data driven models to predict temperature set points for a CAF on an HDGL. The main contributions relate to the description of a methodology for evaluating models, the comparative evaluation of 15 cases and the final recommendations




**8 Box plots of RMSE (left, validation; right, testing): whiskers cover those samples that are outside interquartile range but are not outliers; crosses represent outliers that are samples at distance of >1·5 times interquartile range; grey dots and grey lines are mean and median of data respectively**

for using ensemble learning as a regression technique. Ensemble methods generally show themselves to be highly efficient and reliable in terms of computation cost, generalisation capacity and ease of determination of the best model configuration. However, comparative evaluation proves that the use of MLP as BL is the best choice for generating the best performing model, i.e. AR MLP, Bg MLP and Dg MLP. Compared to single MLP, MLP based EM achieves similar accuracy in predicting temperature set points for types of coils stored in the modelling DB but a lower generalisation error with coils not previously processed. This demonstrates that EMs are universal in nature and better suited for working with low quality and sparse DBs from industrial processes. Finally, the extra cost in computation time required for these models is not so restrictive if it is considered that they can be easily parallelised and plant engineers can also avoid continually having to change the model for new products thanks to its higher overall capacity and stability.

## Acknowledgements

The authors are grateful for financial support provided by the European Union via project no. RFS-PR-06035, by the University of La Rioja via grant FPI-2012 and for support provided by the Autonomous Government of La Rioja under its 3er Plan Riojano de I+D+I via project FOMENTA 2010/13.

