Impact of Data Quality on Predictive Accuracy of ANFIS based Soft Sensor Models

S. Jassar, *Student Member, IEEE*, Z. Liao, *Member, ASHRAE*, L. Zhao, *Senior Member, IEEE*

***Abstract*— Soft sensor models are used to infer the critical process variables that are otherwise difficult, if not impossible, to measure in broad range of engineering fields. For Adaptive Neuro-Fuzzy Inference System (ANFIS) based soft sensor models, when the prediction accuracy is analyzed, it is assumed that both the data used to train the model and the testing data to make predictions are free of errors. But rarely a data set is clean before extraordinary effort having been put to clean the data. This paper investigates the impact of data quality on the prediction accuracy of an ANFIS based soft sensor model that is designed to estimate the average air temperature in distributed heating systems. The average air temperature is estimated based upon the available information, including solar radiation (Q_{sol}), energy used by boiler (Q_{in}) and external temperature (T_0). For this problem, with the measurement errors caused by reading and equipment of all three variables, it is not unusual to have some uneven patterns in data set which will decrease the model accuracy. TANE, an algorithm for finding functional dependencies for large databases, is applied to clean the data. The sensor model output using the cleaned sample data is presented and compared with the results obtained with raw data. The results show that the overall prediction accuracy is improved with the use of clean training and testing data sets.**

***Index Terms*— ANFIS-GRID, Data quality, Inferential control scheme, Soft sensor**

I. Introduction

Soft sensing allows difficult to measure process parameters to be inferred from other easily made measurements (Tham et al., 1991). All soft sensors are based on an inferential modeling module that represents the dynamics between the inputs, or easily measurable variables, and the output, or undetectable variables. Listed below are some commonly used approaches for the development of the inferential modeling module:

- Physical Model
- Neural Network
- Fuzzy Logic
- Adaptive Neuro-Fuzzy Inference System (ANFIS)

Recent research demonstrates the use of ANFIS for the development of soft sensor model for the estimation of average air temperature in a distributed heating system (Jassar et al., 2009). The estimated temperature provides a closed-loop boiler control scheme (see the feedback loop through dashed line in Fig. 1), as in the absence of economic and technically reliable method for measuring the overall comfort level in the built environment, the boilers are normally controlled to maintain the supply water temperature (see the solid feedback loop in Fig. 1).

For ANFIS based soft sensor models, when estimation/prediction accuracy is concerned, it is assumed that both the data used to train the model and the testing data to make estimations are free of errors (Klein and Rosin, 1999). But rarely a data set is clean before extraordinary effort having been put to clean the data. For this problem of average air temperature estimation, with the measurement errors caused by reading and equipment of the input variables to the model, it is not unusual to have some uneven patterns in the data set. This paper is aiming to analyze the impact of data quality of both training and testing data sets on the prediction accuracy of the developed model. The paper is organized as follows. Section 2 will discuss the development of ANFIS based soft sensor model. Impact of data quality on ANFIS performance is analyzed in section 3. Results are presented in Section 4. Finally, conclusion and future scope of the research is given.

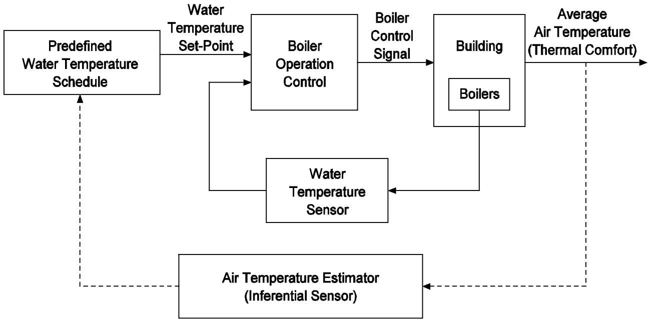

Predefined Water Temperature Schedule
Water Temperature Set-Point
Boiler Operation Control
Boiler Control Signal
Building
Boilers
Average Air Temperature (Thermal Comfort)
Water Temperature Sensor
Air Temperature Estimator (Inferential Sensor)

Fig. 1 Block diagram representation of closed-loop boiler control scheme.

