CHAPTER - II

METHODOLOGICAL ASPECTS OF ANFIS AND STATISTICAL MODELLING

2.1. INTRODUCTION

The untreated and partially treated industrial effluents are the major sources of water and land pollution (Muller et al., 2007). The Biological Oxygen Demand (BOD) and the Chemical Oxygen Demand (COD) are prime pollution parameters of effluents. The standard five days BOD test and few hours COD test are time consuming. Performing the test for BOD requires significant time and commitment for preparation and analysis. A test is used to measure the amount of oxygen consumed by these organisms during a specified period of time (usually 5 days at 20°C). This is referred to as a BOD₅ measurement (Delzer and Mckenzie 1999). Many computational intelligence models are being increasingly employed for the prediction of BOD and COD such as ANFIS, ANN, etc., These models can be described as mathematical methodologies which explain relations between cause (input data) and effects (output data) irrespective to the process and without the need for making assumptions considering the nature of the relations. Therefore, many investigations have been made to decrease the time required for determining these parameters (Oliveira-Esquerre, 2004, Dogan, 2009, Honggui and Junfei 2009). The conventional statistical model and the new ANFIS modelling (kalyanaraman et al., 2006; Areerachakul 2012) can help in fast estimation of BOD and COD values.

In the present work, ANFIS modelling and statistical modelling are used to predict the BOD and COD values from industrial effluents. ANFIS modelling
is purely an empirical approach. Since the ANFIS modelling incorporates the Fuzzy logic theory and Artificial Neural Network architecture, a brief description of basic concepts of fuzzy logic and neural networks are presented in the following sections.

2.2. FUZZY LOGIC

Fuzzy Logic meanwhile is chosen mainly due to its capability to make decisions in an environment of imprecision, uncertainty and incompleteness of information and partiality of truth. (Zadeh, 2008). Fuzzy logic is currently being tested in real environmental problems (McKone and Deshpande 2005; Ocampo-Duque et al., 2006). Its success is mainly due to its closeness to human perception and reasoning, as well as its intuitive handling and simplicity, which are important factors for handling of imprecise data (Romano et.al., 2004). Due to these advantages, complex water-related environmental problems can be addressed easily with fuzzy logic (Sylaios, et.al 2008; Koutroumanidis et.al.; 2009, Boskidis et.al, 2012). Fuzzy logic methods carry out a mapping that allows the rule to have more variability than the two state exactness of binary logic (Vinod Kumar and Joshi 2005).

Based on the nature of fuzzy human thinking, Lofti Zadeh originated the “fuzzy logic” or “fuzzy set theory”, in 1965. He is considered as father of fuzzy theory, to analyse complex and non-liner systems.

*As complexity rises, precise statements lose meaning and meaningful statements lose precision.*

- Lotfi Zadeh

**Fuzzy logic** is a multi-level logic i.e. it uses a ‘multi-valued set’ where degree of membership of a parameter is represented by a number between 0 and 1. With fuzzy logic the transition between sets is gradual and small changes in input
values results in a more graceful change in the model output. With fuzzy sets, human concepts like small, young, big, warm, hot, etc. can be translated into a form usable by computers.

Fuzzy logic is a convenient way to map an input to space to an output space. Fuzzy logic does a good job of trading off between significant precision. Fuzzy logic (FL) provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. Fuzzy logic starts with the concept of a fuzzy set. A fuzzy set describes the relationship between an uncertain quantity, membership function, which ranges between 0 and 1 (Rezaei, 2012).

Conventional methods are good for simpler and linear problems while fuzzy systems are suitable for complex and non-linear problems.

**Justification for fuzzy theory**

- The real world is too complicated and most real systems are non-linear to precise descriptions to be obtained. Therefore approximation (or fuzziness) must be introduced to obtain a reasonable model.
- Human knowledge becomes increasingly important. We need a theory to formulate human knowledge in a systematic manner and put it into engineering systems, together with mathematical models and sensory measurements. Essentially a fuzzy system transforms human knowledge base into a mathematical formula.

