APPLICATION OF FUNCTION APPROXIMATIONS TO RESERVOIR ENGINEERING

A Thesis
Submitted to the Faculty of Graduate Studies and Research
in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy
in
Petroleum Systems Engineering
University of Regina

By
Saber Khaled Elmabrouk
Regina, Saskatchewan

February, 2012
Copyright 2012: S.Elmabrouk.
Saber Khaled Elmabrouk, candidate for the degree of Doctor of Philosophy in Petroleum Systems Engineering, has presented a thesis titled, *Application of Function Approximations to Reservoir Engineering*, in an oral examination held on January 18, 2012. The following committee members have found the thesis acceptable in form and content, and that the candidate demonstrated satisfactory knowledge of the subject material.

External Examiner: Dr. Ali Al-Khafaji, ACTS Consulting

Co-Supervisor: Dr. Ezeddin Shirif, Petroleum Systems Engineering

Co-Supervisor: Dr. Rene Mayorga, Industrial Systems Engineering

Committee Member: Dr. Amr Henni, Industrial Systems Engineering

Committee Member: Dr. Koorosh Asghari, Petroleum Systems Engineering

Committee Member: Dr. Malek Mouhoub, Department of Computer Science

Chair of Defense: Dr. Dongyan Blachford, Faculty of Graduate Studies & Research

*Not present at defense*
ABSTRACT

The need for function approximations arises in many branches of engineering, and in particular, petroleum engineering. A function approximation estimates an unknown function, which then finds an underlying relationship within a given set of input-output data. In this study, a Multiple Regression Analysis (MRA), Artificial Neural Network (ANN), and Least Squares Support Vector Machine (LS-SVM) are applied to address some of the most three important ongoing challenges in reservoir engineering. The three ongoing challenges are reservoir fluid properties in the absence of PVT analysis, average reservoir pressure and oil production prediction.

1. The PVT properties of crude-oil such as the bubble point pressure ($P_b$), oil formation volume factor ($B_{ob}$), dissolved GOR ($R_{sob}$) and stock tank vent GOR ($R_{ST}$) play a key role in calculating reserves as well as for identification of reservoir characteristics. The properties are traditionally determined from laboratory analyses of reservoir oil samples. In the absence of experimental analysis, empirical correlations or models can be used to estimate reservoir fluid properties. In this study, MRA, ANN and LS-SVM are applied to develop PVT models with which to estimate the $P_b$, $B_{ob}$, $R_{sob}$ and $R_{ST}$. Unlike the present PVT models, the proposed models can be applied in a straightforward manner by using direct field data. Additional correlations or experimental analyses are unnecessary.

2. Insight into average reservoir pressure ($P_{avg}$), and its change over time, plays a critical role in reservoir development. However, in order to determine the $P_{avg}$, the well is shut-in for a build up test, resulting in loss of production. In high permeability reservoirs, this may not be a significant issue, but in medium to low permeability reservoirs, the
shut-in period during the entire test may last several weeks before a reliable reservoir pressure can be estimated. This loss of production, and cost of monitoring the shut-in pressure, is often unacceptable. It is of great practical value if the $P_{\text{avg}}$ can be obtained from the historical production and reservoir pressure data without having to shut-in the well. Three different models (BP-ANN, ANN-GA and LS-SVM) are obtained to predict and interpolate $P_{\text{avg}}$ without closing the producing wells. The results indicate the proposed approach has the ability to accurately interpolate and predict current average reservoir pressure by employing historical production data.

3. The prediction of oil reservoir production performance has been an on-going challenge for engineers. It is an essential component of petroleum reservoir management. Traditionally, numerical simulations and decline curve analysis have been used to predict a reservoir’s future performance based on its current and past performance. Reservoir simulation is very time consuming and offers a non-unique solution with a high degree of uncertainty. On the other hand, decline curve analysis fits the observed production rates of individual wells, groups of wells or an entire reservoir using mathematical functions to predict future production by extrapolating the declining function. In this study, ANN and LS-SVM are presented to predict the performance of oil production within water injection reservoirs. The historical production and injection data are used as inputs. The approach can be categorized as a new and rapid method with reasonable results. Another application of these models is that it can be utilized to find the most economical scenario of water injection to maximize ultimate oil recovery. This method could be a new window for fast simulators. It has reasonable accuracy, requires little data and can forecast quickly.
Forest, I would like to express my sincere gratitude to my supervisors Dr. Ezeddin Shirif and Dr. Rene Mayorga for the continuous support of my PhD study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study.

A very special thanks goes out to Dr. Amr Henni, without whose motivation and encouragement I would not have made it this far. He provided me with direction, technical support and became more of a mentor and friend, than a professor. I doubt that I will ever be able to convey my appreciation fully, but I owe him my eternal gratitude. I would like to thank the other members of my committee, Dr. Koorosh Ashagari for the assistance he provided and taking time out from his busy schedule and Dr. Malek Mouhoub from the Department of Computer Science for his valuable suggestions and discussions regarding my thesis study.

I gratefully acknowledge the Waha Oil Company for giving permission to use some of production data in chapters VI, VII, VIII, IX, and X. Moreover, I would also like to thank my family for the support they provided me through my entire life and in particular, I must acknowledge my wife and kids, without whose love, and encouragement I would not have finished my PhD.

In conclusion, I recognize that this research would not have been possible without the financial assistance of Faculty of Graduate Studies and Research, and the Faculty of Engineering and Applied Science, University of Regina (Teaching Assistantships, Graduate Research Scholarships). I express my gratitude to them.
This thesis is dedicated
to
my wife and kids
who have always stood by me
and
dealt with all of my absence from many family occasions
with a smile.
Without my wife’s support, and gentle prodding,
I would still be trying
to decide a topic for my thesis.
# TABLE OF CONTENTS

## CHAPTER I

### Introduction
1.1 Background .............................................................................................................. 1
1.2 Function Approximation .......................................................................................... 3
  1.2.1 Regression Analysis ......................................................................................... 3
  1.2.2 Artificial Neural Network ................................................................................. 4
  1.2.3 Least squares support vector machines ............................................................ 4
1.3 Key Ongoing Challenges ....................................................................................... 5
  1.3.1 Why Reservoir Fluid Properties? ..................................................................... 5
  1.3.2 Why Average Reservoir Pressure? ................................................................... 7
  1.3.3 Why Reservoir Oil Production Prediction? ...................................................... 8
1.4 Goals of the Thesis ............................................................................................... 10
1.5 Chapter by Chapter Overview .............................................................................. 11
1.6 References ............................................................................................................. 14

## CHAPTER II

Prediction of Bubble Point Pressure, Bubble Point Oil Formation Volume Factor in the Absence of PVT Analysis by Utilizing Multiple Regression Analysis

2.1 Introduction ............................................................................................................ 16
2.2 Literature Review ................................................................................................ 19
2.3 Newly Developed $P_b$ and $B_{ob}$ Correlations ...................................................... 25
  2.3.1 Bubble point pressure correlation................................................................... 26
  2.3.2 Bubble Point Oil FVF Correlation .................................................................. 28
2.4 Correlation Verification and Validation ............................................................. 28
2.5 Comparison with Other Correlations .................................................................. 35
2.6 Summary and Conclusions ................................................................................. 39
2.7 References ........................................................................................................... 41

## CHAPTER III

Prediction of Bubble Point Solution Gas/Oil Ratio in the Absence of PVT Analysis

3.1 Introduction .......................................................................................................... 44
3.2 Literature Review .................................................................................................. 46
3.3 Newly Developed Correlations ............................................................................. 48
  3.3.1 PVT Data ........................................................................................................ 49
  3.3.2 Bubble Point Solution Gas/Oil ratio, $R_{sob}$ Correlation .................................. 50
  3.3.3 Stock-Tank Vent Gas/Oil Ratio, $R_{ST}$ Correlation .......................................... 53
Figure 3.2 - Estimated Rsob from Eq. (3.7) vs. Actual Rsob (431 Data Points)

Comparison with Other Correlations

3.5 Conclusions

3.6 References

CHAPTER IV
Development of PVT Models Utilizing ANNs

4.1 Introduction

4.2 Review of Recent PVT Neural Network Models

4.3 PVT Datasets

4.4 Bubble point Pressure ANN Model

4.4.1 Bubble Point Pressure ANN Model Evaluation

4.4.2 Contribution of Input Parameters

4.4.3 Deployment Test

4.5 Bubble Point Oil FVF ANN Model

4.5.1 Graphical Error Analysis

4.5.2 Deployment Test

4.6 Bubble Point Solution Gas/Oil ratio (Rsob) ANN Model

4.6.1 Model Design

4.6.2 Model Evaluation

4.6.3 Model Deployment

4.7 Stock-Tank Vent Gas Oil Ratio ANN Model

4.7.1 Stock-Tank Vent Gas/Oil Ratio ANN Model Evaluation

4.8 Conclusions

4.9 References

CHAPTER V
Development of PVT Models Utilizing LS-SVM Approach

5.1 Introduction

5.2 LS-SVM for Function Approximation

5.3 Model Parameter Selecting

5.4 Newly Developed PVT LS-SVM Models

5.5 PVT Data Preparation

5.5.1 Accuracy of the Bubble Point Pressure LS-SVM Model

5.6 Bubble Point Oil Formation volume Factor, B_{ob} LS-SVM Model

5.6.1 Accuracy of the Bubble Point Oil FVF LS-SVM Model

5.7 Bubble Point Solution Gas/Oil Ratio LS_SVM Model
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.7.1</td>
<td>Accuracy of the Bubble Point Solution Gas/Oil Ratio LS-SVM Model</td>
<td>184</td>
</tr>
<tr>
<td>5.8</td>
<td>Stock-Tank Vent Gas/Oil Ratio LS-SVM Model</td>
<td>185</td>
</tr>
<tr>
<td>5.8.1</td>
<td>Accuracy of the R&lt;sub&gt;ST&lt;/sub&gt; LS-SVM Model</td>
<td>198</td>
</tr>
<tr>
<td>5.9</td>
<td>Conclusions</td>
<td>198</td>
</tr>
<tr>
<td>5.10</td>
<td>References</td>
<td>202</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>204</td>
</tr>
<tr>
<td>6.2</td>
<td>Artificial Neural Networks (ANNs)</td>
<td>205</td>
</tr>
<tr>
<td>6.3</td>
<td>The Proposed Model</td>
<td>206</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Reservoir Production History and Pressure Data</td>
<td>208</td>
</tr>
<tr>
<td>6.3.2</td>
<td>Data Preparation</td>
<td>213</td>
</tr>
<tr>
<td>6.3.3</td>
<td>The Architecture of a Neural Network</td>
<td>216</td>
</tr>
<tr>
<td>6.4</td>
<td>Average Reservoir Pressure ANN Model</td>
<td>219</td>
</tr>
<tr>
<td>6.5</td>
<td>Estimation of Average Reservoir Pressure</td>
<td>232</td>
</tr>
<tr>
<td>6.6</td>
<td>Conclusions</td>
<td>235</td>
</tr>
<tr>
<td>6.7</td>
<td>References</td>
<td>236</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>238</td>
</tr>
<tr>
<td>7.2</td>
<td>Genetic Algorithm</td>
<td>239</td>
</tr>
<tr>
<td>7.3</td>
<td>A hybrid Genetic Algorithm</td>
<td>239</td>
</tr>
<tr>
<td>7.4</td>
<td>BP-ANN-GA Model Building</td>
<td>241</td>
</tr>
<tr>
<td>7.4.1</td>
<td>Evolving PB-ANN weights using GA</td>
<td>242</td>
</tr>
<tr>
<td>7.5</td>
<td>Prediction of Average Reservoir Pressure</td>
<td>244</td>
</tr>
<tr>
<td>7.6</td>
<td>Conclusion</td>
<td>244</td>
</tr>
<tr>
<td>7.7</td>
<td>References</td>
<td>250</td>
</tr>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>251</td>
</tr>
<tr>
<td>8.2</td>
<td>Historical Production Data</td>
<td>252</td>
</tr>
<tr>
<td>8.3</td>
<td>LS-SVM Model</td>
<td>255</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Modeling Process</td>
<td>256</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

Figure 2.1 – Fluid flow of Oil and Gas from the Reservoir to the Separator and Stock-Tank (Two-Stage Separation) ................................................................. 18
Figure 2.2 – Error Distribution for Bubble point Pressure Correlation ............... 31
Figure 2.3 – Error Distribution for $B_{ob}$ Correlation ........................................... 32
Figure 2.4 – A 45° Straight Line Cross Plot for Bubble point Pressure Correlation .... 33
Figure 2.5 – A 45° Straight Line Cross Plot for Bubble point Oil FVF Correlation .... 34
Figure 2.6 – Evaluation and Validation of Bubble Point Oil FVF Correlation .......... 38
Figure 3.1 – Residual Distributions of $R_{sob}$ Obtained from Eq. (3.7) (431 Data Points) .... 55
Figure 3.3 – Evaluation Test; Estimated vs. Actual $R_{sob}$ (355 Data Points) ............ 59
Figure 3.4 – Evaluation Test; Relative Error Distributions (355 Data Points) .......... 60
Figure 3.5 – Validation Test; Actual vs. Estimated $R_{sob}$ (49 Data Points) ............... 64
Figure 3.6 – Validation Test; Relative Error Distributions (49 Data Points) .......... 65
Figure 4.1 – Cross Plot of Modeled Bubble point Pressure on Training Dataset versus Laboratory PVT Values ........................................................................ 87
Figure 4.2 – Cross-plot of Modeled Bubble point Pressure on Validation Dataset versus Laboratory PVT Analysis ................................................................. 88
Figure 4.3 - Cross-plot of Modeled Bubble point Pressure on Testing Dataset versus Laboratory PVT Analysis ................................................................. 89
Figure 4.4 - Cross-plot of Modeled Bubble point Pressure on Overall Points versus Laboratory PVT Analysis ................................................................. 90
Figure 4.5 – Deployment Test: Cross-plot of Modeled Bubble point versus Laboratory PVT Analysis .................................................................................. 92
Figure 4.6 – $B_{ob}$ BP-Model Architecture, [4-5-4-1] ............................................... 97
Figure 4.7 (a, b) – Graphical Error Analysis on Training Dataset ......................... 105
Figure 4.7 (c, d) – Graphical Error Analysis on Training Dataset ......................... 106
Figure 4.7 (e, f) – Graphical Error Analysis on Training Dataset ......................... 107
Figure 4.8 (a, b) – Graphical Error Analysis on Validation Dataset ....................... 108
Figure 4.8 (c, d) – Graphical Error Analysis on Validation Dataset ....................... 109
Figure 4.8 (e, f) – Graphical Error Analysis on Validation Dataset ....................... 110
Figure 4.9 (a, b) – Graphical Error Analysis on Testing Dataset ............................. 111
Figure 4.9 (c, d) – Graphical Error Analysis on Testing Dataset ............................. 112
Figure 4.10 (a, b) – Graphical Error Analysis on Overall Dataset .......................... 114
Figure 4.10 (c, d) – Graphical Error Analysis on Overall Dataset .......................... 115
Figure 4.10 (e, f) – Graphical Error Analysis on Overall Dataset .......................... 116
Figure 4.11 – Deployment Test of $B_{ob}$ BP-ANN Model ....................................... 119
Figure 4.13 – Cross-plot of Modeled $R_{sob}$ on Training Dataset vs. Laboratory PVT Values ........................................................................................................... 130
Figure 4.14 – Cross-plot of Modeled $R_{sob}$ on Validation Dataset vs. Laboratory PVT Values ........................................................................................................... 131
Figure 4.15 – Cross-plot of Modeled $R_{sob}$ on Testing Dataset vs. Laboratory PVT Values ........................................................................................................... 132
Figure 4.16 – Cross-plot of Modeled $R_{sob}$ on Overall Data Points vs. Laboratory PVT Values ........................................................................................................... 133
Figure 5.19 - Contour Plot of the Optimization Error of Cross Validation of $R_{ST}$ when Optimizing the Parameters $\gamma \left[ \exp(3.864) \right]$ and $\sigma^2 \left[ \exp(0.298) \right]$. The Black Square Indicates the Selected Optimal Settings............................ 195

Figure 5.20 – Performance of the Optimal $R_{ST}$ LS-SVM Model in the Training Dataset ................................................................. 196

Figure 5.21 – Error Distribution Resulting from $R_{ST}$ LS-SVM Model .................. 197

Figure 5.22 – Performance of $R_{sob}$ Calculated from $R_{ST}$ LS-SVM Model ............ 200

Figure 5.23 – Performance of $R_{sob}$ Calculated from $R_{ST}$ LS-SVM Model using Testing Dataset................................................................. 201

Figure 6.1 – Impact on the Decline of an Average Reservoir Pressure According to the Oil Production Rate.......................................................... 209

Figure 6.2 – Impact on the Decline of an Average Reservoir Pressure According to the Gas Production Rate ............................................................... 210

Figure 6.3 – Impact on the Decline of an Average Reservoir Pressure According to the Water Production Rate............................................................... 211

Figure 6.4 – Impact on the Decline of an Average Reservoir Pressure According to the Number of Producers............................................................. 212

Figure 6.5 – Actual Average Reservoir Pressure versus Modeled ANN Values for Training Dataset........................................................................ 225

Figure 6.6 – Actual Average Reservoir Pressure versus Modeled ANN Values for Validation Dataset................................................................. 226

Figure 6.7 – Actual Average Reservoir Pressure versus Model ANN Values for Testing Dataset........................................................................ 227

Figure 6.8 – Actual Average Reservoir Pressure versus Modeled ANN Values for Overall Dataset................................................................. 228

Figure 6.9 – Error Distribution for Training Dataset............................................... 230

Figure 6.10 – Error Distribution for Overall Dataset ............................................... 231

Figure 6.11 – ANN model performance .................................................................... 233

Figure 7.1 – Modeled Versus Field Average Reservoir Pressure .................................. 247

Figure 8.1 - Contour Plot of the Optimization Error (MSE) for LS-SVM Regression Model of Calculating Average Reservoir Pressure, The Black Square Indicates the Selected Optimal Settings of $\gamma \left[ \exp(2.837) \right]$ and $\sigma^2 \left[ \exp(1.418) \right]$........................................ 260

Figure 8.2 – Histogram of the Errors of the LS-SVM Average Reservoir Pressure Model .................................................................................................. 262

Figure 8.3 – Performance of the Optimal LS-SVM Model in the Training Dataset ...... 263

Figure 8.4 – Performance of the LS-SVM Model. Both training and Testing Dataset are Scattered around the $y = x$ Line indicating No Overfit Existing in the Model ...... 264

Figure 8.5 – LS-SVM Model Performance. The Plot can be used to interpolate the missing average reservoir pressure values. ......................................................... 265

Figure 9.1– Reservoir Oil Production Performance from Sept. 1978 to July 2010...... 275

Figure 9.2– ANN Model Output versus the Actual Reservoir Production (1st Dataset) 279

Figure 9.3– Performance of the ANN Model on Training Dataset ............................. 280

Figure 9.4– Performance of the ANN Model on Validation Dataset ........................... 281

Figure 9.5– Performance of the ANN Model on Testing Dataset ............................... 282
Figure 9.6 – Prediction of average reservoir oil production (one-year-ahead from Aug. 2009 to Jul. 2010) ................................................................................................................................. 284
Figure 10.1 - Contour Plot of the Optimization Error for LS-SVM Regression M1 when Optimizing the Parameters. The red square indicates the selected optimal Settings of \( \gamma = \exp(3.53) \) & \( \sigma^2 = \exp(4.21) \) ................................................................................................................. 293
Figure 10.2 – History Matching of LS-SVM M1 .................................................................................................................. 294
Figure 10.3 - Cross Plot of Modeled Oil Production Rates vs. Field Values for M1 .... 295
Figure 10.4 - Flow Chart of Load Forecasting ................................................................................................................................. 297
Figure 10.5 – Cross Plot of Predicted Oil Production Rates vs. Field Values for M1 ... 298
Figure 10.6 - Contour Plot of the Optimization Error for M2 Model when Optimizing the Parameters. The red square indicates the selected optimal settings of \( \gamma = \exp(7.201) \) & \( \sigma^2 = \exp(5.532) \) .................................................................................................... 300
Figure 10.7 – History Matching of M2 ........................................................................... 302
Figure 10.8 – Correlation between Modeled and Field Oil Flow Rates for M2 ............. 303
Figure 10.9 – Forecasting Oil Production against Field Values for M2 Model ............. 304
Figure 10.10 – Illustrates the Comparison of ARE between the Two Models in Forecasting Data Domain ....................................................................................... 306
Figure A.1 - Different Features in Residual Plots, Where the Residuals Plotted versus fitted values (x-axis). After Draper and Smith (1998)............................................... 320
Figure A.2 - Further Different Features in Residual Plots.................................................................................................................. 322
Figure A.3 - Histogram of the Residuals. The Residuals Close Normal Distribution.... 324
Figure A.4 - Tests of the Normality. The Histogram and the Normal Probability Plot in (c and e) Show that the Error Terms are Normally Distributed........................................ 325
Figure A.5 - Box Plot ...................................................................................................... 327
Figure A.6 - ANOVA Table ........................................................................................... 331
Figure A.7 - Different Features in Correlation Coefficient ............................................ 333
Figure A.8 - Probability Density Function Graph of a T-distribution. The P-value is the sum of the two shaded areas under the curve. ...................................................... 336
Figure B.1- Schematic of Typical Multilayer Neural Network Model. It consists of an input layer with four neurons \( (x_{p1}, x_{p2}, x_{p3}, \) and \( x_{p4} \)), one hidden layer with five neurons \( (h_{n1}, h_{n2}, h_{n3}, h_{n4}, \) and \( h_{n5} \)) and an output layer with single neuron, \( y \) .... 346
Figure B.2- A Neural Net Perceptron ............................................................................. 350
Figure B.3- Illustration of the Early-stopping Rule. The trained network is saved at the point defined by the early-stopping criterion, and training continues exclusively to check whether the error will fall again. This will ensure that the increase in the validation error is not a temporary event. .................................................................................. 352
Figure B.4- (a) Hypothetical Surface in Weight Space with Many Local Minimums. (b) Cross-section of a Hypothetical Error Surface in Weight Space............................... 356
Figure C.1 - Illustration of the Key Ingredient of LS-SVM: Transformation \( \Phi \) of the input data to the feature space................................................................................................. 360
Figure C.2 – LS-SVM Modeling Procedure........................................................................ 364
Figure C.2 – One-iteration of a 3-fold Cross Validation Algorithmic Pattern .............. 366
LIST OF TABLES

Table 2.1 - Range of Data in Group A ................................................................. 27
Table 2.2 - Quantitative Statistical Error Analysis .............................................. 30
Table 2.3 - Data Range of Group B ................................................................. 36
Table 2.4 - Comparison of Proposed Bubble point Oil FVF Correlation ............. 37
Table 3.1 - Range of Data Points for $R_{sob}$ Correlation (431 Data Points) ...... 51
Table 3.2 - Range of Data points for $R_{ST}$ Correlation (355 Data Points) ...... 52
Table 3.3 - Quantitative Statistical Analysis of the Present Correlations ........ 54
Table 3.4 - Statistical Analysis of Evaluation Test (355 Data Points) ............... 58
Table 3.5 - Range of Data used for Validation Test (49 Data Points) .............. 62
Table 3.6 - Statistical Accuracy of Validation Test (49 Data Points) ............. 63
Table 4.1 - Ranges of Datasets used in Training, Validation, Testing and Deployment $P_b$ ANN Model ................................................................. 81
Table 4.2 - Comparison of $P_b$ Network Architectures - The 5 Best Architectures .... 84
Table 4.3 - Average Error during Training, Validation and Testing $P_b$ Model .. 85
with Various Training Algorithms (Quasi-Newton (Q-N), Conjugate Gradient Descent (CGD), Levenberg-Marquardt (L-M) and Limited Memory Quasi-Newton (LM Q-N)) ................................................................. 85
Table 4.4 - Ranges of $P_b$ ANN Model Output with Statistical Error Analysis .... 86
Table 4.5 - Statistical Error Analysis for Deployment Test ................................ 93
Table 4.6 - Ranges of Datasets used in Training, Validation, Testing and Deployment of the $B_{ob}$ ANN Model ................................................................. 95
Table 4.7 - Comparison of $B_{ob}$ Network Architectures - The 5 Best Architectures .... 98
Table 4.8 - Average Error during Training, Validation and Testing of Architecture [4-5-4-1] with Various Training Algorithms and Input/Output Activation Functions. (Illustrates the Effect of Choice of Training Algorithm and Activation Function) 103
Table 4.9 - Comparison between the Four Best Network Models ..................... 104
Table 4.10 - Statistical Error Analysis for Deployment Test ......................... 120
Table 4.11 - Ranges of Datasets used in Training, Validation, Testing and Deployment $R_{sob}$ ANN Model ................................................................. 121
Table 4.12 - Comparison of $R_{sob}$ Network Architectures - The 5 Best Architectures .... 124
Table 4.13 - Average Error during Training, Validation and Testing $R_{sob}$ Model with Various Training Algorithms (Quick Propagation (QP), Quasi-Newton (Q-N), Conjugate Gradient Descent (CGD), and Limited Memory Quasi-Newton (LM Q-N)) ................................................................. 127
Table 4.14 - Statistical Error Analysis of $R_{sob}$ ANN Model Output ............... 128
Table 4.15 - Statistical Error Analysis for Deployment Test ......................... 140
Table 4.16 - Comparison of $R_{ST}$ Network Architectures - The Best Architectures ...... 143
Table 4.17 - Average Error during Training, Validation and Testing $R_{ST}$ Model with Various Training Algorithms (Quasi-Newton (Q-N), Conjugate Gradient Descent (CGD), and Limited Memory Quasi-Newton (LM Q-N)) ................................................................. 144
Table 4.18 - Ranges of $R_{ST}$ ANN Model Output with Statistical Error Analysis .... 145
Table 5.1 - Ranges of PVT Data .................................................................. 166
Table 5.2 - Statistical Error analysis of $P_b$ LS-SVM Model ....................... 169
Table 5.3 - Statistical Error Analysis of $B_{ob}$ LS-SVM Model ...................... 178
Table 5.4 – Statistical Error Analysis of $R_{sob}$ LS-SVM Model .............................................. 188
Table 5.5 – Ranges of PVT Data used in $R_{ST}$ LS-SVM Model .............................................. 194
Table 5.6 – Error of Analysis of $R_{sob}$ Obtained by Adding LS-SVM $R_{ST}$ and Field $R_{SP}$ 199
Table 6.1 - Range of Overall Data................................................................................... 214
Table 6.2 – Range of Data Used in Training Process....................................................... 215
Table 6.3 - Network Architecture Parameters ................................................................. 217
Table 6.4 - Top Five Architectures - The best results are emphasized in bold................. 218
Table 6.5 - Error Analysis of Architecture [4-10-1] under Various Training Algorithms ................................................................................................................................. 220
Table 6.6 - Error analysis of Architecture [4-8-3-1] under Various Training Algorithms ................................................................................................................................. 221
Table 6.7 - Error Analysis of Architecture [4-8-3-1] under Various Training Algorithms ................................................................................................................................. 222
Table 6.8 - Error Analysis of Architecture [4-8-1-4-1] under Various Training Algorithms ................................................................................................................................. 223
Table 6.9 - Error Analysis of Architecture [4-7-6-4-1] under Various Training Algorithms ................................................................................................................................. 224
Table 6.10 – Statistical Error Analysis of the Proposed Model........................................... 229
Table 6.11 – Model Prediction Performance .................................................................. 234
Table 7.1 – BP-ANN-GA Model Parameters ................................................................. 246
Table 7.2 – Statistical Error Analysis of the Proposed Model ........................................ 248
Table 7.3 – Model Prediction Performance .................................................................. 249
Table 8.1 - Average Reservoir Pressure Data; Testing Dataset is emphasized in Bold ................................................................................................................................. 254
Table 8.2 – Statistical Error Analysis of the Proposed Model ........................................ 266
Table 8.3 – LS-SVM Model Forecasting Performance .................................................... 267
Table 9.1 – Error Analysis of the Proposed Model ......................................................... 283
Table 10.1 – Statistical Error Analysis of M1 and M2 LS-SVM Models .................... 305
LIST OF APPENDICES

A.1 Introduction ........................................................................................................................................ 315
A.2 Simple Linear Regression Model ........................................................................................................ 316
A.3 Correlation and Regression Modeling in the Oil and Gas Industry ...................................................... 316
A.4 Correlation and Regression Modeling Building Procedure ................................................................. 317
A.5 Multiple Linear Regression Model Assumptions ..................................................................................... 318
A.5.1 Linearity - Assumption (1) ................................................................................................................ 318
A.5.2 Independence of Error terms - Assumption (2) ................................................................................ 319
A.5.2.1 Standard Error of Estimate ............................................................................................................. 320
A.5.3 Normality - Assumption (3) .............................................................................................................. 321
A.5.3.1 Outliers, Leverage and Influential Points in Regression ................................................................ 322
A.5.4 Multicollinearity - Assumption (4) .................................................................................................. 328
A.5.4.1 Coefficient of Determination $R^2$ ................................................................................................ 328
A.5.4.2 Adjusted $R^2$ ................................................................................................................................. 329
A.5.4.3 Correlation Coefficient among Variables, Cor$(X_1,X_2)$ ............................................................ 330
A.6 Stepwise Regression (Variables Selection): ............................................................................................ 331
A.7 T-statistics and P-value ....................................................................................................................... 332
A.8 Analysis of Variance (ANOVA Table) .................................................................................................. 333
A.9 Comparing Models ............................................................................................................................... 334
A.10 References ........................................................................................................................................ 335

A.10.1 Regression Analysis ....................................................................................................................... 336
A.10.2 Regression Analysis in Seismic ........................................................................................................ 337
A.10.3 Regression Analysis in Petrophysical .............................................................................................. 338
A.10.4 Regression analysis in Natural Gas Engineering ........................................................................... 339
A.10.5 Regression analysis in Relative Permeability ............................................................................... 339
A.10.6 Regression Analysis in Reservoir Recovery Factor ...................................................................... 340
A.10.7 Regression Analysis in Drilling Engineering .................................................................................. 340
A.10.8 Regression analysis in Flow in Porous Media .............................................................................. 341

B.1 Introduction ....................................................................................................................................... 342
B.2 Backpropagation Neural Network ...................................................................................................... 343
B.3 Mathematic Description of Backpropagation Algorithm ..................................................................... 344
B.4 Design of Neural Networks Architecture ............................................................................................ 345
B.4.1 Number of Input Variables .............................................................................................................. 345
B.4.2 Data Pre-processing ......................................................................................................................... 346
B.4.3 Type of Neural Network Architecture ............................................................................................ 346
B.4.4 Number of Hidden Layers and Hidden Nodes per Hidden Layer .................................................. 347
B.4.5 Activation Function .......................................................................................................................... 348
B.4.6 Type of Optimization Technique .................................................................................................. 349
B.5 Selections of the Best Architecture .................................................................................................... 349
B.6 Model Verification ............................................................................................................................... 350

C.1 Introduction ....................................................................................................................................... 351
C.2 LS-SVM for Function Approximation ................................................................................................... 352
C.3 LS-SVM Regression Modeling Procedure ............................................................................................ 353
C.4 Leave-one-out Cross-validation for LS-SVM ...................................................................................... 354
C.5 Grid-search Technique ......................................................................................................................... 355
C.6 Statistical Measures for the Evaluation of Prediction Performance ................................................... 356

xvi
## NOMENCLATURE

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>ANN-GA</td>
<td>Artificial Neural Network with Genetic Algorithm</td>
</tr>
<tr>
<td>AE</td>
<td>Absolute Error</td>
</tr>
<tr>
<td>ARE</td>
<td>Absolute Percent Relative Error, %</td>
</tr>
<tr>
<td>BOPD</td>
<td>Barrels Oil Per Day</td>
</tr>
<tr>
<td>BP-ANN</td>
<td>Back-Propagation Artificial Neural Network</td>
</tr>
<tr>
<td>CGD</td>
<td>Conjugate Gradient Descent algorithm</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>GOR</td>
<td>Gas-Oil Ratio, SCF/STB</td>
</tr>
<tr>
<td>Q-N</td>
<td>Quasi-Newton algorithm</td>
</tr>
<tr>
<td>Qg</td>
<td>Average reservoir gas production</td>
</tr>
<tr>
<td>Qo</td>
<td>Average reservoir oil production</td>
</tr>
<tr>
<td>Qw</td>
<td>Average reservoir water production</td>
</tr>
<tr>
<td>L-M</td>
<td>Levenberg-Marquardt algorithm</td>
</tr>
<tr>
<td>LM Q-N</td>
<td>Limited Memory Quasi-Newton algorithm</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>LS-SVM</td>
<td>Least Squares Support Vector Machine</td>
</tr>
<tr>
<td>MARE</td>
<td>Mean Absolute Relative Error</td>
</tr>
<tr>
<td>RE</td>
<td>percent relative error</td>
</tr>
<tr>
<td>RMARE</td>
<td>Root Mean Absolute Relative Error</td>
</tr>
<tr>
<td>MRE</td>
<td>Mean Relative Error</td>
</tr>
<tr>
<td>MRA</td>
<td>Multiple Regression Analysis</td>
</tr>
<tr>
<td>SD</td>
<td>Standard Deviation</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>PVT</td>
<td>Pressure Volume Temperature</td>
</tr>
<tr>
<td>STB</td>
<td>Stock Tank Barrel</td>
</tr>
<tr>
<td>SCF</td>
<td>Standard Cubic Foot</td>
</tr>
</tbody>
</table>
psi  Pounds per Square Inch
psia Pounds per Square Inch Absolute
ln Natural Logarithm
API API Stock Tank Oil Gravity, degree
$B_{ob}$ Bubble point oil formation volume factor, bbl/STB
$P_{avg}$ Average reservoir pressure, psi
$p_b$ Bubble point pressure, psia
$P_R$ Reservoir pressure, psi
$P_{SP}$ Separator pressure, psia
$R^2$ Correlation coefficient, 100%
$R_{sob}$ Bubble point solution gas oil ratio, SCF/STB
$R_{so}$ Solution gas oil ratio, SCF/STB
$R_{sFb}$ Flash bubble point solution GOR, SCF/STB
$R_{sDb}$ Differential bubble point solution GOR, SCF/STB
$R_{SP}$ Separator gas oil ratio, SCF/STB
$R_{ST}$ Stock-tank vent gas oil ratio, SCF/STB
$T_R$ Reservoir temperature, °F
$T_{SP}$ Separator temperature, °F
$T_{ST}$ Stock-tank temperature, °F
$\gamma$ LS-SVM regularization parameter
$\gamma_g$ Gas specific gravity, air = 1
$\gamma_{gSP}$ Separator gas specific gravity, air =1
$\gamma_{gST}$ Stock-Tank gas specific gravity, air = 1
$\gamma_{gTotal}$ Total gas specific gravity, air = 1
$\gamma_{oST}$ Stock-Tank oil specific gravity, water = 1
$\sigma^2$ kernel parameter for LS-SVM
CHAPTER I

INTRODUCTION

1.1 Background

Function approximation means simulating a mathematical function with a model such as regression analysis, Artificial Neural Networks (ANNs) and Support Vector Machine (SVM). The need for function approximations arises in many branches of engineering, and in particular, petroleum engineering. The field of petroleum engineering is filled with partial differential equations, nonlinear systems of equations, and uncertainty analysis.

Many problems in the oil and gas industry are extremely difficult, if not impossible, to solve using the existing analytical tools, and the same problems prove to be difficult to solve numerically. As an example, the partial differential equation (Equation 1.1) that describes the constant flow of a fluid in reservoir rocks, and is independent of fluid type is:

\[ \frac{1}{r} \frac{\partial}{\partial r} \left( r \frac{\partial p}{\partial r} \right) = \frac{\partial}{\partial t} (\Phi \rho) \]  \hspace{1cm} (1.1)

However, to derive this equation, several assumptions are made to significantly reduce the complexity of the problem. The assumptions are:

1. Darcy flow exists, i.e., no turbulence effects exist.
3. Gravitational effects are negligible.

4. Reservoir is homogeneous and isotropic (spatially invariant).

5. Permeability and porosity are constant.

6. Flow is isothermal (of constant temperature).

Due to the restrictions imposed by the aforementioned assumptions, it would appear that the diffusivity equation would never apply to actual reservoirs. If it is assumed this is true, and if the limitations are kept in mind, the results obtained using Equation (1.1) are within acceptable engineering accuracies (Chen at all, 2006). Thus, even when the exact representation is known, an approximation may yield a sufficiently accurate solution. The need for function approximations arises in almost all branches of the oil and gas industry.

