MODELLING VISCOSITY BELOW BUBBLE POINT PRESSURE USING GROUP METHOD OF DATA HANDLING (GMDH): A COMPARATIVE STUDY

HARUN BIN AB. RAHMAN

A dissertation submitted to the Graduate Studies Programme in

partial fulfillment as a requirement for the Degree of

BACHELOR OF ENGINEERING (HONS)

PETROLEUM ENGINEERING

UNIVERSITI TEKNOLOGI PETRONAS

BANDAR SERI ISKANDAR

PERAK DARUL RIDZUAN

SEPTEMBER 2013

CERTIFICATION OF DISSERTATION

TITLE OF DISSERTATION

MODELLING VISCOSITY BELOW BUBBLE POINT PRESSURE USING GROUP METHOD OF DATA HANDLING (GMDH): A COMPARATIVE STUDY

I, HARUN BIN AB. RAHMAN ,

hereby declare that the dissertation is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at UTP or other institutions.

_____	_____
Signature of Author	Signature of Supervisor
Harun Bin Ab. Rahman	Dr. Mohammed Abdalla Ayoub Mohammed
Date: _____	Date: _____

ACKNOWLEDGEMENT

All praise to Allah the Almighty the most generous, the most compassionate, the most merciful and peace is upon his prophet Muhammad. Firstly, I would like to express my heartiest gratitude to UniversitiTeknologiPetronas for providing me the opportunity to conduct my very own research project as part of the curriculum structure in my field of engineering course. I owe my sincerest gratitude also to my project supervisor, Dr. Mohammed Abdalla Ayoub Mohammed for his continuous teaching, guidance and words of encouragement that have made this work possible despite my shortcomings. He has provided me with full support, academically and morally throughout my project with his knowledge and patience through timely meetings and discussions. I would like to credit the success of my project to his amazing support and without him; definitely this project would not have been accomplished. One could simply not wish for a better supervisor than he is.

I would also like to extend my appreciation to my colleagues and batch mates for offering me with moral support throughout this period of this project. Special appreciation to my closest colleagues Mr. Haziq and Mr. Dewistiel for their amazing support. They have provided me with helpful insights in developing the GMDH model. Honestly, I could never thank these people enough for their contributions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Below the bubble point pressure, the amount of gas dissolved in the oil increases as the pressure is increased. This causes the in-situ oil viscosity to decrease significantly. Knowledge of viscosity below bubble point is essential to many areas in the petroleum industry including reservoir and fluid production and recovery, and upgrading and transporting produced fluids. However, prediction of this parameter is difficult below bubble point pressure as the liquid undergoes a significant change in composition. These crude oils exhibit regional trends in chemical composition that categorize them as paraffinic, naphthenic, or aromatic. Because of the differences in composition, correlations developed from regional samples that are predominantly of one chemical base may not provide satisfactory results when applied to crude oils from other regions. Although some correlations show modest tolerance to assist prediction in other regions, getting accurate results with acceptable value of errors remains questionable.

The application of GMDH is not only restricted in reservoir engineering. It is critical in many areas which include accounting and auditing, finance, marketing, organizational behaviour, economics, military systems and medicine. They have several advantages compared with conventional neural networks. It has the ability to automatically organize multilayered neural networks by using the heuristic self organization method. In the GMDH-type neural networks, many types of neurons, which are polynomial type, sigmoid function type, and radial basis function type can be used to organize neural network architectures and optimum neuron architectures are selected so as to fit the complexity of the nonlinear system. The recent advancement in Soft Computing (SC) called Group Method of Data Handling (GMDH) type of Neural Networks will be able to provide a more intelligent platform for predicting viscosity below bubble point pressure with an outstanding correlation coefficient.

This paper seeks to develop a new viscosity correlation below bubble point pressure using data points taken from international oil fields. The correlation will be mapped against other existing correlations from the literature using trend analysis to verify its performance. A theoretical justification of the developed correlation will be presented. The correlation is expected to be valid for all types of crude oils within the range of data used in the study.

A series of statistical and graphical analysis relative to existing correlations will be initiated once the correlation has been formulated to provide a numerical insight on its accuracy. The comparison will validate the reliability and relevance of the proposed model to predict the viscosity below bubble point pressure.

CHAPTER 1: INTRODUCTION

1.0 BACKGROUND

Viscosity is the measure of the resistance to flow exerted by a fluid. Oil viscosity is a strong function of temperature, pressure and the amount of dissolved gas in oil. It varies depending on its origin type and nature of its chemical composition, particularly the polar components. Normally, viscosity values are obtained at reservoir temperature, which is a constant value. When laboratory data is unavailable, empirical correlations are sought for temperature values other than reservoir, although its accuracy and reliability vary based on the range of data at hand. Sampling and viscosity measurement methods are the main reasons for the inaccessibility of viscosity at other value of temperature. Therefore, intermolecular interactions can take place and cause a gradation of viscosity among light, heavy, and extra heavy crude oils and bitumen. Ultimately, developing a comprehensive model of viscosity that incorporates data from different regions becomes a very challenging task.

Below the bubble point pressure, the amount of gas dissolved in the oil increases as the pressure is increased. This causes the in-situ oil viscosity to decrease significantly. Above the bubble point pressure, oil viscosity increases minimally with increasing pressure as shown in Figure 1. As reservoir pressure drops below bubble point, a significant change in the composition will occur. The dissolved gas will evolve from the crude and steal the smaller molecules in it, leaving the remaining crude with larger, more complex molecules. The change in composition will cause an increase in the viscosity until dead oil viscosity is reached at atmospheric pressure.

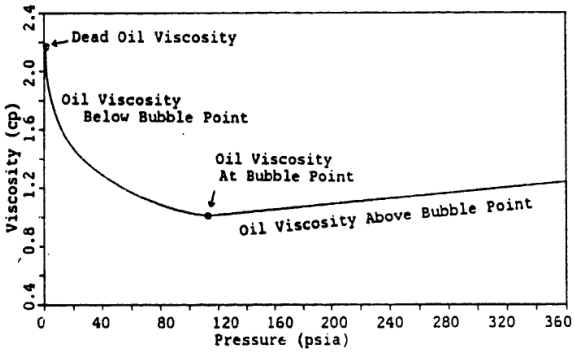


Figure 1: Viscosity versus Pressure plot for crude oils

Generally, there are four viscosity correlations available:

1. Empirical methods
2. Corresponding states methods
3. EOS-based viscosity models
4. Group contribution methods

Numerous viscosity correlations have been proposed over the years. Nonetheless, most of them incorporated data from a specific region which failed to gain popularity for viscosity prediction in other regions due to the complexity of crude oil composition. Generally, oil viscosity correlation method can be classified either a black oil or compositional method. Black oil correlations are formulated from on hand measured variables by fitting of an empirical equation. Relevant variables may include temperature, oil API gravity, solution gas oil ratio, bubble point pressure, specific gas gravity, and dead oil gravity. Examples are Beal (1946), Beggs & Robinson (1975), Chew & Connally (1959) and Khan et al (1987). The compositional method is empirical and/or semi empirical correlations and is derived mostly from the principle of corresponding states and extensions. It mostly uses some parameters excluding those employed in the black oil type such as reservoir fluid composition, pour point temperature, molar mass, normal boiling point, critical temperature and acentric factor of components. Lohrenz (1964) and Little & Kennedy (1968) are among the researchers who opted for this method.

In general, there are three main disadvantages posed by the numerous available viscosity correlations. These are the factors which contribute to the notion that there has yet to be standard method for estimating viscosity in the oil industry. These disadvantages are concluded based on intensive comparison from the literature:

1. Accuracy of a correlation strongly depends on the range and region the data were taken from. Most correlations are developed at low pressure range due to unavailability of data in the literature.
2. As the viscosity of liquid phase is estimated by using different correlations, a smooth transition in the near critical region cannot be achieved.
3. Some correlations have to estimate a parameter such as density, formation volume factor or solution gas oil ratio which will later be used in their viscosity correlations. Hence, separate correlation to obtain that particular parameter is required.

GMDH is used in such fields as data mining, knowledge discovery, prediction, complex systems modeling, optimization and pattern recognition. GMDH algorithms are characterized by inductive procedure that performs sorting-out of gradually complicated polynomial models and selecting the best solution by means of the so-called external criterion. In order to find the best solution GMDH algorithm considers various component subsets of the base function called partial models. Coefficients of these models are estimated by the least squares method. GMDH algorithm gradually increase the number of partial model components and find a model structure with optimal complexity indicated by the minimum value of an external criterion. This process is called self-organization of models.

An artificial neural network is a biologically inspired computational model that consists of processing units and connections between them with coefficients bound to the connections, which constitute the neuronal structure, as well as of training and recall algorithms attached to the structure. Neural networks can be trained by a set of examples of data and therefore represent extensive parallelism, robustness and approximate reasoning, which are important in dealing with ambiguous data. This is exactly why the application of Group Method of Data Handling (GMDH) is considered useful for the development of this correlation (refer Figure 2).

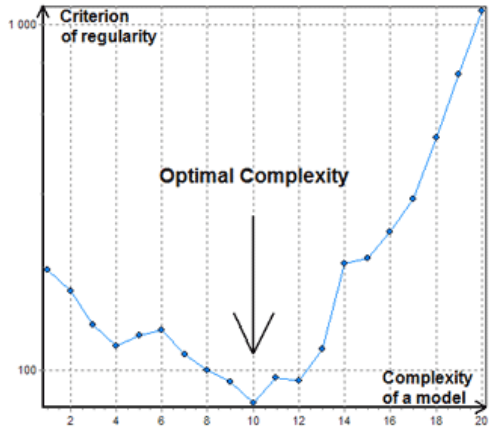


Figure 2: A typical distribution of minimal values of criterion of regularity for Combinatorial GMDH models with different complexity

The main objective of this research is to propose a simple model to predict oil viscosity at reservoir conditions as a function of a number of easily determined physical properties using Group Method of Data Handling (GMDH). This research will avoid costly experimental testing and reduce uncertainty and errors in viscosity determination for various viscosity related oil and gas industry engineering applications.

1.1 PROBLEM STATEMENT

1.1.1 PROBLEM IDENTIFICATION

In many cases, the only information available is taken from oil samples. Both statistical and graphical techniques have PVT analysis of an oil sample are simple and readily been employed to evaluate these equations as compared to measurable parameters such as gas relative density, to other published crude oil viscosity correlations oil API gravity, and gas-oil ratio. Direct viscosity measurements or complete compositional analyses of crude oils are expensive. Therefore, empirical viscosity correlations, which are the functions of these readily measurable PVT properties, are used to estimate oil viscosity.

The earliest work in viscosity estimation dates back to 1866. Since then, numerous formulations have been developed by researchers all over the world as seen in the literature. However, these correlations are geographically dependant and governed strongly by the selection of the range of data used. For example, Labedi (1982) developed correlations for dead oil viscosity, viscosity at bubble point pressure and under-saturated oil viscosity using crude oil data from Libya. Hence, his correlations are only accurate for viscosity prediction for oil fields in Libya. Application in other regions will return low error tolerance.

Human brains process intelligent information by logical reasoning and common sense while artificial neural networks have the ability to learn from data and adapt to the environment. The advancement of technology such as Soft Computing is not available when many of these correlations were being formulated. Therefore, their works are subjected to human error.

1.1.2 SIGNIFICANCE OF THE PROJECT

Researchers have yet to develop correlations for viscosity prediction using data points from international oil fields. Currently, there is no standard correlation used in the industry to determine viscosity below bubble point. This is because they are evidently geographically dependent which may not be representative to oil fields in a specific region. By utilizing GMDH method to generate this model, a viscosity below bubble point pressure with high degree of confidence and precision can be formulated. Consequently, many complex petroleum engineering problems can be solved successfully.

1.2 OBJECTIVES

1. To develop a correlation using GMDH type of neural networks for estimating viscosity below bubble point pressure.
2. To validate the newly developed correlations testing using a set of experimental data from oil fields at reservoir conditions through trend analysis.
3. To compare the newly developed correlations with other existing correlations in the industry by using statistical error analysis.
4. To exploit GMDH method to produce accurate correlation as a function of small number of easily determined variables.

1.3 VIABILITY OF THE PROJECT

The project will weighted more on research project which will eliminate the need to design a prototype. The key to successful execution of this project lies in the ability in collecting data, studying the software and analyzing its reliability relative to other correlations. It will consume most of the time given in executing the project. Apart from that, less concern will be on the cost and budget allocation for the project as most of the data and resources (software) are provided by the project supervisor.

CHAPTER 2: LITERATURE REVIEW

2.0 OVERVIEW

According to Oloso et al (2009), the knowledge of oil viscosity is crucial in petroleum engineering computations for simulating reservoirs, evaluating reserves, forecasting production, designing production facilities and transportation systems. Soft Computing (SC) techniques were introduced to improvise on the time consuming and costly laboratory experiments and empirical derivation of correlations. Alomair et al. (2012) stated that there are two key types of correlations available for the prediction of oil viscosity. The first type utilizes the readily accessible oil field data, such as reservoir temperature, API gravity, solution gas oil ratio, saturation pressure and reservoir pressure. The second type is empirical and/or semi empirical correlations which use some parameters other than those used in the first type such as reservoir fluid composition, pour point temperature, molar mass, normal boiling point, critical temperature and acentric factor of components.

2.1 THE EFFECT OF DATA ABUNDANCY

Many correlations for the purpose of estimating oil viscosity of different kinds of oil have been developed but only few of them are specifically developed to predict viscosity below bubble point pressure. Beal (1946) designed four graphical correlations for predicting viscosity at different reservoir pressures, in which only one of them was geared up to estimate viscosity below bubble point pressure. For below bubble point correlation, he included 351 data sets from 29 oil fields in the United States. Similar to Beal (1946), Beggs & Robinson (1975) and Chew & Connally (1959) were able to determine the viscosity at and below the bubble point pressure as a function of dead oil viscosity (μ_{od}) and solution gas oil ratio (GOR) at a given pressure. However, the latter two correlations utilized a larger number of data sets as compared to Beal (1946). Beggs & Robinson (1975) and Chew & Connally (1959), using 2073 (from 600 oil fields) and 457 (from US, Canada and South America) data sets respectively, discovered that the relation between log ($\mu_{b)}$ and log ($\mu_{od)}$ is a straight line between at a fixed solution GOR. For verification, Sutton & Farshad (1984) tested the correlations of Beggs & Robinson (1975) and Chew & Connally (1959) using data of 285 data points from the Gulf of Mexico and concluded Beggs and Robinson (1975)'s correlation yielded the

most accurate results for oil viscosity below bubble point pressure with average relative error of only 1.83% and standard deviation of 27.25%. Sutton & Farshad (1984) further concluded that Chew and Connally (1959)'s correlation tends to over predict oil viscosity while Beggs & Robinson (1975)'s correlation showed an opposite trend. Their validation works proved that the more data points used to formulate a correlation, the higher its accuracy. This discovery will be the foundation for this research.

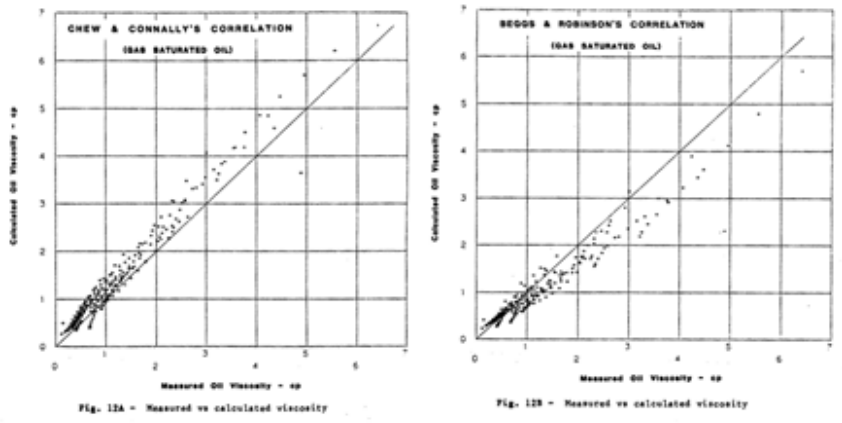


Figure 3: tton & Farshad (1989) stated that Beggs & Robinson's correlation is more accurate than Chew & Connally's

2.2 THE EFFECT OF GEOGRAPHICAL FACTOR

Khan et al (1987) presented a comparison of their correlations against other available correlations using 75 data sets obtained from 62 Saudi Arabian fields. For viscosity above and below bubble point pressure, the correlating variables were solution gas oil ratio, gas relative density and oil API gravity. A total of 1691 data sets were used to develop his correlation for predicting viscosity below bubble point pressure. His model scored an average absolute percent error of 5.157% and a standard deviation of 7.201%. Khan et al (1987) employed non linear multiple least square and least absolute regression analyses to develop his correlation. The model was validated by comparing it to the correlations from Beggs & Robinson (1975) and Chew & Connally (1959). The comparison yielded an expected result where data points from Beggs and Robinson's correlation are nearly similar to the Khan et al (1987)'s newly developed model but exhibit some more scattering. For Chew and Connally's correlation, most of the data points fall above the perfect correlation lines an overall effect of over-estimation. Although Beggs & Robinson (1975) failed to mention the location where his oil samples

were taken, it is highly likely that the differing accuracy between the three correlations is contributed by geographical factor.