**Observations of fuzzy logic**

- Fuzzy logic is conceptually easy to understand
- It is flexible
- It is tolerant to imprecise data
- It can model non-linear functions of arbitrary complexity
- It can be built on top of the experience of experts
- It can be blended with conventional control techniques
- It is based on natural language.

Fuzzy sets

Everything is vague to a degree you do not realize till you have tried to make it precise.

- Bertrand Russell

Fuzzy logic starts with the concept of a fuzzy set. A fuzzy set is a set without crisp, clearly defined boundary. It can contain elements with only a partial degree of membership. In fuzzy logic, the truth of any statement becomes a matter of degree.

Membership functions

A membership function (MF) is a curve that defines how each point in the input space is mapped to a membership value (or degree of membership) between 0 and 1. The only condition, a membership function must really satisfy, is that it must vary between 0 and 1. The function itself can be an arbitrary curve, whose shape can be defined as a function that suits us from the point of view of simplicity, convenience, speed and efficiency.

The fuzzy logic toolbox built in Matlab version 5.3 packages (User’s Guide, 1998) includes 11 built in membership function types. These are built from several basic functions, linear function, the Gaussian distribution function, the sigmoid curve, quadratic and cubic polynomial curves.
Logical operators

Fuzzy logic is a superset of standard Boolean logic. If one keeps the fuzzy values at their extremes of 1 (completely true) and 0 (completely false), standard logical operations will hold. AND, OR and NOT logical operations are used while framing the rules.

IF-THEN rules

Fuzzy systems are knowledge-based or rule-based systems. The heart of a fuzzy system is knowledge based consisting of IF-THEN rules.

The starting point of constructing a fuzzy system is to obtain a collection of fuzzy IF-THEN rules from human experts or based on domain knowledge (i.e. knowledge obtained by previous experiences). The next step is to combine these rules into a single system.

Fuzzy sets and fuzzy operators are the subjects and verbs of fuzzy logic. The IF-THEN rules are used to formulate the conditional statements of fuzzy logic. A single fuzzy IF-THEN rules is of the form,

\[ \text{IF } x \text{ is } A \text{ THEN } y \text{ is } B \]

Where A and B are linguistic variables defined by fuzzy sets on the ranges of x and y.

The ‘IF x is A’ part of the rule is called ’antecedent’ or ‘premise’ (or simply input).

The “THEN y is B” part is called the ‘consequent’ or ‘conclusion’ (or simply output).
Fuzzy systems

The basic configuration of a fuzzy system is shown in Fig. 2.1

![Fuzzy Rule Base](image1)

**Fig. 2.1. Basic configuration of fuzzy system**

**Fuzzy inference system [FIS]**

Fuzzy inference systems are also known as fuzzy-rule-based systems, fuzzy models, fuzzy associative memories (FAM) or fuzzy controllers when used as controllers. Basically a fuzzy inference system is composed of five fundamental blocks are given in Fig. 2.2.

![Fuzzy Inference System](image2)

**Fig. 2.2. Fuzzy inference system**
**Fuzzifier**

The real world input to the fuzzy system is applied to the fuzzifier. In fuzzy literature, this input is called crisp input since it contains precise information about the specified information about the parameter. The fuzzifier converts this precise quantity to the form of imprecise quantity like ‘large’, ‘medium’, ‘high’, etc. with a degree of belongingness to it. Typically, the value ranges between 0 and 1.

**Knowledge base**

The main part of the fuzzy system is knowledge based in which both rule base and database are jointly referred. The database defines the membership functions of the fuzzy sets used in the fuzzy rules where as the rule base contains a number of fuzzy if-then rules.

**Decision making unit**

The inference system or the decision-making unit performs the inference operations on the rules. It handles the way in which the rules are combined.