Another example in reservoir engineering is a type of complicated data, formally defined by \( y = f(x, t) \) in which \( x \) and \( y \) are vectors and \( t \) denotes time. The oil and gas production rates change as a function of the passing of time and, therefore, can be viewed as a function of time and spatial location for certain wells and production operations. If the function \( y = f(x, t) \) is approximated on the basis of known data, the production prediction could be then performed at any time. In general, a function approximation problem asks that a function among a well-defined class that closely matches a target function (value of the function or output) in a task-specific way be selected. However, the task of function approximation is to find an estimate of the unknown function which then finds the underlying relationship from given input-output data.
1.2 Function Approximation

Consider a given set of training samples \( \{x_k, y_k\}_{k=1}^{N} \), in which \( x_k \) is the input vector (predictors) and \( y_k \) is the corresponding target value (output or response) for sample \( k \). The goal of function approximation is to learn the underlying relation between the input and the target value. Following the retrieval of this information, a function approximation implements a mapping \( x \rightarrow \hat{y}(x) \) that can be evaluated for any \( x \). The performance of a function approximation is measured by means of the approximation error \( e_k \) which is defined as:

\[
e_k = y_k - \hat{y}(x_k)
\]

Appendixes A, B and C summarize the theory of three learning mechanisms, namely MRA, ANN and LS-SVM and the appropriate way of applying each method. In all cases, the approximation target is a function of the input variables called “a model”. Thus, function approximations include any techniques with which to model and analyze several variables searching for a relationship between a dependent variable and one or more independent variables. Regression analysis, artificial neural networks, and least squares support vector machine modeling techniques, for example, can be applied.

1.2.1 Regression Analysis

Regression analysis is a statistical method which employs algebraic formulas to estimate the value of a continuous random variable (dependent variable or output) using the value independent variables (inputs). It could be simple or multiple regressions. A simple regression is one in which a single independent variable (just one input) is used to
determine a dependent variable. The MRA method learns a relation between a set of indicators \( f_k(x), k = 1\ldots n \) and the output \( \hat{y} \) is

\[
\hat{y} = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \ldots + \alpha_n f_n(x)
\] (1.3)

In this equation, the \( \alpha \)'s are the \( n \) parameters which receive a value during the training by minimizing the summed squared approximation error over all examples. The indicators can be non-linear functions of the input vectors.

**1.2.2 Artificial Neural Network**

Artificial Neural Network (ANN), Or Neural Network (NN), is a nonlinear statistical data modeling tool. It consists of an interconnected group of artificial neurons, and it processes information using a connectionist approach to computation. The popularity of the ANN method is growing rapidly in the oil and gas industry. The multilayer feed forward neural network is one of the most widely used forms of ANN and is capable of modeling the unknown input/output relation of a variety of complex systems.

**1.2.3 Least squares support vector machines**

Least Squares Support Vector Machines (LS-SVM) are least squares versions of Support Vector Machines (SVM), a class of kernel based learning methods used for regression analysis and classification. In LS-SVM the solution is found by solving a set of linear equations instead of a convex quadratic problem for classical SVMs.
1.3 Key Ongoing Challenges

Key Ongoing Challenges highlight the area of reservoir engineering that describes the key individual challenges that affect reservoir management and their ability to implement new strategies. The narratives in this brief focus on three key sets of issues, i.e. reservoir fluid properties, average reservoir pressure and oil production prediction.

This research regarding the application of MRA, ANN and LS-SVM, in order to address several of the most important ongoing challenges in reservoir engineering, is innovative because it presents a very valuable tool to petroleum engineering. This work is poised to support the oil and gas industry, especially in the areas of reservoir management and field development.

1.3.1 Why Reservoir Fluid Properties?

The physical and chemical properties of reservoir oils vary considerably and are dependent upon the concentration of the various types of hydrocarbons and minor constituents present in the well. Reservoir fluid studies are essentially based on Pressure-Volume-Temperature (PVT) analysis. This analysis consists of a series of laboratory procedures designed to provide the values of the reservoir fluid properties as required in material balance calculations, well test analysis, reserve estimates, inflow performance calculations, and numerical reservoir simulation.

In the absence of such experimental analysis, empirical experimental Pressure-Volume-Temperature (PVT) correlations or artificial neural network models can be used
to estimate reservoir fluid properties. The reasons for using empirical PVT correlations could be due to one or more of the following:

- economic issues
- poor sample quality due to non-representative fluid, human error during sampling or field transfers
- insufficient sample volume to obtain a complete analysis and/or
- errors in laboratory analysis

A literature survey has shown that very little attention is paid to estimating reservoir fluid properties in the absence of experimental Pressure-volume-temperature (PVT) lab analysis. Until recently, not one specific correlation has been published to directly estimate the bubble point pressure ($P_b$) in the absence of PVT analysis and, at the moment, only one published correlation is available to estimate the bubble point of oil FVF ($B_{ob}$) directly in the absence of PVT analysis (Labedi 1990). The majority of the $P_b$ pressure and $B_{ob}$ correlations cannot be applied directly as the correlations of bubble point solution gas/oil ratio ($R_{sob}$) must be known along with gas specific gravity ($\gamma_g$) as part of the input variables, both of which are rarely measured field parameters. However, $P_b$ and $B_{ob}$ correlations are essentially based on the relationship illustrated in Equations 1.4 and 1.5, respectively. $R_{sob}$ and $\gamma_g$ can be obtained either experimentally or estimated from correlations. However, the oil and gas industry offers numerous correlations with which to estimate $R_{sob}$. The basic assumption of $R_{sob}$ correlations is expressed in Equation 1.6:
\[ P_b = f_1 \left( R_{sob}, \gamma_g, API \text{ gravity, } T_R \right) \]  \hspace{1cm} (1.4)

\[ B_{sob} = f_2 \left( R_{sob}, \gamma_g, API \text{ gravity, } T_R \right) \]  \hspace{1cm} (1.5)

\[ R_{sob} = f_3 \left( P_b, \gamma_g, API \text{ gravity, } T_R \right) \]  \hspace{1cm} (1.6)

Accordingly, in order to apply those correlations in the absence of PVT analysis, the \( P_b/R_{sob} \) and \( \gamma_g \) must be known in advance (they are rarely measured as field data).

1.3.2 Why Average Reservoir Pressure?

Insight into average reservoir pressure and its change in production (time), plays a critical role in reservoir development and management, economic evaluation, obtaining a hydrocarbon in place, computing rock and fluid characteristics, reservoir material balance calculations, pressure maintenance projects, surface and subsurface facility designs and predicting reservoir behaviour. In particular, almost every well intervention requires knowledge with respect to average reservoir pressure.

Traditionally, average reservoir pressure is obtained via a pressure build-up test which measures long-term built-up pressure when the producing wells are shut in. The test is initiated after the well is opened to a constant flow rate over an extended period of time. This will cause the reservoir pressure to drawdown due to the withdrawal of fluids from the well. Afterward, the well is shut-in to allow the pressure to build up. During the re-building of pressure, the pressure is recorded as function of time. Over an extended period of time, the build-up pressure reaches average reservoir pressure. This time period depends mainly upon the reservoir permeability and thus, may result in extensive time
consumption in low permeability reservoirs. The resulting pressure build-up is then analysed to determine reservoir permeability, flow efficiency, and average reservoir pressure. Currently, the only available method with which to obtain average reservoir pressure is to conduct an extended build-up test and it should be updated periodically because reservoir pressure changes as fluids (oil, gas and water) are released from the reservoir. However, a significant economic impact is caused by shutting in the producing wells during the entire build-up test.

1.3.3 Why Reservoir Oil Production Prediction?

The prediction of oil reservoir production performance has been an on-going challenge for engineers. It is an essential component of petroleum reservoir management. Traditionally, numerical simulations and decline curve analysis have been used to predict a reservoir’s future performance based on its current and past performance.

When an oil field is first discovered, a reservoir model is constructed utilizing a geological model. Once the oil field enters the production stage, many changes take place in the reservoir and thus, the changes must be reflected in the reservoir model. Usually, simulator programs are used to generate a model in order to simulate the reservoir. In reservoir simulation, the simulator takes a set of reservoir parameter values as inputs and returns a set of fluid flow information as outputs. The outputs are usually a time series over specific period of time. This time series is then compared with the historical production data to evaluate their match. Experts modify the input model parameters and re-run the simulation model. This process (history matching) continues until the model output rates match the historical production rates. History matching is the process of
updating reservoir descriptor parameters to reflect such changes, based on production data. The process of history matching is an inverse problem. In this problem, a reservoir is a black box with unknown parameter values. Given the rock, fluid properties and other production information collected from the field, the task is to identify these unknown parameter values such that the reservoir model gives flow outputs matching production data. Since the inverse problem does not have a unique solution, a large number of models with good history matching can be achieved. In other words, in the reservoir simulation process, more than one combination of reservoir parameter input values can give the same output. As a result, history matching is a challenging task for the following reasons:

- Highly non-linear flow equations with a large number of assumptions and input parameters
- Inverse problem (non-unique solution), where more than one model can produce acceptable outputs
- Inadequate results with high uncertainty in production forecasting
- Time consuming

On the other hand, decline curve analysis fits the observed production rates of individual wells, groups of wells or an entire reservoir using mathematical functions to predict future production by extrapolating the declining function. The basis of decline curve analysis is to match past production performance with a model, assuming production continues to follow the historical trend. In some cases, production decline data do not follow a particular model but crossover the entire set of curves. In addition,
decline curve analysis does not take into account opening or closing intervals and varying water or gas injection rates.

1.4 Goals of the Thesis

This thesis intends to solve the aforementioned challenging dilemmas by applying three different types of function approximation:

- MRA, ANN and LS-SVM methods are applied in order to develop four novel models with which to estimate the bubble point pressure \( (P_b) \), the bubble point oil formation volume factor \( (B_{ob}) \), bubble point solution gas/oil ratio \( (R_{sob}) \), and stock-tank vent gas/oil ratio \( (R_{ST}) \). These new models can be applied in a straightforward manner by using direct field data. Additional correlations or experimental analyses are unnecessary.

- A new alternative technique to predict and interpolate average reservoir pressure without shutting in the producing wells. BP-ANN, BP-ANN-GA and LS-SVM models are performed to interpolate the missing values, forecast and to predict current average reservoir pressure.

- Design a neural network and least square support machine models to predict oil reservoir production performance taking into account number of production and injection wells in service and varying water injection rates. The proposed models can serve as a practical and robust reservoir management tools.
1.5 Chapter by Chapter Overview

The core of this thesis is divided into nine chapters along with the Introduction (Chapter 1) and Conclusions (Chapter 11). Each chapter deals with the application of one type of the aforementioned function approximation techniques. Figure 1.1 illustrates the general structure of this thesis. The content of each chapter is organized as follows:

**Chapter 2** provides four multiple regression analysis models with which to estimate bubble point pressure and the bubble point formation volume factor. The correlations can be applied in the absence of PVT analysis to overcome the limitations faced by the published PVT correlations. Until recently, however, there has not been one specific correlation published to directly estimate bubble point pressure in the absence of PVT analysis. However, one published correlation (labedi 1990) to estimate oil formation volume factor at bubble point pressure directly in the absence of such laboratory analysis does exist.

**Chapter 3** extended from Chapter 2 and proposes another two novel empirical correlation to estimate the bubble point solution gas/oil ratio, and the stock tank vent gas/oil ratio. However, the majority of the available PVT empirical equations are used to estimate the solution gas/oil ratio at various temperatures rather than the reservoir temperature in order to design surface equipment and to study reservoir inflow performance. Since these correlations required prior knowledge of bubble point pressure and gas gravity, it is difficult to apply them in the absence of PVT analysis. Bubble point pressure and gas gravity can be obtained from either PVT analysis or can be estimated using correlations. They are rarely measured in the field.
Chapter 4 covers the use of ANN as an alternative technique to estimate the same three reservoir fluid properties and the stock-tank vent gas/oil ratio that are addressed in Chapter 2 and 3. The four ANN models are proposed as a function of readily available field data with no further requirement of additional correlations or experimental PVT analysis.

Chapter 5 describes a LS-SVM approach to design four models in order to predict the same fluid properties mentioned in Chapters 2 and 3 which would increase accuracy and improve the function approximation of reservoir fluid properties.

Chapter 6 introduces a new technique to interpolate and estimate current average reservoir pressure. A feedforward backpropagation neural network model was established to map the relationship which controls reservoir oil, gas and water production performance in order to interpolate and estimate the current average reservoir pressure without closing the producing wells. This method is suitable for constant and variable flow rates. After interpolating, predicting and/or obtaining the current average reservoir pressure, this method can be utilized in almost all reservoir and production engineering studies.

Chapter 7 extends the work presented in Chapter 6 to a more generalization model. Since the backpropagation network is a gradient descent method, it tends to become stuck in the local minima of the error surface and thus, does not find the global minimum. A backpropagation with a genetic algorithm is proposed in this chapter for tuning the link switches and weight parameters between the input layer and hidden layer. The main advantage of evolutionary computation is that it performs a global exploration of the
search space avoiding being trapped in local minima which usually happens with local search procedures.

Chapter 8 presents a LS-SVM model, designed as an alternative method to model the reservoir pressure performance. Using this approach, the current average reservoir pressure can be interpolated and the current average reservoir pressure can be estimated without shutting-in the well.

Chapter 9 clarifies how ANN, as a function approximation, has the ability to predict reservoir oil production with impressive accuracy. In this way, the history matching problem which associates to traditional simulation techniques can be overcome. In addition, unlike traditional decline curve analysis, the designed ANN model can take into account the number of producers and injectors in service and the varying water injection rates. The training ANN model can serve as a practical and robust reservoir production management tool.

Chapter 10 presents an application of LS-SVM regression to predict reservoir oil production. The novelty in this chapter, however, is the introduction of LS-SVM as new alternative tool to predict and interpolate average reservoir pressure without closing the producing wells.

Additionally, four Appendixes are placed at the end of the thesis.

- **Appendix A** introduces the traditional multiple regression analysis technique. In this Appendix, there is a brief overview of the building of a regression model.

- **Appendix B** explains the plan of building a robust ANN model.
• Appendix C addresses the basic steps of applying a LS-SVM regression model.
• Appendix D describes the methods of error analysis and interprets the model’s accuracy.

1.6 References

Labedi, R.M., 1990. Use of Production Data to Estimate the Saturation Pressure, Solution GOR, and Chemical Composition of Reservoir Fluids. SPE 21164 presented at the SPE Latin American Petroleum Conference held in Rio de Janeiro, October 14-19.
Figure 1.1: Structure of the thesis. Arrows suggest the reading order of the chapters. The main contributions are indicated in dashed box.
CHAPTER II

PREDICTION OF BUBBLE POINT PRESSURE, BUBBLE POINT OIL FORMATION VOLUME FACTOR IN THE ABSENCE OF PVT ANALYSIS BY UTILIZING MULTIPLE REGRESSION ANALYSIS

2.1 Introduction

The physical and chemical properties of reservoir oils vary considerably and are dependent upon the concentration of the various types of hydrocarbons and minor constituents present in the reservoir. Reservoir fluid studies are essentially based upon Pressure-volume-temperature (PVT) analysis. This analysis consists of a series of laboratory procedures designed to provide the values of the reservoir fluid properties as required in material balance calculations, well test analysis, reserve estimates, inflow performance calculations, and numerical reservoir simulation.

Ideally, reservoir fluid properties should be determined from laboratory studies with respect to live oil samples collected from the bottom of the wellbore or from the surface. Standard reservoir PVT fluid studies are designed to simulate the simultaneous fluid flow of oil and gas from the reservoir to the surface. The production path of reservoir fluids from the reservoir to the surface is simulated in the laboratory at reservoir temperature. During this process, the bubble point pressure $P_b$ is measured. Likewise, the oil volumes
and the amount of gas released are measured and used to determine the oil formation volume factor (oil FVF) $B_o$, the gas formation volume factor (gas FVF) $B_g$, solution gas/oil ratio $R_s$, oil compressibility $C_o$, oil density $\rho_o$, gas specific gravity $\gamma_g$, and gas compressibility factor, $z$ as functions of pressure. This is illustrated in Figure 2.1. Moreover, another PVT laboratory procedure is designed to measure oil viscosity as a function of pressure. PVT correlations are ultimately required in order to estimate reservoir fluid properties at temperatures other than the reservoir temperature. However, the PVT analysis is usually carried out at the reservoir temperature only. Reservoir fluid properties, at various temperatures other than the reservoir temperature, are required to design surface operation equipment and to study reservoir inflow performance operations. In such cases, even though PVT analysis is available, the required reservoir fluid properties must be estimated from the correlations.

In the absence of such experimental analysis, empirical PVT correlations can be used to estimate reservoir fluid properties. Reasons for using empirical PVT correlations could be:

- economic issues
- poor sample quality due to non-representative fluid, human error during sampling or field transfers
- insufficient sample volume to obtain a complete analysis and (4) errors in laboratory analysis
Figure 2.1 – Fluid flow of Oil and Gas from the Reservoir to the Separator and Stock-Tank (Two-Stage Separation)
Until recently, not one specific correlation has been published to directly estimate bubble point pressure in the absence of PVT analysis and, at the moment, there is just one published correlation available with which to estimate the bubble point oil FVF directly in the absence of PVT analysis (Labedi, 1990). The majority of the published bubble point pressure and bubble point oil FVF correlations cannot be applied directly because the correlations require prior knowledge of the bubble point solution GOR and gas specific gravity as part of the input variables. Both of which are rarely measured field parameters. Solution GOR and gas specific gravity can be obtained either experimentally or estimated from correlations. In this chapter, multiple regression analysis technique is applied in order to develop two novel correlations with which to estimate the bubble point pressure and the bubble point oil FVF. These new correlations can be applied in a straightforward manner by using direct field data. Separator GOR, separator pressure, stock-tank oil gravity and reservoir temperature are the only key parameters required to predict bubble point pressure and bubble point oil FVF. Additional correlations or experimental analyses are unnecessary.

2.2 Literature Review

PVT correlations are ultimately required in order to estimate reservoir fluid properties at temperatures other than the reservoir temperature. This is due to reservoir fluid properties being found in the PVT analysis report at the reservoir temperature only. Reservoir fluid properties, at various temperatures other than the reservoir temperature, are required to design surface operation equipment and to study reservoir inflow
performance operations. In such cases, even though PVT analysis is available, the required reservoir fluid properties must be estimated from the correlations.

Several correlations already exist within the oil and gas industry to estimate bubble point pressures ($P_b$) and bubble point oil FVF ($B_{ob}$) of reservoir oils. The correlations are essentially based on the assumption that $P_b$ and $B_{ob}$ are strong functions of bubble point solution GOR ($R_{sob}$), reservoir temperature ($T_R$), gas specific gravity ($\gamma_g$), and stock-tank oil specific gravity ($\gamma_{oST}$). Bubble point solution GOR can be obtained as the sum of the stock-tank vent GOR ($R_{ST}$) (seldom conducted field measurement) and the measured separator GOR ($R_{SP}$). This is valid only if the $R_{SP}$ and $R_{ST}$ are measured while the reservoir pressure is above the bubble point pressure. Sometimes, the sum of the two producing GORs is called “flash bubble point solution GOR” ($R_{sFb}$) or “total GOR”, Equation 2.1. Some correlations use differential bubble point solution GOR ($R_{sDb}$) rather than $R_{sFb}$. Examples include the correlations developed by Borden and Rzasa (1950), Knopp and Ramsey (1960), Vasques and Beggs (1980), Al-Marhoun (1988), Dokla and Osman (1992), Elsharkawy and Alikhan (1997), Almehaideb (1997), Hanafy et al. (1997), McCain et al. (1998), Velarde et al. (1999), Boukadi et al. (2002), Gharbi and Elsharkawy (2003) and Mazandarani and Asghari (2007). Others preferred to use flash bubble point solution GOR ($R_{sFb}$) e.g. Standing (1947), Lasater (1958), Tahmani (1967), Glaso (1980), Macary and Batanony (1992), Al-Marhoun (1992), Frashad et al. (1996), Petrosky and Farshad (1998) and Ikiensikimana and Oboja (2009).

$$R_{sob} = R_{sFb} = R_{SP} + R_{ST}$$ 

(2.1)
Moreover, several correlations developed by Labedi (1990), Rollins et al. (1990), Dokla and Osman (1992), Macary and Batanony (1992), Velarde et al. (1999), Petrosky and Farshad (1998) and McCain et al. (1998) use flash separator gas specific gravity ($\gamma_{gSP}$), while still others such as Standing (1947), Borden and Rzasa (1950), Lasater (1958), Elsharkawy and Alikhan (1997), Glasso (1980), Mazandarani and Asghari (2007) use total gas specific gravity. Other correlations use weight average specific gas gravity of the separator and stock-tank vent gas. Among the researchers are Al-Marhoun (1988), Frashad et al. (1996), Al-Marhoun (1997), Al-shammasi (1999), Hemmati and Kharrat (2007) and Ikiensikimama and Oboja (2009). Initially, some methods require adjusting gas gravity to separator pressure at 100 psig before they can be used in the correlations e.g. Vasquez and Beggs (1980). Others require compositional analysis as input variables e.g. Elsharkawy (2003).

Accordingly, in order to apply bubble point pressure and bubble point oil FVF correlations in the absence of PVT analysis, there must be prior knowledge of the bubble point solution GOR ($R_{sob}$), differential bubble point GOR ($R_{Db}$) or stock-tank vent GOR ($R_{ST}$) in addition to separator gas specific gravity ($\gamma_{gSP}$), total gas specific gravity ($\gamma_{gTotal}$), or weight average specific gas gravity. The aforementioned parameters are rarely measured using field data. A literature survey has shown little attention is paid to predicting reservoir fluid properties in the absence of PVT analysis. Currently, not one specific correlation has been published to directly estimate the bubble point pressure in the absence of PVT analysis and, at the moment, there is just one published correlation.
available to estimate the bubble point oil FVF (Labedi, 1990) directly in the absence of PVT analysis.

Labedi (1990) proposed a bubble point pressure correlation (Equation 2.2) based upon separator GOR \( R_{SP} \), separator temperature \( T_{SP} \), separator gas specific gravity \( \gamma_{gSP} \), stock-tank API gravity and reservoir temperature \( T_R \). Obviously, Labedi’s bubble point pressure correlation cannot be used directly because of the separator gas gravity. To work around the problem, Labedi proposed another correlation to estimate separator gas gravity (Equation 2.3). In the same work, Labedi proposed a new correlation to estimate bubble point oil FVF (Equation 2.4). Presently, this is the only existing correlation in the literature which can be applied directly to obtain \( B_{ob} \) in the absence of PVT analysis.

\[
p_b = \frac{6.0}{\gamma_{gSP}} \left[ \frac{R_{SP}^{0.6714} (T_R / \text{API})^{0.7097} T_{SP}^{0.0893}}{10^{0.00007995 R_{SP}}} \right] \quad (2.2)
\]

\[
\gamma_{gSP} = 0.0755 \times 10^{1.1938 / \sqrt{R_{SP}}} \text{API}^{0.0621} \left[ \frac{T_{SP}}{R_{SP}} \right]^{-0.0659} \quad (2.3)
\]

\[
B_{ob} = 0.9976 + 0.5273 \times 10^{-3} R_{SP} + 0.26636 \times 10^{-7} \text{API} \times P_{SP} (T_R - 60)
+ 0.16982 \times 10^{-4} \text{API} (T_R - 60) \quad (2.4)
\]

McCain (1991) provided guidance on the application of the PVT correlations. To estimate bubble point pressure and bubble point oil FVF, he suggested Sanding (1977)
correlations (Equations 2.5 and 2.6) in conjunction with Rollins *et al.* (1990) stock-tank vent GOR correlation, Equation 2.7.

\[ p_b = 18.2 \left( \frac{R_{sb}}{\gamma_{gTotal}} \right)^{0.83} \left( \frac{10}{10} \right)^{0.0091 T_p} - 1.4 \]  

(2.5)

\[ B_{ob} = 0.9759 + 12 \times 10^{-5} \left[ R_s \left( \frac{\gamma_{gTotal}}{\gamma_{oST}} \right)^{0.5} + 1.25 T \right]^{1.2} \]  

(2.6)

\[ \log R_{ST} = 0.3818 - 5.506 \log \gamma_{oST} + 2.902 \log \gamma_{gSP} + 1.327 \log P_{SP} - 0.7355 \log T_{SP} \]  

(2.7)

In short, McCain recommended the following steps to estimate bubble point pressure and bubble point oil FVF with the limitation of the separator temperature not exceeding 140 °F and the reservoir temperature not exceeding 325 °F:

1. estimate stock-tank GOR \((R_{ST})\) from Equation 2.7
2. add the estimated \(R_{ST}\) to field separator GOR (Equation 2.1) to obtain a bubble point solution GOR \((R_{sob})\) and
3. estimate the \(p_b\) and \(B_{ob}\) from Standing correlations (Equations 2.5 and 2.6, respectively). That is, use the \(R_{sob}\) obtained from step 2, and the \(\gamma_{gSP}\) instead of total gas specific gravity \((\gamma_{gTotal})\)
Nevertheless, this approach is impractical because the $\gamma_{SP}$ is a rarely measured field data and is usually based upon a gas composition analysis. In 2003, Valko and McCain revisited reservoir oil bubble point pressure. They derived new correlations to estimate bubble point pressure (Equation 2.8) and stock-tank vent GOR (Equation 2.9). The bubble point correlation is a function of $R_{sob}$, API, $\gamma_{SP}$ and $T_{R}$. The stock-tank vent GOR correlation is a function of $P_{SP}$, $T_{SP}$ and API.

$$\ln P_b = 7.475 + 0.713z + 0.0075z^2$$ (2.8)

where $z = \sum_{n=1}^{4} z_n$ and

$$z_n = C0_n + C1_n \text{VAR}_n + C2_n \text{VAR}_n^2 + C3_n \text{VAR}_n^3$$

<table>
<thead>
<tr>
<th>$n$</th>
<th>VAR</th>
<th>C0</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\ln R_{sob}$</td>
<td>-5.48</td>
<td>-0.0378</td>
<td>0.281</td>
<td>-0.0206</td>
</tr>
<tr>
<td>2</td>
<td>API</td>
<td>1.27</td>
<td>-0.0449</td>
<td>4.36*10^{-4}</td>
<td>-4.76*10^{-6}</td>
</tr>
<tr>
<td>3</td>
<td>$\gamma_{SP}$</td>
<td>4.51</td>
<td>-10.84</td>
<td>8.39</td>
<td>-2.34</td>
</tr>
<tr>
<td>4</td>
<td>$T_{R}$</td>
<td>-0.7835</td>
<td>6.23*10^{-3}</td>
<td>-1.22*10^{-5}</td>
<td>1.03*10^{-8}</td>
</tr>
</tbody>
</table>

$$\ln R_{ST} = 3.955 + 0.83z - 0.024z^2 + 0.075z^3$$ (2.9)

where $z = \sum_{n=1}^{3} z_n$ and

$$z_n = C0_n + C1_n \text{VAR}_n + C2_n \text{VAR}_n^2$$

<table>
<thead>
<tr>
<th>$n$</th>
<th>VAR</th>
<th>C0</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\ln P_{SP}$</td>
<td>-8.005</td>
<td>2.7</td>
<td>-0.161</td>
</tr>
<tr>
<td>2</td>
<td>$\ln T_{SP}$</td>
<td>1.224</td>
<td>-0.5</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>API</td>
<td>-1.587</td>
<td>0.0441</td>
<td>-2.29*10^{-5}</td>
</tr>
</tbody>
</table>
Just as with the McCain (1991) approach, the work of Valko and McCain (2003) has a practical limitation in the absence of PVT analysis because of $\gamma_{gSP}$.

### 2.3 Newly Developed $P_b$ and $B_{ob}$ Correlations

The main objective of this chapter is to overcome the limitations faced by previous correlations by building regression models using directly measured field parameters as input variables in order to estimate $P_b$ and $B_{ob}$. Two correlations are proposed as functions of four readily available field parameters ($R_{SP}$, $P_{SP}$, $\gamma_{oST}$ and $T_R$). By employing the four parameters, engineers can estimate the bubble point pressure and the bubble point oil FVF for crude oil straightforwardly in the absence of PVT analysis.

The PVT data used in this work were obtained from PVT analysis at two-stage and single-stage flash separation tests. A total of 118 reservoir fluid studies (476 data points) were collected from various Libyan oil fields in the Sirte Basin (AGOCO, Sirte Oil, Waha Oil, Millita Oil and Gas, Repsol Oil, and Harouge Oil, 1971-1995). The majority of the data points are taken from two-stage flash separation tests (355 data points). In the single-stage separation test, the separator pressure is atmospheric pressure and the stock-tank vent GOR value is equal to zero.

In order to study the validity of the proposed correlations, the 476 data points were randomly divided into two groups. Group A includes a total of 413 data points. Group B data (62 data points) was used to test the validity of the newly developed correlations. The range of values of Group A are presented in Table 2.1.
2.3.1 Bubble point pressure correlation

Numerous models were tried as regression equations. Equation 2.10 was found to be very accurate. The natural logarithm of bubble point pressure was regressed against the natural logarithms of separator GOR, separator pressure, stock-tank oil gravity and reservoir temperature.

\[ P_b = R_{SP}^{0.683} P_{SP}^{0.18} \gamma_{ST}^{-4.98} T_R^{0.658} \]  

(2.10)

Since the representativeness of a PVT study greatly depends upon sampling conditions, the first and most important operation, before running a complete reservoir fluid study, is to check the sample validity and select the most representative sample. The bottom-hole sample, used for PVT study, is selected according to the results obtained during the verification of the samples. It is worth mentioning that the proposed bubble point pressure correlation can be used to select the most representative sample with which to run a PVT analysis.
Table 2.1 - Range of Data in Group A

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bubble Point Pressure, $P_b$</td>
<td>psia</td>
<td>1734.4</td>
<td>1111.9</td>
<td>121</td>
<td>4244</td>
</tr>
<tr>
<td>Oil FVF, $B_o$</td>
<td>bbl/STB</td>
<td>1.3513</td>
<td>0.2035</td>
<td>1.064</td>
<td>1.795</td>
</tr>
<tr>
<td>Separator Pressure, $P_{SP}$</td>
<td>psia</td>
<td>119.58</td>
<td>126.8</td>
<td>14.7</td>
<td>735</td>
</tr>
<tr>
<td>Separator GOR, $R_{SP}$</td>
<td>SCF/STB</td>
<td>352.2</td>
<td>249.4</td>
<td>10</td>
<td>1256</td>
</tr>
<tr>
<td>Stock-Tank Oil Gravity, $\gamma_{ST}$</td>
<td>water=1</td>
<td>0.84052</td>
<td>0.02506</td>
<td>0.7999</td>
<td>0.921</td>
</tr>
<tr>
<td>Reservoir Temperature, $T_R$</td>
<td>°F</td>
<td>201.66</td>
<td>48.22</td>
<td>100</td>
<td>277</td>
</tr>
</tbody>
</table>
2.3.2 Bubble Point Oil FVF Correlation

Usually, the oil FVF obtained from a differential vaporization test should be adjusted using flash separation oil FVF to properly approximate a combination liberation system. However, at bubble point pressure, oil FVF ($B_{ob}$) is equal to flash separation oil FVF ($B_{oFb}$). Accordingly, by using a multiple regression analysis technique, the $B_{ob}$ was correlated as a function of $P_{SP}$, $R_{SP}$, $\gamma_{oST}$ and $T_R$. After trying many models, the following model was found to be a very good prediction equation of bubble point oil FVF.

$$B_{ob}=1.6624 + 0.000512R_{SP} + 0.00015P_{SP} – 0.802\gamma_{oST} + 0.000501T_R$$  (2.11)

2.4 Correlation Verification and Validation

Verification and validation of the correlation are the most important steps in the development process. Both quantitative and graphical analyses of the residuals are used to verify the accuracy of the proposed correlations. Quantitative error analysis is determined in terms of correlation coefficient ($R^2$), Standard Deviation (SD), average, minimum and maximum of the Error (E), Relative Error (RE) and Absolute Relative Error (ARE). Table 2.2 summarizes the quantitative statistical error analysis for the proposed correlations. The analysis shows a small error and a small standard deviation with an appreciably high correlation coefficient for both correlations.

Following the estimation of a regression model, the graphical error analysis was carried out by analyzing the residuals. The residual distributions for the bubble point pressure and the bubble point oil FVF correlations are presented in Figures 2.2 and 2.3,
respectively. Both figures show that the error is normally distributed, and it can be concluded that both correlations satisfy the normality assumption. Figures 2.4, and 2.5 display the computed values from the regression models versus the experimental values. It is evident from both figures that the points are scattered around the $y = x$ line.
Table 2.2 - Quantitative Statistical Error Analysis

<table>
<thead>
<tr>
<th>Statistical Criterion</th>
<th>$P_b$ Model</th>
<th>$B_{ab}$ Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$, %</td>
<td>95.67</td>
<td>96.3</td>
</tr>
<tr>
<td>SD</td>
<td>435.6, psia</td>
<td>0.0287, bbl/STB</td>
</tr>
<tr>
<td>Average E</td>
<td>17.72, psia</td>
<td>0.0, bbl/STB</td>
</tr>
<tr>
<td>Minimum AE</td>
<td>-2112.8, psia</td>
<td>-0.06411, bbl/STB</td>
</tr>
<tr>
<td>Maximum AE</td>
<td>1172.8, psia</td>
<td>0.11087, bbl/STB</td>
</tr>
<tr>
<td>Average RE, %</td>
<td>2.83</td>
<td>0.038</td>
</tr>
<tr>
<td>Minimum RE, %</td>
<td>-54.37</td>
<td>-8.399</td>
</tr>
<tr>
<td>Maximum RE, %</td>
<td>120.82</td>
<td>5.255</td>
</tr>
<tr>
<td>Average ARE, %</td>
<td>16.757</td>
<td>1.6874</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.05</td>
<td>0.0122</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>120.816</td>
<td>8.3989</td>
</tr>
</tbody>
</table>
Figure 2.2 – Error Distribution for Bubble Point Pressure Correlation
Figure 2.3 – Error Distribution for $B_{ob}$ Correlation
Figure 2.4 – A 45° Straight Line Cross Plot for Bubble Point Pressure Correlation
Figure 2.5 – A 45° Straight Line Cross Plot for Bubble point Oil FVF Correlation
2.5 Comparison with Other Correlations

As mentioned previously, not one specific correlation has been published so far to directly estimate bubble point pressure in the absence of PVT analysis. Earlier correlations first require estimates of $R_{sob}$ and $\gamma_{SP}$ (obtained from secondary correlations or from experiments) before they can be implemented. As a result, none of the published bubble point correlations could be compared with the proposed bubble point correlation in this chapter. However, the proposed $B_{ob}$ correlation was subjected to evaluation and validation. Its accuracy was tested solely against Labedi’s correlation because Labedi’s bubble point oil FVF is presently the only published correlation available in the literature to estimate bubble point oil FVF directly in the absence of PVT analysis. The Group B data set (62 data points), as described in Table 2.3, was used in this test. However, the data points were not switched in the model derivation process. SD, average, minimum and maximum values of E, RE and ARE with a SD and 45 degree line cross plot were used as comparative criteria. Figure 2.6 compares the behaviour of the proposed $B_{ob}$ regression model to Labedi (1990). The Figure shows that the produced model provides further reliable results. The majority of the points, estimated using the proposed correlation, fall very close to the 45 degree line with less RE, less average RE and less average ARE. Table 2.4 demonstrates the statistical analysis of this comparison.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Mean</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bubble Point Pressure, $P_b$</td>
<td>psia</td>
<td>1695</td>
<td>1196</td>
<td>189</td>
<td>4244</td>
</tr>
<tr>
<td>Oil FVF, $B_{OFb}$</td>
<td>bbl/STB</td>
<td>1.3202</td>
<td>0.1806</td>
<td>1.0640</td>
<td>1.674</td>
</tr>
<tr>
<td>Separator GOR, $R_{SP}$</td>
<td>SCF/STB</td>
<td>425</td>
<td>320.2</td>
<td>21</td>
<td>1009.8</td>
</tr>
<tr>
<td>Separator Pressure, $P_{SP}$</td>
<td>psia</td>
<td>72.17</td>
<td>57.63</td>
<td>14.7</td>
<td>228</td>
</tr>
<tr>
<td>Stock-Tank Oil Gravity, $\gamma_{ST}$</td>
<td>water=1</td>
<td>0.83286</td>
<td>0.01835</td>
<td>0.7999</td>
<td>0.8927</td>
</tr>
<tr>
<td>Reservoir Temperature, $T_r$</td>
<td>°F</td>
<td>202.44</td>
<td>45.28</td>
<td>100</td>
<td>275</td>
</tr>
</tbody>
</table>
Table 2.4 - Comparison of Proposed Bubble point Oil FVF Correlation

<table>
<thead>
<tr>
<th>Statistical Criterion</th>
<th>This Study Eq. (2.11)</th>
<th>Labedi 1990 Eq. (2.4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error SD, bbl/STB</td>
<td>0.02322</td>
<td>0.02793</td>
</tr>
<tr>
<td>Average E, bbl/STB</td>
<td>0.00412</td>
<td>0.00418</td>
</tr>
<tr>
<td>Maximum E, bbl/STB</td>
<td>0.05286</td>
<td>0.07064</td>
</tr>
<tr>
<td>Minimum E, bb/STB</td>
<td>-0.05351</td>
<td>-0.0589</td>
</tr>
<tr>
<td>Average RE, %</td>
<td>0.335</td>
<td>0.241</td>
</tr>
<tr>
<td>Maximum RE, %</td>
<td>3.635</td>
<td>4.757</td>
</tr>
<tr>
<td>Minimum RE, %</td>
<td>-3.548</td>
<td>-4.4165</td>
</tr>
<tr>
<td>Average ARE, %</td>
<td>1.412</td>
<td>1.617</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>3.635</td>
<td>4.757</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.028</td>
<td>0.036</td>
</tr>
</tbody>
</table>
Figure 2.6 – Evaluation and Validation of Bubble Point Oil FVF Correlation
2.6 Summary and Conclusions

- A review of the literature shows little attention has been paid to the prediction of reservoir fluid properties in the absence of PVT analysis. Currently, not one specific correlation has been published to directly estimate the bubble point pressure in the absence of PVT analysis and, presently, there is just one published correlation (Labedi 1990) to estimate bubble point oil FVF directly in the absence of PVT analysis.