Table 1: Khan et al (1987)'s correlation is more accurate than Beggs & Robinson (1975) and Chew & Connally (1959)'s respective works by validation using data points from Saudi Arabian fields

	Beggs& Robinson	Chew &Connally	Khan (This study)
Average Percent Relative Error	-4.262	-41.236	-0.991
Average Absolute Percent Relative Error	25.526	46.882	5.157
Standard Deviation	37.411	70.218	7.201
Correlation Coefficient	0.771	0.474	0.994

2.3 THE EFFECT OF TREND ANALYSIS

Kahn (1987) also adopted the same non linear multiple least square and least absolute regression analyses method while also using the exact same 75 data sets obtained from 62 Saudi Arabian fields as Khan et al (1987). Firstly, he studied the relationship between the independent and dependent variables. Secondly, a model was chosen which best correlated against the experimental values. The least square and least absolute regression coefficients thus obtained were fixed one after the other to the nearest rounded or fraction values to formulate the final correlations. Eventually, he tested two models to estimate the viscosity value below bubble point but ultimately discovered that the combination of the two models was proven to be most representative. Previous research by Khan et al (1987) and Kahn (1987) both used non linear multiple regression analysis to achieve their correlations. Kahn managed to find and validate his viscosity below bubble point model through trend analysis and a series of statistical and graphical analyses conducted.

$$\mu_b = \mu_{ob}\left(P/P_b\right)^a$$

$$\mu_b = \mu_{ob} e^{a(P-P_b)}$$

$$\mu_b = \mu_{ob}\left(P/P_b\right)^{a_1} e^{a_2}(P - P_b)$$

The above formulas were the early three models developed by Khan (1987). Kahn (1987)'s combined the first 2 models to arrive to their final correlation which indicated a model adequacy of 99%.

2.4 THE EFFECT OF DATA RANGE

A correlation to predict the viscosity of light crude oils below bubble point pressure by Isehunwa et al (2006) used data from 400 reservoirs from the Niger Delta and yielded average absolute relative percentage error of 3.25% and standard deviation of 0.97. As opposed to Khan et al (1987) and Kahn (1987), Isehunwa et al (2006) applied linear partial correlation coefficient technique to establish simple correlations between viscosity (μ), pressure (P), temperature (T), oil specific gravity (SG) and solution GOR. This model is unique because it does not require the knowledge of dead oil viscosity as generally agreed by other correlations. Isehunwa et al (2006) proved that their model was more accurate than Khan et al (1987) to determine viscosity below bubble point for Niger Delta fields. However, like most correlations for predicting viscosity out there, his correlation suffers from the limitation of the range of data used. Outside the range specified, the correlation is prone to error.

Table 1: The ranges of the data used in this study

PVT Property	Range
Bubble Point Oil Viscosity	0.03 – 9.1 cp
Oil Viscosity above Bubble Point	0.08 – 43.00 cp
Oil Viscosity below Bubble Point	0.0241 – 35.52 cp
Bubble Point Pressure	300.3 – 6593 psia
Solution Gas Oil Ratio	42.9 – 19149 scf/STB
Relative Oil Density	0.8 – 0.94
Temperature	124 – 289 °F
Pressure	299 – 9407 psia

Figure 4: Isehunwa et al (2006)'s correlation is only valid within the range of data used.

2.5 THE LIMITATIONS OF COMPOSITIONAL MODEL

In the early years, many empirical equations describing the effects of composition, temperature and pressure on viscosity were developed. Little & Kennedy (1968) developed new equations for predicting the viscosity of complex hydrocarbon systems with an average absolute deviation of 9.9% using 1006 data points. Lohrenz et al (1964) designed a calculations procedure to determine viscosity in compositional material balance computations at for reservoir liquids and gases and was validated through comparison of 260 different reservoir oils which produced average absolute deviation of 16%. However, the applicability of a majority of these compositional correlations is limited to very low pressures and to a small number of systems. Most of them, when applied to complex hydrocarbon systems, are of little value. Little & Kennedy (1968) claimed that the lack of utility of the majority of equations results from the fact that they were developed to show the separate effect of temperature, pressure or composition on viscosity, but not to predict the viscosity as a function of all three variables. Lohrenz et al (1964) mentioned that the oil-gas viscosity ratio is always used as a multiplier with the relative permeability ratio in compositional material balance calculations. Since the relative permeability ratio is subject to large uncertainties, the accuracy requirement of the viscosity predictions is not severe where even average deviations of ± 25 per cent is considered acceptable. On top of being expensive and time consuming, viscosity prediction using composition poses many limitations which are why it is not recommendable for this research.

2.6 NEURAL NETWORK

Soft computer techniques have become more popular among researchers due to its ability to recognize non linear relationships within the available data. Hajizared (2007) stated that neural network for reservoir engineering involves initial data which is split into 3 sections which are training, validation and test. Training data are presented to the network during training and the network is adjusted according to its error. The more data you use to train your intelligent systems, the better result you get in the performance of your system. In order to make the existing model more reliable and precise the prediction model is built again, training the model with more data points. Training automatically stops when generalization stops improving. Validation data will measure network generalization and to halt training when generalization stops improving. Test data have no effect on training and so provide an independent measure

of network performance during and after training. His work confirms that more data should be used during the training phase to yield better output. Nonetheless, it must be noted that feeding the network with too many data points during training phase will create spontaneous memorization by the system and will in turn impinge on the accuracy of the output generated.

2.7 GROUP METHOD OF DATA HANDLING

Ayoub et al (2007) constructed an ANN model for Pakistani crude oil below bubble point pressure as a function of P, reservoir temperature, bubble point pressure (P_b), oil formation volume factor (Z), solution GOR, gas specific gravity and API gravity. Viscosity from 99 sets of differential liberation data covering a wide range of P, T and oil density were used for validation purpose. Their work dealt with gradient based optimization procedure. This process repeats layer by layer, until each node in the network has received an error signal that describes its relative contribution to the total error. The process is then continued in an iterative, parallel manner, where the variables with least contribution to the final output will be removed. With correlation coefficient of 99.3%, the regression analysis based ANN model which used two hidden layers neural network, outperforms correlations by Khan et al (1987) and Labedi (1982). In confirming the correlation, Ayoub et al (2007) used both graphical (crossplots) and statistical error analysis (average percent relative error, average absolute percent relative error, minimum and maximum absolute percent relative error, root mean square error, standard deviation and correlation coefficient). However, there is only one limitation in their study which is few number of data points being used. Moreover, Ayoub et al (2007) found that correlation coefficients will be less if limited number of data and low range of variables is applied. Ayoub et al (2007)'s work authenticates the choice of using GMDH for this research while consolidating the statement that more data points should be used.

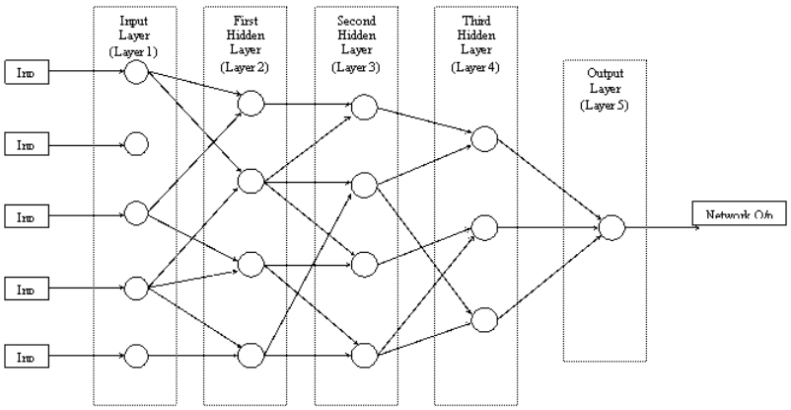


Figure 5: Fully connected network with three hidden layers and out layer

The earlier researches are proven to still be prevalent and applicable in determining the value of viscosity below bubble point pressure. However, there are still rooms of improvement on the studies taking into account the more advanced technology i.e. Soft Computing (SC) which is able to distinguish pattern of errors present in data samples.

CHAPTER 3: METHODOLOGY

3.0 FLOW PROCESS

The methodology of the research is explained in the following flow chart. This methodology explains the flow of the research for the whole project duration (FYPI&FYII). In other words, this methodology will be the guideline, to ensure the research to be executed in a manageable approach in term of time, cost, and feasibility of the research itself.

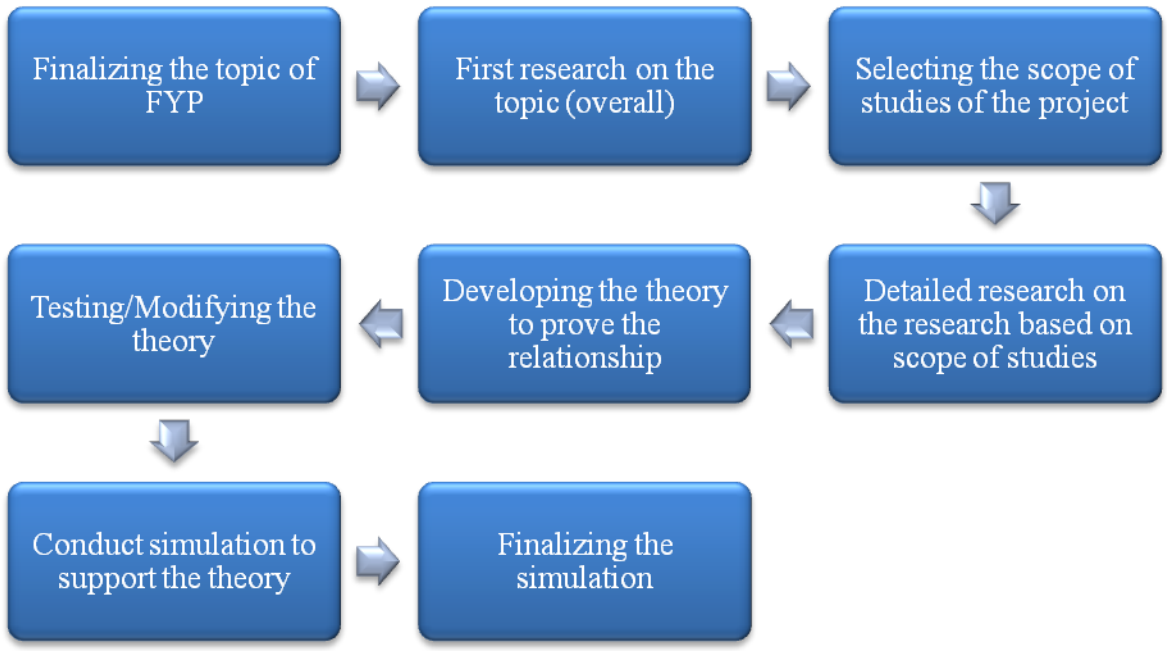


Figure 6: Project Flow Chart

3.1 OVERVIEW

The time given to complete the research is approximately 8 months and several steps as demonstrated in the research methodology below.

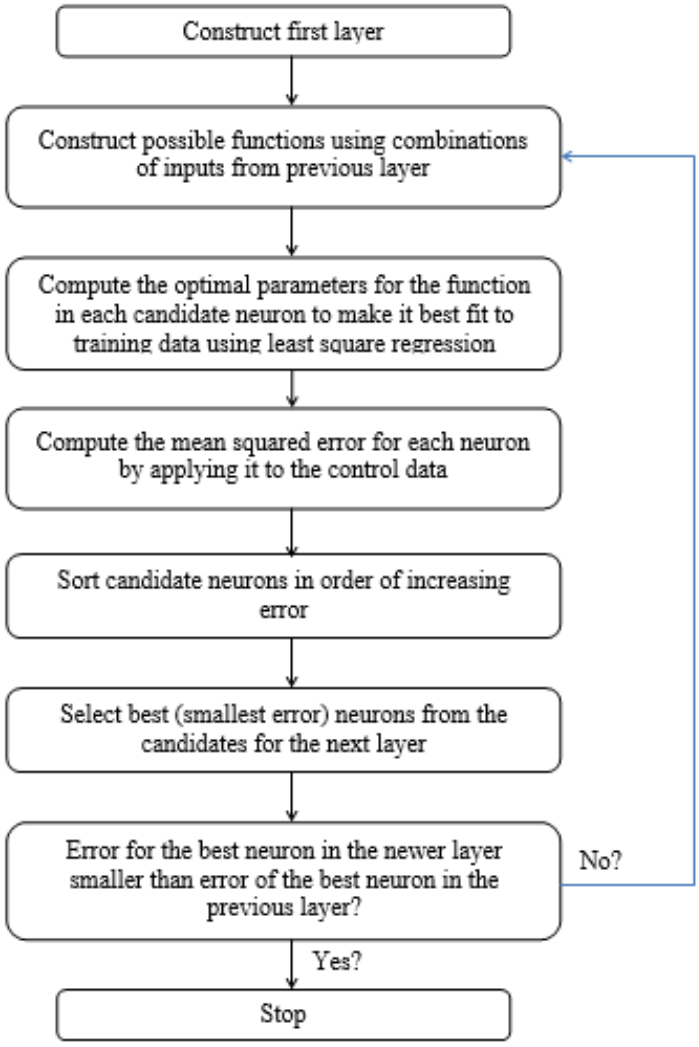


Figure 7: Research Methodology

In the early research development phase, the activities are mainly focusing in background research. Background study has been carried out using extensive materials found on SPE website and other sources. These articles are summarized and the linkage between the data gathered is noted. From the literature, many related information which include definition of viscosity below bubble point and its importance to oil and gas industry, the advantages and drawbacks of existing viscosity correlations and the physical parameters utilized for modeling viscosity. The application of Group Method Data Handling type neural network and how this method can improve the correlation have also been identified through this study.

By linking the different existing correlations available with regard to the topic at hand, the literature review provides an outlook on the gaps to be filled by the current research. It is established that some correlations are more accurate and industry-friendly than the others. The performance of the newly formulated correlations will be compared against the best available models used in the industry.

3.2 DATA COLLECTION

Data points from known oil fields are crucial to this research. Since the data acquired will be used to train, test and validate the correlation, it is pertinent that it fulfills the requirements as per listed below:

I. The data variables are direct function of viscosity
II. The data are abundant and extensive
III. The data are generated from a number of petroleum producing geographical regions
IV. The data are of respectable range
V. The sets of data are consistent with one another
VI. The data are obtained or readily converted at standard conditions

Examples of data variables include temperature, pressure, oil specific gravity, gas specific gravity, gas liquid ratio, water specific gravity, and oil flow rate and oil density.

3.3 FUNDAMENTALS OF GMDH

According to Madala & Ivakhnenko (1994), the Group Method of Data Handling is a combinatorial multi-layer algorithm in which a network of layers and nodes is generated using a number of inputs from the data stream being evaluated. The Group Method of Data Handling (GMDH) was first proposed by Alexy G. Ivakhnenko. The GMDH network topology has been traditionally determined using a layer by layer iterative process based on a pre-selected criterion of what constitutes the best nodes at each level. The traditional GMDH method is based on an underlying assumption that the data can be modeled by using an approximation of the Volterra Series or Kolmorgorov-Gabor polynomial as shown in equation.

$$y = a_0 + \sum_{i=1}^{m} a_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij} x_i x_j + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} a_{ijk} x_i x_j x_k \ldots \quad \text{(A.01)}$$

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & \ldots & x_{1M} \\ x_{21} & x_{22} & \ldots & \ldots & x_{2M} \\ \ldots & \ldots & \ldots & x_{ij} & x_{iM} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{N1} & x_{N2} & \ldots & \ldots & x_{NM} \end{bmatrix} y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ \ldots \\ y_{N1} \end{bmatrix}$$

Where,

$X = (x_1 . x_2 \ldots)$

$Y = the\ corresponding\ output\ value$ (A.02)

$a = vector\ of\ weights\ and\ coefficients$

For simplicity, equation A.01 may be replaced by a system of partial polynomial as displayed in equation A.03.

$$y = G\left(x_i, x_j\right)$$

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i^2 + a_4 x_j^2 + a_5 x_i x_j \quad \text{(A.03)}$$

Where,

$i, j = 1, 2, \ldots, M;\ i \neq j$

3.3.1 BASIC TECHNIQUE

The inductive algorithm follows several systematic steps to finally model the inherent relationship between the input and output target. Data sample of N observations and M independent variables directly related to the system is required. In inductive learning algorithms, it is important to efficiently partition the data into parts because the efficiency of the selection criteria depends largely on this step. The data is split into training set (A) and checking set (B), where ($N = N_A + N_B$).