Manuscript received May 22, 2009. The work presented in this paper is partially funded by National Sciences and Engineering Research Council of Canada (NSERC) with research project reference number as: 313375-07, and is carried out at Ryerson University, Toronto, Canada. The support of this organization is gratefully acknowledged.

S. Jassar is with the Department of Electrical and Computer Engineering, Ryerson University, 350 Victoria Street, M5B2K3, Toronto, Canada. (corresponding author's phone: 416-979-5000 ext 4223; e-mail: sjassar@ ryerson.ca).

Z. Liao is the Associate Professor with the department of Architectural Sciences, Ryerson University, 350 Victoria Street, M5B2K3, Toronto, Canada. (e-mail: zliao@ryerson.ca).

L. Zhao is the Associate Professor with the Department of Electrical and Computer Engineering, Ryerson University, 350 Victoria Street, M5B2K3, Toronto, Canada. (e-mail: lzhao@ ee.ryerson.ca).

II. ANFIS Based Soft Sensor Model

New AI techniques, which is known as “Soft Computing”, integrates powerful artificial intelligence methodologies such as neural networks and fuzzy inference systems. While fuzzy logic performs an inference mechanism under cognitive uncertainty, neural networks posses exciting capabilities such as learning, adaption, fault-tolerance, parallelism and generalization. Since Jang (1993) proposed ANFIS, its applications are numerous in various fields, including engineering, management, health, biology and even social sciences.

ANFIS is a multi-layer adaptive network-based fuzzy inference system. An ANFIS consists of a total of five layers to implement different node functions to learn and tune parameters in a fuzzy inference system (FIS) structure using a hybrid learning mode. In the forward pass of learning, with fixed premise parameters, the least squared error estimate approach is employed to update the consequent parameters and to pass the errors to the backward pass. In the backward pass of learning, the consequent parameters are fixed and the gradient descent method is applied to update the premise parameters. Premise and consequent parameters will be identified for membership function (MF) and FIS by repeating the forward and backward passes. ANFIS has been widely used in prediction problems and other areas.

ANFIS based soft sensor model developed in this research infers the average air temperature, T_{avg}, from three easily measurable variables. The three variables are external temperature, T_0, solar radiation, Q_{sol}, and energy consumed by the boilers, Q_{in} (Liao and Dexter, 2004). The FIS structure is generated by Grid partitioning method.

Grid partition divides the data space into rectangular sub-spaces using axis-paralleled partition based on pre-defined number of MFs and their types in each dimension. The wider application of grid partition in FIS generation is blocked by the curse of dimensions. The number of fuzzy rules increases exponentially when the number of input variables increases. For example, if there are averagely m MFs for each input variable and a total of n input variables for the problem, the total number of fuzzy rules is m^n. It is obvious that the wide application of grid partition is threatened by the large number of rules. According to Jang, grid partition is only suitable for cases with small number of input variables (e.g. less than 6). In this research, the average air temperature estimation problem has three input variables. It is reasonable to apply the grid partition to generate FIS structure, ANFIS-GRID. Fig. 2 shows the model structure for ANFIS-GRID.

Gaussian type MFs, as shown in Fig. 3, is used for characterizing the premise variables. Each input has four MFs and the rule-base has 64 rules.

The developed structure is trained using hybrid learning algorithm. The parameters associated with MFs changes through training process. The shape of MFs also changes after training. This concept is clearly visible from the shape of MFs for T_0 in Fig. 3. The shape of MFs for other two variables, Q_{in} and Q_{sol}, is not clearly changed after training process, but associated parameters have changed significantly.