- The majority of the available PVT correlations are used to estimate the reservoir oil properties at various temperatures rather than reservoir temperature in order to design surface operation equipment and to study reservoir inflow performance. However, due to the fact the correlations require prior knowledge of parameters such as solution GOR and specific gas gravity; it is difficult to apply them in the absence of PVT analysis. Since these parameters are rarely measured in the field, this study proposed correlations which can be applied straightforwardly in the absence of PVT analysis. Thus, there is no need for additional correlations. The only required input parameters are separator GOR, separator pressure, stock-tank oil gravity and reservoir temperature.

- Both quantitative and graphical analyses of the residuals were investigated in order to verify the accuracy of the proposed correlations. The newly proposed bubble
point pressure correlation provides prediction values with an average percent relative error of 2.83%, an absolute average absolute percent relative error of 16.757% and a correlation coefficient of 95.67%. The proposed bubble point oil FVF correlation provides prediction values, with an average percent relative error of 0.038%, an absolute average percent relative error of 1.6874% and a correlation coefficient of 96.3%. The residuals of both correlations are normally distributed, indicating that both models describe the data well.

- None of the previously published bubble point correlations were compared to the proposed bubble point correlation in this study. This is because prior knowledge of some of the data which are rarely measured in the field (bubble point solution GOR and gas specific gravity) is required.

- Of the 476 data points used in the development of the new correlations, 62 were used to validate and evaluate the accuracy of the proposed bubble point oil FVF correlation. Its accuracy was tested against Labedi’s correlation. The newly proposed oil FVF correlation provides better predictions and higher accuracy than Labedi’s correlation.

- Since the representativeness of a PVT study greatly depends upon sampling conditions, the first and most important operation, before running a complete reservoir fluid study, is to check the validity of the samples. The bottom-hole sample, used for the PVT study, is selected according to the results obtained during the verification of sample validity. It is worth mentioning, the proposed bubble point pressure correlation can be used to select the most representative sample with which to run a PVT analysis.
2.7 References

AGOCO, Sirte Oil, Waha Oil, Millita Oil and Gas, Repsol Oil, and Harouge Oil Companies (1971-1995). Reservoir Fluid Studies (PVT Analysis).


using regional correlations. SPE 37295 presented at the 1997 SPE international symposium on oilfield chemistry held in Houston, Texas, 18-20 February.


Labedi, R.M., 1990. Use of Production Data to Estimate the Saturation Pressure, Solution GOR, and Chemical Composition of Reservoir Fluids. SPE 21164 presented at the SPE Latin American Petroleum Conference held in Rio de Janeiro, October 14-19.


Mazandarani, M.T. and Asghari, S.M., 2007. Correlations for predicting solution gas-oil ratio, bubble point pressure and oil formation volume factor at bubble point of Iran crude oils. European Congress of Chemical Engineering (ECCE-6) Copenhagen, (Sep.) 16-20


CHAPTER III

PREDICTION OF BUBBLE POINT SOLUTION GAS/OIL RATIO IN THE ABSENCE OF PVT ANALYSIS

3.1 Introduction

Reservoir fluid studies or Pressure-Volume-Temperature (PVT) are designed to simulate the simultaneous fluid flow of oil and gas from the reservoir to the surface. This process is illustrated in Figure 2.1 with notation of the used variables to enrich the understanding. In absence of such analysis, empirical correlations can be applied. Several published correlations with which to estimate the bubble point pressure and bubble point oil formation volume factor of reservoir oils require the value of the bubble point solution gas/oil ratio to be one of the input variables. Consequently, engineers resort to an additional correlation in order to estimate this value. As well, the majority of published bubble point solution gas/oil ratio correlations are functions of bubble point pressure and gas gravity which can be obtained either experimentally (pressure-volume-temperature, PVT analysis) or estimated from the correlations. Thus, it is difficult to apply the correlations in the absence of PVT analysis.

A bubble point solution gas/oil ratio ($R_{sob}$) is a very important parameter in reservoir and production engineering calculations. The solution gas/oil ratio is the amount of gas
dissolved in the oil at any pressure per unit volume of oil both measured at standard conditions. The solution gas/oil ratio increases with pressure until the bubble point pressure is reached, after which it is a constant and is said to be a bubble point solution gas/oil ratio. A bubble point pressure ($P_b$) for reservoir oil is defined as the pressure at which gas first begins to exit a solution at a constant temperature. The important parameters are usually obtained experimentally (PVT analysis). Additionally, it is used as a basic parameter in many PVT correlations in order to estimate other fluid properties such as bubble point pressure, $P_b$ and the bubble point oil formation volume factor ($B_{ob}$). Numerous correlations, offered in the oil and gas industry, to estimate reservoir oil $P_b$ and $B_{ob}$ require a value of $R_{sob}$ as an input variable. On the other hand, several correlations presently exist to obtain the $R_{sob}$. Examples include Standing (1947), Vasques and Beggs (1980), Petrosky and Farshad (1993), Frashad et al. (1996), Elsharkawy and Alikhan (1997), Boukadi et al. (2002), and most recently, Hemmati and Kharrat (2007), Mazandarani and Asghari (2007) and Khamehchi et al. (2009). The basic assumption of $R_{sob}$ correlations is expressed as a function of the following variables:

$$R_{sob} = f_1 (P_b, \gamma_g, \text{API gravity}, T_R)$$  \hspace{1cm} (3.1)

The correlations are essentially based on the assumption that $R_{sob}$ is a strong function of bubble point pressure ($P_b$), gas specific gravity ($\gamma_g$), stock-tank oil API gravity and reservoir temperature ($T_R$). Accordingly, in order to apply the aforementioned correlations in the absence of PVT analysis, there must be previous knowledge of the $P_b$
and $\gamma_g$. Parameters $P_b$ and $\gamma_g$ are rarely measured as field data and therefore, engineers resort to additional correlations to reach an estimation. On the other hand, most $P_b$ correlations are essentially based on the relationship illustrated in Equation (3.2) which includes Standing (1947), Lasater (1958), Vazquez and Beggs (1980), Glaso (1980), Al-Marhun (1988), McCain (1990), Dokla and Osman (1992), Macary and El-Batanoney (1992), Petrosky and Farshad (1993), Omar and Todd (1993), De Ghetto and Villa (1994), Kartoatmodjo and Schmidt (1994), Frashad et al. (1996), Almehaideb (1997), Hanafy et al. (1997), Boukadi et al. (1999), Velarde et al. (1999), Al-Shammasi (1999), Dindoruk and Christman (2001), Mehran et al. (2006), and most recently, Hemmati and Kharrat (2007), Ikiensikimama and Ogboja (2009) and Moradi et al. (2010). However, the foresaid correlations are ultimately required in order to estimate $P_b$ and $R_{sob}$ at various temperatures other than $T_R$ in order to design surface and subsurface facilities. This is due to reservoir fluid properties being found in the PVT analysis reports at $T_R$ only. In such cases, even though PVT analysis is available, the required reservoir fluid properties must be estimated from the correlations.

$$P_b = f_2(R_{sob}, \gamma_g, API \text{ gravity}, T_R)$$  (3.2)

### 3.2 Literature Review

A review of the literature indicates little attention has been given to the prediction of reservoir fluid properties in the absence of PVT analysis. However, in most cases, $R_{sob}$
correlations are derived by mathematically solving the $P_b$ correlation for a $R_{sob}$. An exception to this precedent are the correlations presented by Vazquez and Beggs (1980), Elsharkawy and Alikhan (1997), Boukadi et al. (2002), Hemmati and Kharrat (2007), Mazandarani and Asghari (2007) and Khamenechi et al. (2009), where the $R_{sob}$ was treated as an independent variable based on Equation (3.1). Nevertheless, the majority of the available $R_{sob}$ correlations are impractical in the absence of PVT analysis, simply because $R_{sob}$ correlations require knowledge of the $P_b$ and $\gamma_g$. As a result, it is difficult to apply the correlations in the absence of PVT analysis.

Furthermore, the $R_{sob}$ can be obtained indirectly by adding the estimated stock-tank vent gas/oil ratio, $R_{ST}$, to the field measured separator gas/oil ratio, $R_{SP}$, as shown in Equation (3.3). This is only valid if the reservoir pressure is above the $P_b$. Accordingly, Rollins et al. (1990) developed a correlation (Equation (3.4)) to estimate the $R_{ST}$ as a function of $\gamma_{oST}$, $\gamma_{gSP}$, $P_{SP}$ and $T_{SP}$. However, this work has a practical limitation in the absence of PVT analysis due to the $\gamma_{gSP}$ which is rarely measured in the field and it is usually based on a gas composition analysis.

\[
R_{sob} = R_{SP} + R_{ST} \tag{3.3}
\]

\[
\ln R_{ST} = 0.2126 - 4.916 \ln \gamma_{oST} + 3.469 \ln \gamma_{gSP} + 1.501 \ln P_{SP} - 0.9213 \ln T_{SP} \tag{3.4}
\]
In 2003, Valko and McCain derived a correlation to estimate the $R_{ST}$ as a function of $P_{sp}$, $T_{sp}$ and $API$. In applying a non-parametric regression analysis method, with 881 data points, they suggest the following empirical Equation (3.5):

$$\ln R_{ST} = 3.955 + 0.83z - 0.024z^2 + 0.075z^3$$  \hspace{1cm} (3.5)

where, $z = \sum_{n=1}^{3} z_n$

$z_1 = -8.005 + 2.7 \ln P_{sp} - 0.161 (\ln P_{sp})^2$

$z_2 = 1.224 - 0.5 \ln T_{sp}$, and

$z_3 = -1.587 + 0.0441 API - 2.29 \times 10^{-5} (API)^2$

Valko and McCain estimated the $R_{sob}$ indirectly using Equation (3.3) and reported an average RE of 0.0% and an average ARE of 5.2%. They proposed an additional correlation (Equation (3.6)) to estimate the $R_{sob}$ directly when the separator conditions are unknown. In estimating the $R_{sob}$, they reported an average RE of 0.0% and an average ARE of 9.9%.

$$R_{sob} = 1.1617 R_{SP}$$  \hspace{1cm} (3.6)

### 3.3 Newly Developed Correlations

This study presents two multiple regression analysis models to predict the $R_{sob}$ in the absence of PVT analysis. Both correlations are proposed as a function of readily available field data with no further requirement of additional correlations or experimental PVT...
analysis. An additional and important application of the proposed stock-tank vent gas/oil rate correlations is to estimate the stock-tank vent gas flow rate.

### 3.3.1 PVT Data

A total of 480 separator tests, measured at the $P_b$, were obtained from 118 PVT reports collected from various Libyan oil reservoirs located in the Sirte Basin. The majority of the data points were taken from a two-stage separation test. Normally, the separation flash test is carried out for two stages of separation (Figure 2.1). The volume of gas liberated at each stage and the volume of remaining liquid are measured. The first-stage separator pressure is generally varied to include at least four possible separator pressures at ambient temperature, and the second stage is generally carried out under stock tank conditions of 14.7 psia (1 atmosphere) at ambient temperature. The stock tank is usually considered to be one stage of separation. However, in single-stage separation, the test is carried out under separator conditions of 14.7 psia at ambient temperature. In this case, $R_{ST}$ is equal to zero.

Prior to the regression analysis, the collected datasets were divided randomly into two sets. The first set of 431 data points was used to construct the regression models, while the second set of 49 data points was used to test the validity and compare the regression models against similar available correlations (Valko and McCain 2003). The first dataset was used to generate the $R_{sob}$ correlation. The range of this data is presented in Table 3.1. Whereas, only 355 two-stage flash separation tests out of the 431 separator tests were used to generate the $R_{ST}$ empirical correlation, since data from the single-stage flash separation tests were excluded. This dataset is described in Table 3.2.
3.3.2 Bubble Point Solution Gas/Oil ratio, $R_{sob}$ Correlation

A novel correlation to estimate the $R_{sob}$ was developed during this study. It was correlated in terms of the $R_{SP}$, $P_{SP}$ and $\gamma_{ST}$ after the forward-backward search algorithm is conducted to select the best subset input variables. Numerous models were attempted as regression equations but the following Equation (3.7) was found to be the most accurate:

$$R_{sob} = -152 + 0.99 R_{SP} + 0.256 P_{SP} + 19.543 \ln(P_{SP}) - 557.2 \ln(\gamma_{ST})$$

(3.7)

The model shows a small average RE of 1.2% and an average ARE of 6.8% with a high $R^2$ of 99.3%. Figure 3.1 shows the histogram plot of the residuals indicating the error terms follow a normal distribution with a mean almost equal to zero, and they satisfy the normality assumption. The estimated $R_{sob}$ was plotted as a function of the actual values in Figure 3.2. The cross-plot shows the precision of the empirical correlation without bias. Table 3.3 summarizes the results of the quantitative analysis of the residuals.
Table 3.1 – Range of Data Points for $R_{sob}$ Correlation (431 Data Points)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{sob}$</td>
<td>SCF/STB</td>
<td>409.3</td>
<td>256.6</td>
<td>17</td>
<td>1256</td>
</tr>
<tr>
<td>$R_{SP}$</td>
<td>SCF/STB</td>
<td>353</td>
<td>250.1</td>
<td>10</td>
<td>1256</td>
</tr>
<tr>
<td>$P_{SP}$</td>
<td>psia</td>
<td>119.35</td>
<td>127.12</td>
<td>14.7</td>
<td>735</td>
</tr>
<tr>
<td>$\kappa_{ST}$</td>
<td>water=1</td>
<td>0.84035</td>
<td>0.02496</td>
<td>0.7999</td>
<td>0.921</td>
</tr>
</tbody>
</table>
Table 3.2– Range of Data points for $R_{ST}$ Correlation (355 Data Points)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Units</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{ST}$</td>
<td>SCFSTB</td>
<td>68.8</td>
<td>51.79</td>
<td>6.3</td>
<td>288.8</td>
</tr>
<tr>
<td>$P_{SP}$</td>
<td>psia</td>
<td>141.17</td>
<td>129.25</td>
<td>29.7</td>
<td>735</td>
</tr>
<tr>
<td>$\gamma_{ST}$</td>
<td>water=1</td>
<td>0.84079</td>
<td>0.02561</td>
<td>0.7999</td>
<td>0.921</td>
</tr>
<tr>
<td>$T_{SP}$</td>
<td>°F</td>
<td>111.19</td>
<td>30.79</td>
<td>73</td>
<td>194</td>
</tr>
</tbody>
</table>
3.3.3  Stock-Tank Vent Gas/Oil Ratio, \( R_{ST} \) Correlation

Usually, the \( R_{SP} \) is a field measurement, while the \( R_{ST} \) is rarely measured in the field. Thus, in order to obtain the \( R_{SB} \) indirectly, we must first estimate the \( R_{ST} \), and add it to the field measured \( R_{SP} \). In this work, a novel correlation to estimate the \( R_{ST} \) was developed. \( P_{SP} \), \( T_{SP} \) and \( \gamma_{O_{ST}} \) are the only independent variables. A total of 355 two-stage separator tests were employed to develop the \( R_{ST} \) correlation. The range of the dataset is illustrated in Table 2.2. The natural logarithm of the \( R_{ST} \) was regressed as a function of the natural logarithms of \( P_{SP} \), \( T_{SP} \) and \( \gamma_{O_{ST}} \) and the following empirical Equation (3.8) was obtained:

\[
R_{ST} = P_{SP}^{1.02} \gamma_{O_{ST}}^{-9.47} T_{SP}^{-0.524}
\]  (3.8)

The model shows small values of average RE and average ARE of 1.1% and 7.6% respectively, with a high \( R^2 \) of 99.2%, indicating the model is sufficient to describe the data.
Table 3.3 – Quantitative Statistical Analysis of the Present Correlations

<table>
<thead>
<tr>
<th>Statistical Criterion</th>
<th>$R_{sob}$ Correlation</th>
<th>$R_{ST}$ Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$, %</td>
<td>99.3</td>
<td>99.2</td>
</tr>
<tr>
<td>Average RE, %</td>
<td>1.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Average ARE, %</td>
<td>6.8</td>
<td>7.6</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>83.9</td>
<td>46.3</td>
</tr>
</tbody>
</table>
Figure 3.1- Residual Distributions of $R_{sob}$ Obtained from Eq. (3.7) (431 Data Points)
Figure 3.2 - Estimated Rsob from Eq. (3.7) vs. Actual Rsob (431 Data Points)
3.4 Comparison with Other Correlations

In order to study the accuracy of the proposed correlations, the estimated $R_{sob}$ values from Equation (3.7) and those obtained from Equation (3.8), using 355 data points described in Table 2.2, were compared with $R_{sob}$ values calculated from Valko and McCain correlations, Equations (3.5) and (2.6). This comparison is illustrated in Figure 3.3. The cross-plot shows the majority of the plotted data points, estimated using our proposed correlations, fall on the $y=x$ line, indicating the two proposed models have a higher accuracy than the Valko and McCain correlations and provide far more reliable results. Moreover, the histogram of the relative error of the proposed correlations, along with the relative error from the Valko and McCain correlations, is plotted in Figure 3.4. The error terms from Equations (3.5), (3.7) and (3.8) satisfy the normality assumption with a mean equal to $-1.958$, $1.261$ and $0.1568$, respectively. However, the relative errors of the $R_{sob}$ values, obtained from Equation (3.6), deviate and do not meet the normal distribution assumption. Table 3.4 summarizes the statistical results of this comparison. It is evident the two proposed correlations give less average RE and average ARE than the Valko and McCain correlations. However, the two proposed correlations have approximately the same accuracy within this range of data.
Table 3.4– Statistical Analysis of Evaluation Test (355 Data Points)

<table>
<thead>
<tr>
<th>Statistical Criterion</th>
<th>This Study Eq. (3.7)</th>
<th>This Study Eq. (3.8)</th>
<th>Valko and McCain Eq. (3.5)</th>
<th>Valko and McCain Eq. (3.6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average RE, %</td>
<td>1.3</td>
<td>-0.2</td>
<td>-2</td>
<td>-10.1</td>
</tr>
<tr>
<td>Average ARE, %</td>
<td>7.3</td>
<td>6.7</td>
<td>7.2</td>
<td>16.0</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>56.1</td>
<td>37.0</td>
<td>72.5</td>
<td>72.7</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Figure 3.3 – Evaluation Test; Estimated vs. Actual $R_{sob}$ (355 Data Points)
Figure 3.4 – Evaluation Test; Relative Error Distributions (355 Data Points)
Moreover, a validation test was carried out using the 49 two-stage flash separation tests described in Table 3.5, against the Valko and McCain correlations. The data points were not switched in the process of correlation derivation. Figure 3.5 illustrates the behaviour of $R_{sob}$, estimated from the two proposed correlations against $R_{sob}$, as estimated from the Valko and McCain correlations. The cross-plot of the estimated values versus experimental values shows the majority of the plotted data points, estimated using our proposed models, fall on the 45 degree line. Our proposed models, as explained by Equations (3.7) and (3.8), show a high $R^2$ of 99.61% and 99.97%, respectively. However, the Valko and McCain modes, expressed by Equations (3.5) and (3.6) show a $R^2$ of 99.27% and 83.49%, respectively. Furthermore, the histogram in Figure 3.6 shows the distribution of the relative errors obtained from Equations (3.5), (3.6), (3.7) and (3.8). The error terms from Equation (3.5) deviate and skew to the right with a mean equal to −5.474, while the error terms from Equation (3.6) are not normally distributed. However, the error terms from Equations (3.7) and (3.8) are satisfying the normality assumption with a mean equal to 2.675 and 0.581, respectively. The statistical results of this validation test are illustrated in Table 3.6. Obviously, the calculated $R_{sob}$, obtained from the two proposed empirical correlations, gives less average RE and average ARE, indicating our proposed correlations have a higher accuracy than the Valko and McCain correlations and provide far more reliable results.
Table 3.5 – Range of Data used for Validation Test (49 Data Points)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{sob}$</td>
<td>473.2</td>
<td>329.9</td>
<td>56</td>
<td>1105.9</td>
</tr>
<tr>
<td>$R_{SP}$</td>
<td>421.6</td>
<td>328.3</td>
<td>21</td>
<td>1009.8</td>
</tr>
<tr>
<td>$R_{ST}$</td>
<td>51.58</td>
<td>31.39</td>
<td>9.8</td>
<td>155</td>
</tr>
<tr>
<td>$P_{SP}$</td>
<td>92.87</td>
<td>35.69</td>
<td>34.7</td>
<td>228</td>
</tr>
<tr>
<td>$T_{SP}$</td>
<td>103.83</td>
<td>27.35</td>
<td>73</td>
<td>158</td>
</tr>
<tr>
<td>$\gamma_{ST}$</td>
<td>0.83324</td>
<td>0.02052</td>
<td>0.7999</td>
<td>0.8927</td>
</tr>
</tbody>
</table>
Table 3.6—Statistical Accuracy of Validation Test (49 Data Points)

<table>
<thead>
<tr>
<th>Statistical Criterion</th>
<th>This Study Eq. (3.7)</th>
<th>This Study Eq. (3.8)</th>
<th>Valko and McCain Eq. (3.6)</th>
<th>Valko and McCain Eq. (3.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average RE, %</td>
<td>-1.273</td>
<td>-0.702</td>
<td>1.119</td>
<td>1.601</td>
</tr>
<tr>
<td>Average ARE, %</td>
<td>3.314</td>
<td>3.471</td>
<td>16.41</td>
<td>3.284</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>19.243</td>
<td>29.509</td>
<td>67.85</td>
<td>36.47</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.003</td>
<td>0.031</td>
<td>0.09</td>
<td>0.034</td>
</tr>
</tbody>
</table>
Figure 3.5 – Validation Test; Actual vs. Estimated $R_{sob}$ (49 Data Points)
Validation Test; Histogram of the Percent Relative Error

This Study Eq (3.7)

This Study Eq (3.8)

Valko & McCain Eq (3.6)

Valko & McCain Eq (3.5)

Mean -1.273
StDev 5.378

Mean -0.7015
SD 6.313

Mean 1.191
SD 22.26

Mean 1.601
SD 6.670

Figure 3.6– Validation Test; Relative Error Distributions (49 Data Points)
3.5 Conclusions

- Two novel and simple empirical correlations, based on a multiple regression analysis to estimate $R_{sob}$ and $R_{ST}$ in the absence of PVT analysis, were proposed as a function of readily available field data. Both correlations can be applied straightforwardly in the absence of PVT analysis. No further correlations or experimental measurements are required.

- The proposed bubble point solution gas/oil ratio correlation was developed based on 480 two-stage and single stage flash separation tests. The majority of the tests are taken from a two-stage flash separation test. In the single-stage separation test, the separator pressure is atmospheric pressure and the stock-tank vent gas/oil ratio is equal to zero. However, the proposed stock-tank vent gas/oil ratio correlation was developed based on 355 two-stage separation tests.

- Both correlations were developed specifically for Libyan oil reservoirs located in Sirte Basin. However, they may be considered for use in other reservoirs around the world.

- Quantitative and graphical analyses of the residuals were investigated to verify the accuracy of the proposed correlations. The present $R_{sob}$ correlation provides calculated values with an average RE of 1.2% and an average ARE of 6.8% with a $R^2$ of 99.3%. The $R_{ST}$ correlation provides calculated values with an average RE of 1.1% and an average ARE of 7.6% with a $R^2$ of 99.2%. The residuals of both
presented correlations satisfy the normal distribution assumption and describe the data well.

- A comparison between the proposed correlations and similar available correlations from the literature indicate both presented correlations render a better approximation.

- The two proposed correlations achieve approximately the same accuracy when predicting $R_{sob}$ values.

- The estimated $R_{sob}$ values from the proposed correlations can be used as a basic input variable in many published PVT empirical correlations in order to estimate other fluid properties such as the $P_b$ and $B_{ob}$ at various temperature rather than reservoir temperature.

- An additional and important application of the proposed stock-tank vent gas/oil ratio correlation is to estimate the stock-tank vent gas flow rate.
3.6 References


Mazandarani, M.T. and Asghari, S.M., 2007. Correlations for predicting solution gas-oil ratio, bubble point pressure and oil formation volume factor at bubble point of Iran crude oils. European Congress of Chemical Engineering (ECCE-6) Copenhagen, (Sep.) pp.16-20


CHAPTER IV

DEVELOPMENT OF PVT MODELS UTILIZING ANNS

4.1 Introduction

The most common and widely used artificial neural networks, ANNs, are found in the literature and are known as the “feedforward neural networks” with a backpropagation training algorithm. This type of neural networks has an excellent computational intelligence modeling scheme in both prediction and classification tasks. In recent years, ANNs has gained popularity in petroleum applications. Many authors in the field of petroleum engineering such as Ali (1994), Kumoluyi and Daltaban (1994), Mohaghegh and Ameri (1994), Mohaghegh (1995), Gharbi and Elsharkawy (1997a,b), Elsharkawy (1998), Varotsis et al. (1999) and Mohaghegh (2000) discussed the applications of ANNs. Presently, ANNs is serving the petroleum industry in predicting PVT correlations (Goda et al. (2003) and Shokir (2004)). Tips and guidelines regarding ANNs modeling procedures are presented in Appendix B.

The main objective of Chapter IV is to realize whether artificial neural networks can be used as

1. an alternative to traditional statistical regression techniques to estimate the most important PVT properties in reservoir engineering such as bubble point
pressure, $P_b$, bubble point oil FVF, $B_{ob}$, bubble point solution GOR, $R_{sob}$ and stock tank vent GOR, $R_{ST}$.

2. a tool to interpolate and model the average reservoir pressure in oil reservoirs

3. an alternative technique to predict reservoir oil production prediction

4.2 Review of Recent PVT Neural Network Models

Several authors have discussed the application of ANNs in petroleum engineering. Presently, the feedforward neural network serves the petroleum industry in predicting PVT Models. Numerous PVT ANN models have been published in the petroleum engineering literature. In this section, some of the efforts related to the PVT ANN models were summary reviewed, while focusing on future required improvements.

Gharbi and Elsharkawy (1997a) published ANN models to estimate $P_b$ and $B_{ob}$ Middle East crude oils. Separate models were used for each property based on Equations 1.4 and 1.5. The models’ architectures were two hidden layers and four nodes in the input layer. The $P_b$ model architecture has eight nodes in the first layer and four nodes in the second, [4-8-4-4]. The $B_o$ model architecture has six nodes in both layers, [4-6-6-1]. Both models were trained using 498 data points collected from the literature and unpublished sources. The models were tested using another dataset of 22 points from the Middle East. The results showed improvement over the conventional correlation methods with at least a 50% reduction in the average error for the $P_b$ and a 30% reduction for $B_{ob}$.

Gharbi and Elsharkawy (1997b) published another ANN model to estimate $P_b$ and $B_{ob}$ for universal use. The two properties were predicted by one model consisting of one
hidden layer of five neurons and the same inputs, [4-5-2]. The study used 5200 datasets collected from all over the world and representing 350 different crude oils. Another set of data, consisting of 234 datasets, was used to verify the model’s results. The results for the universal NN model showed the $P_b$ average error was lower than that of the conventional correlations for both training and test data. The $B_{ob}$ on the other hand, was better than the conventional correlations in terms of correlation coefficient. The average error for the NN model is similar to the conventional correlations for training data and higher for test data than the best performing conventional correlation.

Elsharkawy (1998) presented a new technique to model the behaviour of crude oil and natural gas systems using a radial-basis-function, RBF neural network model. Two RBF models can predict $B_o$, $R_{oso}$, $\mu_o$, saturated $\rho_o$, undersaturated $C_o$, and liberated $\gamma_g$. The study used the differential PVT data of 90 samples for training and another 10 samples for testing the model. The architecture of the RBF network model is arranged in four layers: an input layer, two hidden layers, and an output layer. The input layer for the two models contains $p_R$, $T_R$, $\gamma_{TS}$, and liberated $\gamma_g$. The output layer for Model #1 contains $R_{oso}$, $B_o$, $\mu_o$, and $\rho_o$. Whereas, the output layer for Model #2 contains undersaturated $C_o$ and liberated $\gamma_g$. The accuracy of the RBF model was compared to all published correlations for training and testing samples. The results show that the proposed model is far more accurate than the correlations in predicting the oil properties. The behaviour of the model in capturing the physical trend of the PVT data was also checked against experimentally measured PVT properties of the test samples. The author concluded that although the model was developed for a specific crude oil and gas system, the idea of using neural
network to model behaviour of reservoir fluid can be extended to other crude oil and gas systems as a substitute to PVT correlations that were developed by conventional regression techniques.

Varotsis et al. (1999) presented a novel approach to predict a complete set of physical properties describing the PVT behaviour of both oils and gas condensates using a non-iterative approach. The ANN oil models are relative volume, \( B_o \), \( \mu_o \), \( \rho_o \), z-factor, GOR and liberated \( \gamma_g \). The input data for the oil ANN models consists of reservoir fluid composition, \( P_b \), \( T_r \), \( \rho_{ob} \), \( \mu_{oD} \), flash molar ratio and flash liquid and gas densities. The model was developed using a database containing full PVT data for 650 reservoir fluids from around the world. Two hidden layers were used. The number of neurons in each layer was varied and the networks were retrained until the optimum structure was obtained.

Al-Shammasi (2001) presented two ANNs models to estimate \( P_b \) and \( B_{ob} \), based on global data. Both models include four nodes in the input layer and two hidden layers, five nodes in the first layer and three nodes in the second layer, \([4-5-3-1]\). He compared their performances to numerical correlations. The ANNs models performed better, but suffer from stability and trend problems.

Osman et al. (2001) presented an ANN model to predict the \( B_{ob} \). The model was developed using 803 published data from the Middle East, Malaysia, Colombia, and the Gulf of Mexico fields. One-half of the data was used to train the ANNs models, one quarter to cross-validate the relationships established during the training process and the remaining one quarter to test the models in order to evaluate their accuracy and trend
stability. With regard to the accuracy of the model, they reported an absolute percent relative of 1.7886%, with the highest 98.78% correlation coefficient.

Al-Marhoun (2002) proposed an ANN models to predict $P_b$ and $B_{ob}$. The models were developed using 283 datasets collected from Saudi reservoirs. He divided the datasets into three groups of 142 data points for training, 71 data points for cross validation of the models and 70 data points to evaluate the accuracy of the models. The $B_{ob}$ model provides prediction values with an absolute average percent error of 0.5116% and a standard division of 0.6625 with a correlation coefficient of 99.89% while the $P_b$ model provides prediction values with an absolute average percent error of 5.8915% and a standard division of 8.6781 with a correlation coefficient of 99.65%.

Goda et al. (2003) designed two models, one to predict $P_b$ and the other to estimate $B_{ob}$. The $P_b$ model based on Equations 1.4 has two hidden layers with each one having ten neurons, [4-10-10-1]. The hidden layers are activated by the log sigmoid transfer function while the output layer is activated by the line transfer function. The model was trained with 160 data points and tested with 20 data sets. However, the $B_{ob}$ model has five inputs connected with two hidden layers. The input nodes are $R_{sob}$, $T_r$, $\gamma_\phi$, $\gamma_g$ and the estimated $P_b$ from the first model. Each one of the two hidden layers has eight neurons, [5-8-8-1] with a log sigmoid transfer function. The output layer is joined with pure linear function.

Shokir et al. (2004) designed two models to estimate $P_b$ and $B_{ob}$. The $P_b$ network architecture was selected as a four-layer network. The input layer receives $R_{sob}$, $T_r$, API gravity and $\gamma_g$. Each of the two hidden layers has ten neurons activated by the log sigmoid transfer function. The network was trained with 160 data points and tested with another
20 data points collected from Middle East oil systems. The $B_{ob}$ network has four layers. The input layer has five neurons for five inputs which are $T_R$, $R_{sob}$, API gravity, $\gamma_g$, and the estimated $P_b$ from the first network.

Rasouli et al. (2008) presented two feedforward BP-NN models for the prediction of $P_b$ and $B_{ob}$. The models trained with 106 experimental data points and were evaluated with nine experimental data points. The models are based on Equations 1.4 and 1.5 has three hidden layers. The $P_b$ model, consists of six nodes in the first hidden layer, ten in the second layer and six in the third layer, $[4−6−10−6−1]$. The $B_{ob}$ model consists of nine nodes in the first hidden layer, twelve neurons in the second layer, and seven neurons in the third layer, $[4−9−12−7−1]$. Input and output layers were activated by the logistic activation function.

Dutta, and Gupta (2010) presented six models for $P_b$, $R_{sob}$, $B_{ob}$ and undersaturated $B_o$, and saturated and undersaturated $\mu_o$ from Indian crude oil, based on samples from various onshore oil fields in western India. The model is based upon MLP using Bayesian Regularization coupled with a hybrid genetic algorithm (GA) as the error minimization technique. The model architectures for the $P_b$, $R_{sob}$, $B_{ob}$ undersaturated $B_o$, saturated $\mu_o$ and undersaturated $\mu_o$ are $[4−6−4−1]$, $[4−6−5−1]$, $[4−9−1]$, $[7−6−1]$, $4−6−4−1$ and $[4−9−1]$, respectively. The transfer function in all the hidden layer neurons is hyperbolic tangent and the output layer is linear. The performance of the new model was compared to other existing correlations, and was found to outperform them in terms of average absolute relative error, correlation coefficient, maximum and minimum absolute relative error, and standard deviation of absolute relative error.
Recently, Olatunji et al. (2010) proposed a new model to predict $P_b$ and $B_{ob}$. The study used a Sensitivity Based Linear Learning Method as a learning technique for two-layer feedforward NNs based on sensitivity analysis that uses a linear training algorithm for each layer. The model inputs, based on Equations 1.4 and 1.5, two datasets (160 database and 283 database), were utilized in both internal validation (training) and external validation (testing) by dividing each into a 70% training set and a 30% testing set.

Most recently, Alimadadi and Fakhri (2011) predicted the $B_o$ and $\rho_o$ properties of Iran crude oil using a committee machine (modular feedforward) with ANNs. The model was trained using 170 PVT data points from 19 undersaturated oil reservoirs in south Iran. Two additional datasets, each with 16 points, were used to validate and test the accuracy of the designed model. The model processes its inputs using two parallel MLP networks, and then recombines their results. The upper and lower MLP both share the same output layer which corresponds to $B_o$ and $\rho_o$. The upper MLP has one hidden layer with 15 nodes and five nodes in the input layer containing $P_b$, $T_{RP}$, API and $R_{so}$. The lower MLP has 19 nodes in the input layer (with the input being the composition mole %), three nodes in the hidden layer, and two nodes in the shared output layer.

A review of the literature shows

- The majority of the studies used feedforward neural networks with backpropagation training algorithms.
- Others preferred to use radial-basis-function and the RBF neural network model.
• The majority of the PVT ANN models are designed to predict $P_b$ and $B_{ob}$ and the input data are based on the assumption that $P_b$ and $B_o$ are strong functions of $R_{sob}$, $T_R$, $\gamma_g$, $\gamma_{ST}$, as described in Equations 1.4 and 1.5.

• None of the previous studies deal with the prediction of the stock tank vent GOR, $R_{ST}$.

• All published PVT ANN models can be used to predict reservoir fluid properties at various temperatures rather than at reservoir temperature, utilization in reservoir inflow performance and the designing of surface operation equipment. However, PVT ANN models require prior knowledge of some input parameters that are rarely measured in the field. As a result, it is difficult to apply them in the absence of PVT analysis.

• Less attention is paid to the prediction of PVT properties in the absence of PVT analysis. Until recently, not one specific ANN model has been published to directly predict PVT properties in the absence of PVT analysis.

However, the main objective of this thesis is to build new ANN models to predict $P_b$, $B_{ob}$, $R_{sob}$, and $R_{ST}$ using directly measured field parameters as nodes in the first network layer. Alyuda Neurointelligent software (2006) was employed to construct the ANN models using feedforward with the BP algorithm as a learning algorithm. The models are shown to be robust, accurate and a quick prediction tool that can be used to estimate reservoir fluid properties in the absence of a PVT laboratory analysis over the entire range of operating conditions and fluid types. The main application of the models is envisioned to be the rapid generation of PVT data based on field measurements.
4.3 PVT Datasets

A complete description of PVT property data is not accessible with respect to some reservoirs in order to determine their fluid characteristics. Thus, it is important to predict their fluid properties using directly measured field data parameters as input variables. To resolve this problem and develop accurate PVT models, the ANN must be exposed to sufficient data points or in the training session, the network model learns and processes only a few patterns and cannot provide a precise prediction and interpolation when it encounters new PVT samples. The objectives of this chapter is to build new ANN models to predict $P_b$, $B_{ob}$, $R_{sob}$, and $R_{ST}$ using directly measured field parameters as nodes in the first network layer.