**Defuzzifier**

The output generated by the inference block is always fuzzy in nature. A real world system will always require the output of the fuzzy system to the crisp or in the form of real world input. The job of the defuzzifier is to receive the fuzzy input and provided real world output. In operation, it works opposite to the input block.

Fuzzy inference is the process of mapping a given input to an output using fuzzy logic through which decisions can be made. Fuzzy inference involves membership functions, fuzzy logic operators and IF-THEN rules. There are two types of fuzzy inference systems that can be implemented in the fuzzy logic toolbox.
1. Mamdhani type

2. Takagi-sugeno type

1. Mamdhani type FIS

Mamdhani’s fuzzy inference method is the most commonly seen fuzzy methodology. It was proposed by Mamdani 1975 by synthesizing a set of linguistic control rules obtained from experimental human operators.

Mamdhan fuzzy model is based on the collections of IF-THEN rules with both fuzzy antecedent and consequent parameters (Mahapatra et.al., 2011, Tosun et.al., 2011, Keshwani et.al., 2008). It is also called a linguistic model because both the antecedent and the consequent are fuzzy propositions. Mamdhan fuzzy model due to its popularity and easily application is the most commonly seen fuzzy methodology (Saberi Nasr, 2012).

2. Takagi-Sugeno type FIS

Sugeno type system can be used to control an inference system in which the output functions are either linear or constant (Jang, 1993). Fuzzy inference system can be constructed using the graphical user interface (GUI) tools provided in the fuzzy logic toolbox or it can be also generated and worked from command line. This model is built with if-then rules that have fuzzy antecedent and functional consequent (Mahapatra et.al., 2011).

2.3. NEURAL NETWORKS (NN)

Another powerful tool for modelling complex and non-linear systems is ‘Neural Network’ technology which was developed in the 1960’s from the analogies to the properties of biological neurons and it it can be used to solve problems that are not amenable to conventional statistical and mathematical methods (Dogan et.al.,
ANNs are developed through the use of computer software to recognize the patterns of the data by training data through supervised learning. The idea of developing ANNs was inspired from biological nervous system in human brain, with the ability to organize its neurons and learn through ‘experience’. ANNs flexibility learning system and adaptive ability allows them to learn from linear and non-linear function. (Anita Talib and Mawar Idati Amat, 2012).

In recent years, Artificial Neural Network (ANN) methods have become increasingly popular for prediction and forecasting in a number of disciplines, including water resources and environmental science. Recently, ANNs have been increasingly applied in modelling water quality. ANNs have been successfully used in hydrological processes, water resources, water quality prediction, and reservoir operation (Suen et.al., 2003). A Neuron refers to the processing element in the artificial neural network (ANN). A neuron is also known as a ‘node’ or ‘unit’. The structures of ANNs are composing of nodes interconnected with each other which represent relationship between each node. The number of nodes is determined and variables can be adjusted chosen to build neural model. The outcomes of the model are based on the nodes, hidden layer and variables chosen for the model.

The most important feature of Neural Networks is their ability to achieve an accurate non-linear mapping from input - output pairs of data without knowing their functional relationship. Neural Networks have been successfully applied to many areas including non-linear system modelling and control applications.

There are three types of neural network models

1. Models of the neurons themselves
2. Models of synaptic interconnections and structure

3. The training or learning rules for updating the connecting weights are stated by Lin and Lee (1996).

The basic structure of a NN is shown in Figure 2.3. Each neuron in a NN, processes the incoming inputs into an output. The output is then linked to other neurons. Generally a NN has three functional layers. They are,

1. Input Layer
2. Output Layer and
3. Hidden Layer

![Schematic structure of neural network](image)

**Fig. 2.3. Schematic structure of neural network**

The information enters the neural network at the input layer then the hidden layer processes these signals through the neural network until they reach the output layer. The weighted sum of the inputs are transferred to the hidden neurons, where it is transformed using an activation function. The outputs of the hidden neurons, in turn, act as inputs to the output neuron where it undergoes another transformation.
ANNs have been vastly used for the past few years in many area of research, including bioinformatics, image analysis, speech recognition and financial forecasting.