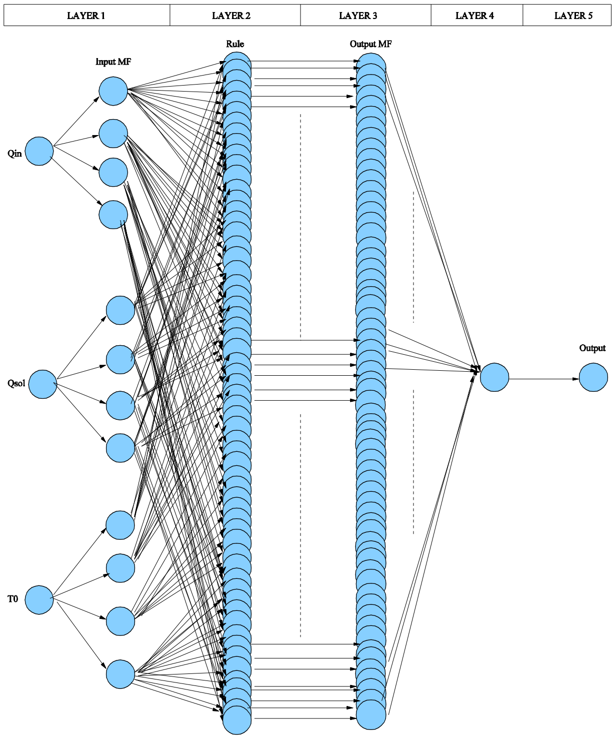
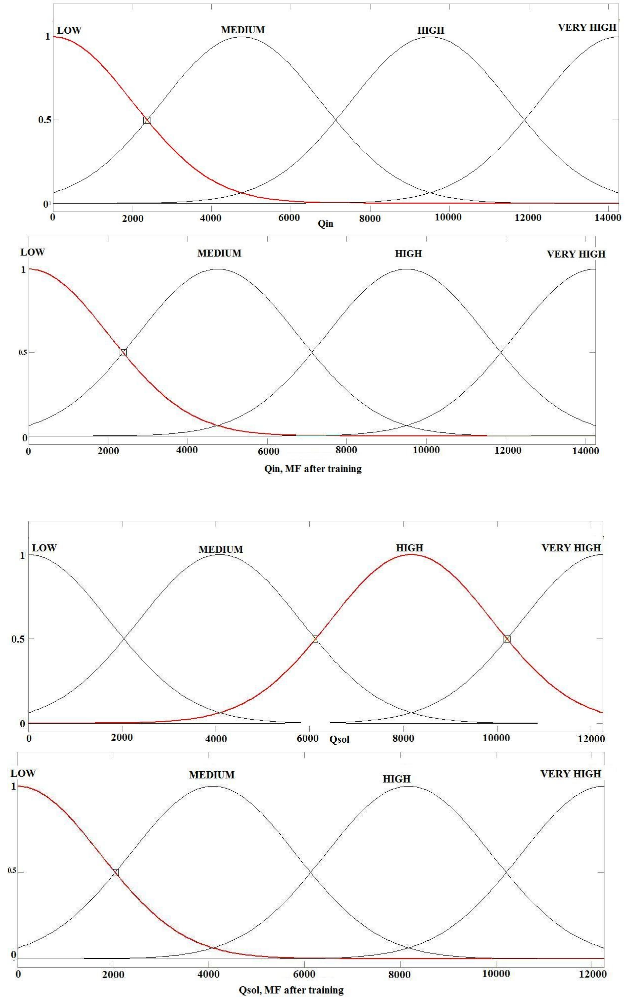


Fig. 2 ANFIS-GRID model structure



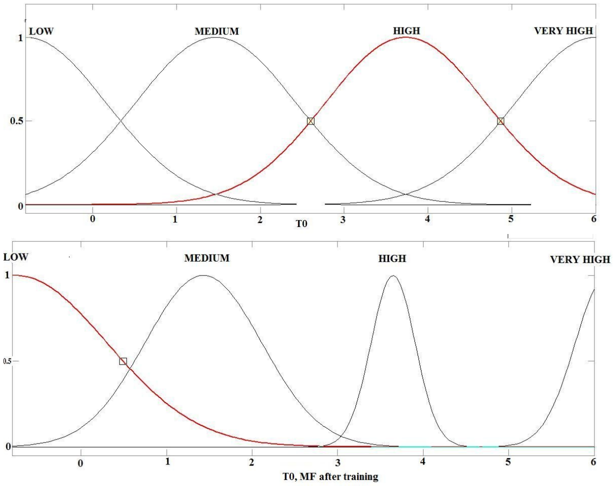


Fig. 3 MFs for Q_{in}, Q_{sol} and T_0 before and after training.