The PVT dataset used in this work contains a total of 118 reservoir fluid studies (PVT analysis) which are the only available studies were collected from several oil fields located in Sirte Based, Libya. Therefore, 493 data points were selected to design the model. However, before the ANNs are carried out, the PVT data was subjected to an investigation to find any outliers and data anomalies. Outliers are the observations that appear to be inconsistent with the remainder of the collected data. A discordant observation is defined as an observation that appears surprising or discrepant to the investigator (Iglewicz, 1983). Analyzing the data detected eight observations that were identified as outliers and rejected as unconvertible for use with the neural network. Typically, the selection of the ANN model is made with the basic cross-validation process. Moreover, to avoid overtraining, which results in poor model performance, the remaining 485 data points were subdivided randomly into four sets, namely, the training
The dataset, the validation dataset, the testing dataset, and the deploying dataset. The validation dataset was used during training to see how the model would perform. Training can be stopped when the performance of the model on the validation dataset gives a minimum error. The testing dataset was used to fine-tune the models. However, numerous ANNs were tried with various architectures and the accuracy of each one on the testing dataset was tested in order to select the competing architecture. The lowest error on the testing dataset is used to select the appropriate model and to indentify the optimum network architecture. Finally, the deployment dataset was used to examine the model for the prediction of new data points.

4.4  Bubble point Pressure ANN Model

The overall data used to model $P_b$ were split into four sets - 286 points (59.95%) to the training set, 14.0% (67 points) to the validation set, 14.0% (67 points) to the testing set and 57 points (12.95%) to the deployment set. The training dataset was used to train the ANN models. A range of each dataset is described in Table 4.1.

Four inputs ($P_{sp}$, $R_{sp}$, $T_R$ and $\gamma_{ST}$) are selected to be incorporated into the input layer. In order to find an efficient network model and analyze the performance, the data must be pre-processed in order to have a small variation of the output data. It is well known that, in theory, the output data can have a large domain of variation. However, if the domain of variation of the output data is large, the ANN model tends to be less stable. The dimensionless output data ranges from 0 to 1, while the input data ranges from $-1$ to 1.
Table 4.1 – Ranges of Datasets used in Training, Validation, Testing and Deployment $P_b$ ANN Model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Overall PVT Dataset</th>
<th>Training Dataset</th>
<th>Validation Dataset</th>
<th>Testing Dataset</th>
<th>Deployment Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{sp}$, psia</td>
<td>109.92</td>
<td>114.10</td>
<td>90.69</td>
<td>111.1</td>
<td>47.41</td>
</tr>
<tr>
<td>$R_{sp}$, SCF/STB</td>
<td>347.5</td>
<td>339.9</td>
<td>374.6</td>
<td>353.1</td>
<td>450.2</td>
</tr>
<tr>
<td>$\gamma_{ST}$, water = 1</td>
<td>0.84027</td>
<td>0.84014</td>
<td>0.84028</td>
<td>0.84078</td>
<td>0.83457</td>
</tr>
<tr>
<td>$T_R$, °F</td>
<td>200.48</td>
<td>200.47</td>
<td>200.64</td>
<td>200.34</td>
<td>209.7</td>
</tr>
<tr>
<td>$P_b$, psia</td>
<td>16989</td>
<td>1689.7</td>
<td>1731</td>
<td>1707</td>
<td>1810</td>
</tr>
</tbody>
</table>

Average | Max | Min | SD | Average | Max | Min | SD | Average | Max | Min | SD | Average | Max | Min | SD |
---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
109.92 | 517 | 14.7 | 101.55 | 114.10 | 517 | 14.7 | 104.54 | 90.69 | 517 | 14.7 | 81.38 | 111.1 | 447 | 14.7 | 106.1 |
347.5 | 1175 | 10 | 246.3 | 339.9 | 1175 | 10 | 238.1 | 374.6 | 1038 | 13 | 274.3 | 353.1 | 1097 | 15 | 253.3 | 450.2 | 109.8 | 21 | 320.9 |
0.84027 | 0.921 | 0.7999 | 0.02489 | 0.84014 | 0.921 | 0.7999 | 0.249 | 0.84028 | 0.92 | 0.8003 | 0.02422 | 0.84078 | 0.919 | 0.7999 | 0.02582 | 0.83457 | 0.8927 | 0.8008 | 0.01767 |
200.48 | 277 | 100 | 47.871 | 200.47 | 277 | 100 | 47.92 | 200.64 | 271.4 | 100 | 49.35 | 200.34 | 275 | 100 | 46.85 | 209.7 | 275 | 134 | 39.29 |
16989 | 4244 | 121 | 1090 | 1689.7 | 4244 | 121 | 10864 | 1731 | 4244 | 152 | 1138 | 1707 | 3582 | 121 | 1072 | 1810 | 4244 | 189 | 1178 |
To specify the number of hidden layers and the number of units in each layer, a trial and error approach was carried out, beginning with one hidden layer and one hidden unit.

Hidden units were then gradually added. A total of 70 architectures were compared with the following network parameters:

- **Hidden layer activation function** = Logistic
- **Output activation function** = Logistic
- **Error function** = Sum-of-squares
- **Hidden layers search range** = 1 to 2 HL
- **First layer hidden units search range** = 1 to 10 nodes
- **Second hidden layer units search range** = 1 to 6 nodes
- **Fitness criteria** = Inverse test error

Table 4.2 shows the five best architectures. At this point, the selected architecture with two hidden layers, nine nodes in the first hidden layer, and four nodes in the second layer was trained using different algorithms. In each case, the architecture was retrained five times with different initial weight randomization. The training algorithm with the lowest average error and highest R² was selected as a best training algorithm. Table 4.3 shows the average and maximum error during training, validation and testing of Architecture [4–9–4–1] under various training algorithms, namely, Quasi-Newton, Conjugate Gradient Descent, Levenberg-Marquardt and Limited Memory Quasi-Newton. It is evident that the Quasi-Newton training algorithm has a higher R² of 99.12 %, and the lowest training, validation and testing error of 77.535, 102.797 and 119.272 respectively.
4.4.1 Bubble Point Pressure ANN Model Evaluation

Both qualitative and graphical analysis of the model error terms was utilized to verify the accuracy of the selected model. Table 4.4 summarizes the quantitative statistical error analysis for the identified $P_b$ ANN model. The analysis shows a small error results on training, validation and testing datasets.

Furthermore, the graphical error analysis was carried out by plotting the predicted values versus actual $P_b$ observation values. Figures 4.1, 4.2, 4.3, and 4.4 show the scatter diagram that compare the modeled $P_b$ with the experimental value for training, validation, testing and overall data points. It is evident that all the figures show the majority of points tended to concentrate in the vicinity of the identity line $y = x$. This indicates excellent agreement between the actual $P_b$ values and the model predicted values.
Table 4.2 – Comparison of $P_b$ Network Architectures - The 5 Best Architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Error, psia</th>
<th></th>
<th></th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training</td>
<td>Validation</td>
<td>Testing</td>
<td></td>
</tr>
<tr>
<td>[4-9-4-1]</td>
<td>79.142</td>
<td>116.597</td>
<td><strong>119.693</strong></td>
<td>99.11</td>
</tr>
<tr>
<td>[4-8-5-4]</td>
<td>72.462</td>
<td>132.506</td>
<td>158.292</td>
<td>99.19</td>
</tr>
<tr>
<td>[4-9-5-1]</td>
<td>85.493</td>
<td>124.296</td>
<td>181.009</td>
<td>98.83</td>
</tr>
<tr>
<td>[4-10-5-1]</td>
<td>59.905</td>
<td>121.312</td>
<td>168.052</td>
<td>99.43</td>
</tr>
<tr>
<td>[4-9-6-1]</td>
<td>77.231</td>
<td>104.626</td>
<td>172.360</td>
<td>99.16</td>
</tr>
</tbody>
</table>
Table 4.3 – Average Error during Training, Validation and Testing \( P_b \) Model

with Various Training Algorithms (Quasi-Newton (Q-N), Conjugate Gradient Descent (CGD), Levenberg-Marquardt (L-M) and Limited Memory Quasi-Newton (LM Q-N))

<table>
<thead>
<tr>
<th>Training Algorithm</th>
<th>( R^2 ) %</th>
<th>Training Error, psia</th>
<th>Validation Error, psia</th>
<th>Test Error, psia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>Max</td>
<td>Average</td>
</tr>
<tr>
<td>Q-N</td>
<td>99.12</td>
<td>77.532</td>
<td>456.831</td>
<td>102.797</td>
</tr>
<tr>
<td>CGD</td>
<td>98.46</td>
<td>104.549</td>
<td>624.501</td>
<td>131.09</td>
</tr>
<tr>
<td>L-M</td>
<td>97.75</td>
<td>135.28</td>
<td>490.169</td>
<td>159.560</td>
</tr>
<tr>
<td>LM Q-N</td>
<td>98.77</td>
<td>88.131</td>
<td>601.371</td>
<td>126.935</td>
</tr>
</tbody>
</table>
Table 4.4 – Ranges of $P_b$ ANN Model Output with Statistical Error Analysis

<table>
<thead>
<tr>
<th></th>
<th>Training output</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual, psia</td>
<td>ANN, psia</td>
<td>AE, psia</td>
<td>ARE</td>
</tr>
<tr>
<td>Average</td>
<td>1689.7</td>
<td>1686.1</td>
<td>130.7</td>
<td>0.126</td>
</tr>
<tr>
<td>Maximum</td>
<td>4244</td>
<td>3858.8</td>
<td>841.5</td>
<td>0.782</td>
</tr>
<tr>
<td>Minimum</td>
<td>121</td>
<td>172.5</td>
<td>1.195</td>
<td>0.0014</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Validation output</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1731</td>
<td>1744</td>
<td>116.4</td>
<td>0.119</td>
</tr>
<tr>
<td>Maximum</td>
<td>4244</td>
<td>3922</td>
<td>495.7</td>
<td>1.023</td>
</tr>
<tr>
<td>Minimum</td>
<td>152</td>
<td>202</td>
<td>0.08</td>
<td>0.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Testing output</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1707</td>
<td>1705</td>
<td>160</td>
<td>0.141</td>
</tr>
<tr>
<td>Maximum</td>
<td>3582</td>
<td>3569</td>
<td>755</td>
<td>0.669</td>
</tr>
<tr>
<td>Minimum</td>
<td>121</td>
<td>202</td>
<td>1.86</td>
<td>0.001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overall output</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1699</td>
<td>1698</td>
<td>133</td>
<td>0.126</td>
</tr>
<tr>
<td>Maximum</td>
<td>4244</td>
<td>3922</td>
<td>841.6</td>
<td>1.023</td>
</tr>
<tr>
<td>Minimum</td>
<td>121</td>
<td>172</td>
<td>0.080</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure 4.1 – Cross Plot of Modeled Bubble Point Pressure on Training Dataset versus Laboratory PVT Values
Figure 4.2 – Cross-plot of Modeled Bubble Point Pressure on Validation Dataset versus Laboratory PVT Analysis
Figure 4.3 - Cross-plot of Modeled Bubble Point Pressure on Testing Dataset versus Laboratory PVT Analysis
Figure 4.4 - Cross-plot of Modeled Bubble Point pressure on Overall Points versus Laboratory PVT Analysis
4.4.2 Contribution of Input Parameters

It is necessary to have some measure of importance to distinguish contributing and non-contributing features. The results of the predictive importance analysis are useful to determine the percentage contribution of the input variables. As for ANNS, the contribution factor is a rough measure of the importance of that variable in predicting the network’s output relative to other input variables in the same network. Identifying the importance or contribution of input parameters of the $P_b$ network is observed. It shows the $R_{sp}$ ranked first with an importance of 62.9% (the network prediction is influenced mostly by $R_{sp}$) and $T_R$ ranked second with an importance of 21.4% while $\gamma_{ST}$ and $P_{sp}$ ranked third and forth with an importance of 8.5% and 7.2%, respectively.

4.4.3 Deployment Test

As a final step in the modelling process, the $P_b$ ANN model is deployed to a new dataset. This is illustrated in Figure 4.5 where the estimated $P_b$, using ANN, was plotted versus the PVT $P_b$ values. The figure shows a tight cloud of points close to the identity line indicating excellent agreement between the actual and model predicted values.

Table 4.5 summarizes the quantitative statistical error analysis for the deployment test. The analysis shows a small error obtained on the deploy datasets. It shows the superiority of the $P_b$ ANN model in performing good generalization.
Figure 4.5 – Deployment Test: Cross-plot of Modeled Bubble Point versus Laboratory PVT Analysis
Table 4.5- Statistical Error Analysis for Deployment Test

<table>
<thead>
<tr>
<th>Deployment test output</th>
<th>Actual, psia</th>
<th>ANN, psia</th>
<th>AE, psia</th>
<th>ARE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>1810</td>
<td>1805</td>
<td>126.6</td>
<td>0.128</td>
</tr>
<tr>
<td>Maximum</td>
<td>4244</td>
<td>3922</td>
<td>695.8</td>
<td>1.013</td>
</tr>
<tr>
<td>Minimum</td>
<td>189</td>
<td>241</td>
<td>2.0</td>
<td>0.004</td>
</tr>
</tbody>
</table>
4.5 Bubble Point Oil FVF ANN Model

Again, the overall PVT dataset (the only available data, 485 points) used for modeling $B_{ob}$ was split into four sets; the training set, the validation set, the testing set and the deployment set. Each was separately used by the ANN for the tasks of training, validation, testing and deploying. 289 points were used for training, 67 points were used for validation, 67 points were used for testing and 57 points were used for the deployment set. A range of each dataset is described in Table 4.6. As described in section 2.3, four inputs ($P_{sp}$, $R_{sp}$, $T_R$ and $\gamma_{oST}$) are selected to incorporate into the input layer and one node into the output layer ($B_{ob}$). Subsequently, the input data were pre-processed and normalized so as to be suitable for neural-network processing. Normalization essentially transforms the input data into a new representation before a network is trained.

With an eye to finding the optimum network design, a trial and error approach was performed, starting with one hidden layer and a number of hidden nodes. A total of 149 architectures are compared to pick the best which displays excellent performance on the testing dataset with the following parameters;

- Hidden layer activation function = Logistic
- Output activation function = Logistic
- Error function = Sum-of-squares
- Hidden layer search range =1 to 2 HL
- $1^{st}$ hidden layer units search range =1 to 15 nodes
- $2^{nd}$ hidden layer units search range =1 to 12 nodes
- Fitness criteria = Inverse test error
Table 4.6 – Ranges of Datasets used in Training, Validation, Testing and Deployment of the $B_{ob}$ ANN Model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Overall PVT Dataset</th>
<th>Training Dataset</th>
<th>Validation Dataset</th>
<th>Testing Dataset</th>
<th>Deployment Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Max</td>
<td>Min</td>
<td>SD</td>
<td>Average</td>
</tr>
<tr>
<td>$P_{sp}$, psia</td>
<td>109.92</td>
<td>517</td>
<td>14.7</td>
<td>101.55</td>
<td>114.10</td>
</tr>
<tr>
<td>$R_{sp}$, SCF/STB</td>
<td>347.5</td>
<td>1175</td>
<td>10</td>
<td>246.3</td>
<td>339.9</td>
</tr>
<tr>
<td>$\gamma_{ST}$, water = 1</td>
<td>0.84027</td>
<td>0.921</td>
<td>0.7999</td>
<td>0.02489</td>
<td>0.84014</td>
</tr>
<tr>
<td>$T_R$, °F</td>
<td>200.48</td>
<td>277</td>
<td>100</td>
<td>47.871</td>
<td>200.47</td>
</tr>
<tr>
<td>$B_{ob}$, psia</td>
<td>1.2835</td>
<td>1.795</td>
<td>1.047</td>
<td>0.14645</td>
<td>1689.7</td>
</tr>
</tbody>
</table>
The comparison indicates none of the networks offered an acceptable error with one hidden layer. Table 4.7 presents the five best architectures of the compared 149 architectures. In general, Architectures [4−5−4−1], [4−8−6−1] and [4−9−6−1] gave similar performances. However, the overall network architecture was determined to be a configuration of one layer of inputs with four nodes, two hidden layers with five nodes in the first hidden layer and four in the second, and one output, [4−5−4−1]. An important step before training the selection architecture model is to determine the optimal training algorithm. The usual strategy is based solely upon a trial-and-error procedure, which may be very time consuming and requires significant user experience. However, the selected architecture was trained by the most commonly used training algorithms including Quasi-Newton BP (Q-N), Conjugate Gradient Descent BP (CGD), Levenberg-Marquardt Bp (L-M) and Limited Memory Quasi-Newton BP (LM Q-N)) with different activation functions. In each case, the architecture was retrained five times with different initial weight randomization (see appendix B section B.4.6). As a result, a total of 16 models were compared and tabulated in Table 4.8. The table presents the networks’ performance in terms of average training error, average validation error and average testing error. The best results of each training algorithm group are emphasized in bold. Thus, the results indicate that the Levenberg-Marquardt algorithm gave the largest average errors and exhibited bad performances in all cases.
Figure 4.6 – $B_{ob}$ BP-Model Architecture, [4-5-4-1]
Table 4.7 – Comparison of $B_{ob}$ Network Architectures - The 5 Best Architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Error, bbl/STB</th>
<th></th>
<th></th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4-5-2-1]</td>
<td>0.017</td>
<td>0.016</td>
<td>0.018</td>
<td>97.53</td>
</tr>
<tr>
<td>[4-3-5-1]</td>
<td>0.019</td>
<td>0.016</td>
<td>0.018</td>
<td>97.04</td>
</tr>
<tr>
<td>[4-5-4-1]</td>
<td>0.014</td>
<td>0.016</td>
<td>0.017</td>
<td>98.4</td>
</tr>
<tr>
<td>[4-8-6-1]</td>
<td>0.015</td>
<td>0.016</td>
<td>0.018</td>
<td>98.20</td>
</tr>
<tr>
<td>[4-9-6-1]</td>
<td>0.015</td>
<td>0.015</td>
<td>0.018</td>
<td>98.2</td>
</tr>
</tbody>
</table>
• With the use of the Q-N, LM Q-N and CGD training algorithms, the resulting average errors are almost similar for all the used activation functions.

• When the hidden layers were activated by the hyperbolic activation function and the output layer was activated by the logistic activation function and trained with the Conjugate Gradient Descent BP algorithm, the model provided the smallest average error of 0.01213 on the training dataset.

• The smallest average error of 0.01409 on the validation dataset is achieved with the network hidden layers being activated by the logistic activation function and the hyperbolic activation function activated the output layer and in turn, was trained with Quasi-Newton.

• The smallest average testing error of 0.01688 was obtained when the network was trained with limited memory Quasi-Newton.

• The highest $R^2$ of 98.76% is provided by training the network with the Conjugate Gradient Descent BP algorithm and activating the hidden layers by hyperbolic and the output layers by the logistic activation layer.

As the testing dataset is used neither in training nor in the validation phase, this set is used to test the final model. Thus, the limited memory Quasi-Newton and Conjugate Gradient Descent training algorithms gave smaller average error of 0.0169 on testing set.

Consequently, it appears to be too difficult to select the best model. However, network Models M1, M5, M9 and M12, in Table 4.8, are selected as the best based on model performance on the testing dataset and this can still be demonstrated by
determining a statistical error analysis to specify the best model among the four models (M1, M5, M9 and M12). Comparisons of average, minimum and maximum of absolute error and absolute percent relative error across the four models are presented in Table 4.9. The table addresses the errors obtained from training, validation, testing and overall datasets. However, the bottom line of comparing models and selecting the best is to observe the absolute error and absolute relative error of the validation and testing dataset in addition to the scatter diagram. The best results of each training model are emphasized in bold. Via the results, the lowest errors in terms of Absolute Error (AE) and Absolute Relative Error (ARE), based on validation and testing datasets can be summarized as:

- **The Validation Dataset,**
  - The Conjugate Gradient Descent BP algorithm (network M12) gives the lowest average AE of 0.0148, lowest average ARE of 0.011291 %, lowest minimum AE of 0.000011, lowest minimum ARE of 0.00009 %, lowest maximum AE of 0.066787 and highest $R^2$ of 97.95% when the hidden layers are activated by the hyperbolic tangent function and the output layer is activated by the logistic function. However, its maximum ARE is 0.0543 %, almost equivalent to the lowest value of 0.04898 obtained by training the network using Quasi-Newton (network M1).
• The Testing Dataset,
  o A maximum regression coefficient of 97.76% was obtained by the Conjugate Gradient Descent BP algorithm when hidden and output layers were activated with the logistic activation function, network M9.
  o A lowest minimum AE and ARE of 0.000159 and 0.000141%, respectively was obtained by training the network with the Conjugate Gradient Descent BP algorithm when the hidden layers were activated with the hyperbolic function and the output layer was activated with the logistic activation function (network M12), which is equivalent to the Conjugate Gradient Descent BP algorithm when hidden and output layers are activated with the logistic activation function, network M9.

• The Overall Dataset (Training + Validation + Testing sets),
  o The lowest minimum and maximum AE and lowest minimum and maximum ARE% are achieved by network M9 in which the network is trained by the Conjugate Gradient Descent BP algorithm and the network layers are activated with the logistic activation function. The values are almost equivalent to those obtained by network M12. Moreover, M9 provides an $R^2$ of 98% on overall data points which is equivalent to the 89.3% obtained by M12.

As a result, from the above discussions, the networks M12 in Table 4.9 gave best performance on validation dataset and overall dataset, while network M9 gave best
performance on testing dataset. Still there are challenges to select the best network model. It needs investigations.

4.5.1 Graphical Error Analysis

In this case, it is too difficult to select the best model from M9 and M12. The error terms were analyzed graphically. $y = x$ line cross-plot, error histogram and probability plot for the error terms resulted from training, validation, testing and overall datasets which are presented in Figures 4.2, 4.3, 4.4 and 4.5, respectively. To choose the best model, the points must be observed and compared where the data points fall, which roughly follows a straight line in the probability plot. The AD value provided with the plot can be used. It measures how well the data follow a particular distribution. The better the distribution fits the data, the smaller the AD value will be.

- **Training set** (Figure 4.7, c and d) shows both M9 and M12 are normally distributed with zero mean. M12 demonstrates a lower SD of 0.01671 compared with 0.01951 for M9. M12 has a smaller AD value. Therefore, M12 is the best training dataset.

- **Validation set** (Figure 4.8, a - f): Both offer same shape and influence.

- **Testing dataset** (Figure 4.9): According to Figure 4.9 (e), Model M9 follows a straight line better than M12 does. M9 has a low AD value and a low SD. Therefore, M9 is a better on testing dataset.

- **Overall points** (Figure 4.10): M12 offers a lower SD while M9 offers a lower AD value.

102
Table 4.8 - Average Error during Training, Validation and Testing of Architecture [4-5-4-1] with Various Training Algorithms and Input/Output Activation Functions. (Illustrates the Effect of Choice of Training Algorithm and Activation Function)

<table>
<thead>
<tr>
<th>Model No.</th>
<th>Activation Function</th>
<th>Average Training Error, bbl/STB</th>
<th>Average Validation Error, bbl/STB</th>
<th>Average Testing Error, bbl/STB</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hidden Output</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Quasi-Newton BP Algorithm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>Logistic Logistic</td>
<td>0.01829</td>
<td>0.01625</td>
<td>0.01828</td>
<td>97.29</td>
</tr>
<tr>
<td>M2</td>
<td>Hyperbolic Hyperbolic</td>
<td>0.01576</td>
<td>0.0145</td>
<td>0.02039</td>
<td>98.02</td>
</tr>
<tr>
<td>M3</td>
<td>Logistic Hyperbolic</td>
<td><strong>0.01424</strong></td>
<td><strong>0.01409</strong></td>
<td>0.01908</td>
<td><strong>98.422</strong></td>
</tr>
<tr>
<td>M4</td>
<td>Hyperbolic Logistic</td>
<td>0.01671</td>
<td>0.01537</td>
<td>0.01914</td>
<td>97.662</td>
</tr>
<tr>
<td><strong>Limited memory Quasi-Newton BP Algorithm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M5</td>
<td>Logistic Logistic</td>
<td>0.01376</td>
<td>0.01602</td>
<td><strong>0.0169</strong></td>
<td><strong>98.31</strong></td>
</tr>
<tr>
<td>M6</td>
<td>Hyperbolic Hyperbolic</td>
<td>0.01772</td>
<td><strong>0.01525</strong></td>
<td>0.01926</td>
<td>97.5</td>
</tr>
<tr>
<td>M7</td>
<td>Logistic Hyperbolic</td>
<td>0.01728</td>
<td>0.01544</td>
<td>0.01933</td>
<td>97.68</td>
</tr>
<tr>
<td>M9</td>
<td>Hyperbolic Logistic</td>
<td>0.01707</td>
<td>0.01531</td>
<td>0.01966</td>
<td>97.79</td>
</tr>
<tr>
<td><strong>Conjugate Gradient Descent BP Algorithm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M9</td>
<td>Logistic Logistic</td>
<td>0.01372</td>
<td>0.01584</td>
<td><strong>0.0169</strong></td>
<td><strong>98.33</strong></td>
</tr>
<tr>
<td>M10</td>
<td>Hyperbolic Hyperbolic</td>
<td>0.01494</td>
<td><strong>0.01456</strong></td>
<td>0.02101</td>
<td>98.24</td>
</tr>
<tr>
<td>M11</td>
<td>Logistic Hyperbolic</td>
<td>0.01777</td>
<td>0.01515</td>
<td>0.01838</td>
<td>97.48</td>
</tr>
<tr>
<td>M12</td>
<td>Hyperbolic Logistic</td>
<td><strong>0.01213</strong></td>
<td>0.0148</td>
<td>0.01746</td>
<td><strong>98.76</strong></td>
</tr>
<tr>
<td><strong>Levenberg-Marquardt BP Algorithm</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M13</td>
<td>Logistic Logistic</td>
<td>0.02044</td>
<td>0.01839</td>
<td><strong>0.02189</strong></td>
<td>97.22</td>
</tr>
<tr>
<td>M14</td>
<td>Hyperbolic Hyperbolic</td>
<td>0.01924</td>
<td>0.0191</td>
<td>0.02413</td>
<td>97.52</td>
</tr>
<tr>
<td>M15</td>
<td>Logistic Hyperbolic</td>
<td><strong>0.01844</strong></td>
<td><strong>0.01809</strong></td>
<td>0.02286</td>
<td><strong>97.68</strong></td>
</tr>
<tr>
<td>M16</td>
<td>Hyperbolic Logistic</td>
<td>0.02168</td>
<td>0.02</td>
<td>0.02272</td>
<td>96.86</td>
</tr>
</tbody>
</table>
### Table 4.9 – Comparison between the Four Best Network Models

#### Model M1
**Quasi-Newton BP Algorithm**
Hidden activating Function = Logistic  
Output activation function = Logistic

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
<th>Overall Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AE</td>
<td>ARE</td>
<td>AE</td>
<td>ARE</td>
</tr>
<tr>
<td></td>
<td>bbl/STB</td>
<td></td>
<td>bbl/STB</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.01655</td>
<td>0.01213</td>
<td>0.01497</td>
<td>0.01122</td>
</tr>
<tr>
<td>Min</td>
<td>0.00001</td>
<td>0.00001</td>
<td>0.00008</td>
<td>0.00007</td>
</tr>
<tr>
<td>Max</td>
<td>0.12892</td>
<td>0.09040</td>
<td>0.06832</td>
<td><strong>0.04898</strong></td>
</tr>
<tr>
<td>$R^2$ %</td>
<td>97.18</td>
<td>97.59</td>
<td>95.74</td>
<td>97.07</td>
</tr>
</tbody>
</table>

#### Model M5
**Limited memory Quasi-Newton BP Algorithm**
Hidden activating Function = Logistic  
Output activation function = Logistic

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
<th>Overall Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AE</td>
<td>ARE</td>
<td>AE</td>
<td>ARE</td>
</tr>
<tr>
<td></td>
<td>bbl/STB</td>
<td></td>
<td>bbl/STB</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.01378</td>
<td>0.01039</td>
<td>0.0159</td>
<td>0.01184</td>
</tr>
<tr>
<td>Min</td>
<td>0.00028</td>
<td>0.00024</td>
<td>0.00011</td>
<td>0.00007</td>
</tr>
<tr>
<td>Max</td>
<td>0.11233</td>
<td>0.07877</td>
<td>0.08618</td>
<td>0.06178</td>
</tr>
<tr>
<td>$R^2$ %</td>
<td>98.27</td>
<td>97.58</td>
<td>96.74</td>
<td>97.69</td>
</tr>
</tbody>
</table>

#### Model M9
**Limited memory Quasi-Newton BP Algorithm**
Hidden activating Function = Logistic  
Output activation function = Logistic

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
<th>Overall Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AE</td>
<td>ARE</td>
<td>AE</td>
<td>ARE</td>
</tr>
<tr>
<td></td>
<td>bbl/STB</td>
<td></td>
<td>bbl/STB</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>0.01368</td>
<td>0.01031</td>
<td>0.01550</td>
<td>0.01158</td>
</tr>
<tr>
<td>Min</td>
<td><strong>0.00000</strong></td>
<td><strong>0.00000</strong></td>
<td>0.00016</td>
<td>0.00013</td>
</tr>
<tr>
<td>Max</td>
<td><strong>0.11129</strong></td>
<td><strong>0.07805</strong></td>
<td>0.07972</td>
<td>0.05714</td>
</tr>
<tr>
<td>$R^2$ %</td>
<td>98.29</td>
<td>97.65</td>
<td><strong>97.76</strong></td>
<td>98</td>
</tr>
</tbody>
</table>

#### Model M12
**Limited memory Quasi-Newton BP Algorithm**
Hidden activating Function = Hyperbolic  
Output activation function = Logistic

<table>
<thead>
<tr>
<th></th>
<th>Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
<th>Overall Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AE</td>
<td>ARE</td>
<td>AE</td>
<td>ARE</td>
</tr>
<tr>
<td></td>
<td>bbl/STB</td>
<td></td>
<td>bbl/STB</td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td><strong>0.01145</strong></td>
<td><strong>0.00852</strong></td>
<td><strong>0.01445</strong></td>
<td><strong>0.01063</strong></td>
</tr>
<tr>
<td>Min</td>
<td>0.00002</td>
<td>0.00002</td>
<td><strong>0.00001</strong></td>
<td><strong>0.00001</strong></td>
</tr>
<tr>
<td>Max</td>
<td>0.08624</td>
<td>0.06048</td>
<td><strong>0.06679</strong></td>
<td>0.0543</td>
</tr>
<tr>
<td>$R^2$ %</td>
<td><strong>98.74</strong></td>
<td>97.95</td>
<td>96.27</td>
<td><strong>98.3</strong></td>
</tr>
</tbody>
</table>
Figure 4.7 (a, b) – Graphical Error Analysis on Training Dataset

**(a)**
 Scatter Diagram
Bob ANN Model # 9 (Training Dataset)

**(b)**
 Scatter Diagram
Bob ANN Model # 12 (Training Dataset)
Figure 4.7 (c, d) – Graphical Error Analysis on Training Dataset

Figure (c) shows the histogram of the error terms for Bob ANN Model # 9, with a mean of -0.00003257, standard deviation of 0.01951, and N = 286. The error terms range from -0.12 to 0.12, and the frequency distribution is skewed slightly to the left.

Figure (d) shows the histogram of the error terms for Bob ANN Model # 12, with a mean of 0.00005602, standard deviation of 0.01671, and N = 286. The error terms range from -0.12 to 0.12, and the frequency distribution is more symmetric compared to Model # 9.
Figure 4.7 (e, f) – Graphical Error Analysis on Training Dataset
Figure 4.8 (a, b) – Graphical Error Analysis on Validation Dataset

Scatter Diagram
Bob ANN Model # 9 (Validation Dataset)

Scatter Diagram
Bob ANN Model #12 (Validation Dataset)
Figure 4.8 (c, d) – Graphical Error Analysis on Validation Dataset

(c) Histogram of the Error Terms
Bob ANN Model # 9 (Validation Dataset)

(d) Histogram of the Error Terms
Bob ANN Model # 12 (Validation Dataset)
Figure 4.8 (e, f) – Graphical Error Analysis on Validation Dataset

Probability Plot of the Error Terms (Validation Dataset)
Bob ANN Model # 9, Normal - 95% CI

Probability Plot of Error Terms (Validation Dataset)
Bob ANN Model # 12, Normal - 95% CI
Figure 4.9 (a, b) – Graphical Error Analysis on Testing Dataset
Figure 4.9 (c, d) – Graphical Error Analysis on Testing Dataset

(c) Histogram of Error Terms
Bob ANN Model # 9 (Testing Dataset)

(d) Histogram of the Error Terms
Bob ANN Model # 12 (Testing Dataset)
Figure 4.9 (e, f) – Graphical Error Analysis on Testing Dataset

(e)

(f)
Figure 4.10 (a, b) – Graphical Error Analysis on Overall Dataset
Figure 4.10 (c, d) – Graphical Error Analysis on Overall Dataset
Figure 4.10 (e, f) – Graphical Error Analysis on Overall Dataset

- **(e)** Probability Plot of Error Terms (Overall Dataset)
  - Bob ANN Model # 9, Normal - 95% CI
  - Mean: 0.0001417
  - SD: 0.02082
  - N: 420
  - AD: 4.997
  - P-Value: <0.005

- **(f)** Probability Plot of Error Terms (Overall Dataset)
  - Bob ANN Model # 12, Normal - 95% CI
  - Mean: 0.0001485
  - SD: 0.01903
  - N: 420
  - AD: 4.749
  - P-Value: <0.005
As result, the graphical analysis provides the same results for the validation set for both models. M12 is the best on training and M9 is the best testing set. Therefore, the final decision is taken based on the ability of a model to perform reliably on any novel data in a given domain. Thus, M9 was considered the best BP-ANN model to predict oil formation volume vector at the bubble point pressure.

4.5.2 Deployment Test

As a final step in the modelling process, the proposed $B_{ob}$ PB-ANN model is deployed to a new dataset. This data consists of 62 data points which were introduced to the trained BP-ANN model (M9) and the corresponding $B_{ob}$'s were obtained. The test is illustrated in Figure 4.11 where the model $B_{ob}$ is plotted versus the laboratory PVT values. It shows a tight cloud of points near the identity line, indicating excellent agreement between the actual $B_{ob}$ PVT data and model predicted values. Table 4.10 summarizes the quantitative statistical error analysis for the deployment test. The analysis shows a small error obtained on the deploy datasets. It shows the superiority of the $B_{ob}$ ANN model to perform good generalization.

4.6 Bubble Point Solution Gas/Oil ratio ($R_{sob}$) ANN Model

Prior to specifying the $R_{sob}$ ANN model architecture, the overall collected PVT data were split into four sets. Fifty-nine percent of the collected PVT data was used to train the network models, 14% was used in the validation process to investigate the model’s performance, another 14% was used to test the model and fine-tune it and the remaining
13% PVT data was used to deploy the selected model. The range of each dataset is described in Table 4.11.

The development of ANN models starts with finding the best network structure to represent the complicated relationship between variables. Possible ANN architectures, starting from a low number of layers and nodes to a higher number of layers and nodes, should be tested to select the best architecture. Successful architectures are those that converge to the target error or reach the possible MSE and exhibit stable performance to new data not included in the training. Topically, the best ANN architecture is the one that provides the lowest training error and testing and validation errors with the highest $R^2$. 
Figure 4.11 – Deployment Test of $B_{ob}$ BP-ANN Model
Table 4.10- Statistical Error Analysis for Deployment Test

<table>
<thead>
<tr>
<th></th>
<th>$B_{ob}$ Deployment Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PVT analysis</td>
</tr>
<tr>
<td></td>
<td>bbl/STB</td>
</tr>
<tr>
<td>Mean</td>
<td>1.320</td>
</tr>
<tr>
<td>Maximum</td>
<td>1.674</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.064</td>
</tr>
</tbody>
</table>
Table 4.11 – Ranges of Datasets used in Training, Validation, Testing and Deployment

\( R_{\text{sof}} \) ANN Model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Overall PVT Dataset</th>
<th>Training Dataset</th>
<th>Validation Dataset</th>
<th>Testing Dataset</th>
<th>Deployment Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Maximum</td>
<td>Minimum</td>
<td>SD</td>
<td>Average</td>
</tr>
<tr>
<td>( P_{sp} ), psia</td>
<td>109.92</td>
<td>517</td>
<td>14.7</td>
<td>101.55</td>
<td>114.10</td>
</tr>
<tr>
<td>( R_{sp} ), SCF/STB</td>
<td>347.5</td>
<td>1175</td>
<td>10</td>
<td>246.3</td>
<td>339.9</td>
</tr>
<tr>
<td>( \gamma_{\text{ST}}, \text{water} = 1 )</td>
<td>0.84027</td>
<td>0.921</td>
<td>0.7999</td>
<td>0.02489</td>
<td>0.84014</td>
</tr>
<tr>
<td>( T_R, , ^\circ\text{F} )</td>
<td>200.48</td>
<td>277</td>
<td>100</td>
<td>47.871</td>
<td>200.47</td>
</tr>
<tr>
<td>( P_b, ) psia</td>
<td>1698.9</td>
<td>4244</td>
<td>121</td>
<td>1090</td>
<td>1689.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.6.1 Model Design

A trial and error attempt was carried out to achieve a well-designed $R_{sob}$ network model architecture. An input vector and the corresponding desired output are considered first. Three inputs ($R_{sp}$, $P_{sp}$ and $\gamma_{ST}$) are selected to incorporate into the input layer. Next, a series of networks with single and multiple hidden layers with different numbers of nodes in each hidden layer were trained with different activation functions and training algorithms. A total of 48 architectures were compared, starting with one hidden layer and one hidden node. Hidden nodes were then gradually added. The following network parameters were used in this process:

- Hidden layer activation function = Logistic
- Output activation function = Logistic
- Error function = Sum-of-squares
- Hidden layers search range = 1 to 3 HL
- 1st layer hidden units search range = 1 to 8 nodes
- 2nd hidden layer units search range = 1 to 5 nodes
- 3rd hidden layer units search range = 1 to 3 nodes
- Fitness criteria = Inverse test error

In the training procedure, the inputs propagated forward through the network to the output vector ($R_{sob}$). The output vector is compared with the desired output (actual $R_{sob}$), and the errors are determined. The errors are then propagated back through the network.
from the output to the input layer. This route is repeated until the errors begin to minimize. The training results showed none of the networks with one or three hidden layers gave an acceptable error. The 5 best network architectures are presented in Table 4.12 where the best results are emphasized in bold. It is evident that Architecture [3−7−4−1] gives the lowest training and testing error, with the highest $R^2$ of 99.27%.