2.4. ADAPTIVE NEURO – FUZZY INFERENCE SYSTEM (ANFIS)

ANFIS derives its name from ‘Adaptive Neuro – Fuzzy Inference System’ (Jang, 1993). ANFIS is so designed technique that can be used in modelling, decision – making and process control applications (Takagi and sugeno, 1985, Kalyanaraman, 2006, Najah et.al., 2010, Folorunsho et.al., 2012). The integration of fuzzy systems and neural networks can combine the merits of both systems and offers a more powerful tool for modelling. A neural fuzzy system is such an integrated system which uses neural network as tool in fuzzy system.

The basic idea behind the ANFIS technique is to provide a method for the fuzzy modelling procedure to learn information from a data set. This makes the FIS to understand the input – output relationship and to compute the number and type of membership functions. This learning process works very similar to ‘neural networks’.

The architecture and learning procedure of ANFIS is a FIS implemented in the framework of adaptive neural networks. ANFIS proposed by Jang (1993) is based on the first – order sugeno fuzzy model that uses either a back propagation algorithm or a hybrid learning algorithm. Although ANFIS is one of the first integrated hybrid Neuro-fuzzy models, surprisingly, it is the best function approximator among the several Neuro-fuzzy models (Abraham and Nath, 2000). Comparing with others, ANFIS has a high speed of training, the most effective learning algorithm and simplicity of the software (Jang and Sun, 1995). Moreover,
ANFIS is faster in convergence and provides better results when applied without any pre-training (Altug et al., 1999).

The parameters associated with the membership functions will change through the learning process. The computation of these parameters is facilitated by a gradient vector, which provides a measure of how well the fuzzy inference system is modelling the input–output relationship for a given set of parameters. Once the gradient vector is obtained, any of several optimization routines could be applied to reduce the error.

The schematic architecture of the ANFIS model is illustrated in Fig. 2.4. It consists of the following key components.

1. Inputs and outputs
2. Database and preprocessor
3. A fuzzy system generator
4. A fuzzy inference system
5. An adaptive NN representing the fuzzy system

The quality of the training database is important for the model to produce correct information about the system. The database should contain adequate and sufficiently large number of data and correct information on the system to enable the model to depict the correct picture about the system. Besides, a more concise training database will significantly reduce the ANFIS training time.
Fig. 2.4. Schematic architecture of the ANFIS model
Methodology

The exclusive feature of ANFIS is that itself - learns and recognizes that exact input - output relationship from the fuzzy inference rules and so acquires enough intelligence in predicting a new situation. The learning process of ANFIS consists of training the FIS to understand the input–output relations using a set of data called ‘Training Data’ (trnData) comprising of input - output parameters of a system under study. Then the validity of the FIS can be tested by using another set of data known as ‘Check Data’ (chkData) which comprises of identical input - output parameters but having different values. ANFIS generates automatically the fuzzy rules and selects the rules with maximum firing strength. Now the FIS is capable of predicting the output of a new environment of inputs. If a third data set called ‘Test Data’ (testData) comprising of input parameters alone is given, the ANFIS will predict the output parameter (User’s Guide, 1998).

The basic idea behind using a checking data set for model validation is that after a certain point in the training, the model begins over fitting the training data set. In principle, the model error for the checking data set tends to decrease as the training takes place up to the point that over fitting begins and then the model error for the checking data suddenly increases.

It is difficult to manually pretreat the data. A fuzzy clustering method is used to automatically carry out this task. Clustering of numerical data forms the basis of many classification and system modelling algorithms. Clustering technique is used to identify natural groupings of data from a large data set and to produce a concise representation of systems behaviour.
The ANFIS is usually started with a prototype fuzzy system, a fuzzy system generator is needed. The software Matlab provides three fuzzy system generators for use i.e. “fuzzy”, “genfis 1” and “genfis2”. Designers, if well versed can directly create a fuzzy system using fuzzy function. The function ‘genfis 1’ examines the data set and then generates a fuzzy system based on the given number and types of membership function. The function ‘genfis 2’ would generate a first – order sugeno fuzzy system based on subtractive clustering of the data set provided.