All the independent variables as presented by the matrix of X in Equation A.02 are taken in pairs of two at a time to produce possible combinations in order to generate a new regression polynomial similar to Equation A.03, where *p* and *q* represent the columns of the X matrix.

$$yi = a_{pq} + b_{pq}x_{ip} + c_{pq}x_{iq} + d_{pq}x_{ip}^2 + e_{pq}x_{ip}^2 + f_{pq}x_{ip}x_{ip} \quad \text{(A.04)}$$

Where,
$p = 1, 2, \dots, M \; p \neq q$
$q = 1, 2, \dots, M \; p < q$
$I = 1, 2, \dots, N$

A set of coefficients of the regression will be calculated for all partial functions by a parameter estimation technique using training data set A and equation A.04.The new matrix C stores the new regression coefficients.

$$C = a_{pq} + b_{pq} + c_{pq} + d_{pq} + e_{pq} + f_{pq}$$

Where,
$p = 1, 2, \dots, M \; p \neq q$ (A.05)
$q = 1, 2, \dots, M \; p < q$
$I = 1, 2, \dots, N$

The number of combinations of pairs for input parameters can be generated by:

$$no.of\ combinations = \frac{M(M-1)}{2} \quad \text{(A.06)}$$

The polynomial at every N data points will be evaluated to calculate a new estimate called z_{pq} as:

$$z_{i,pq} = a_{pq} + b_{pq}x_{ip} + c_{pq}x_{iq} + d_{pq}x_{ip}^2 + e_{pq}x_{ip}^2 + f_{pq}x_{ip}x_{iq} \text{(A.07)}$$

The process continues in an iterative manner until all the pairs are evaluated accordingly in order to generate a new regression pairs that will be stored in a new

matrix called Z matrix. The new generation of regression pairs represents new improved variables with better predictability than the original set of data X as shown by equation A.08.

$$Z = \{z_{ij}\}\begin{cases} i = 1, 2, \dots, N \\ j = 1, 2, \dots, {}^{M(M-1)}/_{2} \end{cases} \quad \text{(A.08)}$$

$$Z = \begin{matrix} z_{11} & z_{12} & \dots & \dots & z_{1,M(M-1)/2} \\ z_{21} & z_{23} & \dots & \dots & \dots \\ \dots & \dots & \dots & z_{ij} & \dots \\ \dots & \dots & \dots & \dots & \dots \\ z_{N1} & z_{N2} & \dots & \dots & Z_{N,M(M-1)/2} \end{matrix} \quad \text{(A.09)}$$

Quality measures of these functions will be computed according to the objective rule chosen using the testing data set B. This can be done through comparing each column of the new generated matrix Z with the dependent variable y. The external criterion somewhere be called regularity criterion (root mean squared values). It is defined as:

$$r_j^2 = \sum_{i-1}^{nt} \frac{(y_i - z_{ij})^2}{(y_i^2)}, j = 1, 2, \dots, {}^{M(M-1)}/_{2} \quad \text{(A.10)}$$

The steps are repeated until the regularity criterion is no longer smaller than the previous layer. The model of the data can be computed by tracing back the path of the polynomials that corresponds to the lowest mean squared error in each layer.

The best measured function will be chosen as an optimal model. If the final result is not satisfied, F number of partial functions will be chosen which are better than all (this is called “freedom of choice”) and do further analysis.

3.3.2 TYPES OF ABDUCTIVE NETWORKS

Various algorithms differ in how they sift partial functions. They are grouped into two types: single-layer and multi-layer algorithms. Combinatorial is the main single-layer algorithm. Multi-layer algorithm is the layered feed forward algorithm. Harmonic algorithm uses harmonics with non multiple frequencies and at each level the output

errors are fed forward to the next level. Other algorithms like multilevel algorithm are comprised of objective system analysis and two-level, multiplicative-additive, and multilayer algorithms with error propagations (Madala & Ivakhnenko, 1994).

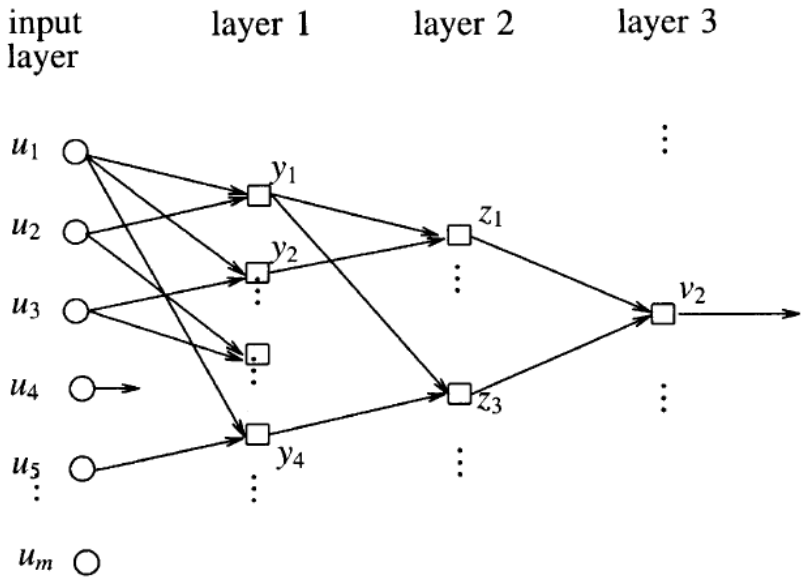


Figure 8: Schematic diagram of a multilayered structure

3.3.3 SELECTION CRITERIA OR OBJECTIVE FUNCTIONS

According to Madala & Ivakhnenko (1994), self-organization modeling embraces both the problems of parameter estimation and the selection of model structure. One type of algorithm generates models of different complexities, estimates their coefficients and selects a model of optimal complexity. The global minimum of the selection criterion, reached by inducting all the feasible models, is a measure of model accuracy. If the global minimum is not satisfied, then the model has not been found. This happens in the following cases:

- The data are too noisy
- There are no essential variables among them
- The selection criterion is not suitable for the given task of investigation
- Time delays are not sufficiently taken into account.

In these cases, it is necessary to extend the domain of sifting until we obtain a minimum.

3.3.4 POLYNOMIAL NEURAL NETWORK

3.3.4.1 Layer Unit

Each layer contains a group of units that are interconnected to the units in the next layer. The weights at each unit are estimated by minimizing the error E. The measure of an objective function is used as the threshold value to make the unit "on" ir "off" in comparison with the testing data N_B which is another part of N and, at the same time, it is considered to obtain the optimum output response. This is used as threshold as well as objectivity measures simultaneously. The outputs of the units in the next layer, that means the output of the K^{th} unit, if it is in the domain of local threshold measure, would become input to some other units in the next level. The process continues layer after layer. The estimated weights of the connected units are memorized in the local memory (Madala & Ivakhnenko, 1994).

3.3.4.2 Multilayer Algorithm

Multilayer network is a parallel bounded structure that is built up based on the type of connection approach given in the basic iterative algorithm with linearized input variables and information in the network flow forward only. Each layer has a number of simulated units depending upon the number of input variables. Two input variables are passed on through each unit.

If there are M input variables, the first layer generates $M_1 = (c_M^2)$ functions. $F1\ (\leq M1)$units are as per the threshold values are made "on" to the next layer. Outputs of these functions become inputs to the second layer and the same procedure is repeated in the second layer. It is further repeated in successive layers until a global minimum on the error criterion is achieved (Madala & Ivakhnenko, 1994).

3.3.5 MATHEMATICAL DESCRIPTION OF THE SYSTEM

The mathematical description of a system can be considered as a nonlinear function in its arguments which may include higher ordered terms and delayed values of the input variables:

$$y = f(x_1, x_2, \cdots, x_1^2, x_2^2, \cdots, x_1x_2, x_1x_3, \cdots, x_{1(-1)}, \cdots, x_{1(-2)}, \cdots),$$

Where, $f(\,)$ is a function of higher degree and y is its estimated output. This can be re-notated as a linearized function by calculating all arguments of x in the following form of full description.

$$\begin{aligned} y &= f(u_1, u_2, \cdots, u_m) \\ &= a_0 + a_1u_1 + a_2u_2 + \cdots + a_mu_m, \end{aligned}$$

Where $u_i, i = 1, 2, \ldots, m$ are the re-notated terms of x, $a_k, k = 0, 1, \ldots, m$ are the coefficients and m is total number of arguments. These m input variables become inputs to the first layer. The partial functions generated at this layer are:

$$\begin{aligned} y_1 &= \nu_{01}^{(1)} + \nu_{11}^{(1)}u_1 + \nu_{21}^{(1)}u_2, \\ y_2 &= \nu_{01}^{(2)} + \nu_{11}^{(2)}u_1 + \nu_{21}^{(2)}u_3, \\ &\cdots \\ y_{M_1} &= \nu_{01}^{(M_1)} + \nu_{11}^{(M_1)}u_{m-1} + \nu_{21}^{(M_1)}u_m, \end{aligned}$$

Where $M_1 (= C_M^2)$ is the number of partial functions generated as the first layer, y_j and $v_{i1}^j, j = 1, 2, \ldots, M_1; i = 0, 1, 2$ are the estimated outputs and corresponding weights of the functions. Let us assume that F_1 functions are selected for the second layer and that there are $M_2 (= C_{F1}^2)$ partial functions generated at the second layer.

$$\begin{aligned} z_1 &= \nu_{02}^{(1)} + \nu_{12}^{(1)}y_1 + \nu_{22}^{(1)}y_2, \\ z_2 &= \nu_{02}^{(2)} + \nu_{12}^{(2)}y_1 + \nu_{22}^{(2)}y_3, \\ &\cdots \\ z_{M_2} &= \nu_{02}^{(M_2)} + \nu_{12}^{(M_2)}y_{F_1-1} + \nu_{22}^{(M_2)}y_{F_1}, \end{aligned}$$

Where $z_j$and $v_{i1}^{j}, j = 1, 2, \dots, M_2, i = 0, 1, 2$ are the estimated outputs and corresponding coefficients of the functions. In the same way, assume that F_2 functions are passed on to the third layer, this means that there are $M_3\left(= C_{F_2}^{2}\right)$partial functions generated in this layer.

$$
\begin{aligned}
v_1 &= \nu_{03}^{(1)} + \nu_{13}^{(1)} z_1 + \nu_{23}^{(1)} z_2, \\
v_2 &= \nu_{03}^{(2)} + \nu_{13}^{(2)} z_1 + \nu_{23}^{(2)} z_3, \\
&\cdots \\
v_{M_3} &= \nu_{03}^{(M_3)} + \nu_{13}^{(M_3)} z_{F_2-1} + \nu_{23}^{(M_3)} z_{F_2},
\end{aligned}
\qquad \text{(A.15)}
$$

Where $v_j$and $v_{i3}^{(j)}$, $j = 1, 2, \dots, M_3$, $i = 0, 1, 2$ are the estimated outputs and corresponding weights of the functions. The process is repeated by imposing threshold levels of $M \geq F1 \geq F2 \geq F3 \geq \dots \geq F1$so that finally a distinctive function is selected at one of the layers. The multilayer network structure with five input arguments and five selected nodes is shown in figure below.

Finally, to get the optimal function in terms of the input arguments, the final model can be traced back as:

$$
\begin{aligned}
v_2 &= f(z_1, z_3) \\
&\equiv f(\, f(y_1, y_2), f(y_1, y_4)\,) \\
&\equiv f(u_1, u_2, u_3, u_5) = f(X).
\end{aligned}
$$

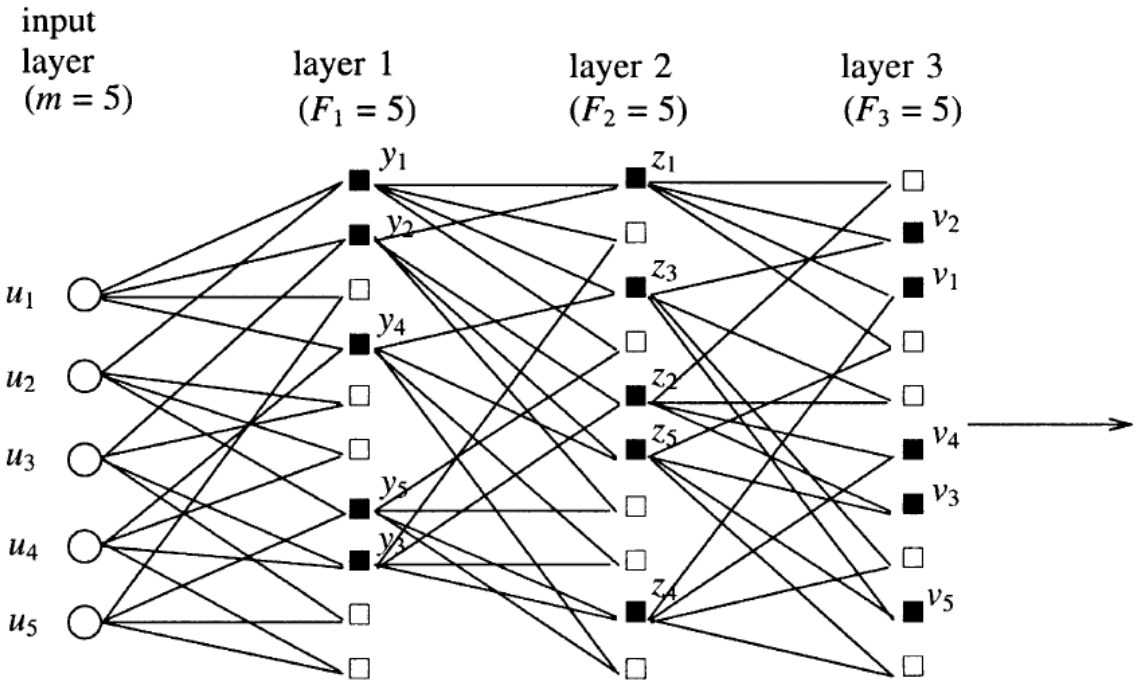


Figure 9: Multilayer Network Structure with 5 Input Arguments and Selected Nodes, reprinted with permission by (Madala & Ivakhnenko, 1994)

3.4 EXPECTED RESULTS

Below the bubble point pressure, viscosity increases due to the liberation of the solution gas which predominate the viscosity decrease due to volumetric expansion of oil. The viscosity will continue to rise until the crude reaches dead oil viscosity at the atmospheric pressure. The table below shows the comparison between various studies in the literature for viscosity below bubble point pressure condition, regardless of the method used. There is a large inconsistency in terms of the type of parameters used and the number of data applied to formulate their correlations, respectively. It is noted that some correlations are more accurate than the others in the literature, in which the list will be further trimmed down to five correlations to be used extensively in the trend analysis phase.

Table 2: Summary of other investigated polynomial GMDH models

Author	Year of Publication	Parameters	Origin	No. of Datapoints	AAPE (%)	STD (%)
Chew &Connally	1959	Dead Oil Viscosity Solution Gas Oil Ratio	USA, Canada & South America	2257	N/A	N/A
Beggs& Robinson	1975	Dead Oil Viscosity Solution Gas Oil Ratio	USA, Canada & South America	2073	1.83	27.25
Labedi	1982	Reservoir Pressure Bubble Point Pressure Bubble Point Viscosity	Nigeria & Angola	31	2.30	24.23
Khan	1987	Reservoir Pressure Bubble Point Pressure Bubble Point Viscosity	Saudi Arabia	1841	0.94	7.20
Petrosky	1990	Dead Oil Viscosity Solution Gas Oil Ratio	Gulf Mexico	864	3.12	19.66
Elsharkawy	1999	Dead Oil Viscosity Solution Gas Oil Ratio	Middle East	254	2.80	25.70
Isehunwa	2006	Bubble Point Viscosity Reservoir Pressure Bubble Point Pressure	Niger Delta	400	3.25	4.45
Ayoub	2007	Reservoir Pressure Reservoir Temperature Oil Density	Pakistan	99	3.23	4.45

Based on the variables used by the literature and its known relationships to viscosity, it is predicted that the finalized variables at layer K may be between reservoir pressure, reservoir temperature, Solution Gas Oil Ration (GOR) and dead oil viscosity. These parameters are acknowledged to have direct relationship with viscosity in general, and are likely to be prominent in shaping the model. The relationships between each of the output parameters with viscosity will be presented through crossplot analysis to provide a better understanding.

A number of input parameters will be fed into the multilayer system, where the comparisons of the combinations of pairs of data take place in an iterative manner. The variable which returns the lowest error value will advance to the next layer for further iterations. It is expected that at layer K, only 2 or 3 direct variables will remain which are regarded as the defining variables in this study.

The newly developed GMDH model is expected to outperform all the correlations listed except for Ayoub et al (2007)'s correlation which employed the Artificial Neural Network (ANN) to design his correlation. According to Ayoub et al (2007), GMDH ranks second to only ANN model in terms of correlation coefficient, AAPE and standard deviation. Although ANN Model possesses greater degree of complexity, the resulting correlation is more accurate.

3.5 TREND ANALYSIS

A trend analysis or synonymous analysis is performed for each generated model to check whether it was justified or not. It helps in understanding the relationship between input and output and increase robustness of the model. It serves as a dominant factor in assessing model building and quality assurance. For this purpose, a different, independent and exclusive set of data is used. This aids in confirming the comparability of the results.

3.6 GRAPHICAL & STATISTICAL ERROR ANALYSIS

To mathematically evaluate the performance of the techniques used and measure the accuracy of the viscosity prediction correlations, graphical and statistical error analysis will be adopted.

3.6.1 GRAPHICAL ANALYSIS

3.6.1.1 CROSSPLOT

It is done by plotting the estimated values against the observedvalues. A 45' straight line is drawn for this purpose. The closer the plotted data points are to this line, the better the correlation.

3.6.1.2 ERROR DISTRIBUTION

Error distribution trend analysis show the error distribution around the zero line to assure that the models have an error trend or not.