Architecture [3−7−4−1], with the lowest average error and the highest $R^2$, was selected as the best. Following this, a selected architecture with two hidden layers, 7 nodes in the first hidden layer, and 4 nodes in the second was trained using different algorithms and activation functions. In each case, the architecture was retrained five times with different random initial weights with concerns over the possibility of converging to a local minimum in weight space. However, a typical neural network might have over two hundred weighs whose values must be found to produce an optimal solution. Once a network settles on a minimum, whether local or global, the learning stops. If a local minimum is reached, the error at the network outputs may still be unacceptably high. If a network stops learning before reaching an acceptable solution, a change in the activation function or in the learning algorithm will often fix the problem.
Table 4.12 – Comparison of $R_{sob}$ Network Architectures - The 5 Best Architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Training Error, SCF/STB</th>
<th>Validation Error, SCF/STB</th>
<th>Testing Error, SCF/STB</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3-6-3-1]</td>
<td>16.05</td>
<td>13.48</td>
<td>12.52</td>
<td>99.14</td>
</tr>
<tr>
<td>[3-7-4-1]</td>
<td><strong>14.27</strong></td>
<td>16.57</td>
<td><strong>11.93</strong></td>
<td><strong>99.27</strong></td>
</tr>
<tr>
<td>[3-8-4-1]</td>
<td>15.45</td>
<td><strong>12.37</strong></td>
<td>13.00</td>
<td>99.14</td>
</tr>
<tr>
<td>[3-5-5-1]</td>
<td>15.95</td>
<td>15.25</td>
<td>13.19</td>
<td>99.11</td>
</tr>
<tr>
<td>[3-7-5-1]</td>
<td>15.62</td>
<td>15.60</td>
<td>12.75</td>
<td>99.11</td>
</tr>
</tbody>
</table>
Figure 4.12 – $R_{sob}$ BP-ANN Model Architecture [3-7-4-1]
Table 4.13 shows the average and maximum error during training, validation and testing of Architecture [4−7−4−1] under various training algorithms, namely, Quick Propagation, Quasi-Newton, Conjugate Gradient Descent, and Limited Memory Quasi-Newton. It is evident that the Limited Memory Quasi-Newton training algorithm had the higher $R^2$, lowest average training, average validation and average testing error, as emphasized in bold.

4.6.2 Model Evaluation

A four-layer network is developed to estimate $R_{sob}$. Both qualitative and graphical analysis of the model’s error terms was utilized to verify the accuracy of the selected model. Table 4.14 summarizes the quantitative statistical error analysis for the identified $R_{sob}$ ANN model. The analysis shows a small AE and ARE on training, validation and testing datasets. The model provided predicted $R_{sob}$ values with an average absolute error of 15.68, an average absolute relative error of 7.25% and a correlation coefficient of 99.2%.
Table 4.13 - Average Error during Training, Validation and Testing $R_{soh}$ Model with Various Training Algorithms (Quick Propagation (QP), Quasi-Newton (Q-N), Conjugate Gradient Descent (CGD), and Limited Memory Quasi-Newton (LM Q-N))

<table>
<thead>
<tr>
<th>Training Algorithm</th>
<th>Training Error, SCF/STB</th>
<th>Validation Error, SCF/STB</th>
<th>Testing Error, SCF/STB</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Max</td>
<td>Average</td>
<td>Max</td>
</tr>
<tr>
<td>QP</td>
<td>17.05</td>
<td>129.85</td>
<td>16.25</td>
<td>66.81</td>
</tr>
<tr>
<td>CGD</td>
<td>16.18</td>
<td>136.02</td>
<td>15.66</td>
<td>59.8</td>
</tr>
<tr>
<td><strong>LM Q-N</strong></td>
<td><strong>13.928</strong></td>
<td><strong>127.81</strong></td>
<td><strong>11.87</strong></td>
<td><strong>73.98</strong></td>
</tr>
<tr>
<td>QN</td>
<td>14.678</td>
<td>133.16</td>
<td>14.39</td>
<td>136.89</td>
</tr>
</tbody>
</table>
Table 4.14 – Statistical Error Analysis of $R_{sob}$ ANN Model Output

<table>
<thead>
<tr>
<th></th>
<th>Training Output</th>
<th>Validation Output</th>
<th>Testing Output</th>
<th>Overall Output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual</td>
<td>ANN</td>
<td>AE</td>
<td>ARE</td>
</tr>
<tr>
<td>SCF/STB</td>
<td>Average</td>
<td>411.66</td>
<td>411.64</td>
<td>16.18</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>17</td>
<td>51.312</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1256</td>
<td>1201.17</td>
<td>136.02</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>413.01</td>
<td>413.10</td>
<td>15.66</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>40</td>
<td>65.53</td>
<td>0.079</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1053</td>
<td>1039.19</td>
<td>59.8</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>436.42</td>
<td>436.02</td>
<td>13.58</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>42</td>
<td>69.54</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1135</td>
<td>1140.65</td>
<td>83.32</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>415.81</td>
<td>415.75</td>
<td>15.68</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>17</td>
<td>51.31</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1256</td>
<td>1201.17</td>
<td>136.01</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Furthermore, the graphical error analysis was carried out by plotting the ANN model predicted values versus laboratory PVT observation $R_{\text{sob}}$ values. Figures 4.13, 4.14, 4.15 and 4.16 show the scatter diagram which compares the actual with the $R_{\text{sob}}$ outputs for training, validation, testing and overall datasets. It is evident that all the figures show the majority of points tended to concentrate in the vicinity of the identity 1:1 line, indicating excellent agreement between the actual and model predicted $R_{\text{sob}}$ values. Moreover, the histogram of the error terms resulting from training, validation, testing, and overall datasets, are presented in Figures 4.17, 4.18, 4.19 and 4.20, respectively. All histograms show the error terms are normally distributed.
Figure 4.13 – Cross-plot of Modeled $R_{sob}$ on Training Dataset vs. Laboratory PVT Values
Figure 4.14 – Cross-plot of Modeled $R_{sob}$ on Validation Dataset vs. Laboratory PVT Values
Figure 4.15 – Cross-plot of Modeled $R_{sob}$ on Testing Dataset vs. Laboratory PVT Values
Figure 4.16 – Cross-plot of Modeled $R_{sob}$ on Overall Data Points vs. Laboratory PVT Values
Figure 4.17 – Histogram of the Residuals of Training Dataset
Figure 4.18 – Histogram of the Residuals of Validation Dataset
Histogram of the Error Terms
RsoB ANN Model (Testing Dataset)

Figure 4.19 – Histogram of the Residuals of Testing Dataset
Histogram of the Error Terms
Rsoh ANN Model (Overall Dataset)

Mean 0.06609
StDev 23.15
N 440

Figure 4.20 – Histogram of the Residuals of Overall Data Points
4.6.3 Model Deployment

Model deployment refers to the application of the model to a new dataset as a final step in modeling development. This step involves the application of the selected model into a new dataset in order to test the generalization of the model. Thus, after a satisfactory $R_{sob}$, the ANN model has been identified as the best model and the model is deployed to a new dataset. This is illustrated in Figure 4.21 where the estimated $R_{sob}$, using the ANN model, was plotted against the actual $R_{sob}$ values. The figure shows a tight cloud of points near the identity line, indicating an excellent agreement between the actual and model predicted values.

Table 4.15 summarizes the quantitative statistical error analysis for the deployment test. The analysis shows a small error obtained on the deploy datasets. It shows the superiority of the $R_{sob}$ ANN model in performing a good generalization.

4.7 Stock-Tank Vent Gas Oil Ratio ANN Model

Three inputs ($P_{SP}$, $T_{SP}$ and $\gamma_{ST}$) are selected to incorporate into the input layer after a forward-backward search algorithm is carried out to select the best subset of input variables. Then, to specify the number of hidden layers and the number of units in each layer, a trial and error attempt was carried out, starting with one hidden layer and one hidden unit. Hidden units were then gradually added. A total of 672 architectures were compared with the following network parameters:
Figure 4.21 – $R_{sob}$ BP-ANN Model Deployment Test
Table 4.15- Statistical Error Analysis for Deployment Test

<table>
<thead>
<tr>
<th>$R_{sob}$ Deployment Test</th>
<th>PVT Analysis</th>
<th>BP-ANN</th>
<th>AE</th>
<th>ARE, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCF/STB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>465</td>
<td>465</td>
<td>10.758</td>
<td>4.573</td>
</tr>
<tr>
<td>Maximum</td>
<td>1106</td>
<td>1114</td>
<td>46.403</td>
<td>38.90</td>
</tr>
<tr>
<td>Minimum</td>
<td>56</td>
<td>78</td>
<td>0.9022</td>
<td>0.1163</td>
</tr>
</tbody>
</table>
Hidden layer activation function = Logistic and hyperbolic tangent
Output activation function = Logistic and hyperbolic tangent
Error function = Sum-of-squares
Hidden layers search range = 1 to 3 HL
1st layer hidden units search range = 1 to 8 nodes
2nd hidden layer units search range = 1 to 5 nodes
3rd hidden layer units search range = 1 to 3 nodes
Fitness criteria = Inverse test error

The results show not one of the networks gave an acceptable error with one hidden layer. However, the use of a three hidden layer network with an activation function of hyperbolic tangent for both hidden and output layers gives better results. Table 4.16 presents the best architectures among the compared 672 architectures. It is evident that the hyperbolic tangent activation function gives the lowest testing error.

One step prior to selecting the best architecture is to determine the optimal training algorithm. The usual strategy is based solely upon a trial-and-error procedure, which could be very time consuming and requires significant user experience. However, Architectures [3-6-5-1-1], [3-8-4-2-1], [3-7-3-3-1], [3-7-4-3-1] and [3-8-5-3-1], which are in bold print in Table 4.15, are trained by the most commonly used training algorithms including Quasi-Newton BP (Q-N), Conjugate Gradient Descent BP (CGD) and Limited Memory Quasi-Newton BP (LM Q-N) (Rahman et al. 2011 and Talib et al. 2009). In each case, the architecture was retrained five times with different initial weights randomization. As a result, a total of 15 models were compared and tabulated in Table 4.17. The table presents the network’s performance in terms of average training error,
average validation error and average testing error. The best results of each training algorithm group are emphasized in bold.

As a result, Architecture [3-8-4-2-1] with three hidden layers, seven nodes in the first hidden layer, four in the second and three nodes in the third was selected as a ANN model to predict the $R_{ST}$. The model gave the best performance on training and testing datasets.

4.7.1 Stock-Tank Vent Gas/Oil Ratio ANN Model Evaluation

Both qualitative and graphical analysis of the model error terms was utilized to verify the accuracy of the selected model. Table 4.18 summarizes the quantitative statistical error analysis for the identified $R_{ST}$ BP-ANN model. The analysis shows a small error in the training results, validation and testing datasets.

Furthermore, the graphical error analysis was carried out by plotting the predicted values versus actual $R_{ST}$ observation values. Figures 4.22, 4.23, 4.24 and 4.25 show the scatter diagram that compare the actual $P_b$ values with the outputs for training, validation, testing and overall data points of the $R_{ST}$ ANN model. It is evident that all the figures show most of the points tended to concentrate in the vicinity of the identity line $y = x$. This indicates excellent agreement between the laboratory PVT $R_{ST}$ values and the modeled values.
Table 4.16 - Comparison of $R_{ST}$ Network Architectures - The Best Architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Activation Function</th>
<th>Error, SCF/STB</th>
<th>$R^2$ %</th>
</tr>
</thead>
<tbody>
<tr>
<td>[3-6-5-2-1]</td>
<td>Logistic</td>
<td>9.54 17.53 19.97</td>
<td>90.93</td>
</tr>
<tr>
<td>[3-7-4-1]</td>
<td>Hyperbolic</td>
<td>9.29 13.31 19.83</td>
<td>91.43</td>
</tr>
<tr>
<td>[3-4-5-1-1]</td>
<td>Logistic</td>
<td>11.19 13.41 16.63</td>
<td>88.46</td>
</tr>
<tr>
<td>[3-6-5-1-1]</td>
<td>Hyperbolic</td>
<td>8.46 11.13 16.19</td>
<td>92.88</td>
</tr>
<tr>
<td>[3-8-4-2-1]</td>
<td>Hyperbolic</td>
<td>10.39 13.54 15.45</td>
<td>90.90</td>
</tr>
<tr>
<td>[3-7-3-3-1]</td>
<td>Hyperbolic</td>
<td>8.72 13.04 16.43</td>
<td>92.72</td>
</tr>
<tr>
<td>[3-7-4-3-1]</td>
<td>Hyperbolic</td>
<td>7.45 11.92 20.27</td>
<td>94.69</td>
</tr>
<tr>
<td>[3-8-5-3-1]</td>
<td>Hyperbolic</td>
<td>8.68 12.46 15.78</td>
<td>92.56</td>
</tr>
</tbody>
</table>
Table 4.17 - Average Error during Training, Validation and Testing $R_{ST}$ Model with Various Training Algorithms (Quasi-Newton (Q-N), Conjugate Gradient Descent (CGD), and Limited Memory Quasi-Newton (LM Q-N))

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Training Algorithm</th>
<th>Average Error, SCF/STB</th>
<th>$R^2%$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training</td>
<td>Validation</td>
</tr>
<tr>
<td>[3-6-5-1-1]</td>
<td>Q-N</td>
<td>13.28</td>
<td>13.19</td>
</tr>
<tr>
<td></td>
<td>CGD</td>
<td>11.16</td>
<td>12.85</td>
</tr>
<tr>
<td></td>
<td>LM Q-N</td>
<td>10.62</td>
<td>13.35</td>
</tr>
<tr>
<td>[3-8-4-2-1]</td>
<td>Q-N</td>
<td>8.60</td>
<td>12.66</td>
</tr>
<tr>
<td></td>
<td>CGD</td>
<td>15.20</td>
<td>13.53</td>
</tr>
<tr>
<td></td>
<td>LM Q-N</td>
<td>13.34</td>
<td>13.73</td>
</tr>
<tr>
<td>[3-7-3-3-1]</td>
<td>Q-N</td>
<td>11.01</td>
<td>11.88</td>
</tr>
<tr>
<td></td>
<td>CGD</td>
<td>12.83</td>
<td>13.64</td>
</tr>
<tr>
<td></td>
<td>LM Q-N</td>
<td>15.11</td>
<td>13.19</td>
</tr>
<tr>
<td>[3-7-4-3-1]</td>
<td>Q-N</td>
<td>8.88</td>
<td>12.86</td>
</tr>
<tr>
<td></td>
<td>CGD</td>
<td>14.47</td>
<td>13.67</td>
</tr>
<tr>
<td></td>
<td>LM Q-N</td>
<td>8.60</td>
<td>12.44</td>
</tr>
<tr>
<td>[3-8-5-3-1]</td>
<td>Q-N</td>
<td>13.85</td>
<td>14.03</td>
</tr>
<tr>
<td></td>
<td>CGD</td>
<td>12.67</td>
<td>13.89</td>
</tr>
<tr>
<td></td>
<td>LM Q-N</td>
<td>12.39</td>
<td>13.45</td>
</tr>
<tr>
<td>SCF/STB</td>
<td>Actual</td>
<td>ANN</td>
<td>AE</td>
</tr>
<tr>
<td>----------------</td>
<td>--------</td>
<td>------</td>
<td>-----</td>
</tr>
<tr>
<td><strong>Training output (R² = 92.25%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>67.07</td>
<td>67.11</td>
<td>8.6</td>
</tr>
<tr>
<td>SD</td>
<td>46.42</td>
<td>44.55</td>
<td>9.16</td>
</tr>
<tr>
<td>Maximum</td>
<td>7</td>
<td>7.6</td>
<td>0.0123</td>
</tr>
<tr>
<td>Minimum</td>
<td>248</td>
<td>246.6</td>
<td>66.06</td>
</tr>
<tr>
<td><strong>Validation output (R² = 82.66%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>67.67</td>
<td>66.48</td>
<td>12.66</td>
</tr>
<tr>
<td>SD</td>
<td>49.54</td>
<td>40.37</td>
<td>13.60</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.3</td>
<td>7.49</td>
<td>0.0954</td>
</tr>
<tr>
<td>Minimum</td>
<td>223</td>
<td>175.70</td>
<td>62.04</td>
</tr>
<tr>
<td><strong>Testing output (R² = 88.30%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>56.52</td>
<td>57.7</td>
<td>10.87</td>
</tr>
<tr>
<td>SD</td>
<td>41.26</td>
<td>44.23</td>
<td>10.59</td>
</tr>
<tr>
<td>Maximum</td>
<td>7</td>
<td>6.49</td>
<td>0.4098</td>
</tr>
<tr>
<td>Minimum</td>
<td>2313</td>
<td>247.48</td>
<td>61.19</td>
</tr>
<tr>
<td><strong>Overall output (R² = 90.04%)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>65.48</td>
<td>65.51</td>
<td>9.61</td>
</tr>
<tr>
<td>SD</td>
<td>46.32</td>
<td>43.99</td>
<td>10.34</td>
</tr>
<tr>
<td>Maximum</td>
<td>6.3</td>
<td>6.45</td>
<td>0.0123</td>
</tr>
<tr>
<td>Minimum</td>
<td>248</td>
<td>247.48</td>
<td>66.06</td>
</tr>
</tbody>
</table>

Table 4.18 - Ranges of $R_{ST}$ ANN Model Output with Statistical Error Analysis
Figure 4.22 – Cross-plot of Modeled $R_{ST}$ on Training Dataset vs. PVT Values

Scatter Diagram
Stock-tank vent gas/oil ratio ANN model (Training Dataset)
Figure 4.23 – Cross-plot of Modeled $R_{ST}$ on Validation Dataset vs. PVT Values
Figure 4.24 – Cross-plot of Modeled $R_{ST}$ on Testing Dataset vs. PVT Values
Figure 4.25 – Cross-plot of Modeled $R_{ST}$ on Overall Data Points vs. PVT Values
Figure 4.26 – Histogram of the Residuals of Training Dataset
Figure 4.27– Histogram of the Residuals of Validation Dataset
Figure 4.28 – Histogram of the Residuals of Testing Dataset
Figure 4.29 – Histogram of the Residuals of Overall Data Points
4.8 Conclusions

• Based on the results of this study, the following findings can be drawn:

• A feedforward Backpropagation Neural Network (BP-ANN) was used to design four new models with which to estimate bubble point pressure, the bubble point oil formation volume factor, bubble point solution gas/oil ratio and stock-tank vent gas/oil ratio.

• The models were established based on 476 PVT flash separation tests collected from the Sirte Basin, Libya.

• Unlike existing PVT models, the models can be applied directly in the absence of PVT analysis. They require on-hand production data with no need of additional laboratory data or other correlations.

• The bubble point BP-ANN model has four inputs, two hidden layers, with nine nodes in the first and four in the second. The model was trained with the Quasi-Newton training algorithm. The entire output layer and the hidden nodes were activated by a logistic activation function.

• The bubble point oil formation volume factor BP-ANN model was trained with the Limited Memory Quasi-Newton training algorithm and activated by the logistic activation function. It has four inputs, two hidden layers with five nodes in the first hidden layer and four nodes in the second hidden layer.

• The bubble point solution gas/oil ratio BP-ANN model was trained with the Limited Memory Quasi-Newton training algorithm. It has three inputs and two
hidden layers. Seven nodes are in the first hidden layer while four nodes are in the second hidden layer.

- The stock-tank vent gas/oil ratio BP-ANN model was trained with the Quasi-Newton training algorithm and activated by the logistic activation function. It has three inputs, three hidden layers, eight nodes in the first layer, four nodes in the second layer, and two nodes in the third layer. All the layers were activated by logistic activation functions.

- In order to build robust BP-ANN models and avoid the models from over fitting, a cross-validation was established during the training process.

- The results show the trained models provided high accuracy on testing dataset. This dataset was not viewed by the networks during the training process.
4.9 References


Heaton, Jeff, 2005. Introduction to neural network with java, Heaton Research, Inc.

Iglewicz, B. and Hoaglin, D. C., 1993 How to Detect and Handle Outliers, American Society for Quality Control, Milwaukee, WI

Krose, B. and Van der Smagt, P., 1996. An Introduction to Neural Networks, University of Amsterdam.


CHAPTER V

DEVELOPMENT OF PVT MODELS UTILIZING LS-SVM APPROACH

5.1 Introduction

The PVT properties of crude-oil such as bubble point pressure, the oil formation volume factor, and dissolved gas-oil ratio play a key role in the calculation of reserves of oil reservoirs as well as for identification of reservoir characteristics. The properties are traditionally determined from laboratory (Pressure-Volume-Temperature, PVT) analyses on samples collected from the bottom of the wellbore or after re-combining the liquid and vapour samples collected from the separator at the surface. However, this process is expensive and the costs are high. Therefore, researchers deduced many empirical formulas and several ANN models which are used to calculate the PVT properties of crude-oil.

The main objective of this chapter is to design a LS-SVM regression model to provide a new method as an alternative technique with which to predict \( P_b \), \( B_{ob} \), \( R_{sob} \) and \( R_{ST} \) in the absence of PVT analysis. In this way, the dilemma that is associated with the published PVT empirical correlations is overcome and eliminated. However, the
deficiencies resulting from these correlations are explained in detail in Chapters II and III.

5.2 LS-SVM for Function Approximation

Function approximation using Least Squares Support Vector Machines (LS-SVM) is relatively new. In recent years, LS-SVM has rapidly developed into an important and effective method of function approximation. However, the current literature review with regard to the application of LS-SVM in the oil and gas industry is quite limited. In 2009, Chun-Xiao developed a LS-SVM model to forecast oil holdup of oil-water two-phase flow. Peng and Yin (2010) applied LS-SVM to establish the void fraction measurement models of oil-gas two-phase flow corresponding to different flow patterns. Cheng et al. (2010) proposed an algorithm which combines Particle Swarm Optimization (PSO) with LS-SVM classification to identify lithology by using well logging data.

Artificial Neural Networks (ANN) can learn the model data by self-adjusting their parameters. Following training, ANN can accurately match with the expected data, and it can quickly and accurately predict the unknown sample data. It can be said that one of the most important features of ANN is its ability to discover the rules and patterns in the data which general observation and many standard statistical methods cannot find especially, when dealing with nonlinear problems. Many researchers recognized that through the use of ANN, we can establish an appropriate model which can accurately predict the PVT properties of crude-oil in petroleum engineering. However, these neural network
modeling schemes suffer from numerous drawbacks (Castillo et al., 2002 and Castillo et al., 2006), such as:

- A limited ability to explicitly identify possible causal relationships and the consumption of time in the development of a backpropagation algorithm, which leads to an overfitting problem and becomes stuck at a local optimum of the cost function.

- The architectural parameters of ANN such as the number and size of hidden layers and the type of transfer function(s) for neurons in the various layers must be guessed in advance.

- The training algorithm parameters were determined based on guessing the initial random weights, learning rate, and momentum.

The Support Vector Machine (SVM) can be used for both regression and calcification problems, just as with ANN. SVM is derived from statistical learning theory and was found in 1995 by Vapnic et al. LS-SVM was proposed by Suykens et al. in 2002 and it is a re-formulation of standard SVM frameworks. In LS-SVM, the nonlinear relationship in the data space is transferred to a linear relationship in the feature space by means of kernels. The model parameters, however, are given by the solution of a system of linear equations. To be properly tuned/optimized, the LS-SVM required just two parameters ($\gamma$ and $\sigma^2$) to minimize the generalization error. Parameter $\gamma$ (regularization parameter) is defined as a trade-off between minimizing the training error and minimizing the complexity of the LS-SVM model. The $\sigma^2$ denotes the RBF kernel parameter and represents the change in the RBF kernel function (width of kernel) (Pandit et al. 2011). It
is a positive real constant. Based on this advantage, the LS-SVM is introduced in this chapter to investigate the capability of LS-SVR regression on modeling \( P_b, B_{ob}, R_{sob} \) and \( R_{ST} \) of crude oil systems in the absence of PVT analysis and to solve some of the above ANNs limitations. The details of the LS-SVM algorithm can be found in Vapnik (1995), Guo et al. (2006) and Chen et al. (2007). Appendix C provides a brief introduction as to how to construct a LS-SVM model. The LS-SVM model can be expressed as:

\[
y(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i) + b
\]  

(5.1)

where the term \( K(x, x_i) \) is the kernel function, \( x_i \) is the input vector, \( \alpha_i \) is Lagrange multiplier called “support value”, and \( b \) is the bias term.

The commonly used kernel functions were linear kernel, polynomial kernel, sigmoid kernel, Radial Basis Function (RBF) kernel. However, compared with other feasible kernel functions, RBF can handle the nonlinear relationships between the spectra and target attributes and is able to reduce the computational complexity of the training procedure and give a good performance under general smoothness assumptions (Wang et al. 2003). Thus, RBF kernel was recommended as the kernel function of the LS-SVM regression models in this thesis.

### 5.3 Model Parameter Selecting

This study employed the RBF as the nonlinear mapping kernel function. It is a common function that is useful in nonlinear regression problems. However, the selecting
of $\gamma$ and $\sigma^2$ is the most significant topic when identifying a non-linear system with the LS-SVM. The regularization parameter ($\gamma$) determines the trade-off when minimizing the training error. The parameter $\sigma^2$ was the bandwidth, and implicitly defined the nonlinear mapping input space to some high dimensional feature space. Consequently, the selection can significantly affect model performance. A grid search method is used at first, and then searching optimization is carried out employing a cross validation method until an optimum parameter pair is obtained. The free LS-SVM toolbox (LSSVM V1.5, Suykens, Leuven and Belgium) was applied to develop LS-SVM models. Appendix III introduces, in brief, a LS-SVM model for function approximation problems.

5.4 Newly Developed PVT LS-SVM Models

Several PVT empirical correlations and several ANN models are available in the literature, but they suffer from a lack of application in the absence of laboratory analysis. This shortcoming and the deficiencies resulting from the correlations are explained in detail in Chapter II and III.

The main objective of this study is to overcome and then eliminate the dilemma by building LS-SVM models using direct measured field parameters as input variables to estimate $P_b$, $B_{ob}$, $R_{osb}$, and $R_{ST}$. By incorporating the four proposed LS-SVM models, engineers can estimate $P_b$, $B_{ob}$, $R_{osb}$, and $R_{ST}$ straightforwardly in the absence of PVT analysis. The $P_b$ and $B_{ob}$ LS-SVM models are proposed as a function of four directly measured field parameters ($R_{sp}$, $P_{sp}$, $\chi_{ST}$ and $T_R$), whereas $R_{osb}$ and $R_{ST}$ LS-SVM models
are proposed as a function of only three direct measured field parameters \((R_{sp}, P_{sp}, T_{SP}, \gamma_{ST})\).

The LS-SVM parameters \(\gamma\) and \(\sigma^2\) must be appropriately selected for the RBF as the kernel function. Immediately following the training process, the evaluation of the quality and capability of the fitted model is applied. The prediction capability was mainly assessed by correlation coefficients \(R^2\) between the actual and the predicted outputs, root mean-squared errors (RMSE), mean, minimum and maximum relative error (RE), mean minimum and maximum Absolute Percent Relative Error (ARE), mean, minimum and maximum of Absolute Error (AE), and Standard Deviation (SD). The best model is the model with the highest \(R^2\) and the smallest other error statistical analysis. Appendix C contains further details regarding the corresponding equations of the statistical parameters. To demonstrate the usefulness of the LS-SVR modeling scheme, a calibration model was developed based on the training dataset, validation dataset, and testing dataset.

### 5.5 PVT Data Preparation

The PVT data used in this study were obtained from two-stage and single-stage flash separation tests. A total of 118 reservoir fluid studies (439 data points) were collected from various Libyan oil fields in the Sirte Basin. The majority of the data points are taken from two-stage flash separation tests. In the single-stage separation test, the separator pressure is atmospheric pressure and the stock-tank vent GOR value is equal to zero.
In order to build robust LS-SVM PVT models, the entire dataset was randomly divided into three subsets; training set of 300 points, validation set of 70 points and testing set of 69 points. Training and validation sets were used to build the LS-SVM models, while testing sat was used to test the accuracy of the newly developed models and not switched in the process of LS-SVM models derivation. The value ranges of training, validation and testing datasets are presented in Table 5.1.

Four directly measured field parameters \( R_{sp} \), \( P_{sp} \), \( \gamma_{ST} \) and \( T_R \) are selected as inputs to generate the \( P_b \) LS-SVM model. To obtain optimal LS-SVM model parameters, a grid search is performed on the basis of leave-one-out cross-validation on the dataset. Grid search techniques with 10-folds cross validation were employed to obtain the LS-SVM model parameters \( (\gamma, \sigma^2) \). The parameter combination \( (\gamma, \sigma^2) \) was tuned simultaneously in a 10×10 grid. The contour plot of the optimization parameters \( \gamma \) and \( \sigma^2 \) for the prediction of bubble point pressure is shown in Figure 5.1. The grids ‘•’ in the first step are 10×10, and the searching step in the first step is large. The natural logarithms of regularization parameter \( \gamma \) and kernel parameter \( \sigma^2 \) were tuned simultaneously. The optimal search area is determined by the error contour line. The grids ‘×’ in the second step are 10×10, and the searching steps in the second step are smaller. The optimal search area is determined based on the first step. The optimal pair of \( (\gamma, \sigma^2) \) for \( P_b \) model was found at the value of 277.88 [exp(5.627)] and 5.79 [exp(1.756)], respectively. The contour plot (Figure 5.1) of the optimization error for LS-SVM regression model is \( P_b \) when optimizing the parameters \( \gamma \) and \( \sigma^2 \). The red square indicates the selected optimal settings.
After the LS-SVM model parameters are tuned, the model was developed and the prediction performance was evaluated. The predicted results of the optimal $P_b$ LS-SVM regression ($LS_{SVM}^{RBF}_{\gamma=277.88, \sigma^2=5.79}$) model for training dataset are shown in Figure 5.2. The figure shows all the points are scattered around the $y = x$ line. It can be concluded that the predicted bubble point values are in good agreement with the bubble point values obtained from PVT analysis. The Root Mean Square Error (RMSE) is 148.4, Mean Absolute Relative Error (MARE) is 8.09% and the Predicted Correlation Coefficient ($R^2$) is 98.27%. The statistical error analysis for the prediction capability, as obtained from the LS-SVM model, is summarized in Table 5.2. As noted in Table 5.2, the proposed model was statistically stable and fitted the data well. Moreover, the histogram of the error distribution for the training dataset is presented in Figure 5.3. It confirms the error is normally distributed.
Table 5.1 – Ranges of PVT Data

<table>
<thead>
<tr>
<th>Laboratory Measurement</th>
<th>Training Dataset (300 separator tests)</th>
<th>Validation Dataset (70 separator tests)</th>
<th>Testing Dataset (69 separator tests)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min.</td>
<td>Mean</td>
<td>Max.</td>
</tr>
<tr>
<td>$P_b$, psia</td>
<td>121</td>
<td>1762</td>
<td>4720</td>
</tr>
<tr>
<td>$B_o$, bbl/STB</td>
<td>1.047</td>
<td>1.293</td>
<td>1.781</td>
</tr>
<tr>
<td>$R_{sob}$, SCF/STB</td>
<td>17</td>
<td>418</td>
<td>1256</td>
</tr>
<tr>
<td>$R_{ST}$, SCF/STB</td>
<td>0</td>
<td>53</td>
<td>248</td>
</tr>
<tr>
<td>$R_{sp}$, SCF/STB</td>
<td>10</td>
<td>365</td>
<td>1256</td>
</tr>
<tr>
<td>$P_{sp}$, psia</td>
<td>14.7</td>
<td>111</td>
<td>517</td>
</tr>
<tr>
<td>$T_R$, °F</td>
<td>12</td>
<td>108</td>
<td>194</td>
</tr>
<tr>
<td>$T_{sp}$, °F</td>
<td>100</td>
<td>202</td>
<td>282</td>
</tr>
<tr>
<td>$\gamma_{oST}$, water =1</td>
<td>0.8</td>
<td>0.84</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Figure 5.1– Contour Plot of the Optimization Error (MSE) for LS-SVM Regression Model of $P_b$ when Optimizing the Parameters $\gamma \left[\exp(5.627)\right]$ and $\sigma^2 \left[\exp(1.756)\right]$. The Red Square Indicates the Selected Optimal Settings.
Figure 5.2 – Performance of Training Dataset for $P_b$ LS-SVM Model. Upper: Calculated $P_b$ vs. Data Observation Order. Lower: Calculated $P_b$ vs. Experimental Values.
Table 5.2 – Statistical Error analysis of $P_b$ LS-SVM Model

<table>
<thead>
<tr>
<th>Error Analysis</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$</td>
<td>98.27%</td>
<td>93.84%</td>
</tr>
<tr>
<td>SD, psia</td>
<td>148.65</td>
<td>279.85</td>
</tr>
<tr>
<td>RMSE, psia</td>
<td>148.4</td>
<td>275.55</td>
</tr>
<tr>
<td>MSE, psia$^3$</td>
<td>22023</td>
<td>75928</td>
</tr>
<tr>
<td>Mean AE, psia</td>
<td>99.62</td>
<td>199.56</td>
</tr>
<tr>
<td>Maximum AE, psia</td>
<td>640.15</td>
<td>758.81</td>
</tr>
<tr>
<td>Minimum AE, psia</td>
<td>0.486</td>
<td>0.499</td>
</tr>
<tr>
<td>Mean ARE</td>
<td>8.09%</td>
<td>14.217%</td>
</tr>
<tr>
<td>Maximum ARE</td>
<td>116.1%</td>
<td>87.33%</td>
</tr>
<tr>
<td>Minimum ARE</td>
<td>0.022%</td>
<td>0.144%</td>
</tr>
</tbody>
</table>
Figure 5.3– Error Distribution Resulting from the Training Dataset for $P_b$ LS-SVM Model
5.5.1 Accuracy of the Bubble Point Pressure LS-SVM Model

The final step in LS-SVM regression model building is to test its accuracy. A testing dataset (69 data points) was used to scan the proposed $P_b$ LS-SVM regression model’s accuracy. The predicted results of the optimal $P_b$ LS-SVM regression ($LS\_SVM^{RBF}_{\gamma=277.88,\sigma^2=5.79}$) model for testing the dataset is shown Figure 5.4. All data points in the figure are scattered around the $y = x$ line. The testing dataset provides $P_b$ with a RMSE of 275.55, a MARE of 14.214%, and a $R^2$ of 93.84. The histogram of the error term distribution confirms normal distribution, as clarified in Figure 5.5. Statistical analysis of the error for the predicted capability on the testing dataset is summarized in Table 5.2. Figure 5.6 shows the performance of the validation and testing datasets of the LS-SVM model for the prediction of $P_b$. Both datasets have similar performances, indicating the proposed $P_b$ LS-SVM model is adequate.
Figure 5.4 – Performance of Testing Dataset for $P_b$ LS-SVM Model. Upper: Calculated $P_b$ vs. Data Observation Order. Lower: Calculated $P_b$ vs. Experimental Values.
Figure 5.5— Error Distribution Resulting From Testing Dataset for $P_b$ LS-SVM Model
Figure 5.6– Performance of Validation/Testing Dataset for $P_b$ LS-SVM Model
5.6 Bubble Point Oil Formation Volume Factor, $B_{ob}$ LS-SVM Model

A LS-SVM model to calculate $B_{ob}$ in the absence of PVT analysis was developed using 10×10 grid search is an error-surface spanned by the model parameters. In this grid search, two steps are carried out. The first step was a search for crude with a large step size presented in the form of (•) and the second step of the specified search is a small step size, presented in the form of (×). The results of the two-step grid search are shown in Figure 5.7. The optimal combination of $(\gamma, \sigma^2)$ was achieved with $\gamma$ equalling 7142.508 and $\sigma^2$ equalling 984.957. The $B_{ob}$ model was developed using the aforementioned parameters and the prediction results for the training dataset are shown in Table 5.3. The model provides $B_{ob}$ with a $R^2$ of 97.11%, a MSE of 6.9×10^{-4}, and a RMSE of 0.0263. The performance of the optimal $B_{ob}$ LS-SVM regression ($LS_{SVM}^{RBF}_{\gamma=7142.508, \sigma^2=984.957}$) model in the training dataset is shown in Figure 5.8. The tightest cloud of points around the $y = x$ line indicates excellent agreement between the experimental and the calculated data values. Moreover, the histogram of the error distribution for the training dataset in Figure 5.9 confirms the error is normally distributed with the mean being equal to zero.