## References

1. M. M. Prieto, F. J. Fernández and J. L. Rendueles: 'Thermal performance of annealing line heating furnace', *Ironmaking Steelmaking*, 2005, **32**, (2), 171–176.
2. M. Schlang and B. Lang: 'Current and future development in neural computation in steel processing', *Control Eng. Pract.*, 2001, **9**, 975–986.
3. J. B. Ordieres, A. González, J. A. González and V. Lobato: 'Estimation of mechanical properties of steel strip in hot dip galvanising lines', *Ironmaking Steelmaking*, 2004, **31**, (1), 43–50.
4. F. T. P. de Medeiros, S. J. X. Noblat and A. M. F. Fileti: 'Reviving traditional blast furnace models with new mathematical approach', *Ironmaking Steelmaking*, 2007, **34**, (5), 410–414.
5. P. J. Laitinen and H. Saxén: 'A neural network based model of sinter quality and sinter plant performance indices', *Ironmaking Steelmaking*, 2007, **34**, (2), 109–114.
6. L. Breiman: 'Bagging predictors', *Mach. Learn.*, 1996, **24**, (2), 123–140.
7. Z. Zhou: 'Encyclopedia of database systems'; 2009, Berlin, Springer.
8. E. Bauer and R. Kohavi: 'An empirical comparison of voting classification algorithms: bagging, boosting and variants', *Mach. Learn.*, 1999, **36**, 105–139.
9. D. Opitz and R. Maclin: 'Popular ensemble methods: an empirical study', *J. Artif. Intell. Res.*, 1999, **11**, 169–198.
10. T. G. Dietterich: 'Machine-learning research: four current directions', *AI Mag.*, 1998, **18**, (4), 97–136.
11. J. H. Friedman: 'Stochastic gradient boosting'; 1999, Standford, Stanford University.
12. K. M. Ting and I. H. Witten: 'Stacking bagged and dagged models', In Proc. 14th International Conference on Machine Learning, Morgan Kaufmann, Burlington, Massachusetts, USA, 367–375; 1997.
13. P. J. Rousseeuw and A. M. Leroy: 'Robust regression and outlier detection'; Hoboken, New Jersey, USA, 1987, Wiley.
14. K. P. Burnham and D. R. Anderson: 'Model selection and inference: a practical information-theoretic approach', New York, USA, 353; 1998, Springer.
15. Y. Wang and I. H. Witten: 'Induction of model trees for predicting continuous classes', Proc. 9th European Conf. on 'Machine learning', Prague, Czech Republic, April 1997, Springer, 128–137.
16. S. Haykin: 'Neural networks: a comprehensive foundation'; Upper Saddle River, New Jersey, USA, 1999, Prentice Hall.
17. S. S. Sahay and P. C. Kapur: 'Model based scheduling of a continuous annealing furnace', *Ironmaking Steelmaking*, 2007, **34**, (3), 262–268.
18. F. J. Martínez-de-Pisón, A. Sanz, E. Martínez-de-Pisón, E. Jiménez and D. Conti: 'Mining association rules from time series to explain failures in a hot-dip galvanizing steel line', *Comput. Ind. Eng.*, 2012, **63**, (1), 22–36.
19. S. R. Yoo, I. S. Choi, P. K. Nam, J. K. Kim, S. J. Kim and J. Davene: 'Coating deviation control in transverse direction for a continuous galvanising line', *IEEE Trans. Control Syst. Technol.*, 1999, **7**, (1), 129–135.
20. N. Yoshitani and A. Hasegawa: 'Model-based control of strip temperature for the heating furnace in continuous annealing', *IEEE Trans. Control Syst. Technol.*, 1998, **6**, (2), 146–156.
21. J.-S. Kim and J.-H. Chung: 'Galvannealing behaviour of high strength galvanized sheet steels', Proc. Galvanised Steel Sheet Forum – Automotive Conf., London, UK, May 2000, IOM Communication Ltd., 103–108.
22. J. Mahieu, M. De-Meyer and B. C. De-Cooman: 'Galvanizability of high strength steels for automotive applications', Proc. Galvanised Steel Sheet Forum – Automotive Conf., London, UK, May 2000, IOM Communication Ltd., 185–198.
23. G. Bloch, F. Sirou, V. Eustache and P. Fatrez: 'Neural intelligent control for a steel plant', *IEEE Trans. Neural Netw.*, 1997, **8**, (4), 910–918.
24. L. Bitschnau and M. Kozek: 'Modeling and control of an industrial continuous furnace', Proc. Int. Conf. on 'Computational intelligence, modelling and simulation', Brno, Czech Republic, September 2009, IEEE, 231–236.
25. Z. Ming and Y. Datai: 'A new strip temperature control method for the heating section of continuous annealing line', Proc. IEEE Conf. on 'Cybernetics and intelligent systems', London, UK, September 2008, IEEE, 861–864.
26. F. J. Martínez-de-Pisón: 'Optimización mediante técnicas de minería de datos del ciclo de recocido de una línea de galvanizado', PhD thesis, University of La Rioja, Logroño, Spain, 2003.
27. Y. Jaluria: 'Numerical simulation of the transport processes in a heat treatment furnace', *Int. J. Numer. Methods Eng.*, 1988, **25**, 387–399.
28. 'Hot dip coating line for ACERALIA', S. A. Drever International, Liege, Belgium, 1998.
29. S. S. Sahay, A. M. Kumar and A. Chatterjee: 'Development of integrated model for batch annealing of cold rolled steels', *Ironmaking Steelmaking*, 2004, **31**, 144–152.
30. C. S. Townsend: 'Closed-loop control of coating weigth on a hot dip galvanizing line', *Iron Steel Eng.*, 1988, **65**, 44–47.
31. R. Mehta and S. S. Sahay: 'Heat transfer mechanisms and furnace productivity during coil annealing: aluminum vs. steel', *J. Mater. Eng. Perform.*, 2009, **18**, (1), 8–15.
32. M. M. Prieto, F. J. Fernández and J. L. Rendueles: 'Development of stepwise thermal model for annealing line heating furnace', *Ironmaking Steelmaking*, 2005, **32**, (2), 165–170.
33. J. Tenner, D. A. Linkens, P. F. Morris and T. J. Bailey: 'Prediction of mechanical properties in steel heat treatment process using neural networks', *Ironmaking Steelmaking*, 2001, **28**, (1), 15–22.
34. D. M. Jones, J. Watton and K. J. Brown: 'Comparison of hot rolled steel mechanical property prediction models using linear multiple regression, non-linear multiple regression and non-linear artificial neural networks', *Ironmaking Steelmaking*, 2005, **32**, (5), 435–442.
35. M. Schlang, B. Feldkeller, B. Lang, T. Poppe and T. Runkler: 'Neural computation in steel industry', Proc. European Control Conf. '99, Session BP-1, Karlsruhe, Germany, August–September 1999, European Union Control Association, 1–6.
36. Y.-Y. Yang, M. Mahfouf and G. Pnoutsos: 'Development of a parsimonious GA-NN ensemble model with a case study for Charpy impact energy prediction', *Adv. Eng. Software*, 2011, **42**, 435–443.
37. M. A. Hassan, M. A. El-Sharief, A. Aboul-Kasem, S. Ramesh and J. Purbolaksono: 'A fuzzy model for evaluation and prediction of slurry erosion of 5127 steels', *Mater. Des.*, 2012, **39**, 186–191.
38. J. Li, H. Feng and S. Li: 'Wavelet prediction fuzzy neural network of the annealing furnace temperature control', Proc. 2011 Int. Conf. on 'Electric information and control engineering', Wuhan, China, April 2011, IEEE, 940–943.
39. K. Agarwal and R. Shivpuri: 'An on-line hierarchical decomposition based bayesian model for quality prediction during hot strip rolling', *ISIJ Int.*, 2012, **52**, (10), 1862–1871.