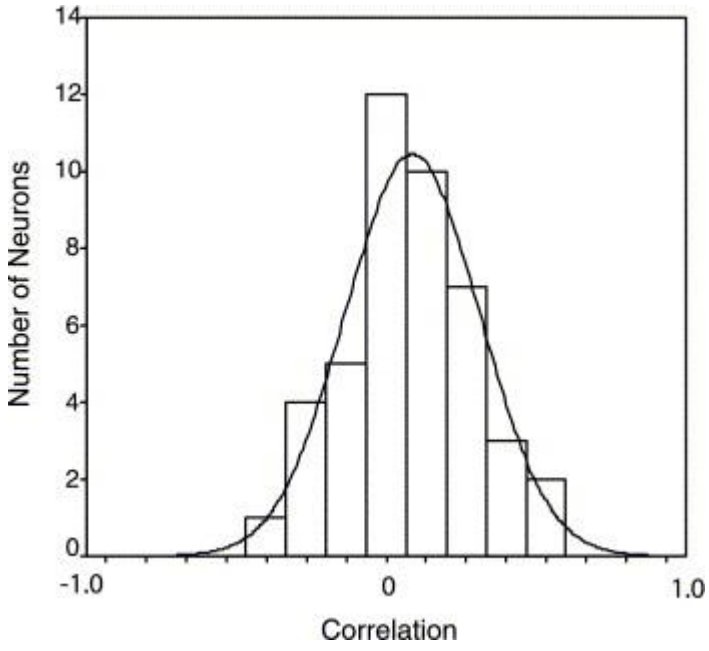


Figure 10: (Sample) Graph of a Histogram Type Error Distribution

3.6.2 STATISTICAL ANALYSIS

3.6.2.1 CALCULATING THE AVERAGE PERCENT RELATIVE ERROR (APE)

It is the measure of relative deviation from the experimental data, defined by:

$$E_r = \frac{1}{n}\sum_{i=1}^{N}\left|E_i\right|$$

Where, E_1 is the relative deviation of an estimated value from an experimental value:

$$E_a = \frac{1}{n}\sum_{i=1}^{N}|E_i|$$

$$E_i = \left[\frac{(\mu)_{\exp} - (\mu)_{est}}{(\mu)_{\exp}}\right]\times 100, \quad \mathrm{i} = 1,2,3,\ldots,\mathrm{n}$$

3.6.2.2 CALCULATING THE AVERAGE ABSOLUTE PERCENT RELATIVE ERROR (AAPE)

This will be the key criterion in statistical error analysis throughout the research. It measures the relative absolute deviation from the experimental values, defined by:

$$E_a = \frac{1}{n}\sum_{i=1}^{n}\left|E_i\right|$$

3.6.2.3 CALCULATING MIN. ABSOLUTE PERCENT RELATIVE ERROR

$$E_{\min} = \min_{i+1}^{n}\left|E_i\right|$$

3.6.2.4 CALCULATING MAX. ABSOLUTE PERCENT RELATIVE ERROR

$$E_{\max} = \max_{i+1}^{n}\left|E_i\right|$$

3.6.2.5 CALCULATING THE ROOT MEAN SQUARE ERROR (RMSE)

It is the measure of data dispersion around zero deviation, defined by:

$$RMSE = \left[\frac{1}{n} \sum_{i=1}^{n} E_i^{\,2} \right]^{0.5}$$

3.6.2.6 CALCULATING THE STANDARD DEVIATION

It is a measure of data dispersion and is expressed as:

$$S = \sqrt{\left[\left(1 \Big/ (m-n-1) \right) \right] \sum_{i=1}^{m} \left[\left\{ \left(\mu_{\exp} - \mu_{est} \right) \Big/ \mu_{\exp} \right\} 100 \right]^2}$$

Where $m-n-1$ is the degree of freedom in multiple regressions analysis. A lower value of standard deviation indicated a smaller degree of scattering.

3.6.2.7 CALCULATING THE CORRELATION COEFFICIENT

It is the degree of success in reducing the standard deviation by regression analysis, defined by:

$$R = \sqrt{1 - \sum_{I=1}^{n} \left[(\mu)_{\exp} - (\mu)_{est} \right]_i^n \Big/ \sum_{I=1}^{n} (\mu)_{\exp} - \overline{\Delta\mu}}$$

Where

$$\overline{\Delta\mu} = \sum_{I=1}^{n} \left[(\Delta\mu)_{\exp} \right]_i$$

"R" values range between 0 and 1. The closer value to 1 represents perfect correlation whereas 0 indicates no correlation at all among the independent variables.

3.7 TOOLS AND EQUIPMENT

The software to be used is the latest software by MATLAB, Inc which is **MATLAB version 7.1.**

3.8 SUMMARY

Firstly, this chapter includes the flow process chart that will be followed closely throughout the duration of the research, where the next crucial step will be collecting ample data sets from available resources and developing MATLAB codes.

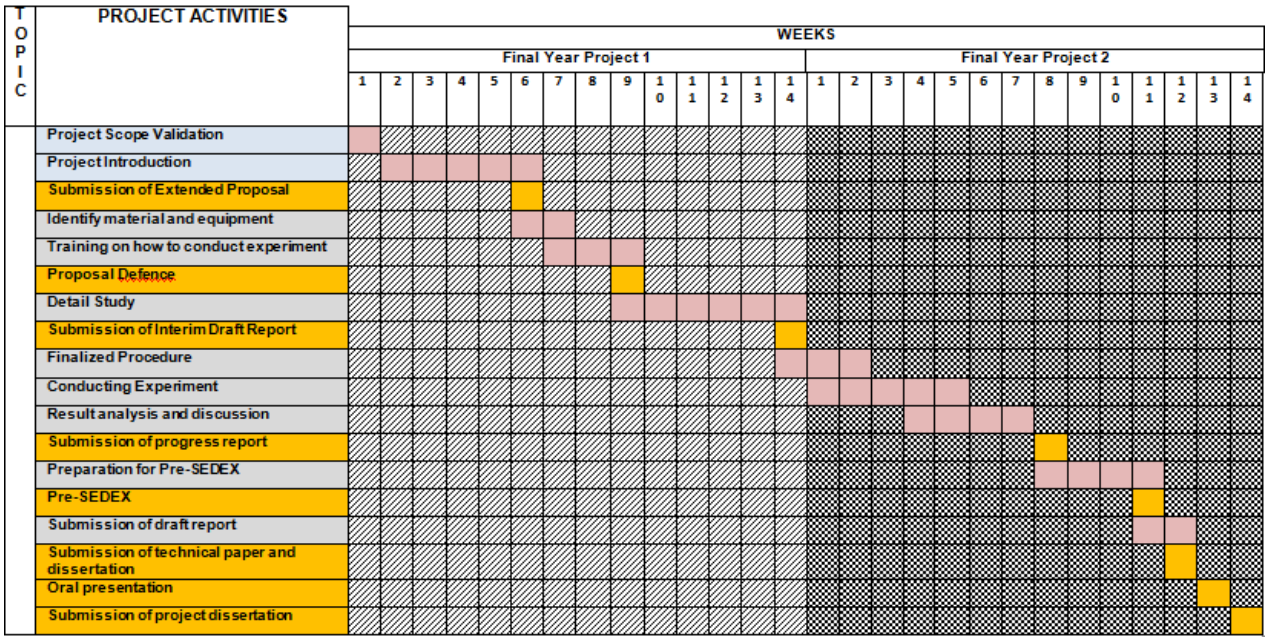
Secondly, the fundamentals and characteristics of GMDH is discussed in this chapter. The proposed algorithm to be used is based on the general form of multilayer structure using Kolmogorov-Gabor polynomial. Inherent relationship between the input and output is calculated through several systematic steps, where the vector of weights and coefficients, a will be estimated accordingly. The data is split into two sets (training set and checking set). The process is repeated in an iterative manner using the independent data sets, X which has been rearranged into combinations of pairs and results in a new matrix, Z. The new matrix, Z is then compared with dependent data sets, y. At this step, regularity criterion (root mean squared values) is generated. The whole procedure is repeated until the regularity criterion breaks its decreasing streak.

In multilayer network of polynomial neural network, each layer is assigned a number of simulated units depending upon the number of input variables where two input variables are passed on through each unit. The output variables that fits a certain threshold values advance to the next layer and used as the input for the next layer. The process is repeated until a global minimum error criterion is achieved.

The relationship between the final output variables is examined through trend analysis where multiple graphs are plotted to compare the contribution of each variable to viscosity. Subsequently, graphical (crossplots and error distribution) and statistical analysis (Standard Deviation, AAPE, Min AAPE, Max AAPE, RMSE and correlation coefficient) is conducted to compare the new model with other existing models. This is especially important to rate the relevance of the new model relative to other popular models.

3.9 GANTT CHART

Project Key Milestones are highlighted in yellow.



CHAPTER 4: RESULTS & DISCUSSION

4.0 SUMMARY OF MODEL'S EQUATION

During the earlier stage of building the polynomial GMDH model, at least 10 different models were tested into the system in order to generate a ratio with the least magnitude of Average Percent Relative Error (APRE), $E_{r.}$

The analysis of the viscosity correlations were carried out on 195 oil samples from Mediterranean Basin, Africa, Persian Gulf and North Sea. The model consists of four layers as follows:

Layer #1 (Number of Neurons: 1)

X_{12}=-0.652767010767115 +1.06009342265199*X_7 +0.000663995099134165*X_5 +5.33323212157755E-09*X_3 -0.00024178827366657 7*X_5*X_7 +0.00016308329221 5454*X_3*X_7 - 1.64291980061468E-07*X_3*X_5 +0.0008471733768505 3*X_7*X_7 +6.04184318853624E-08*X_5*X_5 +1.65781153502158E-08*X_3*X_3

Layer #2 (Number of Neurons: 2)

X_{23}= -0.286914841948905 +1.48736462186417*X_{12} +0.186805403793799*X_9 - 1.05427584253201*X_7 +0.0089012661919065 4*X_9*X_{12} -0.057050145193426*X_7*X_{12} - 0.0226014331265954*X_7*X_9 +0.0104209305331836*X_{12}*X_{12} +0.00108874666165677*X_9*X_9 +0.0757142049291589*X_7*X_7

X_{24} = = 0.192815618078532 +1.3313542815966*X_{12} -0.476007761527864*X_7 - 0.00019875893222538*X_5 -0.010177187634852*X_7*X_{12} -0.000222083378327133*X_5*X_{12} +0.0003249039326361*X_5*X_7 +0.0028842984237607 1*X_{12}*X_{12} +0.00825764645613237*X_7*X_7 +1.75993210060825E-08*X_5*X_5

Layer #3 (Number of Neurons: 1)

X_{34}= 0.227930453472679 +1.50161096690017*X_{24} -0.583847148646646*X_{23} +0.0538054207396279*X_9 +1.70514990554592*X_{23}*X_{24} -0.0182453114967788*X_9*X_{24} +0.0191662756981936*X_9*X_{23} -0.846492143004566*X_{24}*X_{24} - 0.858930797736622*X_{23}*X_{23} -0.00034897903434167 9*X_9*X_9

Layer #4 (Number of Neurons: 1)

$y = 2.25752963264635 + 1.45649748275306*X_{34} - 0.640573852389375*X_7 - 0.111156291009249*X_1 - 0.00895711849166917*X_7*X_{34} - 0.0271446484729675*X_1*X_{34} + 0.0354881821458055*X_1*X_7 + 0.00232409013584509*X_{34}*X_{34} + 0.00793527395260086*X_7*X_7 + 0.00129397022983533*X_1*X_1$

Where,

X_1 = API Gravity

X_3= Reservoir Pressure, psia

X_5= Bubble Point Pressure, psia

X_7 = Bubble Point Viscosity, cp

X_9 = Dead Oil Viscosity, cp

4.1 TREND ANALYSIS FOR THE GMDH MODEL

A trend analysis was conducted for every model's run to check for the physical accuracy of the model. Depending on the final parameters involved in the estimation of viscosity below bubble point pressure that was obtained which are reservoir pressure, dead oil viscosity, bubble point viscosity, bubble point pressure and API Gravity. The relationship of these parameters with viscosity below bubble point was investigated to certify their physical compatibility with each other.

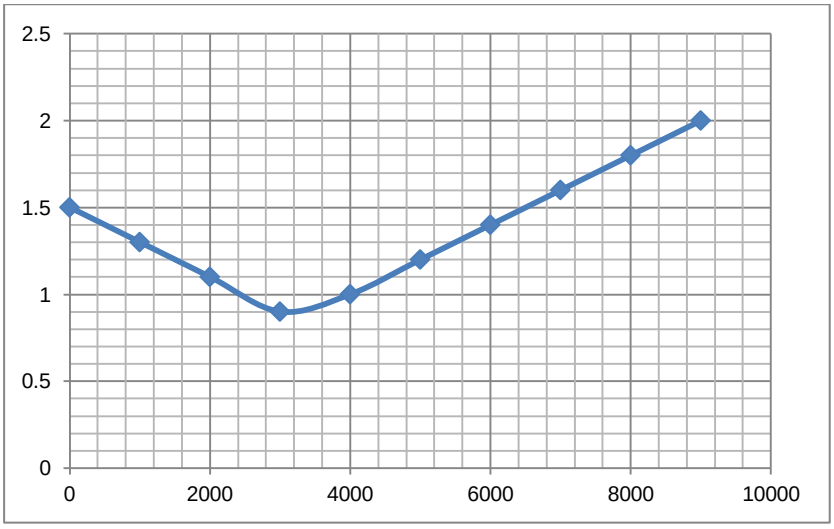


Figure 11: Effect of Reservoir Pressure on Viscosity below Bubble Point

The relationship between the reservoir pressure and viscosity below bubble point was also plotted as shown in Figure 11. As predicted, the proposed GMDH model was able to accurately determine the correct phenomenon for viscosity curve when plotted against pressure.

Figure 12 shows the effect of Oil API on the viscosity below bubble point. This can be achieved by plotting all range of API values against viscosity below bubble point. The model was able to generate the expected trend where pressure viscosity is known to have an inverse relationship with API.

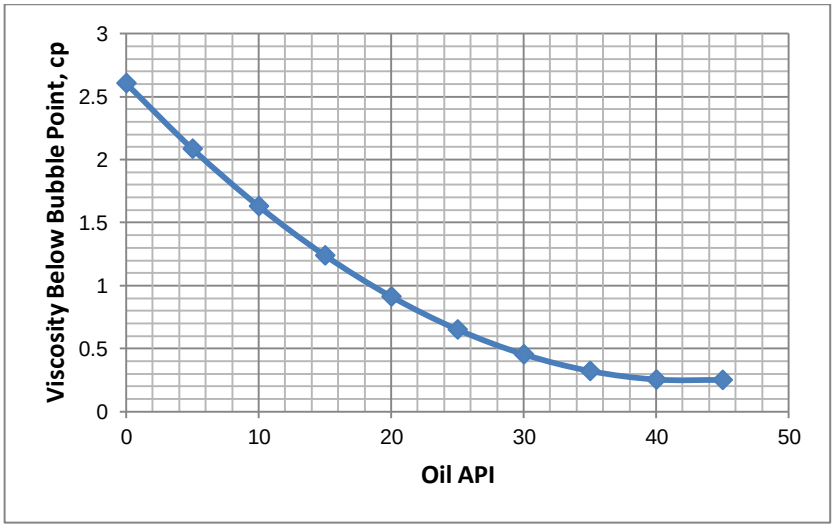


Figure 12: Effect of Reservoir Pressure on Viscosity below Bubble Point

Figure 13 shows the effect of bubble point pressure on viscosity below bubble point. As expected, the model was able to correctly determine the plot to explain the nearly inverse relationship between these two parameters.

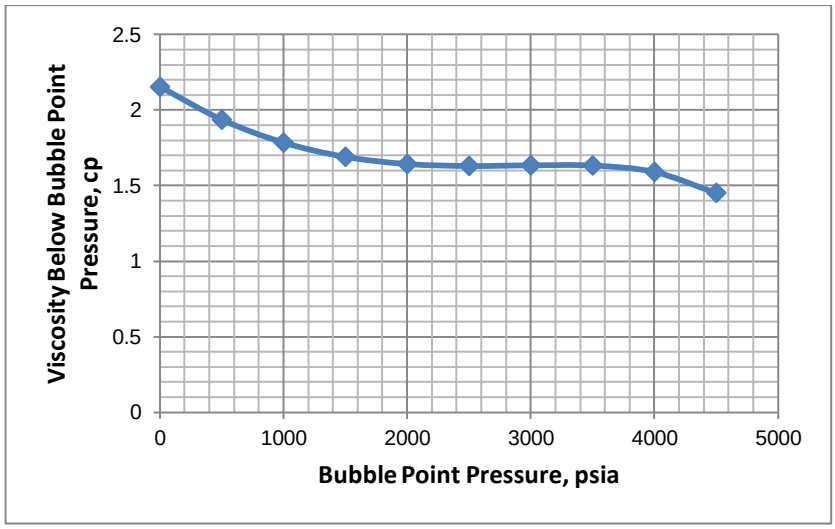


Figure 13: Effect of Bubble Point Pressure on Viscosity below Bubble Point

The bubble point viscosity versus viscosity below bubble point points shows a linear relationship as seen in Figure 14. Viscosity below bubble point is directly proportional to the bubble point viscosity as predicted by the proposed GMDH Model.

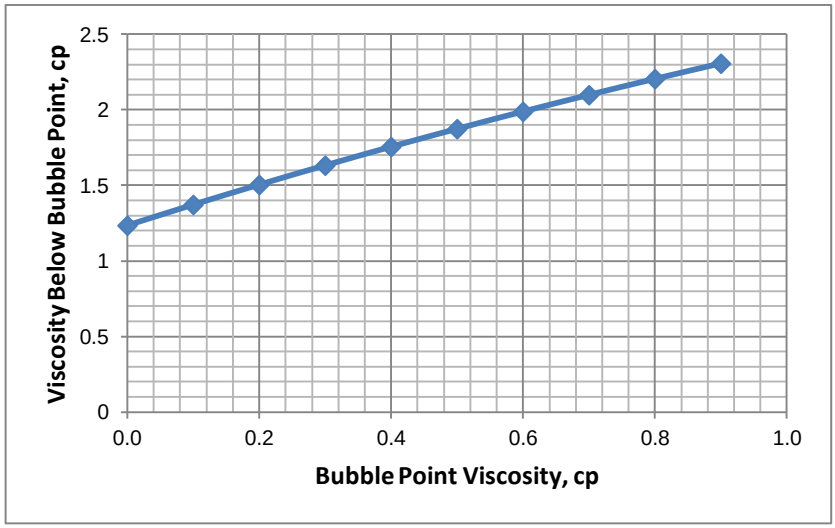


Figure 14: Effect of Bubble Point Viscosity on Viscosity below Bubble Point

Figure 15 shows the relationship between dead oil viscosity and viscosity below bubble point that was predicted by the Polynomial GMDH Model. It shows that viscosity below bubble point is the square root of dead oil viscosity.