5.6.1 Accuracy of the Bubble Point Oil FVF LS-SVM Model

Once the parameters of $\gamma$ and $\sigma^2$ were settled, there was no need to implement the grid search step. The testing dataset can be used to inspect model accuracy. This inspection provides $B_{ob}$ with a $R^2$ of 96.94%, a MSE of 7.08×10^{-4} and a RMSE of 0.0266 as illustrated in Table 5.3 and Figure 5.10. The figure shows the majority of the calculated points fall too close to the $y = x$ line. Furthermore, the histogram of the error terms in
Figure 5.11 show the errors are normally distributed with mean being equal to zero. Figure 5.12 illustrates the scatter diagram of the modeled $B_{ob}$ using validation and testing datasets versus experimental values. Again, this figure shows all the points of both datasets scattered around the $y = x$ line, indicating excellent agreement between the experimental and the calculated $B_{ob}$ values.

5.7 Bubble Point Solution Gas/Oil Ratio LS_SVM Model

When using LS-SVM, three crucial problems required solving, including the determination of the optimal input feature subset, proper kernel function, and the optimal kernel parameters. As mentioned previously, the RBF kernel was recommended as the kernel function of the LS-SVM in this work. In addition, proper parameter setting played a crucial role in building a good LS-SVM regression model with high prediction accuracy and stability. Again, a two-step grid search technique with leave-one-out cross validation was employed to obtain the optimal combination of $(\gamma, \sigma^2)$. Leave-one-out cross-validation was used to avoid over fitting problems in the selection of optimal LS-SVM parameters. Grid search is a minimization procedure based on exhaustive search in a limited range. In each iteration, 10 points are left out, and fits a model to the other data points. The performance of the model is estimated based on the 10 points left out. This procedure is repeated for each 10 points.
Figure 5.7 - Contour Plot of the Optimization Error (MSE) for LS-SVM Regression Model of $B_{ob}$ when Optimizing the Parameters $\gamma \left[ \exp(8.874) \right]$ and $\sigma^2 \left[ \exp(6.893) \right]$. The Black Square Indicates the Selected Optimal Settings.
Table 5.3 – Statistical Error Analysis of $B_{ob}$ LS-SVM Model

<table>
<thead>
<tr>
<th>Error Analysis</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$, %</td>
<td>97.11</td>
<td>96.94</td>
</tr>
<tr>
<td>SD, bbl/STB</td>
<td>0.0263</td>
<td>0.0267</td>
</tr>
<tr>
<td>RMSE, bbl/STB</td>
<td>0.0263</td>
<td>0.0266</td>
</tr>
<tr>
<td>MSE, (bbl/STB)$^2$</td>
<td>$6.9 \times 10^{-4}$</td>
<td>$7.1 \times 10^{-4}$</td>
</tr>
<tr>
<td>Mean AE, bbl/STB</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Maximum AE, bbl/STB</td>
<td>0.112</td>
<td>0.079</td>
</tr>
<tr>
<td>Minimum AE, bbl/STB</td>
<td>$4.8 \times 10^{-5}$</td>
<td>$4.5 \times 10^{-5}$</td>
</tr>
<tr>
<td>Mean ARE, %</td>
<td>1.53</td>
<td>1.49</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>8.48</td>
<td>4.4</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.004</td>
<td>0.004</td>
</tr>
</tbody>
</table>
Figure 5.8 – Performance of the Optimal $B_{ob}$ LS-SVM Model in Training Dataset
Figure 5.9– Error Distribution Resulting from Training Dataset for $B_{ob}$ LS-SVM Model
Figure 5.10– Performance of the Optimal $B_{ob}$ LS-SVM Model in Testing Dataset
Figure 5.11 – Error Distribution Resulting from Testing dataset for $B_{0b}$ LS-SVM Model
Figure 5.12—Performance of Validation/Testing Dataset for $B_{ob}$ LS-SVM Model
The results of this two-step grid search are shown in Figure 5.13. The first step grid search was for a crude search with a large step size, presented in the form of (●), and the second step for the specified search with a small step size, is presented in the form of (×). The optimal search area is determined by an error contour line. Following the grid search process, the optimal combination of \( \gamma \) and \( \sigma^2 \) would be achieved for the LS-SVM models. Finally, all the different estimates of the performance are combined. For each combination of \( \gamma \) and \( \sigma^2 \) parameters, MSE was calculated and the optimum parameters were selected which produced smaller MSE. The black square in Figure 5.13 indicates the selected optimal combination of \( \gamma \) and \( \sigma^2 \). The optimal pair of \((\gamma, \sigma^2)\) was found at the values of 2936154.38 and 7865.36, respectively. In order to assess the relative robustness of the proposed LS-SVM model, calculated \( R_{sob} \) was plotted against experimental values in Figure 5.14. In the upper plot, the model estimating curve is very close to the actual data, indicating that the LS-SVM model has a strong ability for regression analysis. The lower plot illustrates the correlation between the LS-SVM model and the experimental values. The correlation coefficient is very close to unity \( (R^2 = 99.25\%) \). Perfect accuracy would result in the data points forming a straight line along the diagonal axis \( (y = x \text{ line}) \). The accuracy analysis of the proposed model is summarized in Table 5.4.

5.7.1 Accuracy of the Bubble Point Solution Gas/Oil Ratio LS-SVM Model

The performance of the \( R_{sob} \) model was evaluated by 69 data points in the testing set. The prediction results are summarized in Table 5.4. The \( R^2 \) and RMSE of the testing set were 98.73\% and 28.7, respectively. Moreover, Figure 5.16 brings the calculated \( R_{sob} \) using the proposed LS-SVM model against the experimental values. The solid line is the
regression line corresponding to the $y = x$. All the points are tight around this line indicating excellent agreement between the experimental and the calculated $R_{sob}$ values. Notably, the errors are random and follow a normal distribution with zero mean, as shown in Figure 5.17. However, if the mean is not zero, it might be that the model is not the right choice for a particular data. Figure 5.18 shows the scatter points of both validation and testing sets match the $y = x$ line indicating no overfitting existed in the model.

5.8 Stock-Tank Vent Gas/Oil Ratio LS-SVM Model

Usually, the $R_{SP}$ is a field measurement, while the $R_{ST}$ is rarely measured in the field. Thus, in order to obtain the $R_{sob}$ indirectly, we must first estimate the $R_{ST}$, and add it to the field measured $R_{SP}$ (Equation 2.1). In this work, a novel LSA-SVM model with which to estimate $R_{ST}$ was developed. $P_{SP}$, $T_{SP}$ and $\gamma_{oST}$ are the only independent variables. A total of 364 two-stage separator tests were used to develop the $R_{ST}$ LS-SVM model. In order to build a robust $R_{ST}$ LS-SVM model, 364 separator tests were randomly split into three subsets, a training set which includes 246 separator tests while validation and testing sets were 59 separator tests each. The value ranges of these sets are summarized in Table 5.5.
Figure 5.13 - Contour Plot of the Optimization Error (MSE) for LS-SVM Regression Model of $R_{sob}$ when Optimizing the Parameters $\gamma [\exp(14.89)]$ and $\sigma^2 [\exp(8.97)]$. The Black Square Indicates the Selected Optimal Settings.
Figure 5.14 – Performance of the Optimal $R_{\text{rob}}$ LS-SVM Model in the Training Dataset
Table 5.4 – Statistical Error Analysis of $R_{sub}$ LS-SVM Model

<table>
<thead>
<tr>
<th>Error Analysis</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$, %</td>
<td>99.25</td>
<td>98.73</td>
</tr>
<tr>
<td>SD, SCF/STB</td>
<td>22.75</td>
<td>28.90</td>
</tr>
<tr>
<td>RMSE, SCF/STB</td>
<td>22.72</td>
<td>28.7</td>
</tr>
<tr>
<td>MSE, (SCF/STB)$^2$</td>
<td>516.1</td>
<td>823.64</td>
</tr>
<tr>
<td>Mean AE, SCF/STB</td>
<td>15.63</td>
<td>19.39</td>
</tr>
<tr>
<td>Maximum AE, SCF/STB</td>
<td>150.44</td>
<td>96.28</td>
</tr>
<tr>
<td>Minimum AE, SCF/STB</td>
<td>0.203</td>
<td>0.292</td>
</tr>
<tr>
<td>Mean ARE, %</td>
<td>6.65</td>
<td>6.755</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>159.7</td>
<td>41.97</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.031</td>
<td>0.0568</td>
</tr>
</tbody>
</table>
Figure 5.15 – Error Distribution Resulting from Training Dataset for $R_{lob}$ LS-SVM Model
Figure 5.16 – Performance of the Optimal $R_{\text{sol}}$ LS-SVM Model in the Testing Dataset
Figure 5.17 – Error Distribution Resulting from Testing Dataset for $R_{sob}$ LS-SVM Model
Figure 5.18 – Performance of Validation/Testing Dataset for $R_{sob}$ LS-SVM Model
A robust $R_{ST}$ LS-SVM regression model with high prediction accuracy and stability is built based on a proper parameter setting. The optimal LS-SVM parameters were obtained by employing cross validation with a grid search. The optimal search area is determined based on the first step. The optimal pair of $(\gamma, \sigma^2)$ was found at a value of $\gamma = 47.6552$ and $\sigma^2 = 1.3477$, respectively. The optimizing process is shown in Figure 5.19. The black square indicates the selected optimal settings.

The calculated $R_{ST}$, using the LS-SVM model, was plotted against the experimental values in Figure 5.20. In the upper plot, the model estimating curve is very close to the actual PVT data, indicating the LS-SVM has a strong ability for regression analysis. The lower plot, illustrates the correlation between the LS-SVM model and the experimental PVT values. Moreover, the histogram regarding the errors of the proposed model is plotted in Figure 5.21. The error terms follow a normal distribution with a mean equal to zero.
Table 5.5 – Ranges of PVT Data used in $R_{ST}$ LS-SVM Model

<table>
<thead>
<tr>
<th>Laboratory Measurement</th>
<th>Training Dataset (246 separator tests)</th>
<th>Validation Dataset (60 separator tests)</th>
<th>Testing Dataset (59 separator tests)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Min.</td>
<td>Mean</td>
<td>Max.</td>
</tr>
<tr>
<td>$P_b$, psia</td>
<td>121</td>
<td>1831</td>
<td>4720</td>
</tr>
<tr>
<td>$B_o$, bbl/STB</td>
<td>1.047</td>
<td>1.291</td>
<td>1.73</td>
</tr>
<tr>
<td>$R_{sob}$, SCF/STB</td>
<td>26</td>
<td>419</td>
<td>1135</td>
</tr>
<tr>
<td>$R_{ST}$, SCF/STB</td>
<td>6.3</td>
<td>65.2</td>
<td>248</td>
</tr>
<tr>
<td>$R_{sp}$, SCF/STB</td>
<td>10</td>
<td>354</td>
<td>1097</td>
</tr>
<tr>
<td>$P_{sp}$, psia</td>
<td>29.7</td>
<td>132</td>
<td>517</td>
</tr>
<tr>
<td>$T_{sp}$, °F</td>
<td>12</td>
<td>110</td>
<td>194</td>
</tr>
<tr>
<td>$T_R$, °F</td>
<td>100</td>
<td>203.6</td>
<td>282</td>
</tr>
<tr>
<td>$\gamma_{ST}$, water =1</td>
<td>0.8</td>
<td>0.84</td>
<td>0.92</td>
</tr>
</tbody>
</table>
Figure 5.19 - Contour Plot of the Optimization Error of Cross Validation of $R_{ST}$ when Optimizing the Parameters $\gamma [\exp(3.864)]$ and $\sigma^2 [\exp(0.298)]$. The Black Square Indicates the Selected Optimal Settings.
Figure 5.20 – Performance of the Optimal $R_{ST}$ LS-SVM Model in the Training Dataset
Figure 5.21 – Error Distribution Resulting from $R_{ST}$ LS-SVM Model
5.8.1 Accuracy of the $R_{ST}$ LS-SVM Model

In order to assess the relative robustness of the LS-SVM regression model for $R_{ST}$ in estimating $R_{sob}$, the calculated $R_{ST}$ values are added to the field $R_{SP}$ (Equation 2.1) and plotted against $R_{sob}$, as obtained from PVT analysis (Figure 5.22). The sold line, in Figure 5.21, is the regression line corresponding to the $y = x$. The figure shows all the points are tight around this line indicating an excellent agreement between the experimental and the calculated $R_{sob}$ values. $R^2$, RMSE and other error analysis are illustrated in Table 5.6. The error analysis demonstrates a high $R^2$ of 99.15% and a low RMSE of 15.76. Figure 5.23 shows the scatter points of the testing set match the $y = x$ line indicating that no over fitting existed in the model.

5.9 Conclusions

This chapter has described the LS-SVM model with which to develop a nonlinear model to predict some PVT properties in the absence of experimental analysis. Four robust LS-SVM models were built to calculate $P_b$, $B_{ob}$, $R_{sob}$ and $R_{ST}$. The models were developed using the RBF kernel function. The model parameters ($\gamma$, $\sigma^2$) were tuned by a two-step grid search with a cross-validation process. The cross-validation process was utilized to avoid the over fitting problem and to evaluate model performance during the calibration stage. The models are obtained by selecting $\gamma$ and $\sigma^2$ parameters that give the lowest error in a smooth area based on 10-fold cross-validation. The performances of LS-SVM model are encouraging and gave satisfactory regression results.
Table 5.6 – Error of Analysis of $R_{sob}$ Obtained by Adding LS-SVM $R_{ST}$ and Field $R_{SP}$

<table>
<thead>
<tr>
<th>Error Analysis</th>
<th>Training Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$, %</td>
<td>99.15</td>
<td>98.86</td>
</tr>
<tr>
<td>SD, SCF/STB</td>
<td>15.8</td>
<td>34.457</td>
</tr>
<tr>
<td>RMSE, SCF/STB</td>
<td>15.76</td>
<td>34.273</td>
</tr>
<tr>
<td>MSE, (SCF/STB)$^2$</td>
<td>248.53</td>
<td>1174.6</td>
</tr>
<tr>
<td>Mean AE, SCF/STB</td>
<td>9.7</td>
<td>21.79</td>
</tr>
<tr>
<td>Maximum AE, SCF/STB</td>
<td>97.23</td>
<td>124.06</td>
</tr>
<tr>
<td>Minimum AE, SCF/STB</td>
<td>0.011</td>
<td>0.0317</td>
</tr>
<tr>
<td>Mean ARE, %</td>
<td>3.828</td>
<td>8.39</td>
</tr>
<tr>
<td>Maximum ARE, %</td>
<td>37.1</td>
<td>55.3</td>
</tr>
<tr>
<td>Minimum ARE, %</td>
<td>0.002</td>
<td>0.006</td>
</tr>
</tbody>
</table>
Figure 5.22– Performance of $R_{sob}$ Calculated from $R_{ST}$ LS-SVM Model
Figure 5.23 – Performance of $R_{\text{sob}}$ Calculated from $R_{\text{ST}}$ LS-SVM Model using Testing Dataset
5.10 References


CHAPTER VI

ESTIMATING AVERAGE RESERVOIR PRESSURE: A NEURAL NETWORK APPROACH

6.1 Introduction

Insight into average reservoir pressure and its change in production (time), plays a critical role in reservoir development and management, economic evaluation, obtaining a hydrocarbon in place, computing rock and fluid characteristics, reservoir material balance calculations, pressure maintenance projects, surface and subsurface facility designs and predicting reservoir behaviour. In particular, almost every well intervention requires insight with respect to average reservoir pressure.

Traditionally, average reservoir pressure is obtained via a pressure build-up test which measures long-term built-up pressure when the producing wells are shut in. The test is started after the well is opened to a constant flow rate over an extended period of time. This will cause the reservoir pressure to drawdown due to the withdrawing of fluids from the well. Afterward, the well is shut-in to allow the pressure to build up. During the re-building of pressure, the pressure is recorded as a function of time. Over an extended period of time, the build-up pressure reaches average reservoir pressure. This time period depends mainly upon the reservoir permeability and thus, may take a prolonged period of
time in low permeability reservoirs. The resulting pressure build-up is then analysed to determine reservoir permeability, flow efficiency, and average reservoir pressure. Currently, the only available method with which to obtain average reservoir pressure is to conduct an extended build-up test. It must then be evaluated using Horner (Horner, 1951) or Miller-Dyes-Hutchinson, MDH (possibly et al., 1950) evaluation procedures. Average reservoir pressure measurements should be updated periodically because reservoir pressure changes as fluids (oil, gas and water) are released from the reservoir. However, a significant economic impact is caused by shutting in the producing wells during the entire build-up test.

6.2 Artificial Neural Networks (ANNs)

ANNs are relatively new computational tools which are extensively utilized to solve many complex engineering problems. ANNs can assist engineers and researchers by addressing many fundamental petroleum engineering problems, as well as specific problems that conventional computing has been unable to solve. Petroleum engineers may benefit from ANNs on occasions when engineering data for designs and interpretations are less than adequate, Mohaghegh (1995). However, a review of the literature reveals that ANNs has been applied successfully to different branches of petroleum engineering in recent years. Examples include Alimadadi (2011), Ardjmandpour et al. (2011), Popa et al. (2011), and Lu and Fleming (2011).

A basic ANNs has three layers of processors - an input layer, an output layer and one hidden layer between them. Each layer has a number of nodes. ANNs and their learning
algorithms and backpropagation (BP) algorithms have been studied extensively in the literature by Bulsari (1995), Veelenturf (1995), Gropal (1998), Mehrotra et al. (2000), Basheer and Hajmeer (2000) and Rabunal and Dorado (2006). A brief overview of the aforementioned studies is presented in Appendix B.

A neural network model has been established in this study to map relationships which control reservoir oil, gas and water production performance in order to interpolate, predict and estimate the current average reservoir pressure without closing the producing wells. This method is suitable for constant and variable flow rates. After interpolating, predicting and/or obtaining the current average reservoir pressure, it can be utilized in almost all reservoir and production engineering studies.

6.3 The Proposed Model

In order to design an effective neural network, it is very important to define the problem in a way that will be addressable by neural networks. The selection of input variables for the ANNs is the key problem in this approach. In other words, the first step in any modeling design procedure consists of reducing the dimension of the input space. Thus, our task is to find a set of inputs that are strongly correlated to the output.

The overall performance of oil reservoirs is largely determined by the nature of the energy (reservoir pressure) available to move the oil to the surface. In petroleum engineering, the term “energy” refers to a reservoir driving mechanism. A reservoir driving mechanism may be defined as natural forces in the reservoir that displace reservoir fluids (oil, gas and water) from the reservoir to the producing wells and up to
the surface. Mainly, there are three common driving mechanism forces (water drive, gas cap drive and solution gas drive) or a combination of these forces which provide the natural energy necessary for oil recovery. Each reservoir is composed of a unique combination of geometric forms, geological rock properties, fluid characteristics, and drive mechanisms. Although no two reservoirs are identical in all aspects, they can be grouped according to the drive mechanism by which they produce. It has been observed that each drive mechanism has certain typical performance characteristics in terms of pressure decline and production rates.

Usually, in water drive reservoirs, the reservoir pressure declines very gradually depend on the aquifer efficiency. The reason for the minimal decline in reservoir pressure is that oil and gas withdrawals from the reservoir are replaced, almost volume for volume, by water encroaching upon the aquifer into the oil zone. Many reservoirs are bounded on a portion, or all of their peripheries, by water bearing rocks called aquifers. The aquifers may appear to be larger than the reservoir they adjoin, and thus, appear infinite for all practical purposes. The aquifers may diminish to the point of being negligible in their effects on reservoir performance. However, in gas cap reservoirs, due to the ability of the previously existing (above the oil) gas cap to expand, the reservoir pressure falls slowly and continuously. The actual rate of pressure decrease is related to the size of the gas cap when compared to the oil volume. On the other hand, the reservoir pressure in the solution gas drive declined rapidly and continuously. This reservoir pressure behavior is attributed to the fact that no external fluids or gas caps are available to provide a replacement to the oil withdrawals. When oil is produced from the reservoir,
reservoir pressure decreases and dissolved gas bubbles emerge from the oil. This gas expands and pushes the reservoir oil to the wellbore and up to the surface.

Accordingly, the decline of average reservoir pressure, at a given time, depends upon changes in production conditions and/or production problems. As a result, the average reservoir pressure can be correlated as a function of the oil, gas and water production rates (production history) and the number of producing wells. Figures 6.1, 6.2 and 6.3 illustrate the impact on the decline of an average reservoir pressure according to oil, gas and water production rates. Figure 6.4 shows the impact on declining average reservoir pressure according to the number of producing wells. Based on the above discussions, four types of data (oil, gas and water production rates and the number of producing wells) are selected to incorporate into the input layer of the network model.

6.3.1 Reservoir Production History and Pressure Data

A Libyan oil field, located in the Sirte Basin, was utilized in this study to develop a model with which to interpolate, predict and estimate current average reservoir pressure. The field, consisting of 60 naturally flowing oil production wells, began production in January, 1970. The daily production rates were 15,884 BPD of oil, 30 BPD of water and 2,621 M SCF/D of gas. The initial reservoir pressure was 1188 psi and the saturation pressure was 267 psi. The most recent average reservoir pressure of 832 psi occurred in May 2009. As of May 2009, the daily field production rates were 11,877 BPD for oil, 81,915 BPD for water and 1,941M SCF/D for gas. A total of 49 average reservoir pressure data points were recorded during more than 30 years of production (January 1979 to May 2009). The range of the dataset is illustrated in Table 6.1.
Figure 6.1 – Impact on the Decline of an Average Reservoir Pressure According to the Oil Production Rate. (unpublished data, personal communication)
Figure 6.2 – Impact on the Decline of an Average Reservoir Pressure According to the Gas Production Rate (unpublished data, personal communication)
Figure 6.3 – Impact on the Decline of an Average Reservoir Pressure According to the Water Production Rate (unpublished data, personal communication)
Figure 6.4 – Impact on the Decline of an Average Reservoir Pressure According to the Number of Producers (unpublished data, personal communication)
6.3.2 Data Preparation

In order to evaluate the model prediction performance after it was properly trained, the last two recorded average reservoir pressures (June 2007 and May 2009) were put aside and not switched in the process of model derivation. The remainder of the data points (47 points) was used to build the network. To avoid over fitting (overtraining), it is necessary to incorporate an early stopping technique. If the validation error increases while the training error steadily decreases, then a situation of over fitting may occur. When the performance of the validation test ceases to improve, the algorithm halts.
Table 6.1- Range of Overall Data (unpublished data, personal communication)

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Max</th>
<th>Min</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{avg}$, psi</td>
<td>1188</td>
<td>820</td>
<td>924</td>
</tr>
<tr>
<td>Qo, BPD</td>
<td>132740</td>
<td>10770</td>
<td>48538</td>
</tr>
<tr>
<td>Qw, BPD</td>
<td>139456</td>
<td>3572</td>
<td>72962</td>
</tr>
<tr>
<td>Qg, MSCF/d</td>
<td>21902</td>
<td>1732</td>
<td>7983</td>
</tr>
<tr>
<td>Producers</td>
<td>60</td>
<td>12</td>
<td>40</td>
</tr>
</tbody>
</table>
Table 6.2 – Range of Data Used in Training Process

<table>
<thead>
<tr>
<th>Model Parameter</th>
<th>Training Set</th>
<th>Validation Set</th>
<th>Testing Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>Min</td>
<td>Avg</td>
</tr>
<tr>
<td>$P_{\text{avg}, \text{psi}}$</td>
<td>1188</td>
<td>833</td>
<td>930</td>
</tr>
<tr>
<td>Qo, BPD</td>
<td>111455</td>
<td>10770</td>
<td>52100</td>
</tr>
<tr>
<td>Qw, BPD</td>
<td>139456</td>
<td>3572</td>
<td>71380</td>
</tr>
<tr>
<td>Qg, M SCF/D</td>
<td>18390</td>
<td>1739</td>
<td>8572</td>
</tr>
<tr>
<td>Producers</td>
<td>55</td>
<td>15</td>
<td>40</td>
</tr>
</tbody>
</table>
Accordingly, the training data points (47 points) are split randomly into three sets which are: training set, 70%, validation set, 15% and testing set, 15%. The validation set (7 points) is used to learn how to decide when to stop and the testing set (7 points) is used to evaluate how well an ANN performs on previously unidentified data. The ranges of the aforementioned datasets are shown in Table 6.2.

6.3.3 The Architecture of a Neural Network

After identifying the model inputs and data preparation, the next step is to specify the number of hidden layers and nodes in each layer. The usual strategy is based solely on trial and error. The attempt began with one hidden layer and one hidden node. Hidden nodes were gradually added. A total of 310 architectures were compared to the network parameters, as presented in Table 6.3. Table 6.4 shows the five best architectures from the compared 310 architectures. Before selecting the best architecture, the optimal training algorithm should be determined. Once more, the usual strategy is based solely on trial-and-error, which can be very time-consuming and requires significant user experience. However, the architectures presented in Table 6.4 are trained by the most common three used training algorithms, namely, Conjugate Gradient Descent BP (CGD), Quasi-Newton BP (Q-N), and Limited Memory Quasi-Newton BP (LM Q-N). In each case, the architecture was retrained five times with different initial weight randomizations. However, one method to avoid local minima is to retrain the network with a different weight initialization (Freeman and Skapura, 1991). This will locate the MSE function at different minima in the error surface, and then choose the acceptable solution.
Table 6.3 - Network Architecture Parameters

<table>
<thead>
<tr>
<th>Network Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden layer activation function</td>
<td>Logistic</td>
</tr>
<tr>
<td>Output activation function</td>
<td>Logistic</td>
</tr>
<tr>
<td>Error function</td>
<td>Sum-of-squares</td>
</tr>
<tr>
<td>Hidden layers search range</td>
<td>From 1 to 3</td>
</tr>
<tr>
<td>1\textsuperscript{st} layer hidden units search range</td>
<td>From 1 to 10 nodes</td>
</tr>
<tr>
<td>2\textsuperscript{nd} hidden layer units search range</td>
<td>From 1 to 6 nodes</td>
</tr>
<tr>
<td>3\textsuperscript{rd} hidden layer units search range</td>
<td>From 1 to 4 nodes</td>
</tr>
<tr>
<td>Fitness criteria</td>
<td>Inverse test error</td>
</tr>
</tbody>
</table>
Table 6.4 - Top Five Architectures - The best results are emphasized in bold

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Error, psi</th>
<th>R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>[4-10-1]</td>
<td>11.664</td>
<td>3.1229</td>
</tr>
<tr>
<td>[4-8-3-1]</td>
<td>12.765</td>
<td>8.8507</td>
</tr>
<tr>
<td>[4-3-4-1-1]</td>
<td>13.769</td>
<td>13.739</td>
</tr>
<tr>
<td>[4-8-1-4-1]</td>
<td>11.272</td>
<td>16.468</td>
</tr>
<tr>
<td>[4-7-6-4-1]</td>
<td>14.187</td>
<td>14.313</td>
</tr>
</tbody>
</table>
As a result, a total of 15 models were compared and the results are tabulated in Tables 6.5 through 6.9. The tables present the network performances of the testing dataset in terms of AE, ARE, and SE. The best results from each training algorithm group are emphasized in bold. Consequently, the Architecture [4–10–1] provides the best results when trained with CGD BP, as presented in Table 6.5.

6.4 Average Reservoir Pressure ANN Model

The proposed reservoir pressure ANN model architecture is multilayered with four input nodes and one hidden layer with ten nodes, [4–10–1]. The network was trained using feedforward backpropagation with a Conjugate Gradient Descent training algorithm. The neurons in the backpropagation used a logistic function as an input and an output activation function.

The proposed model provides low minimum errors in training, validation, testing and the overall dataset which is illustrated in Figures 6.5, 6.6, 6.7 and 6.8, respectively. The Figures show all the points are scattered around the $y = x$ line, indicating excellent agreement between the actual and estimated average reservoir pressure points. The statistical parameters for the prediction capability, as obtained from the network, are summarized in Table 10. The proposed network provides prediction values of average reservoir pressure with an average AE of 11.028 psi and an average ARE of 1.194%, indicating the model describes the data very well. Moreover, the error distribution for the training and overall datasets is presented in Figures 9 and 10 and both figures show the error is normally distributed with a mean approaching zero.
Table 6.5 - Error Analysis of Architecture [4-10-1] under Various Training Algorithms

<table>
<thead>
<tr>
<th>Training Algorithms</th>
<th>Criterion</th>
<th>AE, psi</th>
<th>ARE%</th>
<th>SE, (psi)$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGD</td>
<td>Average</td>
<td>19.28</td>
<td>2.0168</td>
<td>693.4</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>17.94</td>
<td>1.8705</td>
<td>880.2</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>0.50</td>
<td>0.0596</td>
<td>0.2552</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>44.62</td>
<td>5.0595</td>
<td>1991.4</td>
</tr>
<tr>
<td>QN</td>
<td>Average</td>
<td>27.4</td>
<td>2.7605</td>
<td>1859.1</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>33.29</td>
<td>3.2673</td>
<td>3643.9</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>0.070</td>
<td>0.0079</td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>99.33</td>
<td>9.9929</td>
<td>9866.3</td>
</tr>
<tr>
<td>LM-QN</td>
<td>Average</td>
<td>34.46</td>
<td>3.6487</td>
<td>2465.2</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>35.74</td>
<td>3.6823</td>
<td>4197.5</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>3.34</td>
<td>0.3799</td>
<td>11.18</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>107.54</td>
<td>10.8187</td>
<td>11564.3</td>
</tr>
</tbody>
</table>
Table 6.6 - Error analysis of Architecture [4-8-3-1] under Various Training Algorithms

<table>
<thead>
<tr>
<th>Training Algorithms</th>
<th>Criterion</th>
<th>AE, psi</th>
<th>ARE%</th>
<th>SE, (psi)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGD</td>
<td>Average</td>
<td>27.18</td>
<td>2.864</td>
<td>1159.8</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>20.52</td>
<td>2.141</td>
<td>1510.7</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>2.76</td>
<td>0.326</td>
<td>7.638</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>59.66</td>
<td>6.236</td>
<td>3559.3</td>
</tr>
<tr>
<td>QN</td>
<td>Average</td>
<td>27.53</td>
<td>2.747</td>
<td>2085.7</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>36.44</td>
<td>3.581</td>
<td>4394.6</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>3.21</td>
<td>0.364</td>
<td>10.30</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>108.78</td>
<td>11.944</td>
<td>11834.36</td>
</tr>
<tr>
<td>LM-QN</td>
<td>Average</td>
<td>44.36</td>
<td>4.758</td>
<td>3414.6</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>38.03</td>
<td>3.950</td>
<td>4846.6</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.225</td>
<td>0.144</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>116.5</td>
<td>11.720</td>
<td>13572</td>
</tr>
</tbody>
</table>
Table 6.7 - Error Analysis of Architecture [4-8-3-1] under Various Training Algorithms

<table>
<thead>
<tr>
<th>Training Algorithms</th>
<th>Criterion</th>
<th>AE, psi</th>
<th>ARE%</th>
<th>SE, (psi)$^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGD</td>
<td>Average</td>
<td>38.83</td>
<td>4.1392</td>
<td>767622</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>37.3</td>
<td>3.7969</td>
<td>217796</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>7.90</td>
<td>0.9328</td>
<td>572074</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>118.82</td>
<td>11.9539</td>
<td>1237794</td>
</tr>
<tr>
<td>QN</td>
<td>Average</td>
<td>43.98</td>
<td>4.4257</td>
<td>3849</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>43.75</td>
<td>4.1608</td>
<td>6425.6</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>3.66</td>
<td>0.4229</td>
<td>13.41</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>129.73</td>
<td>13.0513</td>
<td>16830</td>
</tr>
<tr>
<td>LM-QN</td>
<td>Average</td>
<td>27.77</td>
<td>2.9179</td>
<td>1890.5</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>33.45</td>
<td>3.342</td>
<td>4314.8</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>2.656</td>
<td>0.3067</td>
<td>7.052</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>108</td>
<td>10.8647</td>
<td>11663</td>
</tr>
</tbody>
</table>
Table 6.8 - Error Analysis of Architecture [4-8-1-4-1] under Various Training Algorithms

<table>
<thead>
<tr>
<th>Training Algorithms</th>
<th>Criterion</th>
<th>AE, psi</th>
<th>ARE%</th>
<th>SE, (psi)²</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGM</td>
<td>Average</td>
<td>44.29</td>
<td>4.6882</td>
<td>4751.7</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>52.8</td>
<td>5.3733</td>
<td>9418</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.140</td>
<td>0.1346</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>160.42</td>
<td>16.1389</td>
<td>25734.7</td>
</tr>
<tr>
<td>QN</td>
<td>Average</td>
<td>33.61</td>
<td>3.4476</td>
<td>2566.5</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>37.90</td>
<td>3.7487</td>
<td>5446.9</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>6.188</td>
<td>0.7146</td>
<td>38.29</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>121.81</td>
<td>12.2546</td>
<td>14837.8</td>
</tr>
<tr>
<td>LM-QN</td>
<td>Average</td>
<td>50.44</td>
<td>4.9274</td>
<td>5963.8</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>58.48</td>
<td>5.2843</td>
<td>10207</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>5.82</td>
<td>0.6727</td>
<td>33.94</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>159.71</td>
<td>14.0344</td>
<td>25508</td>
</tr>
</tbody>
</table>
Table 6.9 - Error Analysis of Architecture [4-7-6-4-1] under Various Training Algorithms

<table>
<thead>
<tr>
<th>Training Algorithms</th>
<th>Criterion</th>
<th>AE, psi</th>
<th>ARE%</th>
<th>SE, (psi)^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGD</td>
<td>Average</td>
<td>23.4</td>
<td>2.445</td>
<td>1043.8</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>22.28</td>
<td>2.268</td>
<td>1794.7</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>2.90</td>
<td>0.3426</td>
<td>8.42</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>70.47</td>
<td>7.0899</td>
<td>4966.5</td>
</tr>
<tr>
<td>QN</td>
<td>Average</td>
<td>34.68</td>
<td>3.6191</td>
<td>2100</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>29.9</td>
<td>3.1814</td>
<td>2688</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>2.714</td>
<td>0.3205</td>
<td>7.367</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>84.12</td>
<td>9.4099</td>
<td>7077</td>
</tr>
<tr>
<td>LMQN</td>
<td>Average</td>
<td>28.26</td>
<td>2.9021</td>
<td>2068.4</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>35.63</td>
<td>3.6052</td>
<td>4278</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>0.042</td>
<td>0.0048</td>
<td>0.0018</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>107.92</td>
<td>10.8577</td>
<td>11648</td>
</tr>
</tbody>
</table>
Figure 6.5 – Actual Average Reservoir Pressure versus Modeled ANN Values for Training Dataset
Figure 6.6 – Actual Average Reservoir Pressure versus Modeled ANN Values for Validation Dataset
Figure 6.7 – Actual Average Reservoir Pressure versus Model ANN Values for Testing Dataset
Scatterplot of Model Output vs Actual Average Reservoir Pressure

Overall Dataset

R² = 96.67%

Figure 6.8 – Actual Average Reservoir Pressure versus Modeled ANN Values for Overall Dataset
# Table 6.10 – Statistical Error Analysis of the Proposed Model

<table>
<thead>
<tr>
<th></th>
<th>AE, psi</th>
<th>ARE%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Training Dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>9.404</td>
<td>1.041</td>
</tr>
<tr>
<td>SD</td>
<td>10.063</td>
<td>1.140</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.056</td>
<td>0.006</td>
</tr>
<tr>
<td>Maximum</td>
<td>45.705</td>
<td>5.067</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Validation Dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>10.9</td>
<td>1.136</td>
</tr>
<tr>
<td>SD</td>
<td>11.686</td>
<td>1.135</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.253</td>
<td>0.149</td>
</tr>
<tr>
<td>Maximum</td>
<td>38.972</td>
<td>3.8472</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Testing Dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>19.278</td>
<td>2.017</td>
</tr>
<tr>
<td>SD</td>
<td>17.939</td>
<td>1.870</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.505</td>
<td>0.06</td>
</tr>
<tr>
<td>Maximum</td>
<td>44.625</td>
<td>5.059</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overall Dataset</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>11.028</td>
<td>1.194</td>
</tr>
<tr>
<td>SD</td>
<td>12.224</td>
<td>1.314</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.0559</td>
<td>0.006</td>
</tr>
<tr>
<td>Maximum</td>
<td>45.705</td>
<td>5.067</td>
</tr>
</tbody>
</table>
Figure 6.9 – Error Distribution for Training Dataset
Figure 6.10 – Error Distribution for Overall Dataset
6.5 Estimation of Average Reservoir Pressure

Once an ANN model is trained to a satisfactory level, it may be applied as an analytical tool with respect to new data. In the majority of cases, the reservoir pressure is not obtainable over long periods of time for several economic reasons. In this particular case, during more than 30 years of production, there are just 49 reservoir pressure readings available. In such a case, interpolation techniques are used to estimate the missing values. Interpolation is a method of obtaining values at positions in between the data points. Figure 6.11 illustrates the results of the ANN interpolation technique. Figure 6.10 can be used to estimate the average reservoir pressure at any time. Furthermore, the proposed model can be used to estimate current average reservoir pressure and to predict average reservoir pressure. Two unknown points, (June 2007 and May 2009) which were put aside and not switched in the process of model derivation, were used to evaluate the proposed model performance. Table 6.11 shows the prediction performance of the proposed model. The network provides average reservoir pressure values with an average AE of three psi and an average ARE of 0.361%. 
Figure 6.11 – ANN model performance
<table>
<thead>
<tr>
<th>Time</th>
<th>Actual, psi</th>
<th>Model, psi</th>
<th>AE, psi</th>
<th>ARE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>June-07</td>
<td>820</td>
<td>825</td>
<td>5</td>
<td>0.6097</td>
</tr>
<tr>
<td>May-09</td>
<td>832</td>
<td>831</td>
<td>1</td>
<td>0.1202</td>
</tr>
</tbody>
</table>
6.6 Conclusions

- A significant economic effect occurs when shutting in producing wells during the entire pressure build-up test. This study presents a new method with which to predict and interpolate the average reservoir pressure without closing the producing wells. This technique is suitable for constant and variable flow rates.
- The data used contained a set of 49 measured average reservoir pressures, collected from a Libyan oil field, located in the Sirte Basin.
- The network consisted of one hidden layer with ten neurons. The input layer received four parameters consisting of oil, gas and water production rates and a number of producers. The network was trained using feed forward backpropagation, employing a Conjugate Gradient Descent training algorithm.
- The results indicate average reservoir pressure can be predicted using neural network models without the wells having to be closed. Nevertheless, the neural network model has the ability to predict and interpolate average reservoir pressure accurately by employing reservoir production data.