A. *Training and Testing Data*

Experimental data obtained from a laboratory heating system is used for training and testing of the developed model (BRE, 1999-2001). The laboratory heating system is located in Milan, Italy. The details of experimental data collection for the four variables, Q_{in}, Q_{sol}, T_0 and T_{avg}, are given by Jassar et al. (2009). The data set used for the training of ANFIS-GRID has 1800 input-output data pairs and is shown in Fig. 4.

The experimental data used for checking the performance of the developed model is shown in Fig. 5. The testing data set has 7132 data pairs, which is large enough as compared to training data set used for the development of the model.

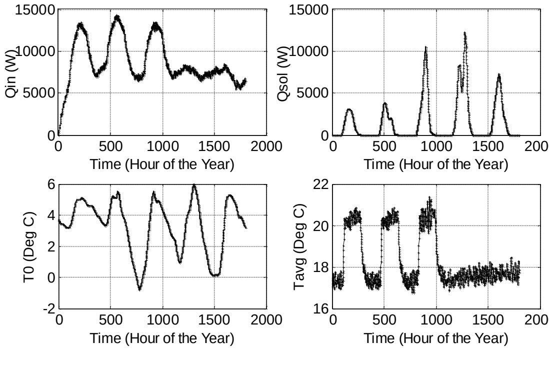


Fig. 4 Training data set (February 2000: day 22 to day 27)

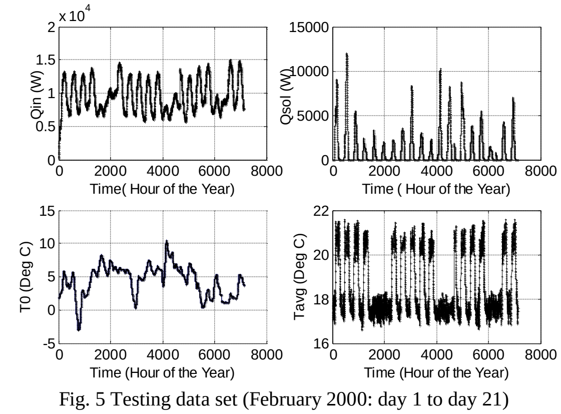


Fig. 5 Testing data set (February 2000: day 1 to day 21)

III. Impact of Data Quality

Data quality is generally recognized as a multidimensional concept. This study is primarily concerned about the data accuracy, defined as conformity between a recorded value and the corresponding actual data value.

Several studies have investigated the effect of data errors on the outputs of computer based models. Bansel et al. (1993) studied the effect of errors in test data on predictions made by neural network and linear regression models. The training data set applied in the research was free of errors. The research concluded that the error size had a statistically significant effect on predictive accuracy of both the linear regression and neural network models. O'Leary (1993) investigated the effect of data errors in the context of a rule-based artificial intelligence system. He presented a general methodology for analyzing the impact of data accuracy on the performance of an artificial intelligence system designed to generate rules from data stored in a database. The methodology can be applied to artificial intelligence systems that analyze the data and generate a set of rules of the form "if X then Y". It is often assumed that a subset of the generated rules is added to the system's rule base on the basis of the measure of the "goodness" of each rule. O'Leary showed that the data errors can affect the subset of rules that are added to the rule base and that inappropriate rules may be retained while useful rules are discarded if data accuracy is ignored.

Wei et al. (2007) analyzed the effect of data quality on the predictive accuracy of ANFIS model. The ANFIS model is developed for predicting the injection profiles in the Daquing Oilfields, China. As the study is using experimentally collected data for training and testing of the ANFIS model, it is not unusual to have some extreme patterns in the data set. The research analyzed the data quality using TANE algorithm. They concluded that the cleaning of data has improved the predictive accuracy of ANFIS model from 78% to 86.1%.