Data formalities in ANFIS

Training data

To start training an FIS using ANFIS, first it is necessary to have a training data (trnData) set that contains desired input/output data pairs of the target system to be modelled. Each row starts with an input vector and is followed by an output value. The number of rows of trnData is equal to the number of training data pairs. Since there is only one output the number of columns of trnData is equal to the number of inputs plus one. In other words any data set used in ANFIS must be a matrix with the input data arranged as vectors in all but the last column. The output data must be in the last column.

Checking data

The checking data (chkData) is used for testing the generalization capability of the fuzzy inference system at each epoch. The checking data has the same format as that of the training data and its elements are generally distinct from those of the training data.

The use of the minimum checking data error epoch to set the membership function parameters assumes:
• The checking data is similar enough to the training data that the checking data error will decrease as the training begins.

• The checking data increases at some point in the training, after which data over fitting has occurred.

**Training data error**

The training error is the difference between the training data output value, and the output of the fuzzy inference system corresponding to the same training data input value, (the one associated with that training data output value). The training error \( \text{trnError} \) records the root mean square error (RMSE) of the training data set at each epoch. Fismat1 is the snapshot of the FIS structure when the training error measure is at its minimum.

**Checking data error**

The checking error is the difference between the checking data output value, and the output of the fuzzy inference system corresponding to the same checking data input value, (the one associated with that checking data output value). The checking error \( \text{chkError} \) records the RMSE for the checking data at each epoch. Fismat2 is the snapshot of the FIS structure when the checking error is at its minimum.

**RMSE**

The RMSE is calculated using the relation

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_i)^2}
\]
where $x_i$ and $\bar{x}_i$ are observed and simulated results set and $n$ is number of observations.

The performance of the ANFIS model can be observed by evaluating the error on the ‘training data’ ($RMSE_{trn}$) as well as on the ‘Check data’($RMSE_{chk}$). Most usually the best model is one which has the lowest RMSE value.

When both the $RMSE_{trn}$ and $RMSE_{chk}$ are very small we can conclude that.

1. The ANFIS has captured the essential components of the input-output relations.

2. The training data contains adequate information about the system being modelled.

ANFIS modelling can be done either using the GUI tools or worked from the command line in the fuzzy logic toolbox available in MATLAB. In the later case, a suitable programme comprising of trnData and chkData is written and executed. Typical program is provided in the next chapter.

**Modelling application**

The fuzzy modelling or fuzzy identification, first explored by Takagi and Sugen (1985) has numerous practical applications like control (Mingzhi Huang *et.al.*, 2012, Najah *et.al.*, 2010), prediction and inference (Hanife Sari *et.al.*, 2012, Mostafa Rezazadeh shirdar *et.al.*, 2011). The ANFIS can be used model the highly non-liner relationship between inputs and outputs and to predict the output for a new set of inputs.
The various functions used in the programs and their purposes are given in Table 2.1.

Table 2.1. Functions used and their purposes

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Function</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>fismat</td>
<td>The name of an FIS, (fuzzy inference system) used to provide ANFIS with an initial set of membership functions for training</td>
</tr>
<tr>
<td>2.</td>
<td>genfis2</td>
<td>The genfis2 is a fast, one-pass method that does not perform iterative optimization</td>
</tr>
<tr>
<td>3.</td>
<td>evalfis</td>
<td>It perform the fuzzy inference calculations</td>
</tr>
<tr>
<td>4.</td>
<td>trnRMSE</td>
<td>trnRMSE is the root mean square error of the system generated by the training data</td>
</tr>
<tr>
<td>5.</td>
<td>chkRMSE</td>
<td>chkRMSE is the root mean square error of the system generated by the checking data</td>
</tr>
<tr>
<td>6.</td>
<td>plotmf</td>
<td>Plot all of the membership functions associated with a given variable</td>
</tr>
<tr>
<td>7.</td>
<td>fismat2</td>
<td>fismat2 is the FIS structure whose parameters are set according to a minimum checking error criterion</td>
</tr>
<tr>
<td>8.</td>
<td>Showrule</td>
<td>It displays the FIS structure</td>
</tr>
</tbody>
</table>