6.7 References


Elmabrouk, S., Mayorga, R. and Shirif, E. 2010 “Artificial Neural Network Modeling for Reservoir Oil Production Prediction.” Paper presented at 5th TOG, Tripoli, Libya Nov.12-14

Gropal, Suchatita, 1998. Artificial Neural Networks for Spatial Data Analysis. [http://www.ncgia.ucsb.edu/giscc/units/u188/u188.html](http://www.ncgia.ucsb.edu/giscc/units/u188/u188.html)


CHAPTER VII

A HYBRID GENETIC ALGORITHM TO ESTIMATE AVERAGE RESERVOIR PRESSURE

7.1 Introduction

Chapter VII is an extension of Chapter VI. A Back-propagation Neural Network (BP-ANN) was established in Chapter VI to interpolate, predict and estimate the current average reservoir pressure ($P_{avg}$). In this chapter a Genetic Algorithm (GA) was employed for the evolution of connection weights in the BP-ANN model, as proposed in Chapter VI, in order to improve upon model performance and support the simultaneous optimization of the connection weights. Some researchers searched the connection weights of ANN using the GA instead of local search algorithms, including a gradient descent algorithm. They suggested that global search techniques, including the GA, might prevent ANN from falling into a local optimum (Gupta & Sexton, 1999). Thus, a combination of Artificial Neural Network (ANN) and the Genetic Algorithm (GA) based optimization, termed the “ANN-GA Model” was proposed to interpolate, predict and estimate average reservoir pressure. In this chapter, the GA is employed.
7.2 Genetic Algorithm

GA is a search algorithm, based upon the survival of the fittest among string structures, used to form a search algorithm (Goldberg, 1989). With respect to a solution for optimization problems, GA has been investigated recently and shown to be effective in exploring a complex space in an adaptive manner, guided by the biological evolution mechanisms of reproduction, crossover and mutation (Adeli and Hung, 1995). GA may avoid falling into minimum when a local search method such as gradient descent is used in ANN. GA performs the search process in four stages: initialization, selection, crossover and mutation (Davis, 1991).

The GA serves as an intelligent search and optimization technique and an adaptation of network weights. GA was proposed as a mechanism to improve the performance of the ANN inferential estimator. It decides the representation of connection weights in the form of binary or real numbers. During the connection weights evolution, the architecture of an ANN is re-defined and fixed. The ANN-GA combination provides powerful regression capabilities with tuning flexibility for either performance or cost-efficiency (Shanthi et al. 2009).

7.3 A hybrid Genetic Algorithm

The backpropagation (BP) algorithm and sigmoid function are used in this model as weights to train the ANN and are optimized by the GA rather than the gradient descent algorithm. The BP learning algorithm is a widely used algorithm but it has a drawback of converging to a set of sub-optimal weights. The GA offers an efficient search method
with respect to complex problem space and can be used as a powerful optimization tool.

The ANN-GA algorithm steps are:

1. Split the data into training validation and testing sets
2. Use the training and validation sets to design a suitable neural network
3. Initialize the populations (connection weights and thresholds)
4. Assign input and output values to ANN
5. Compute hidden layer values
6. Compute output values
7. Compute the error
8. If the error is acceptable, proceed to Step 12
9. Select parents of the next generation and apply a genetic operator (crossover and mutation)
10. Proceed to Step 6
11. Train the neural network with selected connection weights
12. Study the performance with the test data

The construction and generic operation of the GA for FFNN is as follows: Given a neural network that is ready to be trained, an initial population of individuals (chromosomes) is generated, each of which codifies a set of values for the weights of the connections and biases of the neural network. Then the fitness of each individual is evaluated, which entails allocating the values of the weights and biases codified by the individual in order to assess the network connections and calculate the mean square error. The set of training patterns are used to conduct this calculation. The error value is the
individual's fitness. Following the initialization stage, the genetic reproduction, crossover and mutation operators are usually applied to output new generations until there is convergence toward a population of individuals that encode the set of weights which minimize the network error.

7.4 BP-ANN-GA Model Building

Training BP-ANN with GA is divided into two parts: the first part is to optimize the weights with GA, and the second part is to train the BP-ANN with the result of GA. The multilayered feedforward BP-ANN, designed in Chapter VI, is used in this study. Four inputs (oil, gas and water production flow rates, and the number of producing wells) and one hidden layer with ten nodes, $[4−10−1]$ were chosen to interpolate, predict and estimate the current average reservoir pressure. The number of hidden nodes is determined using a trial and error approach. The main goal of this study is to establish a more accurate ANN model to interpolate and estimate current average reservoir pressure by applying a GA to optimize the connection ANN weights. GA may avoid falling into minimum when a local search method is used in the ANN model, as proposed in Chapter VI. Learning in a neural network involves modifying the weights and biases of the network in order to minimize the cost function. The cost function always includes an error term (a measure of how close the network's predictions are to the training set).

The field data described in Table 6.2 was used to build the BP-ANN-GA model. Again, the last two recorded average reservoir pressures (June 2007 and May 2009) were put aside to use in the prediction process. However, they not switched in the process of
model derivation. The remaining data points were split randomly into three sets: training, validation and testing datasets. The validation and testing sets consist of 7 points while the training set consists of 33 data points.

Based on trial and error, as explained in Chapter VI, the BP-ANN model architecture is chosen to be multilayered with four input nodes and one hidden layer with ten nodes, [4–10–1]. The neurons in the backpropagation used a logistic function as an input and output activation function. The only difference between the model proposed in Chapter VI and this model is that this model was trained using feedforward backpropagation with a hybrid GA algorithm to locate the global minimum of the MSE rather than the Conjugate Gradient Descent training algorithm.

7.4.1 Evolving PB-ANN weights using GA

ANN is initialized with random weights using GA with each weight being between −1.0 to +1.0. This study requires two sets of vectors of weights. The first vector set consists of connection weights between inputs to the hidden layer. The second vector set consists of connection weights between the hidden layers to the output layer. This study uses 4 input features and 10 hidden elements. The weights in the ANN are encoded in such a way that each weight is between −1.0 to +1.0. At this point, weights are assigned to each link. The network is trained with the training data to calculate the fitness of each chromosome and the mean square error (MSE) was returned. In this method, GA searches several sets of weight vectors simultaneously. Unlike the ANN model presented in Chapter VI, the training is conducted by the GA search instead of the Conjugate Gradient Descent
algorithm method. In this application, each string or chromosome in the population represents the weight and bias values of the network. The initial population is randomly generated. By selecting suitable parameters such as selection criteria, probability of crossover, probability of mutation, initial population, etc., for the GA, high efficiency and performance can be achieved. The objective function is minimization of the MSE. The considered fitness function is the minimum MSE and it is computed by recalling the network. After achieving the fitness values of all chromosomes, they are ranked based on the best fitness values. Half of the best ranked population is selected for the production of offsprings for the next generation. This half population undergo cross over with cross-over probability. Again, this will be mutated to give a new offspring, with mutation probability, which is combined with the selected best population to form a new population for the next generation. Table 7.1 shows the PB-ANN-GA parameters used in this study.

Basically, the performance of weight evolution using GA depended upon the number of populations and generations. If the parameters are set too low, the evolution may converge to an immature solution. However, the larger number of populations and generations would require a longer computation time for convergence. In this study, the number of individuals in the population was chosen as 100 and the number of generations used to evolve the solution is 20. Meanwhile, the Generational GA evaluates the fitness of the current population before creating an entirely new population to evaluate. The probability of crossover on average is chosen (0.7) and the mutation rate is chosen (0.033).
Eventually, the chromosome with the minimum error from the last generation of the GA algorithm was chosen as the initial weight of BP-ANN to conduct BP-ANN training according to the field data. The field average reservoir pressure and modeled values of BP-ANN-GA are compared. It shows the mean AE is 7.517 psi and mean ARE is 0.817% in the training dataset with a regression coefficient of 98.42%. The validation dataset provides a regression coefficient of 96.7%, whereas the testing dataset provides 94.38%. Table 7.2 shows the statistical error analysis of the modeled average reservoir pressure for training validation and testing datasets. Figure 6.1 shows the field reservoir pressure versus the modeled output for a given training, validation and testing data points.

### 7.5 Prediction of Average Reservoir Pressure

The two unknown points, (June 2007 and May 2009, which were put aside and not switched in the process of model derivation, were used to evaluate the proposed model prediction performance. The oil, gas and water production rates and the number of producing well values between June 2007 and May 2009 are used as the input vector to the trained BP-ANN-GA model and then the output average reservoir pressure was matched to the field values. The modeled values matched the field average reservoir pressure. Table 7.3 summarizes the prediction results.

### 7.6 Conclusion

In an attempt to improve the performance of the proposed BP-ANN model (proposed in Chapter VI), a new BP-ANN model, trained with GA (BP-ANN-GA), was established. In order to optimize the initial weights of BP-ANN, a hybrid algorithm of BP-ANN-GA
was introduced in this chapter. The algorithm was applied to the same oil field data used in Chapter VI in order to model the reservoir pressure performance. In comparing with BP-ANN (Chapter VI), this case study demonstrates its application with the results showing a slight improvement in training and validation datasets. Meanwhile, it gives similar results when testing and prediction data sets are used.
Table 7.1 – BP-ANN-GA Model Parameters

<table>
<thead>
<tr>
<th><strong>Network Parameters</strong></th>
<th><strong>Description</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural network architecture</td>
<td>[4-10-1]</td>
</tr>
<tr>
<td>Transfer function</td>
<td>logistic</td>
</tr>
<tr>
<td>Training dataset</td>
<td>33</td>
</tr>
<tr>
<td>Validation dataset</td>
<td>7</td>
</tr>
<tr>
<td>Testing dataset</td>
<td>7</td>
</tr>
<tr>
<td>Search method</td>
<td>Genetic algorithm</td>
</tr>
<tr>
<td>Population size</td>
<td>100</td>
</tr>
<tr>
<td>Fitness function</td>
<td>MSE</td>
</tr>
<tr>
<td>Number of generations</td>
<td>20</td>
</tr>
<tr>
<td>Probability of crossover</td>
<td>0.7</td>
</tr>
<tr>
<td>Probability of mutation</td>
<td>0.033</td>
</tr>
<tr>
<td>Random number seed</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 7.1 – Modeled Versus Field Average Reservoir Pressure
Table 7.2 – Statistical Error Analysis of the Proposed Model

<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>AE, psi</th>
<th>ARE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>7.517</td>
<td>0.817</td>
</tr>
<tr>
<td>SD</td>
<td>7.942</td>
<td>0.869</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.0425</td>
<td>0.005</td>
</tr>
<tr>
<td>Maximum</td>
<td>36.933</td>
<td>4.095</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Validation Dataset</th>
<th>AE, psi</th>
<th>ARE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>10.4</td>
<td>1.122</td>
</tr>
<tr>
<td>SD</td>
<td>9.368</td>
<td>0.979</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.923</td>
<td>0.108</td>
</tr>
<tr>
<td>Maximum</td>
<td>28.305</td>
<td>2.794</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Testing Dataset</th>
<th>AE, psi</th>
<th>ARE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>19.278</td>
<td>2.017</td>
</tr>
<tr>
<td>SD</td>
<td>17.94</td>
<td>1.87</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.505</td>
<td>0.06</td>
</tr>
<tr>
<td>Maximum</td>
<td>44.625</td>
<td>5.059</td>
</tr>
</tbody>
</table>
Table 7.3 – Model Prediction Performance

<table>
<thead>
<tr>
<th>Time</th>
<th>Field $P_{\text{avg, psi}}$</th>
<th>ANN-GA $P_{\text{avg, psi}}$</th>
<th>$AE$, psi</th>
<th>ARE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>June-07</td>
<td>820</td>
<td>830</td>
<td>10</td>
<td>1.219</td>
</tr>
<tr>
<td>May-09</td>
<td>832</td>
<td>832</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
7.7 References


CHAPTER VIII

ESTIMATING AVERAGE RESERVOIR PRESSURE: A LS-SVM APPROACH

8.1 Introduction

Insight into average reservoir pressure, and its change over time, plays a critical role in reservoir development. In particular, almost every oil well intervention requires insight with respect to average reservoir pressure. However, in order to determine the average reservoir pressure, the well is shut-in, resulting in loss of production. This is regarded as a pressure build-up test. In high permeability reservoirs, this may not be a significant issue, but in medium to low permeability reservoirs, the shut-in period during the entire build-up test may last several weeks before a reliable reservoir pressure can be estimated. This loss of production opportunity, as well as the cost of monitoring the shut-in pressure, is often unacceptable. It is of great practical value if the average reservoir pressure can be obtained from the historical production and reservoir pressure data without having to shut-in the well. It is clear that the production rate of a well is a function of many factors such as permeability, viscosity, thickness etc. Also, the rate is directly related to the driving force in the reservoir, i.e. the difference between the average reservoir pressure and the flowing pressure.
Chapter V clarifies how the ANNs have the ability to estimate and interpolate average reservoir pressure with good accuracy. The purpose of the work in this chapter, however, is to introduce a LS-SVM as a new alternative tool to predict and interpolate average reservoir pressure without closing the producing wells.

Recently Support Vector Machine (SVM) has garnered support in replacing ANN because of their predominance in global optimization but they are only one solution. LS-SVM is a new type of SVM which was introduced by Suykens et al. (2002). LS-SVM uses equality constraints instead of inequality constraints and a least squares error term in order to obtain a linear set of equations in the dual space. However, they all face the same underlying question - how to set the key parameters which may affect the performance of the algorithm. In the LS-SVM average reservoir pressure model, a two-step grid search with cross-validation on the training dataset was used to search for the best parameter $\gamma$ and $\sigma^2$.

8.2 Historical Production Data

The historical monthly reservoir production data used in this study, from March 1997 to July 2010, is related to an oil reservoir located in the Sirte Basin, Libya. This oil reservoir which has 60 naturally producing wells began production in January, 1970. Generally, during the life of an oil reservoir, cumulative production is recorded, and average reservoir pressures are periodically measured. However, only 49 measurements of average reservoir pressure were obtained during the time period from January 1970 to May 2009, as illustrated in Table 8.1. In order to evaluate the LS-SVM model prediction
performance after it was properly trained, the last two recorded average reservoir pressures (June 2007 and May 2009) were put aside and not switched in the process of model derivation. The ranges of the dataset are illustrated in Table 6.1
Table 8.1 - Average Reservoir Pressure Data (unpublished data, personal
communication); Testing Dataset is emphasized in Bold

<table>
<thead>
<tr>
<th>Date</th>
<th>P_{avg}</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec-70</td>
<td>1081</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-70</td>
<td>1138</td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Sep-70</td>
<td>1134</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-71</td>
<td>1072</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-71</td>
<td>1053</td>
<td>Training</td>
</tr>
<tr>
<td>Sep-72</td>
<td>1040</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-72</td>
<td>1006</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-73</td>
<td>996</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-73</td>
<td>1004</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-74</td>
<td>986</td>
<td>Training</td>
</tr>
<tr>
<td><strong>Dec-74</strong></td>
<td><strong>994</strong></td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Mar-75</td>
<td>1072</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-75</td>
<td>1013</td>
<td>Training</td>
</tr>
<tr>
<td>Mar-76</td>
<td>949</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-76</td>
<td>927</td>
<td>Training</td>
</tr>
<tr>
<td>Mar-77</td>
<td>898</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-77</td>
<td>907</td>
<td>Training</td>
</tr>
<tr>
<td><strong>Dec-77</strong></td>
<td><strong>894</strong></td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Sep-78</td>
<td>841</td>
<td>Training</td>
</tr>
<tr>
<td>Mar-79</td>
<td>833</td>
<td>Training</td>
</tr>
<tr>
<td>Sep-79</td>
<td>854</td>
<td>Training</td>
</tr>
<tr>
<td>Sep-80</td>
<td>861</td>
<td>Training</td>
</tr>
<tr>
<td>Mar-82</td>
<td>874</td>
<td>Training</td>
</tr>
<tr>
<td>Oct-82</td>
<td>844</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-83</td>
<td>904</td>
<td>Training</td>
</tr>
<tr>
<td><strong>Dec-84</strong></td>
<td><strong>880</strong></td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Jun-85</td>
<td>874</td>
<td>Training</td>
</tr>
<tr>
<td>Dec-85</td>
<td>881</td>
<td>Training</td>
</tr>
<tr>
<td>Apr-86</td>
<td>876</td>
<td>Training</td>
</tr>
<tr>
<td>Oct-86</td>
<td>866</td>
<td>Training</td>
</tr>
<tr>
<td>Apr-87</td>
<td>863</td>
<td>Training</td>
</tr>
<tr>
<td><strong>Feb-88</strong></td>
<td><strong>882</strong></td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Mar-89</td>
<td>894</td>
<td>Training</td>
</tr>
<tr>
<td>Feb-90</td>
<td>930</td>
<td>Training</td>
</tr>
<tr>
<td>Jul-90</td>
<td>908</td>
<td>Training</td>
</tr>
<tr>
<td>Mar-91</td>
<td>902</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-92</td>
<td>876</td>
<td>Training</td>
</tr>
<tr>
<td>Apr-94</td>
<td>934</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-97</td>
<td>865</td>
<td>Training</td>
</tr>
<tr>
<td>Sep-98</td>
<td>853</td>
<td>Training</td>
</tr>
<tr>
<td><strong>Dec-99</strong></td>
<td><strong>866</strong></td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Mar-01</td>
<td>852</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-02</td>
<td>852</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-03</td>
<td>859</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-04</td>
<td>853</td>
<td>Training</td>
</tr>
<tr>
<td><strong>Jun-05</strong></td>
<td><strong>847</strong></td>
<td><strong>Testing</strong></td>
</tr>
<tr>
<td>Jun-06</td>
<td>842</td>
<td>Training</td>
</tr>
<tr>
<td>Jun-07</td>
<td>820</td>
<td>Forecasting</td>
</tr>
<tr>
<td>May-09</td>
<td>832</td>
<td>Forecasting</td>
</tr>
</tbody>
</table>
An issue of concern in evaluating the approximation ability of a predicting model is to ensure that it has no overfitting on the training dataset and to obtain good model performance. In general, as with any LS-SVM problem, if the model parameters ($\gamma$ and $\sigma^2$) are not well selected, the predicted results will not be sufficient and the generalization ability will be poor. Tuning the model parameters is usually completed by minimizing the estimated generalization error. The generalization error is a function that measures the generalization ability of the constructed models, i.e., the ability to predict correctly the performance of an unknown sample. In the present methodology, a holdout technique is used to estimate the generalization error. Therefore, the 47 data points are separated into two sets, called “the training set” (40 points) and “the testing set” (7 points). The testing data points are emphasized in bold in Table 8.1. The LS-SVM model is constructed using only the training set and is then tested using the test dataset. The test data are completely unknown to the estimator. The errors it makes are accumulated to give the mean test set error, which is used to evaluate the model.

8.3 LS-SVM Model

To obtain satisfactory LS-SVM model performances, two model parameters ($\gamma$, $\sigma^2$) must be carefully chosen. The higher level parameters are usually referred to as “hyperparameters”. In general, in any LS-SVM problem, if the hyper parameters of the model are not well selected, the predicted results will not be adequate and the generalization ability will be poor. Tuning of the hyper parameters is usually conducted by minimizing the estimated generalization error. The generalization error is a function that measures the generalization ability of the constructed models, i.e., the ability to predict correctly the
performance of an unknown sample. In the present methodology, the holdout technique is used to estimate the generalization error.

The holdout method is the simplest method of cross validation. The data set is separated into two sets, called the “training set” and the “testing set”. The LS-SVM is constructed using only the training set. Then, the LS-SVM is asked to predict the output values for the data in the testing set. This testing dataset is completely unknown to the model (it has never observed the output values). The errors it makes are accumulated as before in order to give the Mean Square Error (MSE), which is used to evaluate the model.

As described in Chapter VI, the decline of average reservoir pressure in oil reservoirs, at a given time, depends upon changes in production conditions and/or production problems. As a result, the average reservoir pressure can be correlated as a function of the oil, gas and water production rates (production history) and the number of producers. Accordingly, four inputs are selected to be the inputs to the LS-SVM average reservoir pressure model. The LS-SVM regression model with RBF kernel function was established. In this method, the two parameters of regularization parameter ($\gamma$) and the kernel parameter ($\sigma^2$) of RBF are crucial to the performance of the LS-SVM model. The former determines the trade off between minimizing the training error and minimizing model complexity and the latter is the bandwidth and implicitly defines the nonlinear mapping from input space to a high dimensional feature space.

8.3.1 Modeling Process

The input layer has four nodes (oil, gas and water production rates and the number of producers) and the output layer has one node (average reservoir pressure). The RBF is
selected as a kernel function. A selected training sample set is used to train the LS-SVM.

Because of the variables in the collected data, each has different levels of magnitudes from each other. Both the input and output data are normalized before the training (all data are scaled between 0 and 1) by using the following formula:

\[ x_i' = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} \quad (7.1) \]

where \( x_i \) is the value of the \( i^{th} \) dimension of the data, \( \min(x_i) \) and \( \max(x_i) \) are the minimum and maximum of the \( i^{th} \) dimension of the data.

The main steps of the modeling process of the LS-SVM average reservoir pressure are as follows:

- **Step 1**: select proper data and split it into training and testing groups.
- **Step 2**: data are pre-processed.
- **Step 3**: chose a kernel type (RBF kernel is used).
- **Step 4**: set an initial guess for regularization parameter (\( \gamma \)) and kernel parameter (\( \sigma^2 \)).
- **Step 5**: apply the second level of Bayesian inference. The parameter \( \gamma \) is estimated on the second level.
- **Step 6**: apply the third level of Bayesian inference. This will estimate the parameter \( \sigma^2 \).
- **Step 7**: tune the \( \gamma \) and \( \sigma^2 \) obtained from Steps 5 and 6 using a two-grid search with cross-validation on the training dataset. The parameter set (\( \gamma, \sigma^2 \)) with the minimal criteria is optimal.
Step 8: Use a training dataset train with the LS-SVM having the tuned values of $\gamma$ and $\sigma^2$.

Step 9: measure the error by calculating MSE, RMSE and $R^2$.

Step 10: test the LS-SVM model using the testing dataset. Calculate MSE, RMSE and $R^2$.

Step 11: use the monthly production historical data from Mach 1970 to Jun 2006 as inputs of the LS-SVM to interpolate the average reservoir pressure and estimate the missing values.

Step 12: use the monthly production data from July 2006 to July 2010 as inputs for the LS-SVM to forecast average reservoir pressure.

8.3.2 Tuning Model Parameters

Two-step grid searching and leave-one-out cross validation was applied to the training dataset in order to realize the global optimization. The main idea behind this method is to find the optimal parameters that minimize the prediction error of the regression LS-SVM model. The prediction error can be estimated by leave-one-out cross-validation. Cross validation was employed with respect to the training data of this work.

The LS-SVR model is carried out using the LS-SVMLab1.5 Toolbox, developed by Suykens (2003) under Matlab.

The four reservoir parameters (oil, gas and water production rates and a number of producing wells) were employed as the input of the LS-SVM model. Figure 8.1 presents the process of optimizing the LS-SVM method parameter $\gamma$ and $\sigma^2$ and in which the symbol ‘•’ denotes the grid dot during the first search while the ‘×’ denotes the grid dots.
in second time period. The optimal combination of $\gamma$ and $\sigma^2$ for average reservoir pressure LS-SVM model was obtained at the values of 17.0586 and 4.1287, respectively.

The error term between calculated and measured average reservoir pressure values was obtained and plotted as a histogram in Figure 8.2. The histogram shows the error term is normally distributed with zero mean. Figure 8.3 shows the scatter plot of the observation average reservoir pressure versus the output from training LS-SVM$_{RBF}$ $\gamma=17.0586, \sigma^2=4.1287$ model. The model shows a correlation coefficient ($R^2$) of 97.29% with a MSE of 207.65 psi and a RMSE of 14.416 psi.
Figure 8.1 - Contour Plot of the Optimization Error (MSE) for LS-SVM Regression Model of Calculating Average Reservoir Pressure. The Black Square Indicates the Selected Optimal Settings of $\gamma [\exp(2.837)]$ and $\sigma^2 [\exp(1.418)]$. 
8.4 Interpolating Average Reservoir Pressure

The trained LS-SVM is positioned as an interpolating tool. It is used to interpolate the average reservoir pressure during a specific period. Interpolation is a method to predict unknown values from data observed at known locations or times. After the LS-SVM model is trained, interpolations are applied to estimate the missing average reservoir pressure values on the basis of the remaining known sample points. In order to analyze the interpolation quality, the errors between observed average reservoir pressure and the interpolated values during the same period of time are calculated. The results show that the MSE is 1510 psi, RMSE is 38.86 psi and the correlation coefficient is 83.6%. Figure 8.4 illustrates the accuracy of the LS-SVM model on the testing dataset. The plot shows all the points remain close to the \( y = x \) line, indicating no over fit exists in the model. Figure 8.5 can be used as a tool to interpolate the missing average reservoir pressure values. The statistical error analysis for training testing and all data points are illustrated in Table 8.2.

8.5 Forecasting Process

The two points of June 2007 and May 2009 (Table 8.1) are used as the input of the LS-SVM model, and then forecasting is carried out. The model shows excellent forecasting precision as illustrated in Table 8.3. The model provides forecasting average reservoir pressure with a mean AE of 8.5 psi and a mean ARE of 1.029%.
Figure 8.2 – Histogram of the Errors of the LS-SVM Average Reservoir Pressure Model
Figure 8.3 – Performance of the Optimal LS-SVM Model in the Training Dataset
Figure 8.4 – Performance of the LS-SVM Model. Both training and Testing Dataset are Scattered around the $y = x$ Line indicating No Overfit Existing in the Model
Figure 8.5 – LS-SVM Model Performance. The Plot can be used to interpolate the missing average reservoir pressure values.
Table 8.2 – Statistical Error Analysis of the Proposed Model

<table>
<thead>
<tr>
<th>Statistical Error Analysis</th>
<th>Training Set</th>
<th>Testing Set</th>
<th>All Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R^2$, %</td>
<td>97.29</td>
<td>83.6</td>
<td>95</td>
</tr>
<tr>
<td>MSE, psi</td>
<td>207.6</td>
<td>1510</td>
<td>393.7</td>
</tr>
<tr>
<td>RMSE, $(\text{psi})^2$</td>
<td>14.416</td>
<td>38.56</td>
<td>19.84</td>
</tr>
<tr>
<td>SD, psi</td>
<td>14.58</td>
<td>38.34</td>
<td>19.92</td>
</tr>
<tr>
<td>Mean AE, psi</td>
<td>10.53</td>
<td>24.94</td>
<td>12.59</td>
</tr>
<tr>
<td>Max AE, psi</td>
<td>37.88</td>
<td>91.82</td>
<td>91.82</td>
</tr>
<tr>
<td>Min AE, psi</td>
<td>0.09</td>
<td>0.045</td>
<td>0.045</td>
</tr>
<tr>
<td>Mean ARE, %</td>
<td>1.14</td>
<td>2.55</td>
<td>1.34</td>
</tr>
<tr>
<td>Max ARE, %</td>
<td>4.2</td>
<td>9.24</td>
<td>9.24</td>
</tr>
<tr>
<td>Min ARE, %</td>
<td>0.008</td>
<td>0.005</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Table 8.3 – LS-SVM Model Forecasting Performance

<table>
<thead>
<tr>
<th>Time</th>
<th>Actual, psi</th>
<th>Model, psi</th>
<th>AE, psi</th>
<th>ARE%</th>
</tr>
</thead>
<tbody>
<tr>
<td>June-07</td>
<td>820</td>
<td>828</td>
<td>8</td>
<td>0.97</td>
</tr>
<tr>
<td>May-09</td>
<td>832</td>
<td>840</td>
<td>9</td>
<td>1.08</td>
</tr>
</tbody>
</table>
8.6 Conclusions

Shutting in the wells during the entire pressure build-up test means a loss of income. The purpose of the work, however, is to introduce a LS-SVM as a new alternative tool to predict and interpolate average reservoir pressure without shutting in the producing wells. This technique is suitable for constant and variable flow rates. The study is based on historical production and reservoir pressure data for an oil reservoir located in Sirte Basin, Libya. The reservoir contained a set of 60 oil wells. There are just 49 measurements of average reservoir pressure from the period of March 1970 to July 2010 (more than 40 years of production). However, average reservoir pressure is measured periodically, resulting in a loss of production. LS-SVM model is performed to interpolate the missing values and to predict current average reservoir pressure.

Oil, water and gas production rates and a number of producers are used as inputs for the LS-SVM prediction model. The optimal model parameters ($\gamma$, $\sigma^2$) are obtained by applying a two-grid search with cross-validation on the training dataset only. The results indicate the LS-SVM approach has the ability to interpolate and predict current average reservoir pressure accurately by employing historical production data.
8.7 References


CHAPTER IX

ARTIFICIAL NEURAL NETWORK MODELING FOR OIL
PRODUCTION PREDICTION

9.1 Introduction

The prediction of oil reservoir production performance has been an on-going challenge for petroleum engineers. It is an essential component to petroleum reservoir management. Traditionally, numerical simulations and decline curve analysis have been used to predict a reservoir’s future performance based on its current and past performance.

Reservoir simulation is a tool used to predict hydrocarbon reservoir performance under various operating strategies. A set of reservoir parameter values are placed in the simulator to obtain a set of fluid flows. They are listed as a time-series over a specified period of time. This time-series is compared with historical data to evaluate their differences. If they are not in agreement, the input parameters are modified and the simulation is repeated until a satisfactory agreement (history matching) is reached between the simulated flow outputs and the historical production data. This history matching is very time consuming and offers a non-unique solution with a high degree of uncertainty.

Decline curve analysis fits the observed production rates of individual wells, groups of wells or a whole reservoir using mathematical functions to predict future production
by extrapolating the declining function. The basis of decline curve analysis is to match past production performance with a model assuming production continues to follow the historical trend. Most decline curve analysis techniques are based on the empirical Arps exponential, hyperbolic and harmonic equations (Arps 1954). In some cases, production decline data do not follow a particular model but crossover the entire set of curves (Camacho and Raghavan 1989). In addition, decline curve analysis does not take into account opening or closing intervals and varying water or gas injection rates.

In recent years, there has been a steady increase in the application of Artificial Neural Networks (ANNs) in engineering. ANNs have been used to address several fundamental and specific problems that conventional computing problems have been unable to solve, especially when the data for engineering design, interpretations and calculations are less than adequate. Lately, neural network was found to be a very suitable technology to solve problems in petroleum and gas engineering. Most recently, Elshafei et al. (2011) presented two neural networks to predict the performance of the multi-stage gas/oil separation plants in crude oil production. The neural networks accept the initial and final pressures and temperatures of each stage and then attempt to utilize the oil composition information to predict the stage gas/oil ratio. Recently, Elmabrouk et al. (2010) established a neural network model that is able to map certain relationships that control previous production performance of the reservoir in order to predict current average reservoir pressure without shutting in the wells. However, the current literature with respect to the prediction of oil reservoir production performance using artificial intelligence techniques is sparse because it is a new area of research. In 2001, Zhong He et al. developed a backpropagation neural network model to predict the production
performance of oil wells with respect to spatial variation and time. The network contains one hidden layer with 10 neurons, an input layer containing 11 neurons and one bias neuron. The 11 inputs include the number of wells, the X and Y coordinates of the wells, the cumulative production at time $t-2$, the cumulative production at time $t-1$, the cumulative production at time $t$, the derivative of cumulative production at time $t$, the shut in switch, the average distance to surrounding wells, average cumulative production of surrounding wells at time $t$ and the cumulative production days at time $t$. The output layer contains one neuron representing the cumulative production at time $t+1$. By comparing the two cases, the results indicate an adequate capacity for short term predictions.

Nguyen and Chen (2005) used Decision Support System (DSS) to model production by incorporating curve fitting and Artificial Neural Network (ANN). DSS is a computer-based system designed to model data and qualify decisions. According to the results, Nguyen and Chen concluded no one model provides a consistently reliable prediction. This is due to a weakness in the DSS and assumptions based on expert opinion. A major system weakness is an assumption that the production rate is proportional to the initial production.

Da Selva et al. (2007) used four data-mining technologies (Specific Knowledge, General Knowledge, Optimal Search and Gols) to build a time series model to predict oil production in petroleum reservoirs. They studied several models with which to forecast production curves by applying historical data from Bonsucceso and Carmopolis Brazilian oil fields. The study shows the Optimal Search approach worked very well and the Gols
approach gave reasonable results, while Specific Knowledge and General Knowledge gave poor results.

The objective of this study is to design a feedforward Artificial Neural Network model with backpropagation neural network model as an alternative technique with which to predict oil reservoir production performance taking into account the number of production and injection wells in service and the varying water injection rates. Historical production data from a Libyan oil field located in the Sirte Basin was used to train the network. The training network is able to serve as a practical and robust reservoir production management tool.

9.2 Oil Production Prediction Neural Networks

The most popular ANN model is the architecturally trained Multi-Layer Perceptron (MLP) model employing the feedforward backpropagation algorithm. The MLP architecture is composed of at least three layers of processing units interconnected via weighted connections. The first layer consists of the input vector and the last layer consists of the output vector. The intermediate layers, called “hidden layers”, represent neural pathways and modify the input data through several weighted connections.

There are three major phases to network training with backpropagation. During the initial phase, the input vector is presented to a network which is activated via the forward pass. This generates a difference (error) between the input of the network and the desired output. During the second phase, the computed output error propagates back through the network (error backward pass). During the third phase, connection weights are corrected by feeding the sum of the squared errors from the output layer back, through the hidden
layers, to the input layer. This process is repeated until the connection weights produce an output which is within a predetermined tolerance of the desired output (Ripley 1996).

The selection of an optimum network model is a difficult task requiring a procedure of trial-and-error (Heaton 2008). Thus, several networks with various (a) numbers of hidden units, (b) training algorithms and (c) activation functions are attempted and a generalization error is estimated for each. The network with the minimum estimated generalization error is chosen.

9.2.1 Oil Reservoir Production Data

An oil field in the Sirte Basin of Libya has 11 oil wells and 3 water injection wells. The field has been in production since September, 1978 when it initially had an average production rate of 3700 BOPD, an initial average reservoir pressure of 2960 psig and a bubble point pressure of 2710 psig. Following 7 months of production, the average reservoir pressure declined to 2665 psig. This rapid decline in reservoir pressure indicates a limited aquifer size and very weak water influx. Pressure maintenance by water injection was initiated in April, 1982 when 3 wells were converted from oil production wells to injection wells and placed on the water injection system. As of December, 2009 the reservoir had produced a cumulative oil production of 38.5 MMSTB, and a cumulative water production of 19.75 MMBW with an average oil production rate of 2316 BOPD and an average water cut of 38%. Figure 9.1 illustrates the reservoir’s oil production performance.
Figure 9.1 — Reservoir Oil Production Performance from Sept. 1978 to July 2010 (unpublished data, personal communication)
9.2.2 Data Pre-Processing

It is not possible to expect to achieve a good model from insufficient data. Therefore, before training the network, the oil field production data are checked and pre-processed to avoid peculiar answers from the trained network model. However, depending upon the problem, there might be special features in the data that could be used in testing data quality. One way to check for quality is to view graphical representations of the data in question, in the hope of selecting a reasonable subset while eliminating problematic areas. As presented in Figure 9.1, the oil field production data are time-dependent and incorporate atypical patterns. Such patterns lead to the deterioration of the model performance. Thus, the primary production period (the end of March 1982) in which water injection was started and the period from April 1982 to January 1995 were excluded as illustrated in Figure 9.1. The remainder of the production data points were then split into two sets before starting to specify the model architecture. The first set (Feb. 1995 to July 2009) was used to build the network model. The second set (Aug. 2009 to July 2010) was used to predict the average reservoir oil production rate.