In this research the experimental data collected from a laboratory heating system is used for training and testing of the developed ANFIS-GRID model. The data collected has some uneven patterns. In this section we will discuss the method, TANE algorithm, to identify those conflicting data pairs and replace those rows of data from the data set with the expected ones.

A. *Functional Dependencies*

The raw data is analyzed using approximate functional dependence mining method. An approximate dependency, or an approximate functional dependency, is a functional dependency that is almost valid with the exception of data tuples. A functional dependency studies the relationship of attributes in one or several tables, and claims that the value of an attribute is uniquely determined by the values of some other attributes. The discovery of functional dependencies in databases leads to useful knowledge and data quality problems.

More formally, a functional dependency over a relation is expressed as $X \rightarrow A$, where $X \subseteq R$ and $A \subseteq R$. The dependency is valid in a given relation r if for all pairs of records t, $u \in r$, following statements hold: if $t(B) = u[B]$ for all $B \in X$, then $t(A) = u[A]$. A functional dependency $X \rightarrow A$ is trivial if $A \in X$. The task in functional dependency mining is to find all minimal non-trivial dependencies that hold in r.

Approximate dependencies arise in many databases when there are natural dependencies between attributes, but some records contain errors and inconsistencies. For example, the relationship between zip code and the combination of city and state in a country. Another example is the social security number (SSN) and a corresponding person residing in the USA. Theoretically, these attributes have consistent relationships, as one person associated with one SSN, and one zip code associated with one combination of city, state in a country. But if errors are somehow introduced, the relationships between these attributes will be violated, which leads to the approximate dependencies.

B. *TANE Algorithm*

The TANE algorithm (Huhtala et al., 1999), which deals with discovering functional and approximate dependencies in large data files, is an effective algorithm in practice. The TANE algorithm partitions attributes into equivalence partitions of the set of tuples. By checking if the tuples that agree on the right-hand side agree on the left-hand side, one can determine whether a dependency holds or not. By analyzing the identified approximate dependencies, one can identify potential erroneous data in the relations.

In this research, relationship of the three input parameters (Q_{in}, Q_{sol}, and T_0) and the average air temperature (T_{avg}) is analyzed using TANE algorithm. For equivalence partition, all the four parameters are rounded off to zero decimal points.

After data pre-processing, four approximate dependencies are discovered, as shown in Table 1. Although all these dependencies reflect the relationships among the parameters, the first dependency is the most important one because it shows that the selected input parameters have consistent association relationship with the average air temperature except a few data pairs, which is a very important dependency for average air temperature estimation.

To identify exceptional tuples by analyzing the approximate dependencies, it is required to investigate the equivalence partitions of both left-hand and right-hand sides of an approximate dependency. It is non-trivial work that could lead to the discovery of problematic data. By analyzing the first dependency, conflicting tuples are identified as some of them are given in Table 2. From Table 2, one can see that detected tuples contain conflicting relationships or associations among parameter, and some of them contain severe ones. For example, as the same parameters in tuples 3 and 4, and tuples 7 and 8, the average air temperature values for these cases bear large difference. These data pairs could create trouble for average air temperature estimation. Based on the data trend, pairs 4, 7, 23, 24 are detected as conflicting tuples and are fixed using appropriate methodology. Table 2 shows only a small part of the total data set. The total data set has 8932 data pairs. For the first approximate dependency from Table 1, 42 conflicting data pairs are present which needs to be fixed for better performance of ANFIS-GRID model.