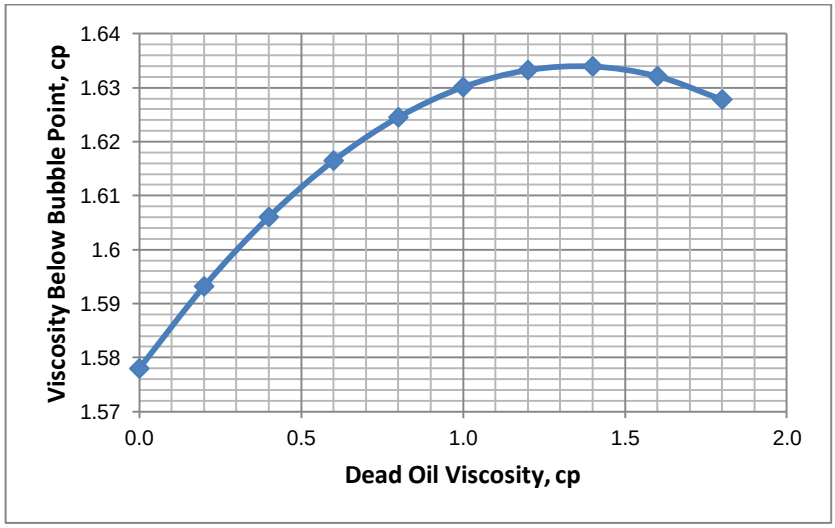


Figure 15: Effect of Dead Oil Viscosity on Viscosity below Bubble Point Pressure:

4.2 GROUP ERROR ANALYSIS OF THE POLYNOMIAL GMDH MODEL AGAINST OTHER INVESTIGATED MODELS

To demonstrate the reliability of the developed model, group error analysis was applied. Average Absolute Relative Error is utilized as a powerful tool for checking the reliability of all empirical correlations as well as Polynomial GMDH Model. By comparing all the investigated correlations and mechanistic models, evaluation of a newly designed correlation can be effective since it is a main criterion in assessing its performance. Average Absolute Relative Error (AAPE) was used in the analysis by grouping input parameter and plotting the resultant AAPE for each set.

Figure 16, 17, 18, 19, 20 present the statistical accuracy of viscosity below bubble point correlation against other investigated models by reservoir pressure, oil API, reservoir pressure, bubble point pressure and dead oil viscosity, respectively.

With the exception of statistical analysis by Oil API which yielded larger AAPE than other investigated models, the other graphs show that the Polynomial GMDH shows competitive value of errors against other models.

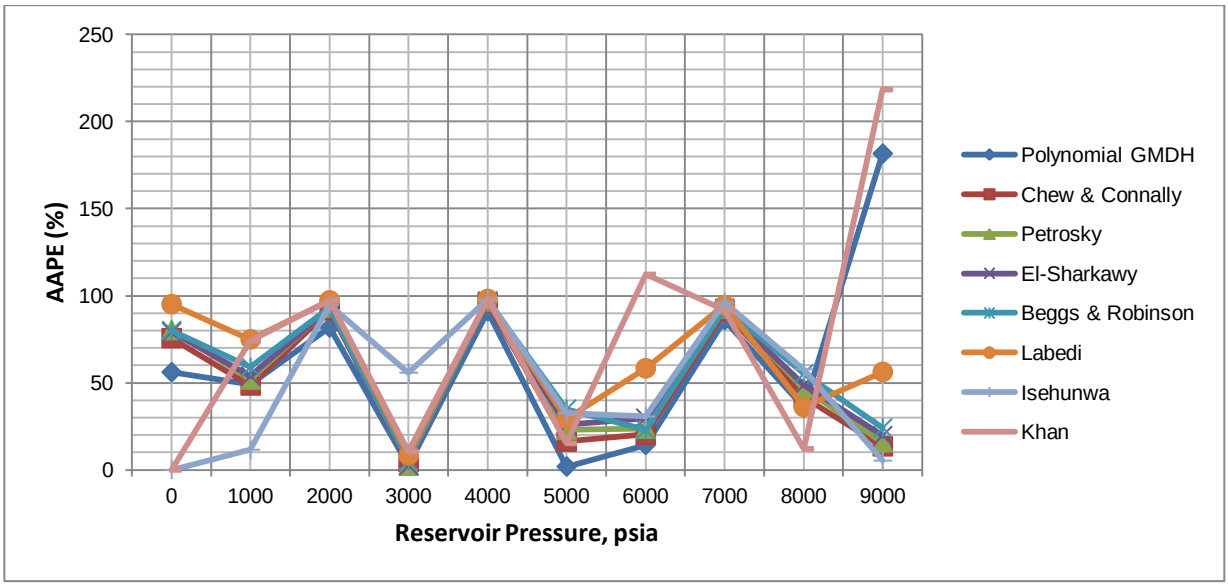


Figure 16: Statistical Accuracy of Viscosity below Bubble Point for Polynomial GMDH and other Investigated Models by Reservoir Pressure

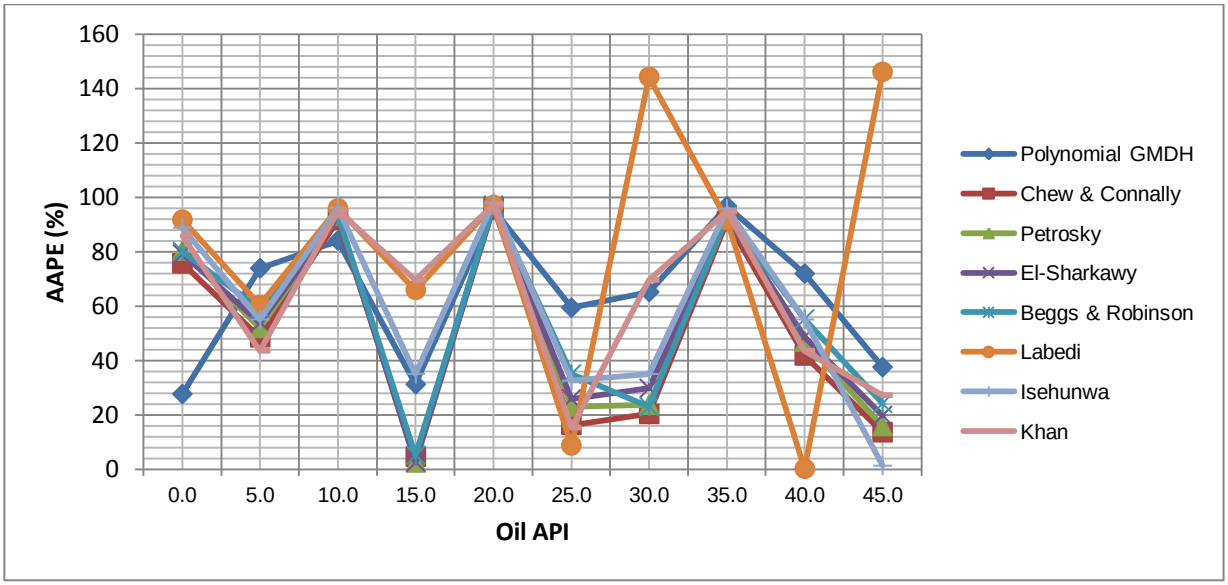


Figure 17: Statistical Accuracy of Viscosity below Bubble Point for Polynomial GMDH and other Investigated Models by Reservoir Pressure

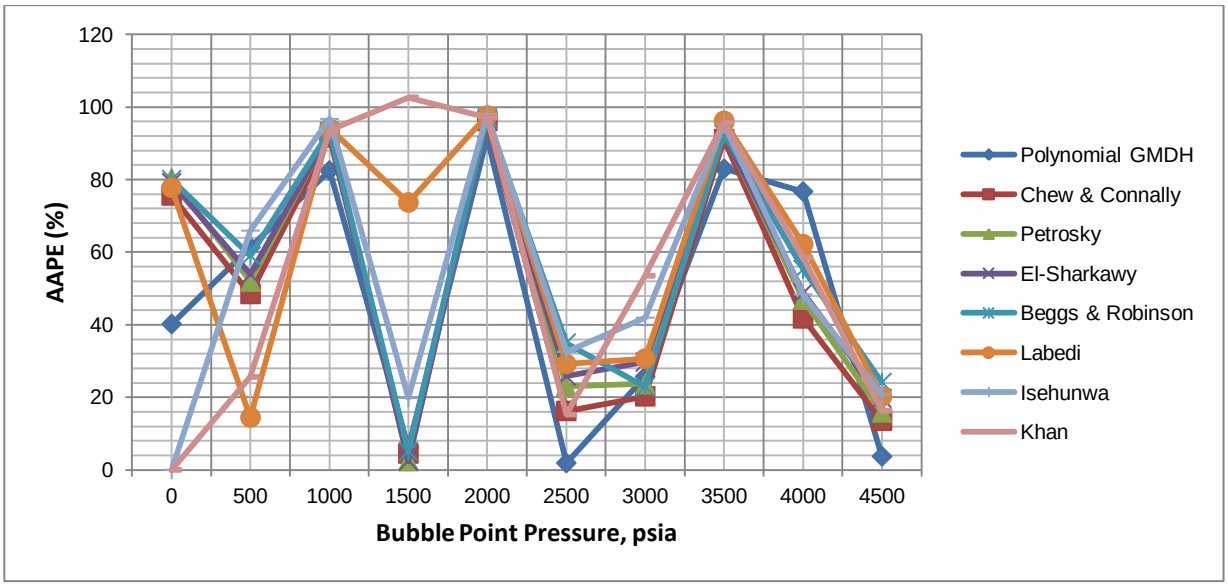


Figure 18: Statistical Accuracy of Viscosity below Bubble Point for Polynomial GMDH and other Investigated Models by Bubble Point Pressure

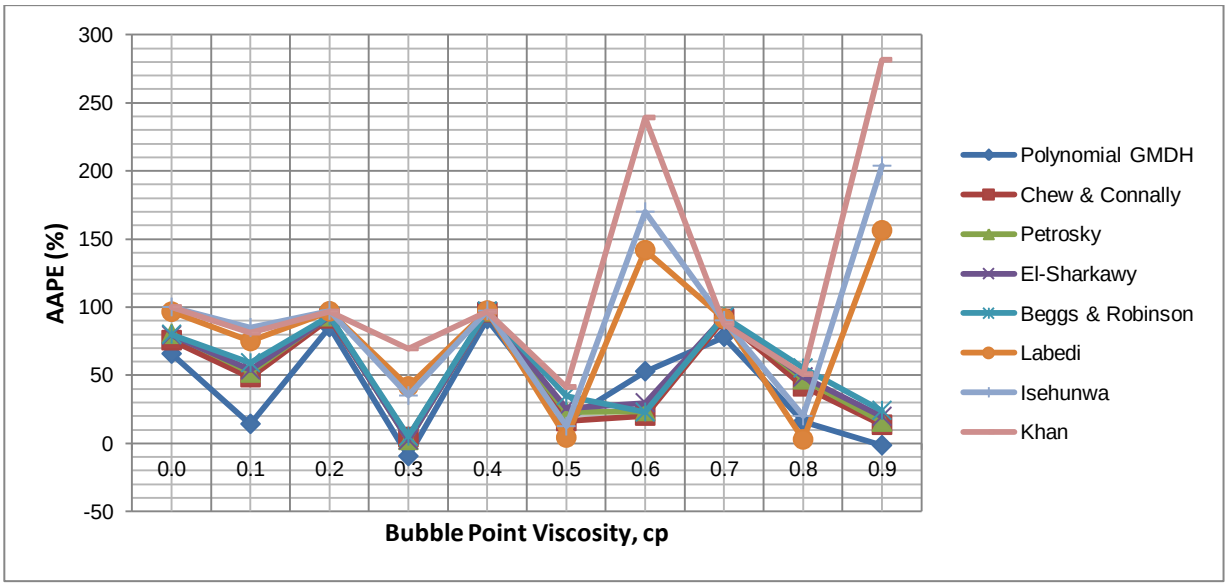


Figure 19: Statistical Accuracy of Viscosity below Bubble Point for Polynomial GMDH and other Investigated Models by Bubble Point Viscosity

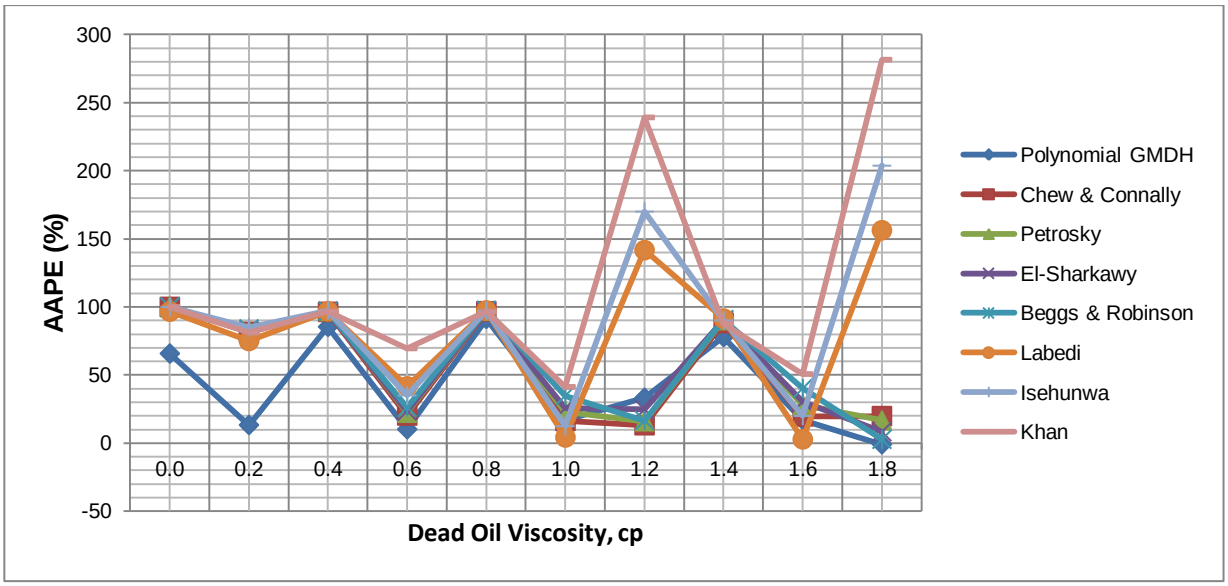


Figure 20: Statistical Accuracy of Viscosity below Bubble Point for Polynomial GMDH and other Investigated Models by Dead Oil Viscosity

4.3 STATISTICAL ERROR ANALYSIS

The summary of statistical comparisons between all data sets (training, validation and testing) of the polynomial GMDH Model for estimating viscosity below bubble point pressure is presented in the table 3 below. For this purpose, the same statistical parameters have been adopted for comparison between different types of models.

Table 3: Statistical Analysis Results of the Polynomial GMDH Model

Parameters / Data Set	E_a	E_r	E_{max}	E_{min}	RMSE	R	STD
Training	17.6088	-0.9746	136.0341	0.0928	0.9716	0.9998	60.2983
Validation	25.4572	-12.5236	102.3683	0.8233	2.7067	0.9984	36.1363
Testing	11.5669	-19.5916	113.4101	0.1838	2.4382	0.9969	26.9877

4.3GRAPHICAL ERROR ANALYSIS

Two graphical analysis techniques have been engaged to envisage the performance of the newly developed GMDH Model and other models at hand. This includes cross plots and error distribution analysis.

4.3.1 CROSSPLOTS OF THE POLYNOMIAL GMDH MODEL

Fig. 21, Fig. 22 and Fig. 23 show the cross plots of the predicted versus experimental viscosity below bubble point pressure for the polynomial GMDH Model where training, validation and testing data sets are plotted separately.