2.5. STATISTICAL MODELLING

Correlation analysis

Correlation is a measure of the relation between two or more variables. The measurement scales used should be at least interval scales but other correlation coefficients are available to handle other types of data.
**Simple linear correlation**

Correlation analysis is a method used to measure the degree of association between two variables. Correlation coefficient is used to find out the significant relationship between the independent and dependent variables of the sample. The most widely-used type of correlation coefficient is $r$ also called linear or product-moment correlation.

Correlation coefficients can range from “-1 to +1”. The value of “-1” represents a perfect negative correlation, while a value of “+1” represents a perfect positive correlation. A value of “0” represents a lack of correlation.

If “$x$” is an independent parameter and “$y$” is a dependent parameter on $x$ then these two are connected by the simple linear expression of the type $y = A + Bx$. A straight line representing this equation has the intercept ‘$A$’ and ‘$B$’ is its slope. Knowing the values of the regression coefficients $A$ and $B$, one can predict the value of $y$ for a given value of $x$. However this prediction will be accurate only when there exist significant correlation between $x$ and $y$ ($r < 0.5$ approx). In cases, where there is no significant correlation, then multiple regression analysis can be carried out to predict the value of $y$ (Draper and Smith, 1998).

**Regression analysis**

The general purpose of multiple regression, the term was first used by Pearson (1978), is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable.

In regression analysis, the dependent variable is a function of the independent variables and the degree of contribution of each variable to the output is represented by the regression coefficients on these variables.
When several independent variables $x_1$, $x_2$, $x_3$... have influence on $y$, the multiple regression analysis of the type $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...$ is more effective in predicting the dependent variable $y$, where $\beta_0, \beta_1, \beta_2 ...$ are regression coefficients.

The regression analysis provides equations for estimating individual values of one variable from the given values of other variables. Regression analysis might find its greatest application in estimating missing data.

To interpret the direction of the relationship between variables one looks at the signs (+ or -) of the regression of $\beta$ coefficients.

If a $\beta$ coefficient is positive then the relationship of this variable with the dependent variable is positive. If the $\beta$ coefficient is negative then the relationship is negative. If the $\beta$ coefficient is equal to zero then there is no relationship between the variables.

**R² value**

The $R^2$ value is an indicator of how well the model fits the data. The value of $R^2$ is 0.5, it indicates the variability of the $y$-values around the regression line is 1-0.5 times the original variance or 50% of the original variability and left 50% residual variability.

**F-test**

The F-distribution test is a method of determining whether there exists a statistical difference between two sets of data. The F-distribution is most commonly used in Analysis of Variance (ANOVA) and the F test (to determine if two variances
are equal). The F-distribution is the ratio of two chi-square distributions, and hence is right skewed. It has a minimum of 0, but no maximum value (all values are positive).

**P – Value (statistical significance)**

The P value represents the probability of error that is involved in accepting our observed result as valid. The value of P represents a decreasing index of the reliability of a result. The p value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected. The smaller the p-value, the stronger the evidence is in favour of the alternative hypothesis. Typically the results that yields P = 0.05 are considered borderline statistically significant but remember that this level of significance still involves a pretty high probability of error (5%). Results that are significant at the P = 0.01 level are commonly considered statistically. Significant and P = 0.005 or P = 0.001 levels are often called ‘highly’ significant (Draper and Smith, 1981).

The application of ANFIS modelling and statistical modelling in the estimation of BOD and COD values of Engineering works industry and Pharmaceutical industry effluents are described in the next chapter.