Table 1. Approximate functional dependencies detected using the TANE algorithm

Index	Approximate dependencies	Number of rows with conflicting tuples
1	$Q_{in}, Q_{sol}, T_0 \rightarrow T_{avg}$	42
2	$Q_{in}, T_0, T_{avg} \rightarrow Q_{sol}$	47
3	$Q_{in}, Q_{sol}, T_{avg} \rightarrow T_0$	43
4	$Q_{sol}, T_0, T_{avg} \rightarrow Q_{in}$	54

Table 2. Conflicting tuples identified by analyzing the first approximate dependency in Table 1

Index	Q_{in}	Q_{sol}	T_0	T_{avg}
1	276	0	2	17
2	5168	283	3	17
3	**6415**	**4576**	**3**	**18**
4	**6412**	**4572**	**3**	**21**
5	12030	8579	5	21
6	12601	4306	6	21
7	**11778**	**8896**	**4**	**21**
8	**10501**	**8875**	**4**	**18**
9	12651	3107	6	21
10	12575	0	5	21
11	9448	0	4	18
12	**5296**	**0**	**3**	**22**
13	10595	0	1	18
14	9384	0	0	18
15	8794	0	-1	18
16	7340	0	-1	18
17	7465	0	-2	18
18	6886	0	-3	17
19	7605	0	-3	18
20	7409	115	-2	18
21	11406	5262	1	21
22	6367	0	8	18
23	**7344**	**0**	**-1**	**21**
24	**7456**	**0**	**-2**	**21**
25	8472	1014	7	17

IV. Results

A. *Model Validation*

The developed ANFIS-GRID model is validated using experimental results (BRE, 1999-2001). The model performance is measured using the following statistical indices:

Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(T_{avg}(i) - \hat{T}_{avg}(i)\right)} \qquad (1)$$

R^2, Coefficient of determination, tells us how much of the experimental variability is accounted for by the estimate model.

$$R^2 = \frac{\sum_{i=1}^{N}\left[\hat{T}_{avg}(i) - \overline{T}_{avg}\right]^2}{\sum_{i=1}^{N}\left[T_{avg}(i) - \overline{T}_{avg}\right]^2}, \qquad (2)$$

For equations (1) and (2), N is the total number of data pairs, $\hat{T}_{avg}$ is the estimated and T_{avg} is the experimental value of average air temperature. $\overline{T}_{avg}$ is the average of experimental data.

B. Results

Initially, ANFIS-GRID model uses the raw data for both the training as well as the testing. Fig. 6 compares ANFIS-GRID estimated average air temperature values with the experimental results. ANFIS-GRID estimated average air temperature values are in agreement with the experimental results, with RMSE 0.56^{0}C. However there are some points at which estimation is not following the experimental results. For example, around 1900-2200 and 5100-5200 hour of the year time, there is a significant difference between estimated and experimental results.

For checking the effect of data quality on ANFIS-GRID performance, the training and testing data sets are cleaned using TANE algorithm. The conflicting data pairs are replaced with the required data pairs. Then the cleaned data set is applied for the training and the testing of ANFIS-GRID model. A comparison of the model output with clean data and the experimental results is shown in Fig. 7.

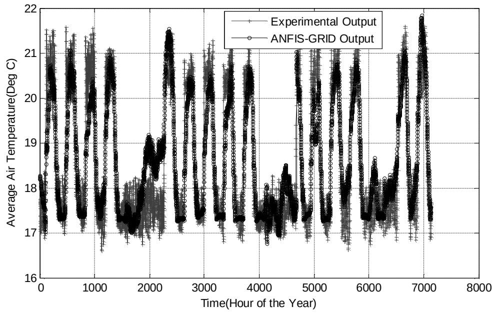


Fig. 6 Comparison of ANFIS-GRID estimated average temperature values with experimental results.

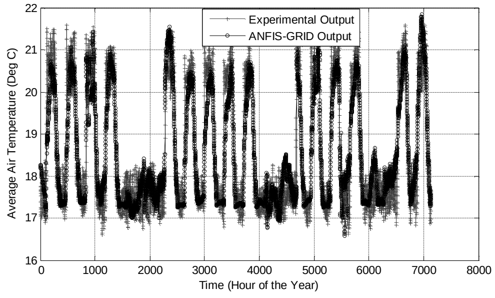


Fig. 7 Comparison of ANFIS-GRID output with clean data and the experimental results.