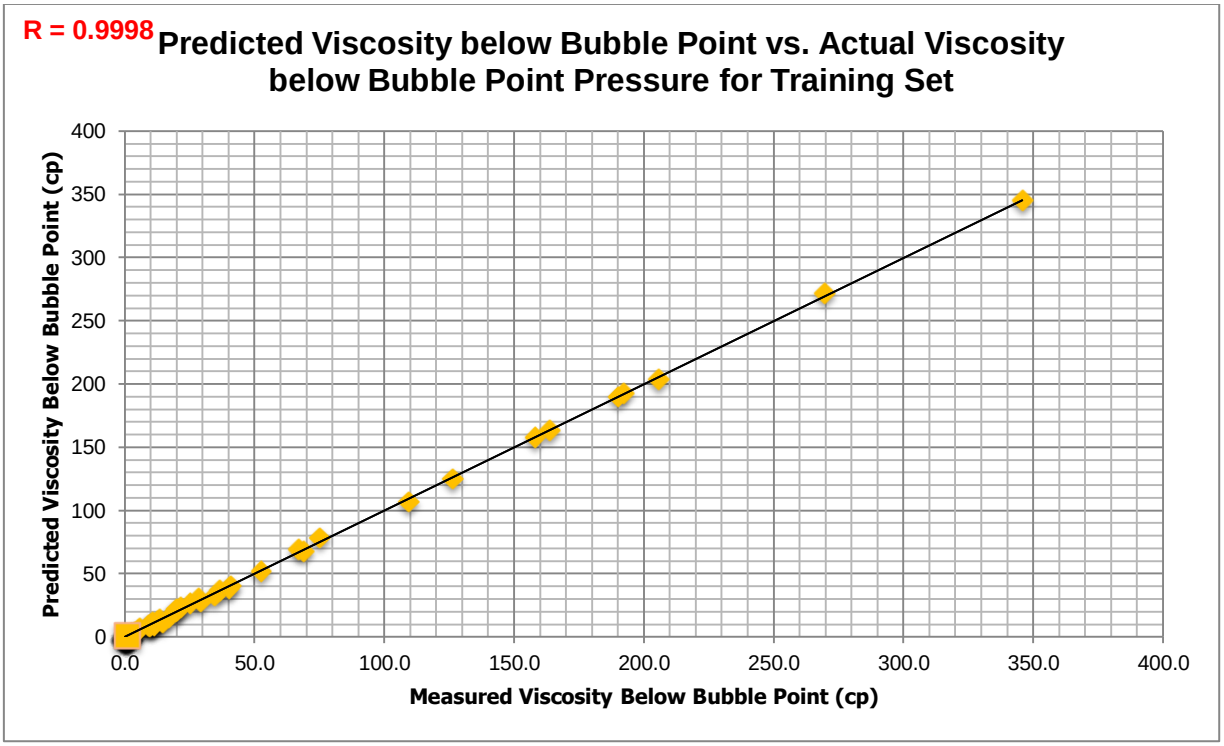


Figure 21: Crossplot of Predicted Viscosity below Bubble Point vs. Actual Viscosity below Bubble Point Pressure for Training Set

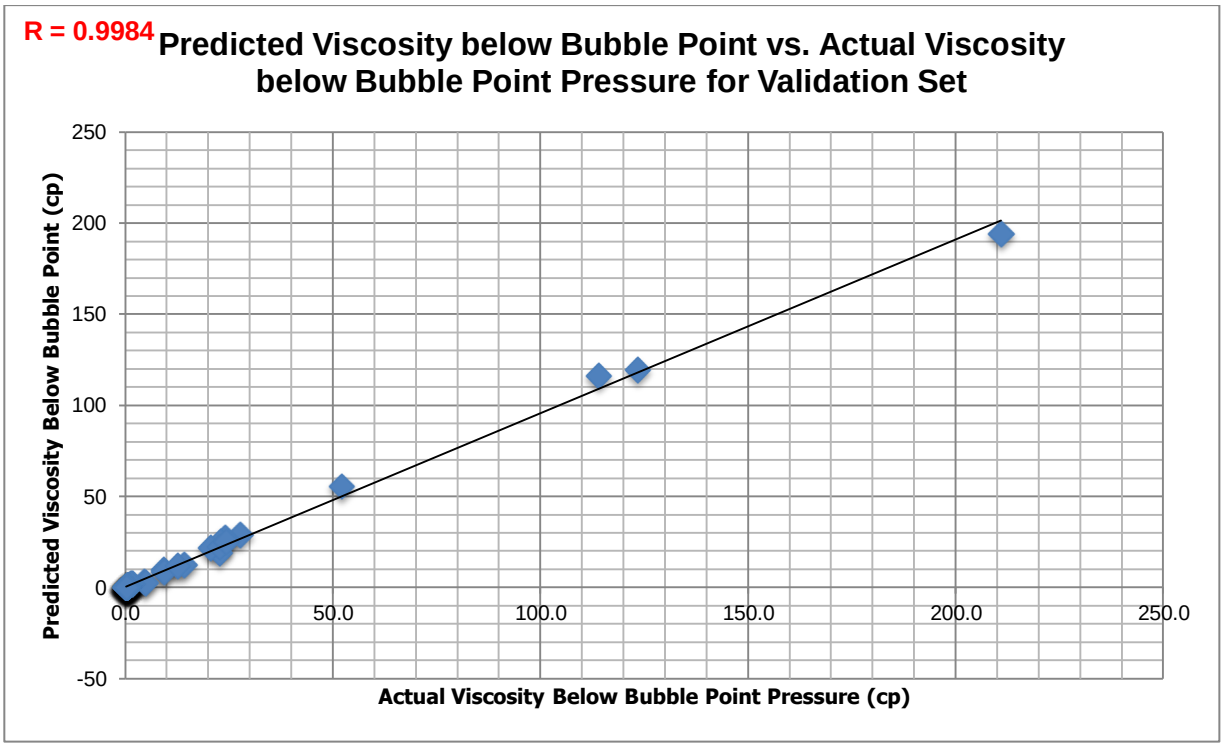


Figure 22: Crossplot of Predicted Viscosity below Bubble Point vs. Actual Viscosity below Bubble Point Pressure for Validation Set

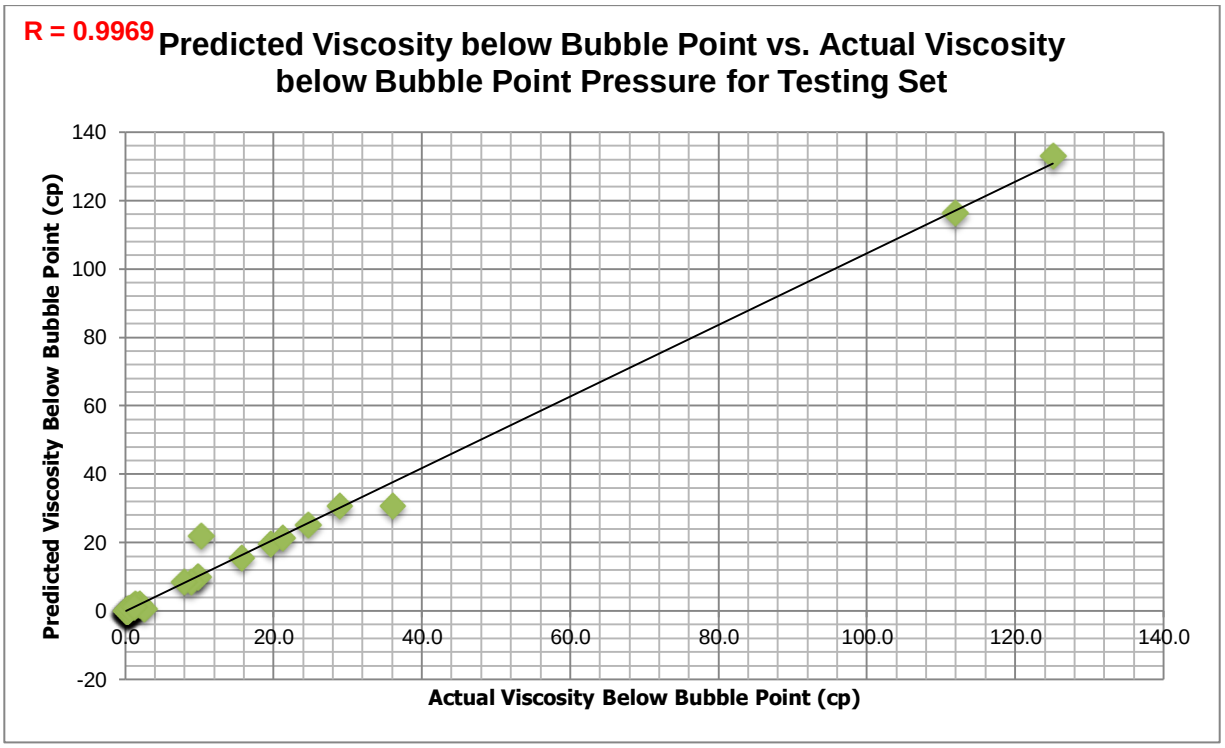


Figure 23: Crossplot of Predicted vs. Actual Viscosity below Bubble Point for Testing Set

Fig. 21 displays a cross plot between the predicted and actual viscosity below bubble point pressure values for training set where a correlation coefficient of 0.9998 was obtained by the GMDH Model. The model shows decent agreement between the actual and estimated values especially at the earlier range between 0.01 – 40.00 centipoise. However, it is noted that correlation coefficient will not be the main criterion for evaluating the performance of the model since it will not give a clear insight on the actual error trend while points under the straight line may be recovered by others under the same line.

Fig. 22 indicates the predicted versus actual viscosity below bubble point pressure for validation set. A correlation coefficient of 0.9984 is recorded for this data set. As previously mentioned in the preceding chapters, validation set is introduced during training of the new model to avoid excessive degree of training. The model displays precise agreement between the actual and estimated values where it is seen how most data points are located very close to the trend line.

Fig. 23 demonstrates a cross plot between the predicted and actual viscosity below bubble point values for the test set created by the model. As expected, the model achieved reasonable correlation coefficient of 0.9969. The correlation coefficient has been achieved

with five input parameters which are dead oil density, reservoir pressure, bubble point viscosity, bubble point pressure and API. In addition, the performance of the model is likely to be improved further if more datasets from wider range of variables are fed into the system. Nevertheless, the excellent correlation coefficient proves that the proposed GMDH model is still very reliable. This will be validated in the later part of the report through comparisons with other investigated models.

The main purpose of this utilizing GMDH in developing a correlation to predict the viscosity below bubble point pressure is to deploy the potential of using soft computing in delivering accuracy and precision which could not be achieved using traditional methods. The process includes finding the most influential input parameters in estimating viscosity below bubble point pressure. The newly developed correlation is deemed competitive against other correlations for viscosity prediction below bubble point pressure for Mediterranean Basin, Africa, Persian Gulf and North Sea oil samples.

The main criterion for evaluating the model's performance is the Average Absolute Percent Relative Error (AAPE) whereby a lower AAPE value indicates that a more accurate model has been produced. The newly developed model achieved an excellent AAPE score of 11.57% which shows that the model is dependable for estimating viscosity below bubble point pressure. Comparisons between the performances of all investigated models with the proposed GMDH Model are provided in Table 4 Comparison between the AAPE for all tested models including the polynomial GMDH Model is provided in Figure 24.

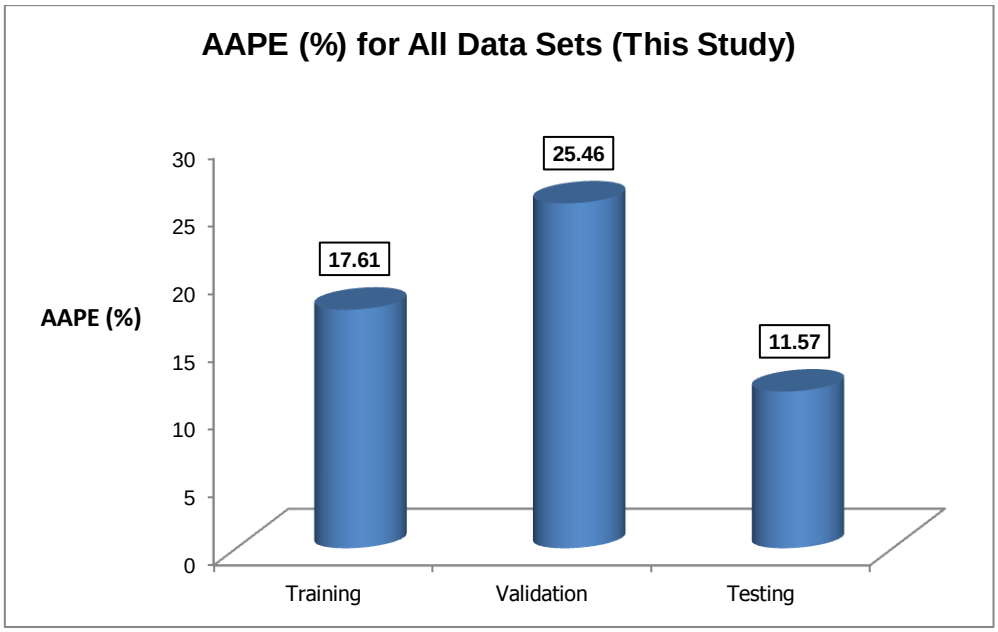


Figure 24: Comparison of AAPE for All Data Sets

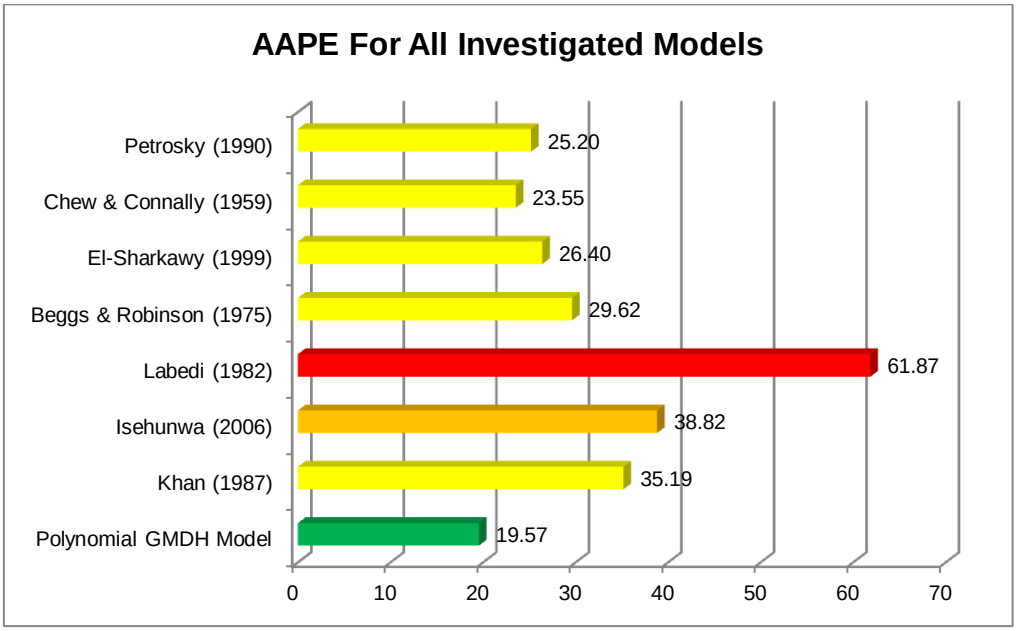


Figure 25: Comparison of Average Percent Relative Errors for the Polynomial GMDH Model against All Investigated Models

In addition to AAPE, other criterions for appraising the model's performance are Standard Deviation, Root Mean Square Error (RMSE), Minimum Absolute Percent Relative Error and Maximum Absolute Percent Error. The GMDH Model failed to provide small maximum absolute percent relative error where a value of 113.4101% is obtained as shown in Figure 26. This shows significant deviation of a unit's predicted value against the actual viscosity below bubble point pressure. If this criterion is opted as the sole criterion to evaluate the model's performance, then the GMDH model will not be considered the best option against other investigated models where it is ranked second behind only Beggs & Robinson (1975). However, this is redeemed by achieving a minute value of minimum absolute percent relative error of only 0.1838%.

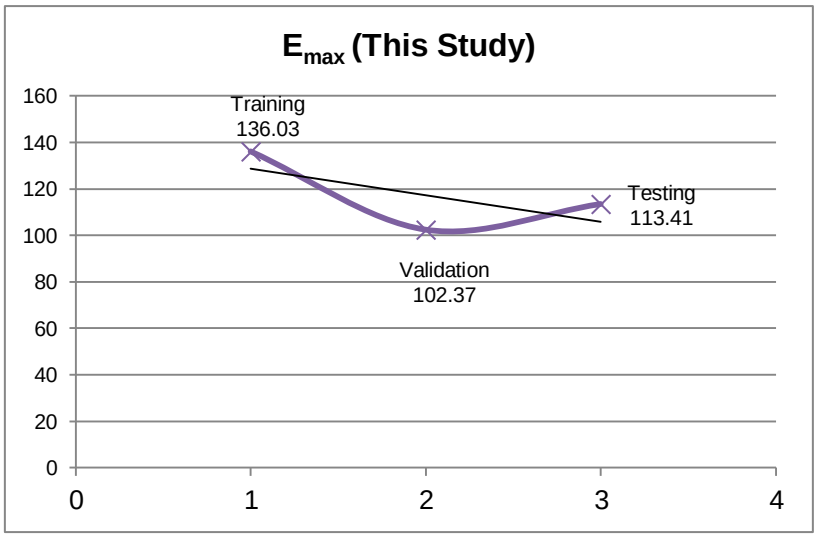


Figure 26: Emax for All Data Sets

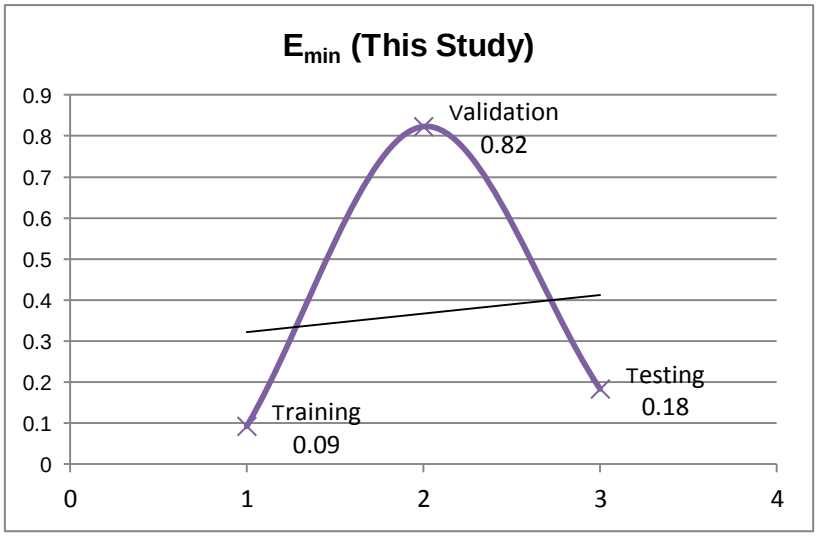


Figure 27: Emin for All Data Sets

On the contrary, if the minimum absolute percent relative error is considered as the only parameter for evaluating a model's performance, the proposed GMDH Model is ranked first in front of other investigated models with a value of 0.1838% as shown in Figure 27 above.