40. Y. Y. Yang, M. Mahfouf and G. Panoutsos: 'Probabilistic characterisation of model error using Gaussian mixture model – with application to Charpy impact energy prediction for alloy steel', *Control Eng. Pract.*, 2012, **20**, (1), 82–92.
41. E. Kuru and L. Kuru: 'Fuzzy inference system controls in hot dip galvanizing lines', Proc. 7th Int. Conf. on 'Electrical and electronics engineering', Bursa, Turkey, December 2011, IEEE, II-400–II-404.
42. Y. Zhang, F.-Q. Shao, J.-S. Wang and B.-Q. Liu: 'Thickness control of hot dip galvanizing coating based on fuzzy adaptive model', *J. Shenyang Univ. Technol.*, 2012, **34**, (5), 576–580 + 590.
43. C. Schiefer, F. X. Rubenzucker, H. P. Jörgl and H. R. Aberl: 'A neural network controls the galvannealing process', *IEEE Trans. Ind. Appl.*, 1999, **35**, (1), 114–118.
44. Y.-Z. Lu and S. W. Markward: 'Development and application of an integrated neural system for an HDCL', *IEEE Trans. Neural Netw.*, 1997, **8**, (6), 1328–1337.
45. A. Pernía-Espinoza, M. Castejón-Limas, A. González-Marcos and V. Lobato-Rubio: 'Steel annealing furnace robust neural network model', *Ironmaking Steelmaking*, 2005, **32**, (5), 418–426.
46. F. J. Martínez-de-Pisón, A. Pernía, E. Jiménez-Macías and R. Fernández: 'Overall model of the dynamic behaviour of the steel strip in an annealing heating furnace on a hot-dip galvanizing line', *Rev. Metal. Madrid*, 2010, **46**, (5), 405–420.
47. M. Mitchell: 'An introduction to genetic algorithms'; 1998, Cambridge, The MIT Press.
48. F. J. Martínez-de-Pisón, F. Alba-Elías, M. Castejón-Limas and J. A. González-Rodríguez: 'Improvement and optimisation of hot dip galvanising line using neural networks and genetic algorithms', *Ironmaking Steelmaking*, 2006, **33**, (4), 344–352.
49. F. J. Martínez-de-Pisón, L. Celorrio, M. Pérez-De-La-Parte and M. Castejón: 'Optimising annealing process on hot dip galvanising line based on robust predictive models adjusted with genetic algorithms', *Ironmaking Steelmaking*, 2011, **38**, (3), 218–228.
50. G. Köksal, N. Batmaz and M. C. Testik: 'A review of data mining applications for quality improvement in manufacturing industry', *Expert Syst. Appl.*, 2011, **38**, (10), 13448–13467.
51. S.-H. Liao, P.-H. Chu and P.-Y. Hsiao: 'Data mining techniques and applications – a decade review from 2000 to 2011', *Expert Syst. Appl.*, 2012, **39**, (12), 11303–11311.
52. F. J. Martínez-de-Pisón, A. V. Pernía, A. González, L. M. López-Ochoa and J. B. Ordieres: 'Optimum model for predicting temperature settings on hot dip galvanising line', *Ironmaking Steelmaking*, 2010, **37**, (3), 187–194.
53. H. Akaike: 'A new look at the statistical model identification', *IEEE Trans. Autom. Control*, 1974, **19**, (6), 716–723.
54. R. J. Quinlan: 'Learning with continuous classes', Proc. 5th Australian Joint Conf. on 'Artificial intelligence', Singapore, November 1992, World Scientific, 343–348.
55. K. Hornik, M. Stinchcombe and H. White: 'Multilayer feedforward networks are universal approximators', *Neural Netw.*, 1989, **2**, (5), 359–366.
56. S. K. Shevade, S. S. Keerthi, C. Bhattacharyya, K. R. K. Murthy and I. Smola: 'Improvements to the SMO algorithm for SVM regression', *IEEE Trans. Neural Netw.*, 2000, **1**, 1188–1193.
57. S. C. Larson: 'The shrinkage of the coefficient of multiple correlation', *J. Educ. Psychol.*, 1931, **22**, (1), 45–55.
58. K. Pearson: 'On lines and planes of closest fit to systems of points in space', *Philos. Mag.*, 1901, **2**, (6), 559–572.
59. RCore Team: 'R: a language and environment for statistical computing'; 2012, Vienna, R Foundation for Statistical Computing.
60. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten: 'The WEKA data mining software: an update', *SIGKDD Explor.*, 2009, **1**, (1).
61. M. Castejón-Limas, J. B. Ordieres Meré, E. P. Vergara, F. J. Martínez-de-Pisón, A. V. Pernía and F. Alba: 'The AMORE package: a MORE flexible neural network package'; 2009, CRAN Repository.
62. K. Hornik, C. Buchta, T. Hothorn, A. Karatzoglou, D. Meyer and A. Zeileis: 'R/Weka interface'; 2011, CRAN Repository.