Table 3. Comparison of results

Model	RMSE (^{0}C)	R^2
ANFIS-GRID using raw data	0.56	0.7831
ANFIS-GRID using cleaned data	0.35	0.8945

Fig. 7 and Table 3 clearly show the effect of data quality on predictive accuracy of ANFIS-GRID model. The RMSE is improved by 37.5% to 0.35^{0}C. RMSE is considered as a measure of predictive accuracy. Less RMSE means less difference between the estimated values and the actual values. Finally, predictive accuracy is improved with decrease in RMSE.

V. Conclusions and Future Scope

ANFIS-GRID based soft sensor model has been developed to estimate the average air temperature in distributed heating systems. This model is simpler than the subtractive clustering based ANFIS model (Jassar et al., 2009) and can be used as the air temperature estimator for inferential control scheme for distributed heating systems (Fig. 1). Grid partition based FIS structure is used as there are only three input variables. The training data set is also large enough as compared to the modifiable parameters of the ANFIS. As the experimental data is used for both the training as well as the testing of the developed model, it is expected that data can have some discrepancies. The discrepancies in the data can be the measurement errors due to reading and equipment. TANE algorithm is used to identify the approximate functional dependencies among the input and the output variables. The most important approximate dependency is analyzed to identify the data pairs with uneven patterns. The identified data pairs are fixed and again the developed model is trained and tested with the cleaned data. Table 3 shows that the RMSE is improved by 37.5%. Therefore, it is highly recommended that the quality of data sets should be analyzed before they are applied in ANFIS based modelling.

Future work can be focused on analyzing the predictive accuracy of ANFIS based soft sensor models by conducting two types of experiments. The experiments will mainly concentrate on the analysis of data quality issues for predictive accuracy of ANFIS based models. In the first experiment, the quality of the training and the testing data sets will be analyzed independently. From the analysis, it can be concluded that if the model performance is more affected by either the training or the testing data set quality. In the second experiment, the effect of the fraction, the percentage of erroneous data items in a data set and amount of errors, the amplitude of error in each pair, can be studied independently.

References

[1] A. Bansal, R. Kauffman, and R. Weitz, “Comparing the modeling performance of regression and neural networks as data quality varies”, *Journal of Management Information Systems*, vol. 10, 1993, pp. 11-32.

[2] BRE, ICITE, “Controller efficiency improvement for commercial and industrial gas and oil fired boilers”, *A craft project*, Contract JOE-CT98-7010, 1999-2001.

[3] Y. Huhtala, J. Karkkainen, P. Porkka, and H. Toivonen, “TANE: an efficient algorithm for discovering functional and approximate dependencies”, *The Computer Journal*, vol. 42, 1999, pp. 100-111.

[4] J.R. Jang, “ANFIS: Adaptive-network-based fuzzy inference system”, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 2,1993, pp. 665-685.

[5] S. Jassar, Z. Liao and L. Zhao, “Adaptive neuro-fuzzy based inferential sensor model for estimating the average air temperature in space heating systems”, *Building and Environment*, vol. 44, 2009, pp. 1609-1616.

[6] B.D. Klein, and D.F. Rossin, “Data errors in neural network and linear regression models: an experimental comparison”, *Data Quality*, vol. 5, 1999, pp 33-43.

[7] Z. Liao and A.L. Dexter, “A simplified physical model for estimating the average air temperature in multi-zone heating systems”, *Building and Environment*, vol. 39, 2004, pp.1013-1022

[8] D. O’Leary, “The impact of data accuracy on system learning”, *Journal of Management Information Systems*, vol. 9, 1993, pp. 83-98.

[9] M.T. Tham, G.A. Montague, A.J. Morris and P.A. Lant, “Estimation and inferential control”, *Journal of Process Control,* vol. 1,1993, pp. 3-14.

[10] M. Wei et al.’ Predicting injection profiles using ANFIS”, *Information Sciences,* vol. 177, 2007, pp. 4445-4461.