Root Mean Square Error (RMSE) is applied to measure the data dispersion around zero deviation. Fig. 28 shows a comparison of RMSE errors for all data sets. Overall, the polynomial GMDH model shows a respectable value of RMSE of only 2.4382which is the lowest against all the investigated models. If RMSE is the main criterion for determining the quality of a model, the polynomial GMDH model is deemed outstanding. Figure 29 shows a comparison of the RMSE for the polynomial GMDH Model against all investigated models.

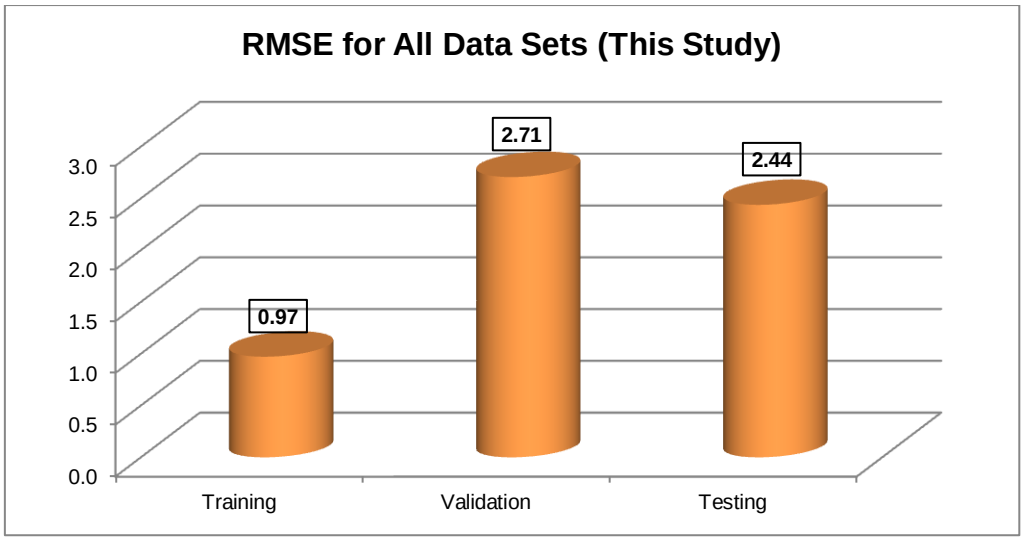


Figure 28: RMSE Recorded for All Data Sets

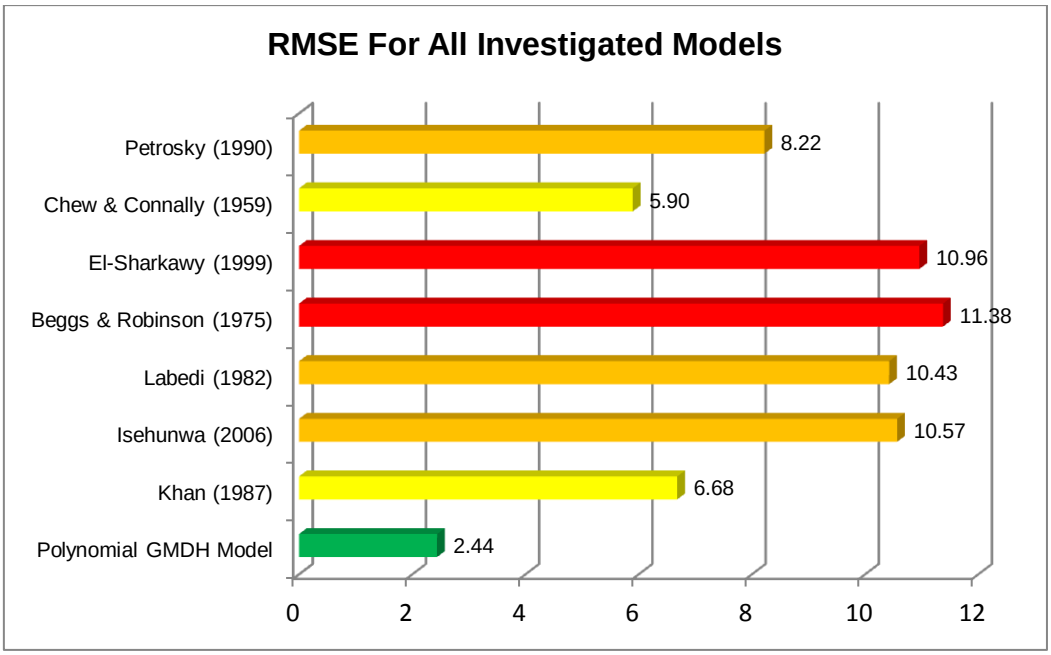


Figure 29: Comparison of Root Mean Square Errors for the Polynomial GMDH Model against all Investigated Models

Fig. 30 shows a comparison of standard deviation for the polynomial GMDH with its smaller counterparts. Standard Deviation (STD) is used to measure the dispersion of a collective set of data from its mean. The testing set recorded a high degree of scattering at 43.72 which affected the overall STD of the polynomial GMDH model. In the meanwhile, Figure 31 shows a comparison of standard deviation for the Polynomial GMDH model against other investigated models. The GMDH Model, however, failed to compete with other models in term of STD where it recorded second worst value at 26.9877, in front of only Khan (1987) at 29.2547. El-Sharkawy displayed the lowest standard deviation against other models at only 16.2759.

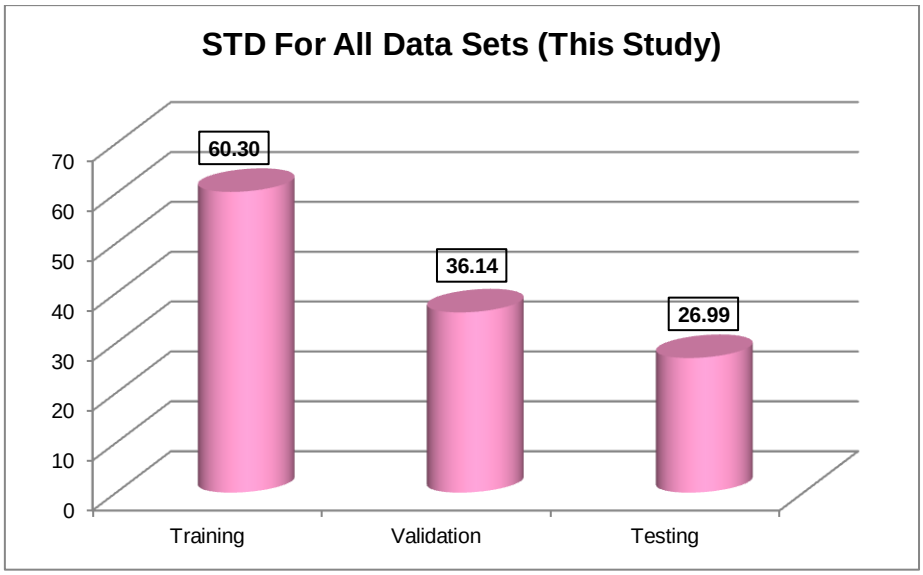


Figure 30: STD Recorded for All Data Sets

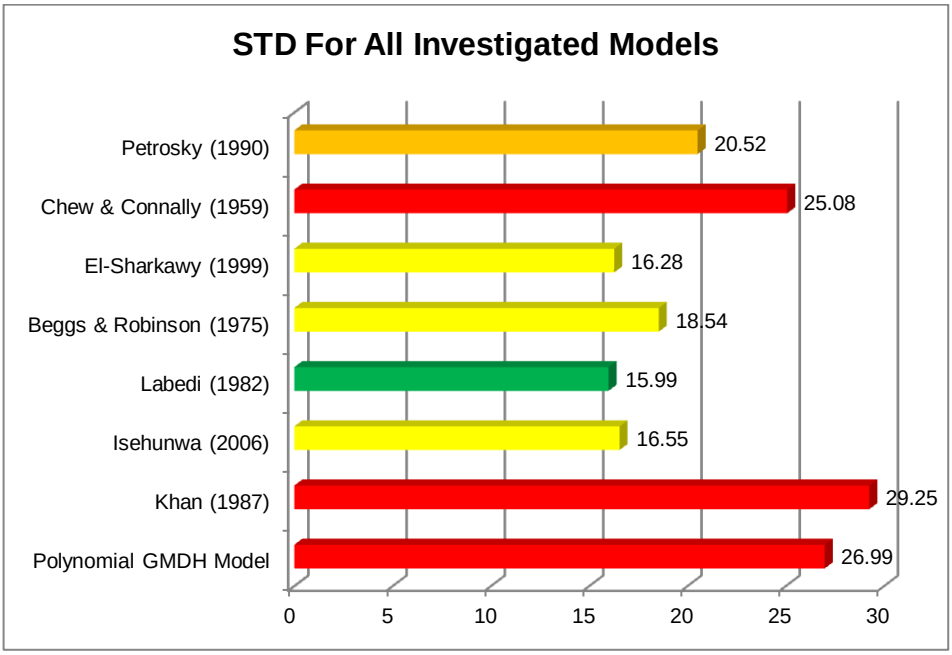
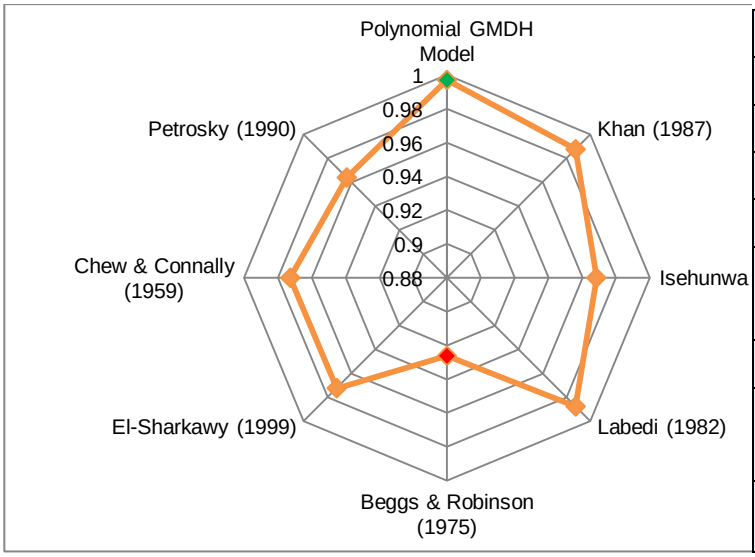


Figure 31: Comparison of Standard Deviation for the Polynomial GMDH Model against All Investigated Models



Model	R
Polynomial GMDH	0.9969
Khan (1987)	0.9875
Isehunwa (2006)	0.9683
Labedi (1982)	0.9876
Beggs& Robinson (1975)	0.9262
El-Sharkawy (1999)	0.9723
Chew &Connally (1959)	0.9727
Petrosky (1990)	0.9639

Figure 32: Comparison of Correlation Coefficients for the Polynomial GMDH Model against All Investigated Models

Figure 32 above shows a comparison of correlation coefficients for GMDH Model against all investigated models. It is noted that the GMDH Model outperformed all other models with the highest correlation coefficient of 0.9969.

Table 4: Statistical Analysis Results of Empirical Correlations, Mechanistic Models, against the Proposed GMDH Model

Parameters / Data Set	E_a	E_r	E_{max}	E_{min}	RMSE	R	STD
Polynomial GMDH Model	19.5669	-19.5916	113.4101	0.1838	2.4382	0.9969	26.9877
Khan (1987)	35.1898	28.7924	838.3845	0.3768	6.6806	0.9875	29.2547
Isehunwa (2006)	38.8210	24.9890	126.7162	2.7039	10.5666	0.9683	16.5471
Labedi (1982)	61.8663	49.3476	752.5419	0.6372	10.4258	0.9876	15.9870
Beggs& Robinson (1975)	29.6158	-15.4709	88.5444	0.6042	11.3775	0.9262	18.5448
El-Sharkawy (1999)	26.3996	-11.6905	108.64827	1.3629	10.9592	0.9723	16.2759
Chew &Connally (1959)	23.5472	8.4717	130.8470	2.0622	5.8960	0.9727	25.0795
Petrosky (1990)	25.1993	-4.8234	118.9796	1.1977	8.2217	0.9639	20.5224

4.3.2 ERROR DISTRIBUTION OF THE POLYNOMIAL GMDH MODEL AGAINST OTHER INVESTIGATED MODELS

Figure 33, Figure 34 and Figure 35 show the error distribution histogram for the polynomial GMDH Model for training, validation and testing sets. A normal distribution curve is added to each of the histogram. Error distribution histogram is analyzed to check model's performance for all the data sets.

Figure 33 shows the error distribution histogram with an added normal distribution curve for the training set of the new polynomial GMDH Model. It shows slight shift to the mean of the errors towards the negative side of the plot at less than 1% indicating that the viscosity is slightly overestimated. It also indicates that most of the total error frequencies are located within the normal distribution curve.

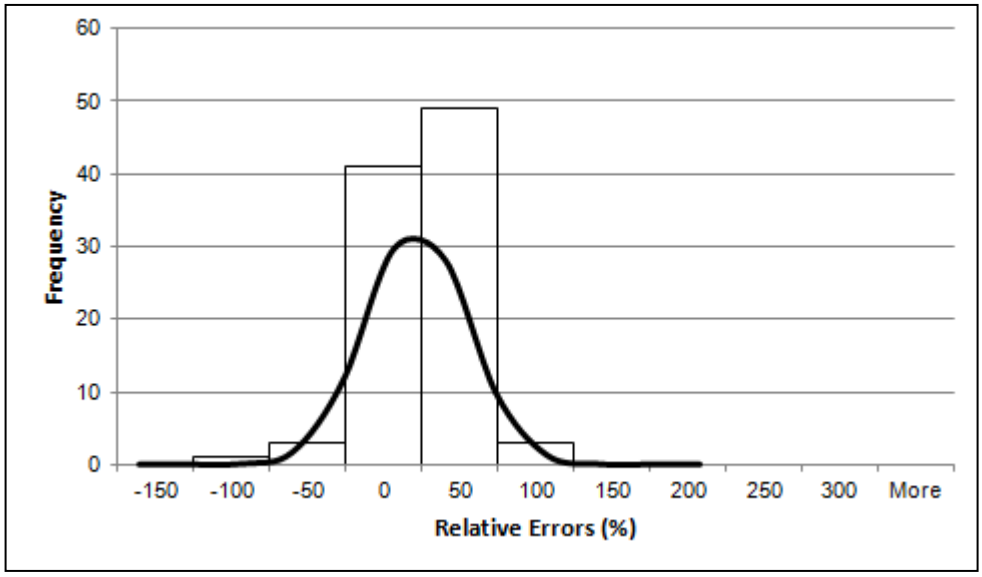


Figure 33: Error Distribution for Training Set

Figure 34 illustrates the error distribution histogram and the normal distribution curve for the validation set of the proposed GMDH Model. It shows a little skewing of the mean of the errors towards the negative side of the plot (about 12%) indicating that the viscosity is also overestimated by this set.

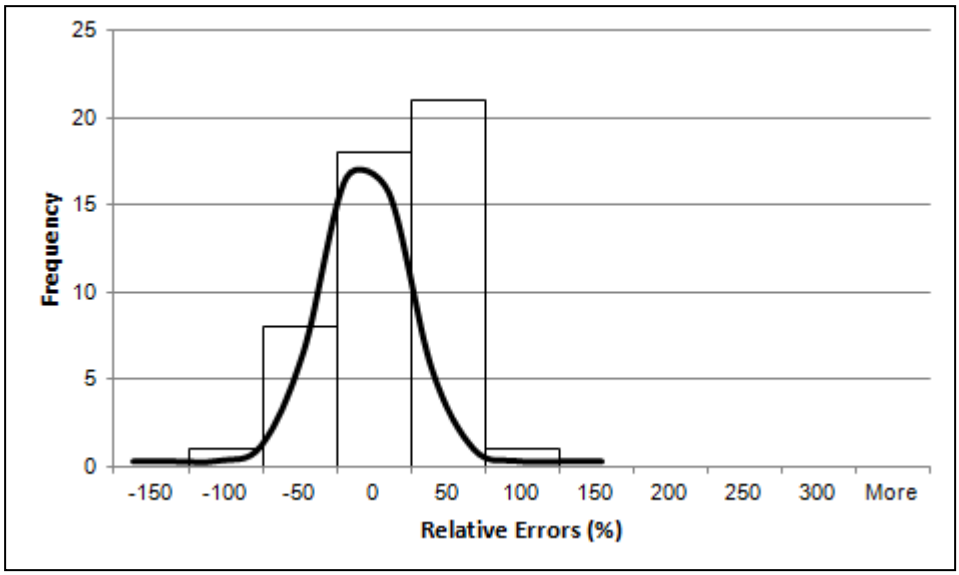


Figure 34: Error Distribution for Validation Set

Figure 35 shows the error distribution histogram and the normal distribution curve for the testing set of the polynomial GMDH Model. The mean of the errors is skewed almost 20% to the left. This also indicated overestimation of viscosity below bubble point pressure for the tested region.

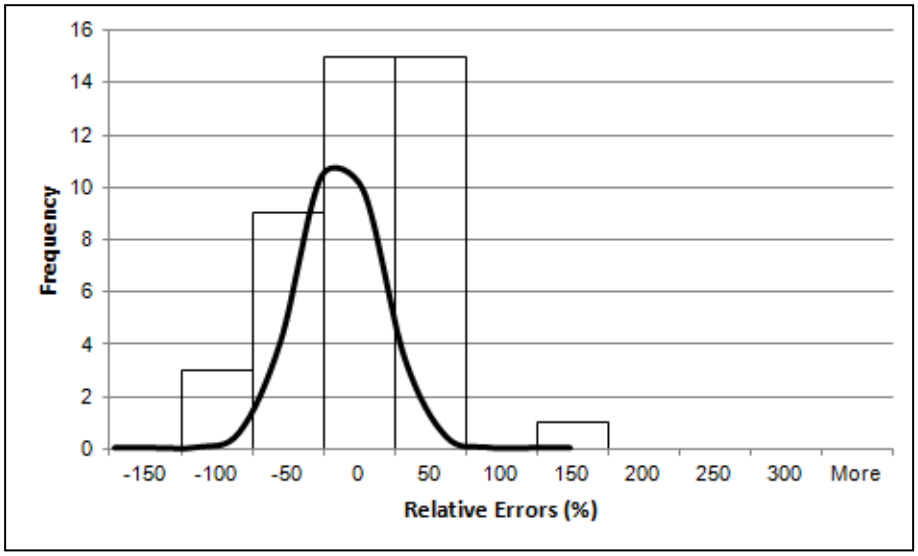


Figure 35: Error Distribution for Testing Set

The (correlation coefficient and the AAPE)and (RMSE and STD) for each model were tabulated next to one another to make the comparison process more comprehensible and clear. Table 5 and Table 6 compared these criterions where a rating has been assigned based on:

- Lower Average Absolute Percent Relative Error but higher correlation coefficient, R
- Lower Root Mean Square Error and Lower Standard Deviation

Table 5: Evaluating Model's Performance by Average Absolute Percent Relative Error and Correlation Coefficient

Parameters / Data Set	E_a	R	Rating
Polynomial GMDH Model	19.5669	0.9969	1
Chew &Connally (1959)	23.5472	0.9727	2
El-Sharkawy (1999)	26.3996	0.9723	3
Petrosky (1990)	25.1993	0.9639	4
Khan (1987)	35.1898	0.9875	5
Isehunwa (2006)	38.8210	0.9683	6
Labedi (1982)	61.8663	0.9876	7
Beggs& Robinson (1975)	29.6158	0.9262	8

Table 6: Evaluating Model's Performance by Root Mean Square Error and Standard Deviation

Parameters / Data Set	RMSE	STD	Rating
Polynomial GMDH Model	2.4382	26.9877	1
Chew &Connally (1959)	5.8960	25.0795	2
Khan (1987)	6.6806	29.2547	3
Petrosky (1990)	8.2217	20.5224	4
Labedi (1982)	10.4258	15.9870	5
Isehunwa (2006)	10.5666	16.5471	6
El-Sharkawy (1999)	10.9592	16.2759	7
Beggs& Robinson (1975)	11.3775	18.5448	8

4.3.3 RESIDUAL ANALYSIS ERROR DISTRIBUTIONS OF THE POLYNOMIAL GMDH MODEL AGAINST ALL INVESTIGATED MODELS

Residual analysis is important to check for a model's consistency and observe the trend in error distribution around the zero line. The relative frequency of deviations between estimated and actual values is shown in Figure 36, Figure 37 and Figure 38 for the Polynomial GMDH Model for each of the data sets which are training, validation and testing.

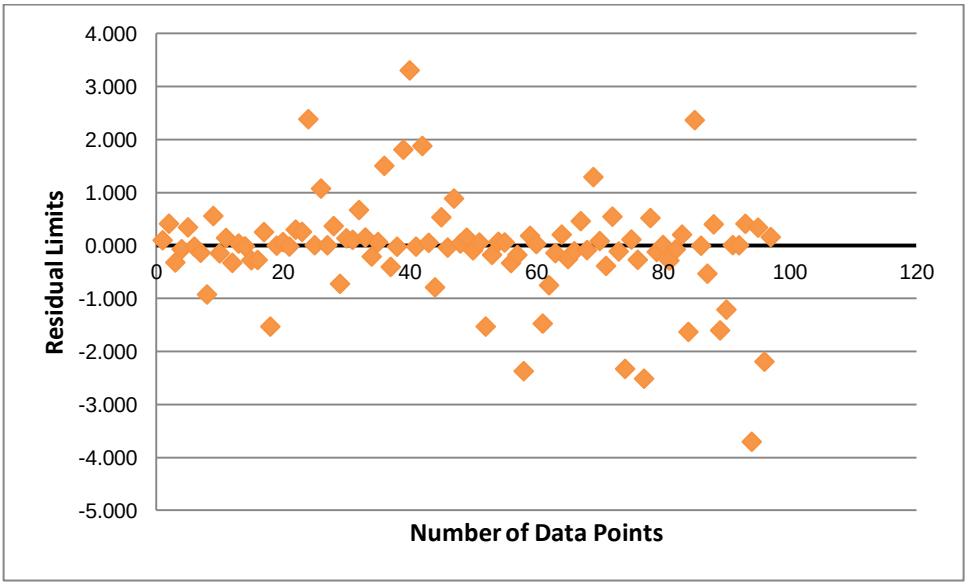


Figure 36: Residual Graph for Training Set

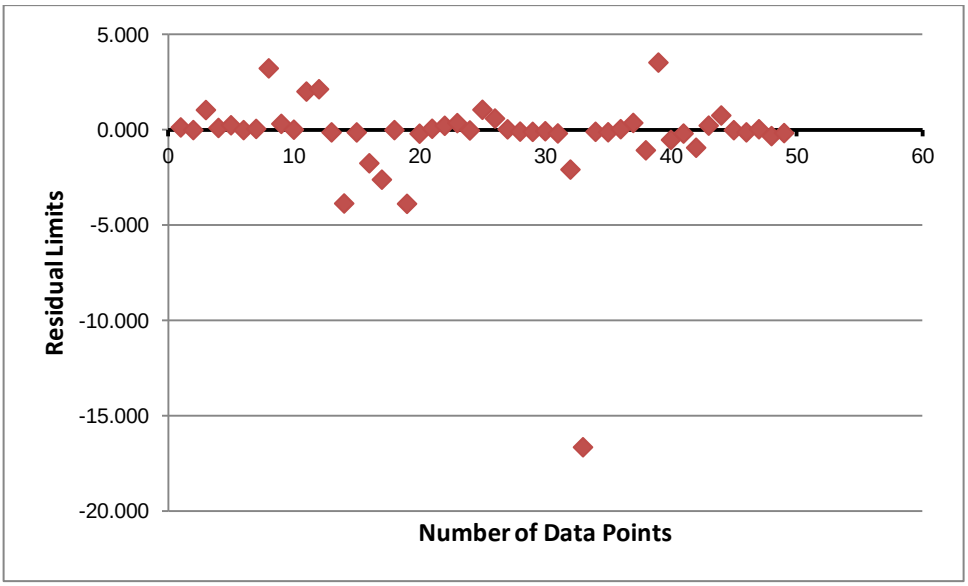


Figure 37: Residual Graph for Validation Set

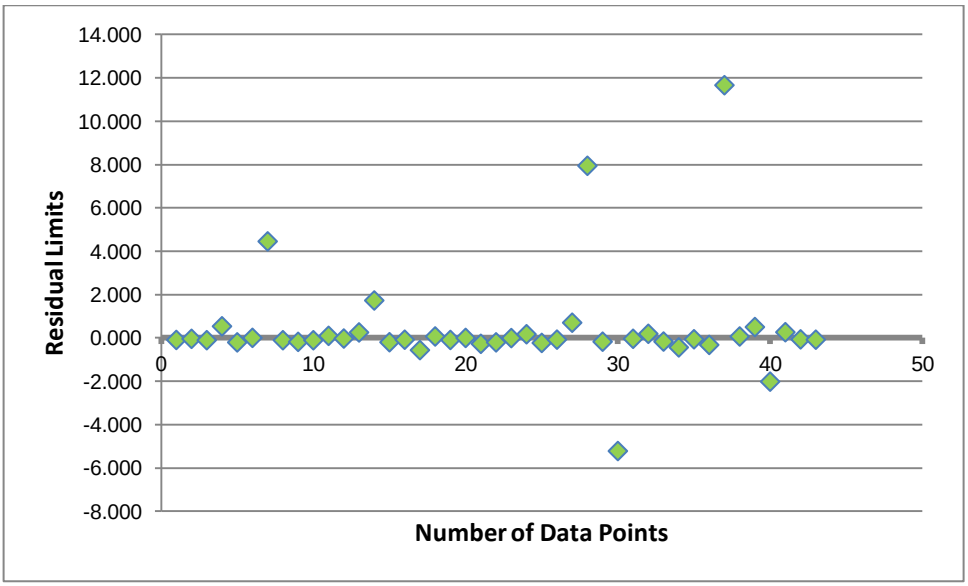


Figure 38: Residual Graph for Testing Set

The Polynomial GMDH Model managed to achieve upstanding consistency in achieving residual limits closer to zero. It demonstrated lower range of residual errors for all its data sets especially in validation and testing set where most of the errors are small and plotted close to the zero line.

The training range achieved a very respectable range in residual analysis from -3.7002 to 3.3105 followed by the validation set at -16.6450 to 3.5253 and testing set at -5.2156 to 11.5563. Residual analysis proves that the proposed GMDH Model is very encouraging as a new correlation to predict viscosity below bubble point pressure.

This shows that GMDH technique can be very assistive for developing a new viscosity correlation at a function of easily determined PVT parameters.

The residual limits of the Polynomial GMDH Model and other investigated models are tabulated in Table 7 below.

Table 7: Residual Limits of the Polynomial GMDH against the Best Investigated Models

Parameters / Data Set	E_{max}	E_{min}
Polynomial GMDH Model	11.6653	-16.6450
Khan (1987)	62.6899	-11.6419
Isehunwa (2006)	26.3744	-174.1110
Labedi (1982)	8.4757	-167.9490
Beggs& Robinson (1975)	4.2669	-189.0919
El-Sharkawy (1999)	4.1871	-228.6519
Chew &Connally (1959)	118.9796	1.1977
Petrosky (1990)	21.0291	-158.9762

Residual analysis on the models reveals that five of the eight models analyzed recorded negative minimum errors of less than -150.0000. These models, Isehunwa (2006), Labedi (1982), Beggs & Robinson (1975), El-Sharkawy (1999) and Petrosky (1999), however, scored relatively low maximum errors. El-Sharkawy achieved the lowest maximum error of only 4.1871 with a tiny lead over the model at second place, Beggs & Robinson (1975). The Polynomial GMDH Model does not rank first in any of the two categories but it achieved consistently low values for both positive and negative errors as opposed to other models with 11.6653 and -16.6450, respectively.

4.4 SUMMARY

This chapter includes comprehensive analysis of the results obtained from the current research. Trend analysis and group error analysis were conducted extensively on the GMDH model and checked for their main input parameters and their respective range. The bottom-line is that statistical and graphical analyses show that the Polynomial GMDH Model is justifiably better than all the investigated models. Average Absolute Percentage Error (AAPE) has been chosen as a main statistical criterion for evaluating the performance of the models. The model recorded AAPE of 19.5669 which is considered good. The GMDH model managed to uncover the most relevant input parameters involved in estimating viscosity below bubble point with a reasonable degree of accuracy which can improve the modeling procedure. Last but not least, the potential of GMDH technique has been explored successfully.

CHAPTER 5: CONCLUSION AND RECOMMENDATIONS

5.0 CONCLUSION

Estimating viscosity below bubble point pressure is considered necessary due to the shortcomings of other existing alternatives such as the expensive and timely PVT laboratory analysis and the difficult wellhead readings conversion to reservoir values. The new polynomial GMDH model was derived as a function of five PVT parameters which are dead oil viscosity, oil API, bubble point pressure, bubble point viscosity and reservoir pressure based on 195 data sets acquired from a number of wells from the Middle East.

Average Absolute Percent Error (AAPE) which has been decided as the main criterion for the evaluation of the model showed that the Polynomial GMDH Model recorded 19.5669 and is ranked first against other investigated models. The correlation also recorded near perfect correlation coefficient of 0.9969, higher than all the other investigated models indicating that the model is highly reliable within the range of the data.

Polynomial GMDH model helps in reducing the problem of dimensionality that lowers the performance of neural network modeling efficiency.

5.1 RECOMMENDATIONS

1. Trend analysis and group error analysis should be conducted to check whether the correlation mirrors experimental conditions.
2. Viscosity correlations are geographic. Therefore, it is advisable that future researches focus on regions which have yet to have a correlation developed from their oil samples such as Malay Basin.
3. Every correlation responds greatly to a certain small range of data out of the total range of the data used to develop it. Efforts should be undertaken to map the best range of every correlation.
4. GMDH is a powerful technique which should be used widely in every job field.

REFERENCES

1. Alomair, O., Ali, M. A. J., Alkoriem, A., and Hamed, Mohamed. 2012. Heavy Oil Viscosity and Density Prediction at Normal and Elevated Temperatures. SPE Paper 163342, presented at SPE Kuwait International Petroleum Conference and Exhibition, Kuwait, 10 – 12 December.

2. Ayoub, M.A., Raja, D.M. and Al-Marhoun, M.A. 2007. Evaluation of below Bubble point Viscosity Correlations & Construction of a NewNeural NetworkModel. Paper SPE 108349, presented at 2007 SPE Asia Pacific Oil & Gas Conference and Exhibition, Indonesia, 30 October – 1 November.

3. Beal, C.: "The Viscosity of Air, Water, Natural Gas, Crude Oil and Its Associated Gases at Oil Field Temperature and Pressures", *Society of Petroleum Engineers Journal* (Feb 1946).

4. Beggs, H.D. and Robinson, J.R.:"Estimating the Viscosity of Crude Oil System", *Society of Petroleum Engineers Journal* (Sep 1975).

5. Chew, J. and Connally, C.A. Jr. 1959. A Viscosity Correlation for Gas–Saturated Crude Oils. SPE Paper 8050, presented at 33rd SPE Annual Fall Meeting of Petroleum Engineers, USA, 5-8 October.

6. De Ghetto, G.,Paone, & Villa, M. 1994. Reliability Analysis on PVT Correlations. Paper SPE 28904, presented at the European Petroleum Conference, UK, 25-27 October.

7. Elsharkawy, A.M &Alikhan, A.A. 1999. Models For Predicting The Viscosity of Middle East Crude Oils. *Fuel*, 891-903.

8. Farshad, F., LeBlanc, J.L. and Garber, J. D, 1996. Empirical PVT Correlations for Colombian Crude Oils. Paper SPE 36105, presented atSPE Fourth Latin American and Caribbean Petroleum Engineering Conference, Tobago, 23-26 April.

9. Hajizadeh, Y. 2007. Intelligent Prediction of Reservoir Fluid Viscosity. Paper SPE 106764, presented at 2007 SPE Production and OperationsSymposium, USA, 31 March – 3 April.

10. Isehunwa, O.S., Olomigoke, O., and Makinde, A.A.2006. A Correlation to Predict the Viscosity of Light Crude Oils. Paper SPE 105983, presented at the 31st Annual SPE International Technical Conference and Exhibition, Nigeria, 31 July – 2 August.

11. Khan, S.A., Al-Marhoun, M.A., Duffuaa, S.O. and Abu-Khamsin, S.A.1987. Viscosity Correlations for Saudi Arabian Crude Oils. PaperSPE 15720, presented at Fifth SPE Middle East Oil Show, Bahrain, 7-10 March.

12. Kahn, S. A.: "Development of Viscosity Correlations for Crude Oils", *Society of Petroleum Engineers Journal* (Aug 1987).

13. Cogger, K. O. (2010). Non Linear Multiple Regression Methods: A Survey and Extensions. *Intell. Sys. Acc. Fin. Mgmt.* 17, 19–39. Retrieved from http://www.interscience.wiley.com

14. Labedi, R.. 1992. Improved Correlations for Predicting the Viscosity of Light Crudes. *J. Pet. Sc& Eng*.8: 221 – 234.

15. Labedi, R.M. "PVT Correlations of the African Crudes," PhD Thesis, Colorado School of Mines (May, 1982)

16. Lohrenz, J., B.C. Bray, and C.R. Clark, Calculating viscosities of reservoir fluids from their composition. JPT, 1964. 10:

p. 1170– 1176.

17. Little, J.E. and Kennedy, H.T. 1966. Calculating the viscosity of hydrocarbon systems with pressure temperature andcomposition. SPE Paper 1589, presented at SPE 41st Annual Fall Meeting, USA, 2-5 October.

18. Oloso, M. A., Khouki, A., Abdulraheem, A., and Elshafei, M. 2009. Prediction of Crude Oil Viscosity and Gas/Oil Ratio Curves Using Recent Advances to Neural Networks. Paper SPE 125360, presented at 2009 SPE/EAGE Reservoir Characterization and Simulation Conference, UAE, 19-21 October.

19. Omar, M.I., Petronas (PRI)., and Todd, A.C. 2009. Development of New Modified Black Oil Correlations for Malaysia Crudes. Paper SPE 25338, presented at 1993 SPE Asia Pacific Oil & Gas Conference & Exhibition, Singapore, 8-10 February.

20. Petrosky, G.E. Jr. and Farshad, F.F. 1995. Viscosity Correlations for Gulf of Mexico Crude Oils. Paper SPE 29468 presented at ProductionOperations Symposium, , U.S.A., 2-4 April.

21. Sutton, R. P., and Farshad, F.F. 1989. Evaluation of Empirically Derived PVT Properties for Gulf of Mexico Crude Oils. Paper SPE 13172, presented at 1984 SPE Annual Technical Conference and Exhibition, US, 16-19 September.

22. Vazquez, M. and Beggs, H.D. 1980. Correlation for Fluid Physical Property Prediction. *J. Pet Tech* 32(6): 968-970.

23. Madala, H.R. and Ivakhnenko, A.G., (1994). Inductive Learning Algorithms for Complex System Modelling. Florida, USA, CRC Press, Inc.