To avoid over fitting and improve the network model generalization, the first dataset was subdivided randomly into training, validation, and testing sets. The training set (67%) is used to compute gradient and update the network weights and biases. The validation set (16.5%) is used to check network performance during the training process. Training can be stopped when the performance of the model on the validation dataset gives a minimum error. The testing set (16.5%) is used to fine-tune models. It is not utilized in training or in validation, but is used to determine the optimum network
architecture, select the appropriate model and assess the performance (generalization) of a fully-specified model.

**9.2.3 The Network Model Architecture**

The key problem in our approach is the determination of how to select as few inputs as possible. In other words, our task is to find an effective set of inputs to capture the main features of our prediction problem. The proposed network was based on practical experience and the necessity to map relationships that control past oil reservoir production performance in order to predict production levels, while taking into account the number of production and injection wells in service and the varying water injection rates. The chosen network has five inputs and one output. The inputs are the average reservoir oil production rate at time \( t \), the average reservoir gas production rate at time \( t \), the water injection rate at time \( t+1 \), the number of oil wells in production at time \( t+1 \) and the number of injection wells at time \( t+1 \). The output is the average reservoir oil production rate at time \( t+1 \). The selection of an optimum network model can be achieved using a trial-and-error approach. Accordingly, the best results were obtained by a network model consisting of 2 hidden layers, 13 nodes in the first layer and 8 in the second layer. The nodes in the hidden and output layers are activated with a logistic function then trained by the Quasi-Newton algorithm which provides a backpropagation error with the lowest sum-squared-error.

**9.3 Results and Discussions**

Figure 9.2 shows a comparison between the network outputs of the average reservoir oil rate versus the actual average reservoir oil rate for each training data point. The average
Absolute Error (AE) of the network output is 381.2 BOPD and the average Absolute Relative Error (ARE) is 10.41%. The network provides results which are very close to the actual average reservoir oil rates. This indicates good agreement between the actual and observed average reservoir oil rates. The performances of the training, validation and testing datasets are presented in Figures 9.3, 9.4 and 9.5, respectively. The statistical parameters used to measure the network’s prediction capability are obtained from the training, validation and testing datasets and summarized in Table 9.1. The proposed network provides prediction values for oil reservoir rates with average AEs of 391 BOPD, 324 BOPD and 392 for the training, validation and testing datasets, respectively. Moreover, average AREs of 10.3%, 9.2% and 12% were obtained from the training, validation and testing datasets, respectively indicating the network describes the data very well.

The second dataset was used to predict reservoir oil production so the robustness and accuracy of our network approach could be studied with respect to predicting oil reservoir production as illustrated in Figure 9.6. The predicted reservoir oil rate values agree with the historical values, indicating the training network can serve as a practical robust reservoir production management tool. The network provides reservoir oil rates with an average AE of 358 and an average ARE of 12%.
Figure 9.2—ANN Model Output versus the Actual Reservoir Production (1st Dataset)
Figure 9.3– Performance of the ANN Model on Training Dataset
Figure 9.4—Performance of the ANN Model on Validation Dataset
Figure 9.5– Performance of the ANN Model on Testing Dataset
Table 9.1 – Error Analysis of the Proposed Model

<table>
<thead>
<tr>
<th></th>
<th>Target BOPD</th>
<th>Output BOPD</th>
<th>AE BOPD</th>
<th>ARE %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall (1st Dataset)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3866</td>
<td>3909</td>
<td>381</td>
<td>10</td>
</tr>
<tr>
<td>SD</td>
<td>1315</td>
<td>1236</td>
<td>348</td>
<td>9</td>
</tr>
<tr>
<td>Minimum</td>
<td>1832</td>
<td>2310</td>
<td>1.42</td>
<td>0.02</td>
</tr>
<tr>
<td>Maximum</td>
<td>8195</td>
<td>6759</td>
<td>1680</td>
<td>39.6</td>
</tr>
<tr>
<td></td>
<td>Training</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3922</td>
<td>3920</td>
<td>391</td>
<td>10</td>
</tr>
<tr>
<td>SD</td>
<td>1347</td>
<td>1224</td>
<td>351</td>
<td>8</td>
</tr>
<tr>
<td>Minimum</td>
<td>1832</td>
<td>2310</td>
<td>1.42</td>
<td>0.02</td>
</tr>
<tr>
<td>Maximum</td>
<td>8195</td>
<td>6644</td>
<td>1602</td>
<td>39.7</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3926</td>
<td>3960</td>
<td>325</td>
<td>9</td>
</tr>
<tr>
<td>SD</td>
<td>1224</td>
<td>1260</td>
<td>374</td>
<td>10</td>
</tr>
<tr>
<td>Minimum</td>
<td>1931</td>
<td>2374</td>
<td>1.84</td>
<td>0.04</td>
</tr>
<tr>
<td>Maximum</td>
<td>6250</td>
<td>6251</td>
<td>1680</td>
<td>39.4</td>
</tr>
<tr>
<td></td>
<td>Testing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3556</td>
<td>3811</td>
<td>392</td>
<td>12</td>
</tr>
<tr>
<td>SD</td>
<td>1210</td>
<td>1263</td>
<td>303</td>
<td>9</td>
</tr>
<tr>
<td>Minimum</td>
<td>1985</td>
<td>2465</td>
<td>9.6</td>
<td>0.25</td>
</tr>
<tr>
<td>Maximum</td>
<td>7038</td>
<td>6759</td>
<td>1163</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td>Prediction (2nd Dataset)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>3177</td>
<td>3509</td>
<td>358</td>
<td>12</td>
</tr>
<tr>
<td>SD</td>
<td>356</td>
<td>256</td>
<td>233</td>
<td>8.3</td>
</tr>
<tr>
<td>Minimum</td>
<td>2676</td>
<td>3131</td>
<td>8.76</td>
<td>0.22</td>
</tr>
<tr>
<td>Maximum</td>
<td>3970</td>
<td>3961</td>
<td>666</td>
<td>24.3</td>
</tr>
</tbody>
</table>
Figure 9.6– Prediction of average reservoir oil production (one-year-ahead from Aug. 2009 to Jul. 2010)
9.4 Conclusions

This study provides a new, alternative method with which to predict oil reservoir production based on historical production data. It demonstrates the artificial neural network technique can be used to predict oil reservoir performance, and can serve as a practical and robust tool with regard to reservoir management.
9.5 References

CHAPTER X

FORECASTING OF OIL PRODUCTION UNDER WATER INJECTION: LS-SVM APPROACH

10.1 Introduction

The art of managing oil and gas reservoirs has progressed over the years. Water injection has proven to be one of the most economical methods of reservoir management. Water injection in reservoir engineering refers to the method wherein water is injected into the reservoir. The technology can be valuable in helping to maintain a pressure that continues to move oil to the surface. Reservoir simulation models are applied to forecast oil production performance within a water injection project.

When an oil field is first discovered, a reservoir model is constructed utilizing a geological model. Once the oil field enters into the production stage, many changes take place in the reservoir and thus, the changes must be reflected in the reservoir model. Usually, simulator programs are used to generate a model which simulates the reservoir. In reservoir simulation, the simulator takes a set of reservoir parameter values as inputs and returns a set of fluid flow information as outputs. The outputs are usually a time series over a specific period of time. This time series is then compared with the historical production data in order to evaluate their match. Experts modify the input model parameters and rerun the simulation model. This process (history matching) continues
until the model output rates match the historical production rates. History matching is the process of updating reservoir descriptor parameters to reflect such changes, based on production data. The process of history matching is an inverse problem. In this problem, a reservoir is a black box with unknown parameter values. Since inverse problems have no unique solution, a large number of models, with good history matching, can be achieved. In other words, in a reservoir simulation process, more than one combination of inputs of reservoir parameter values gives the same output. As a result, history matching is a challenging task for the following reasons:

- Highly non-linear flow equations with a large number of assumptions and input parameters
- Inverse problem (non-unique solution), where more than one model can produce acceptable outputs
- Inadequate results with high uncertainty in production forecasting
- Time consuming

In this study, a novel method of utilizing LS-SVM regression is presented to predict the performance of oil production within water injection reservoirs. In this approach, historical production and injection data are used as inputs. The approach can be categorized as a new and rapid method with reasonable results. Another application of this model is that it can be utilized to find the most economical scenario of water injection to maximize ultimate oil recovery. This method could be a new window of fast simulators. It has reasonable accuracy, requires little data and can forecast quickly.
10.2 Description of the Production and Injection Data

The historical production and injection data have been obtained from a small oil field located in the Sirte Basin of Libya. The oil field initially had 14 oil wells. It has been in production since September, 1978 when it initially had an oil production of 3700 BOPD, an initial average reservoir pressure of 2960 psig and a bubble point pressure of 2710 psig. Following 7 months of production, the average reservoir pressure declined to 2665 psig. This rapid decline in reservoir pressure indicates a limited aquifer size and very weak water influx. Pressure maintenance by water injection was initiated in April, 1982 when 3 wells were converted from oil production wells to injection wells and placed on the water injection system to maintain the reservoir pressure that continues to move oil to the surface. The field’s operations were shutdown in February 1999 and returned to production in October 1999. Oil and gas production with water injection rates were used from February 1995 to July 2010 in order to design the LS-SVM regression model. The shut-in period was omitted from this dataset.

As a first step, the field production and injection dataset are split into two subsets, namely, the “training dataset” and the “testing dataset”. The training set was enclosed the first 162 months (February 1995 to July 2009). This set was used to train the model and tune the model parameters $\gamma$ and $\sigma^2$. The testing set was enclosed the last 12 months (August 2009 to July 2010).

10.3 LS-SVM Models

The selection of input features for the LS-SVM is the key problem in this approach. In other words, the first step in any modeling design procedure is selecting the most
impact input parameters and at the same time reducing the dimension of this input space. Thus, our task is to find a set of inputs that are strongly correlated to the output. Therefore, based on an engineer’s judgment, parameters which are believed to have an impact on the oil production are selected. Every reservoir, however, is unique because of the myriad of geological and fluid dynamic characteristics and thus, the production can vary drastically from one reservoir to another. These variations make it difficult to simply predict the amount of oil it will require to produce from a particular reservoir. The second challenge in modeling is to obtain good model performance. Two LS-SVM regression parameters must be chosen carefully. Choosing the parameters is usually done by minimizing the estimated generalization error. The generalization error is a function that measures the generalization ability of the constructed models. The above mentioned field data were used to create two LS-SVM models with which to predict a field oil production rate under water injection. Thereinafter, the best forecasting performance will be chosen to predict the performance of oil production under water injection reservoirs and to find the most economical scenario of water injection to maximize ultimate oil recovery. The two models employed Radial Basis Function (RBF) as a kernel function. The regularization parameter ($\gamma$) and the RBF kernel parameter ($\sigma^2$) are tuned with two-step grid search on the basis of 10-fold cross-validation on the training set.

A two-step grid search technique with 10-fold cross validation on the training dataset was employed to obtain the optimal combination of ($\gamma$, $\sigma^2$) for both LS-SVM models. Grid search is a minimization procedure based upon an exhaustive search in a limited range. During each iteration of this algorithm, 10 points are left and fit a model on the other data points. The performance of the model is estimated based on the 10 points left
out. This procedure is repeated for each 10 points. The first step grid search was for a crude search with a large step size, presented in the form of (●), and the second step for the specified search was a small step size, presented in the form of (×). The optimal search area is determined by an error contour line. Following the grid search process, the optimal combination of $\gamma$ and $\sigma^2$ would be achieved for the LS-SVM models. Finally, the different estimates of the performance are combined. For each combination of $\gamma$ and $\sigma^2$ parameters, MSE was calculated and the optimum parameters were selected when they produced smaller MSE.

10.4 LS-SVM Model 1, (M1)

This model has six inputs (oil and gas production rates at time $t$, water injection rate at time $t$, producers at time $t$, producers at time $t+1$, water injection rate at time $t+1$) and one output (oil production rate at time $t+1$). The training data set was used to tune the model parameters while the testing dataset used to analyze the forecasting model capability.

The optimal hyper-parameters combination is determined through two-step grid search with 10-fold cross-validation on training dataset. The optimal hyper-parameters $\gamma$ and $\sigma^2$ are 34.1142 and 67.3738 respectively. The parameter combination ($\gamma$, $\sigma^2$) was tuned simultaneously in a 10×10 grid. Figure 10.1 shows the contour plot of optimization the LS-SVM parameters. The grids ‘●’ in the first step is 10×10, and the searching step in the first step is large. The natural logarithms of regularization parameter $\gamma$ and RBF kernel parameter $\sigma^2$ were tuned simultaneously. The optimal search area is determined by the error contour line. The grids ‘×’ in the second step is 10 × 10, and the searching steps in the second step are smaller. The optimal search area is determined based on the first
step. The optimal pair of \((\gamma, \sigma^2)\) for LS-SVM oil production forecasting model is indicated by red square in the Figure 10.1.

After the LS-SVM model parameters are tuned, the model was developed and the forecasting performance is evaluated. The predicted results of the optimal LS-SVM regression for training dataset are shown in Figure 10.2 and 10.3. The history matching is presented in Figure 10.2. It can be conclude that there is a good match between the modeled and field oil flow rates. In addition, the cross plot (Figure 10.3) illustrates the agreement between the field and modeled oil production rates. It shows that the regression coefficient \((R^2)\) is 89.05% mean absolute relative error 8.93%. The statistical error analysis for the reservoir oil forecasting capability, as obtained from the LS-SVM model, is summarized in Table 10.1. As can be seen from the proposed model is statistically stable and fitted the data well.
Figure 10.1 - Contour Plot of the Optimization Error for LS-SVM Regression M1 when Optimizing the Parameters. The red square indicates the selected optimal Settings of $\gamma = \exp(3.53)$ & $\sigma^2 = \exp(4.21)$
Figure 10.2 – History Matching of LS-SVM M1
Figure 10.3 - Cross Plot of Modeled Oil Production Rates vs. Field Values for M1

Scatter plot of LS-SVM output vs. reservoir oil production

LS-SVM $^{RBF}$

$\gamma = 34.1142, \sigma = 67.3738$

$R^2 = 89.05\%$

Training dataset  $y = x$ line
10.4.1 Prediction Performance of M1

To assess the prediction accuracy, the proposed model utilized to undertake forecasting tasks. The last 12 months of field production period from August 2009 to July 2010 (testing dataset) is used as the input of the trained LS-SVM regression model and the perfection is carried out. The flowchart of load forecasting is shown in Figure 10.4. The correlation between the predicted oil production rates and the field values is illustrated in Figure 10.5. The plot explains the performance of the model prediction process. It provides forecasting oil production rates with a RMSE of 404.78 BOPD and mean ARE of 11.58.214%. The statistical analysis of the error for the predicted capability on testing dataset is summarized in Table 10.1.

10.5 LS-SVM Model 2, (M2)

In an attempt to improve the effectiveness of the model prediction, different input features subset was selected, as a now prediction LS-SVM model (M2). The input features of M2 are oil production rate at time \( t-1 \) and \( t \), gas production rate at time \( t \), water injection rate at time \( t \) and \( t+1 \), producers at time \( t \) and \( t+1 \). The target is oil production at time \( t+1 \). Thus, seven input features of LS-SVM with RBF kernel function used in designing this model.
Figure 10.4 - Flow Chart of Load Forecasting
Figure 10.5 – Cross Plot of Predicted Oil Production Rates vs. Field Values for M1
A two-step grid search technique with 10-fold cross validation was employed to obtain the model parameters ($\gamma$, $\sigma^2$). The result of the optimal search area is determined by error contour line in Figure 10.6. After the process of grid search, the optimal combination of $\gamma$ and $\sigma^2$ would be achieved for the LS-SVM models. The red square in the Figure 10.6 indicates the selected optimal combination of $\gamma$ and $\sigma^2$. The optimal pair of ($\gamma$, $\sigma^2$) was found at the value of 1341.4872 and 252.5918 respectively.

In order to assess the relative robustness of this model, the calculated oil production rates of training dataset were plotted against the field values in Figure 10.7. It shows the modeled oil flow rates are matching the field data. Figure 10.8 illustrated the correlation between the LS-SVM model output and the field oil production rates. The model provides oil rates with a correlation coefficient $R^2$ of 90.56%, a RMSE of 402.08 BOPD and a mean ARE of 8.22%. The statistical accuracy analysis of the proposed model is summarized in Table 10.1. It shows stable model and fit the data well.
Figure 10.6 - Contour Plot of the Optimization Error for M2 Model when Optimizing the Parameters. The red square indicates the selected optimal settings of $\gamma = \exp(7.201)$ & $\sigma^2 = \exp(5.532)$.
10.5.1 Prediction Performance of M2

The model’s prediction performance is evaluated by employing the testing dataset. This dataset was used as input for the trained LS-SVM model then the predicted oil production rates matched up to the field values. Figure 10.9 illustrated the forecasting accuracy performance. The model provides forecasting values of oil production rates with a mean ARE of 10.79% and a RMSE of 380.17.
Figure 10.7 – History Matching of M2
Figure 10.8 – Correlation between Modeled and Field Oil Flow Rates for M2
Figure 10.9 – Forecasting Oil Production against Field Values for M2 Model
### Statistical Error Analysis of M1 and M2 LS-SVM Models

<table>
<thead>
<tr>
<th>Statistical Error Analysis</th>
<th>LS-SVM M1 6 inputs</th>
<th>LS-SVM M2 7 inputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2, % )</td>
<td>Training 89.05</td>
<td>Forecasting 20.48</td>
</tr>
<tr>
<td>( \text{MSE, (BOPD)}^2 )</td>
<td>188150 163520</td>
<td>161670 144530</td>
</tr>
<tr>
<td>( \text{RMSE, BOPD} )</td>
<td>433.76 404.78</td>
<td>402.08 380.17</td>
</tr>
<tr>
<td>( \text{SD, BOPD} )</td>
<td>435.00 355.03</td>
<td>403.34 347.36</td>
</tr>
<tr>
<td>Mean AE, BOPD</td>
<td>326.76 358.28</td>
<td>299.3 333.43</td>
</tr>
<tr>
<td>Max AE, BOPD</td>
<td>1543 631.57</td>
<td>1601 635.2</td>
</tr>
<tr>
<td>Min AE, BOPD</td>
<td>3.582 49.8</td>
<td>1.38 20.29</td>
</tr>
<tr>
<td>Mean ARE, %</td>
<td>8.93 11.58</td>
<td>8.22 10.79</td>
</tr>
<tr>
<td>Max ARE, %</td>
<td>49.25 23.06</td>
<td>43.72 23.19</td>
</tr>
<tr>
<td>Min ARE, %</td>
<td>0.078 1.255</td>
<td>0.031 0.511</td>
</tr>
</tbody>
</table>
Figure 10.10 – Illustrates the Comparison of ARE between the Two Models in Forecasting Data Domain
10.6 Results and Discussions

Figure 10.10 illustrates the comparison of ARE between the two models in forecasting data domain. The final results of forecasting precision among these models can be judged based on MSE, RMSE and MARE. The lowest MSE, RMSE and MARE are found with M2 in training and forecasting stages. The model M2 provides prediction oil flow rates with a MSE of 144530 BOPD, a RMSE of 380 BOPD and a MARE of 10.79%. Table 10.1 summarizes the assessment results for both models.

10.7 Conclusions

This study applied a novel approach to predict oil production in oil fields under water injection. Because the selection of input features for the LS-SVM is the key problem in this approach, two LS-SVM models with different input features were compared. The first model had six inputs while there were seven inputs in the second model. The case study involving a small oil field under water injection located in the Sirte Basin of Libya demonstrated that the input features of Model 2 are superior to those of the Model 1. The LS-SVM with the inputs of oil production rate at time \( t-1 \) and \( t \), gas production rate at time \( t \), water injection rate at time \( t \) and \( t+1 \), and producers at time \( t \) and \( t+1 \) is robust in oil production prediction.

In summary, this approach provides rapid forecasting of oil production performance of oilfield under water injection with reasonable accuracy. The approach can be instructed by utilizing the historical production and injection data. It is suitably employable to find the most economical injection rates by applying various water
injection scenarios. Meanwhile, it could be a new window for fast simulators. It has reasonable accuracy, requires little data and is able to forecast quickly.
CHAPTER XI

CONCLUSIONS AND RECOMMENDATIONS

11.1 Conclusions

This thesis puts foreword there different function approximation tools with different methodologies to address the following three ongoing challenges in reservoir engineering.

1. Estimating reservoir fluid properties in the absence of PVT analysis

2. Forecasting, interpolating and estimating the current value of average reservoir pressure.

3. Oilfield production prediction

The results of the three used functions approximations provided predicted values with acceptable magnitude of errors by the oil industry. Based on the results of this study, the following conclusions can be obtained:

For Estimating Reservoir Fluid Properties the Results Showed:

- A review of the literature shows little attention has been paid to the prediction of reservoir fluid properties in the absence of PVT analysis. Currently, not one specific model has been published to directly estimate the bubble point pressure in the absence of PVT analysis and, presently, there is just one published empirical correlation (Labedi 1990) to estimate bubble point oil FVF and one correlation (Valko and McCain 2002) to estimate stock-tank vent gas oil ratio directly in the absence of PVT analysis.
• The majority of the available PVT Models are used to estimate the reservoir oil properties at various temperatures rather than reservoir temperature in order to design surface operation equipment and to study reservoir inflow performance. However, due to the fact the models require prior knowledge of parameters such as solution GOR and specific gas gravity, it is difficult to apply them in the absence of PVT analysis. Since these parameters are rarely measured in the field, this study proposed correlations which can be applied straightforwardly in the absence of PVT analysis. Thus, there is no need for additional correlations. The only required input parameters are separator GOR, separator pressure, stock-tank oil gravity and reservoir temperature.

• Four novel empirical correlations, four ANN models and four LS-SVM models to estimate $P_b$, $B_{ob}$, $R_{sob}$ and $R_{ST}$ in the absence of PVT analysis, were proposed as a function of readily available field data. All the proposed PVT models can be applied straightforwardly in the absence of PVT analysis. No further correlations or experimental measurements are required.

• The proposed PVT MRA, ANN and LS-SVM models were developed based on 118 reservoir fluid studies collected from various Libyan oil fields in the Sirte Basin.

• The estimated $R_{sob}$ and $R_{ST}$ values from the proposed models can be used as a basic input variable in many PVT imperial correlations in order to estimate other fluid properties such as the $P_b$ and $B_{ob}$.

• The bubble point and oil FVF models have four inputs while bubble point GOR and stock-tank vent GOR models have three inputs.
• The results show the proposed PVT models provided high accuracy on testing dataset. This dataset was not viewed by the models during the building process.

• Since the representativeness of a PVT study greatly depends upon sampling conditions, the first and most important operation, before running a complete reservoir fluid study, is to check the validity of the samples. The bottom-hole sample, used for the PVT study, is selected according to the results obtained during the verification of sample validity. It is worth mentioning, the proposed bubble point pressure models can be used to select the most representative sample with which to run a PVT analysis.

• An additional and important application of the proposed stock-tank vent gas/oil ratio correlation is to estimate the stock-tank vent gas flow rate.

**For Forecasting, Interpolating and Estimating the Current Values of Average Reservoir Pressure it was observed that:**

• Since the average reservoir pressure is measured during a pressure build up test while the well is shut-in, a significant economic effect occurs in producing wells during the entire build up test. Moreover, average reservoir pressure is measured periodically, resulting in a loss of production. This study introduces three function approximation techniques as new alternative tools to predict and interpolate average reservoir pressure without shutting in the producing wells. BP-ANN, BP-ANN-GA and LS-SVM models are performed to interpolate the missing values, forecast and to predict current average reservoir pressure.
This case study is based on historical production and reservoir pressure data for an oilfield located in Sirte Basin, Libya. The reservoir contained a set of 60 oil wells. There are just 49 measurements of average reservoir pressure from the period of March 1970 to July 2010 (more than 40 years of production).

The three proposed models have the ability to predict and interpolate average reservoir pressure accurately by employing oil, water and gas production rates and a number of producers are used as inputs for the three proposed models.

For Oil Production Prediction the Results Demonstrated that:

- This study shows the ANN and LS-SVM approximation techniques provide a new alternative method with which to predict oil reservoir production based on historical production data.

- A case study involving a small oil field under water injection located in the Sirte Basin of Libya demonstrates the proposed approach can be used to predict oil reservoir performance, and can serve as a practical and robust tool with regard to reservoir management.

- Since the selection of input features is the key problem in this approach, the following input features of oil production rate at time \( t-1 \) and \( t \), gas production rate at time \( t \), water injection rate at time \( t \) and \( t+1 \), and producers at time \( t \) and \( t+1 \) provides rapid forecasting of oil production performance of oilfield under water injection with reasonable accuracy.

- This approach is suitably employable to find the most economical injection rates by applying various water injection scenarios. Meanwhile, it could be a new
window for fast simulators. It has reasonable accuracy, requires little data and is able to forecast quickly.

In general, ANN and LS-SVM approximation techniques can serve as a practical, cost-effective and robust tool for oilfield production management.

11.2 Recommendations

The following recommendations are derived from this study:

1. Regarding the Estimating Reservoir Fluid Properties Study:
   - Since the proposed PVT models based on two-stage separation test, these models may be re-studied by using Three-stage separation test.
   - Because the proposed PVT models were developed specifically for Libyan oil reservoirs located in Sirte Basin, there testing accuracy are needed to consider for use in other Libyan Basins and/or other reservoirs around the world.

2. Regarding Forecasting, Interpolating and Estimating the Current Values of Average Reservoir Pressure Part:
   - Additional testing on other oil reservoirs is required to check the robustness of the approach and extend it for gas reservoirs.
   - In order to address the reservoir pressure distribution within the reservoir, predict the average reservoir pressure by utilizing historical data for individual oil well instead of entire reservoir is recommended.

3. Regarding oil production prediction Study
• The developed models for oil production prediction were successfully applied. Further studies are needed for other oilfield to test the robustness of the proposed approach and extend it for gas reservoirs.

• Incorporate the optimizing techniques with this approach to help petroleum engineers find out an optimal water injection rates and/or to modify the existing water injection policy to increase oil production and decrease reservoir damage

• Predict the oil production behaviour of individual producers instead of entire reservoir.
Appendix A

Correlations and Regression Modeling

A.1 Introduction
In any system in which variable quantities change, it is of interest to examine the effects that some variables exert or appear to exert on others. Often there exists a functional relationship which is too complicated to describe in simple terms. In this case, it may approximate to this functional relationship by some simple mathematical function, which contains the appropriate variables and which approximates to the true function over some limited ranges of the variables involved. Variables may be related by some sort of mathematical equation, while the equation might be physically meaningless, and it might be a linear or nonlinear model. This model can be generated by what is called “regression analysis”.

Regression analysis is a theoretically simple method with which to investigate functional relationships among variables. It is used to explain or model the relationship between a single variable Y, called “the response”, output or dependent variable, and one or more predictor, input, independent or explanatory variable, $X_1$, ..........., $X_p$, where p is the number of variables. When p is equal to 1, it is called “simple regression”, (building a model with two variables, Y and X). When p is greater than 1, it is called “multiple regression” or sometimes “multivariate regression”. Regression analysis is one technique to determine the model of the process $Y = f(X_p)$. When there is more than one Y, it is called “multivariate multiple regression” which is not to be covered in this work. However, regression analyses have several possible objectives including:

1. Prediction of future observations
2. Assessment of the effect of, or relationship between, explanatory variables on the response
3. A general description of data structure

In fact, there may be a simple regression between variables in most engineering processes. However, this is the exception rather than the rule. Often there exists a function relation which is too complicated to describe in simple forms. The model may be linear or nonlinear. When a model is said to be linear or nonlinear, this refers to linearity or non linearity in the parameters. The general form of the multiple linear models with n predictor variables is

$$Y = \beta_0 + \beta_1 X + \epsilon$$  \hspace{1cm} (A.1)

where $\beta_0$ and $\beta_1$, are constants called “the model regression coefficients” (or regression parameters), and $\epsilon$ is the random error or residual term. The random error $\epsilon_i$ is the term which accounts for the variation of the $i$th response variable $Y_i$ away from the linear
predictor $\beta_0 + \beta_1 X_i$ at the point $X_i$. In other words, residual term is an estimate of experimental errors obtained by subtracting the observed responses from the predicted responses. The predicted response is calculated from the model. That is,

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 x_i, \quad i = 1, \ldots, n$$ (A.2)

Before discussing the multiple regression models in detail, a short description of a simple linear regression model is given.

**A.2 Simple Linear Regression Model**

A simple linear regression model is defined as the straight-line relationship between a response variable $Y$ and a single predictor variable $X$. As mentioned previously, regression involves one predictor variable called a “multiple regression”. When a linear relationship exists between the explanatory $X$ variable and the response $Y$ variable, the least-squares regression line is used to fit the data and make the predictions. The simple linear regression for the probabilistic model could be described as per Equation A.1. It is assumed that in the range of the observations studied, the linear equation (A.1) provides an acceptable approximation to the true relation between $Y$ and $X$. In other words, $Y$ is approximately a linear function of $X$, and the residual, and $\varepsilon$ measures the discrepancy in that approximation. The coefficient $\beta_1$, called “the slope”, may be interpreted as the change in $Y$ for the unit change in $X$. The coefficient, $\beta_0$ called “the constant coefficient” or “intercept”, is the predicted value of $Y$ when $X = 0$. One very general linear form for the model would be

$$Y = f(X_1, X_2, X_3) + \varepsilon$$ (A.3)

where $f$ is an unknown function and $\varepsilon$ is the error in this representation which is an additive in this instance. As there is usually not enough data to attempt to estimate $f$ directly, it is usually assumed that it has some more restricted form, perhaps linear as in

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$ (A.4)

**A.3 Correlation and Regression Modeling in the Oil and Gas Industry**

Multiple regression analysis is an essential tool to analysis data. In the oil industry, regression analysis is one of the most widely used statistical tools because it provides a conceptually simple method with which to investigate functional relationships among variables. A very long list of examples of regression analysis applications exists in the petroleum industry. A partial list would include its application to reservoir & production engineering, drilling operations and exploration.

In production engineering, regression analysis was applied to develop models in multiphase flow in production pipes encountered in oil and gas operations. An example is Beggs and Brill (1973), Gray (1975) and Farshad et al. (2001) used regression analysis to correlate multiphase flow in uphill, downhill, horizontal, inclined and vertical flows. Another study applied multiple regression analysis to describe the critical and sub-critical flow behaviour of oil and gas through wellhead chokes (Fortunati, 1972 and Guo et al.,
Moreover, Salama (2000) and Yi et al. (2005) used regression analysis to produce models for sand production management.

In drilling engineering, regression analysis is applied to establish models to optimize drilling operations, such as Al-Betairi et al. (1988) and to optimize drilling costs, Kaiser (2007). Others established regression models for lifting capacity and penetration rate, as for example, Yidan, et al. (1995), Bilgesu, et al. (1997) and Altindag (2004).


Many studies addressed regression analysis in different areas of reservoir engineering. Several studies used regression analysis to model the relative permeability, as in App and Mohanty (2002). They presented a regression technique for history matching relative permeability in near-critical gas condensate systems. Further applications of regression analysis in reservoir engineering were employed in fluid flow in porous media. Some studies used regression analysis to generate two and three-phase permeability models such as Lin(1982), Fulcher et al (1985), Watson et al. (1988), Tan(1995), Grimstad et al (1997), Mohamed and Koederitz (2001) and Machad et al. (2008). Many attempts have been made to establish a relationship between porosity and permeability from core analysis data and/or log data by using multiple regression analysis with examples being Swanson (1981), Wendt et al. (1986), Schere (1987), Motiee (1991), Pittman (1992), Yao and Holditch (1993), Balan et al. (1995), Mohaghegh et al. (1995) and Rezaee et al. (2006).

However, a very long list of regression models to develop PVT properties of reservoir fluids have been offered in the petroleum engineering literature. A comprehensive literature review of PVT correlation will be addressed in Chapters 2 and 3.

A.4 Correlation and Regression Modeling Building Procedure
The general purpose of multiple regression analysis is to learn more about the relationship between several independent or predictor variables and a dependent or criterion variable. The multiple linear regression model is an extension of a simple linear regression model which incorporates two or more explanatory variables in a prediction equation for a response variable. It is necessary to be very careful when using regression analysis. Regression analysis is not just intended to take data, fit a model, and then evaluate the fit using statistics such $R^2$ and be finished. The regression analysis approach is much broader. It should view regression analysis as a set of data analytic techniques that examine the interrelationships among a given set of variables. The emphasis is not on
formal statistical tests and probability calculations. It should be used as an informal analysis directed towards uncovering patterns in the data. Before discussing the procedure of regression model building, regression assumptions will be addressed.

A.5 Multiple Linear Regression Model Assumptions
There are four principal assumptions which justify the use of linear regression models for the purposes of prediction:

1) **Linearity**: Linear stochastic relationship

\[ Y_i = b_0 + b_1X_{i1} + b_2X_{i2} + \cdots + b_kX_{ik} + \varepsilon_i \]

for \( i = 1, 2, \ldots, n \)

2) **Independence of Error Terms**: Residuals are independent of one another and have a constant variance versus the dependent and any independent variables.

3) **Normality**: Residuals have a normal distribution with a zero mean.

4) **Multicollinearity**: There is no exact linear relationship among independent variables. They are linearly independent.

Generally, residuals are estimates of experimental error obtained by subtracting the observed responses from the predicted responses. The predicted response is calculated from the chosen model. Examining residuals is a key part of all statistical modeling. Close observation discloses whether the assumptions are reasonable and if the model choice is appropriate.

A.5.1 Linearity - Assumption (1)
Multiple linear regression analysis requires that the relationship between the independent and dependent variables be linear. In other words, in a multiple linear regression model, it is assumed that each of the explanatory variables \( X_1, X_2, \ldots, X_p \) affects the mean of the response in a linear way. If a linear model is fitted to data which are nonlinearly related, the predictions are likely to be seriously in error, especially when they are extrapolated beyond the range of the sample data.

**How to detect**: nonlinearity is usually most evident in a plot of observed versus predicted values or a plot of residuals versus predicted values, which are a part of standard regression output. The points should be symmetrically distributed around a diagonal line in the former plot or a horizontal line in the latter plot. Careful observation will disclose possible evidence of a "bowed" pattern, indicating that the model makes systematic errors whenever it is making unusually large or small predictions.

**How to fix**: consider applying a nonlinear transformation to the dependent and/or independent variables if a transformation seems appropriate. For example, if the data are strictly positive, a log transformation may be feasible. Another possibility to consider is adding another independent variable. For example, if there is regressed \( Y \) on \( X \), and the graph of residuals versus predicted suggest a parabolic curve, then, it may make sense to regress \( Y \) on both \( X \) and \( X^2 \). Generally, if the multiple linear model is not adequate, a transformation model may be adequate. It is often more fruitful to investigate the effects.
produced by transformations of the independent variables, or by transformation of the independent variable, or by both. There are many possible transformations, and different transformations could be undertaken within the same model.

The purpose of using transformation techniques is to be able to use a regression model of a simple form or to reduce a complicated model rather than a more complicated model in the original variables. However, a nonlinear model can be divided into two types which can be called “intrinsically linear” and “intrinsically nonlinear” models. If a model is intrinsically linear, it can be expressed by a suitable transformation of the variables in standard linear model form (Equation A.1). If a nonlinear model cannot be expressed in the form of Equation A.1, then it is intrinsically nonlinear. Some examples for intrinsically linear models are:

1) **Power Model**: \( y = a x_1^b x_2^c x_3^d \epsilon \) By taking the logarithms of both sides, the above equation converts into a linear form of Equation A.1.

2) **Exponential Model**: \( y = \exp(b_0 + b_1x_1 + b_2x_2) \epsilon \) or \( y = \exp(b_0 + b_1x_1) b_2x_2 \epsilon \). By taking the natural logarithms of both sides, the model converts into a linear form.

3) **Reciprocal Model**: \( y = 1 / (b_0 + b_1x_1 + b_2x_2 + \epsilon) \) To convert the model into a linear form, the reciprocals of both sides are taken.

4) A more complicated exponential model: \( y = 1 / (1+ \exp(b_0 + b_1x_1 + b_2x_2 + \epsilon)) \) To convert this model into linear form, reciprocals are taken, 1 is subtracted and the natural logarithm of both sides is then taken.

**A5.2 Independence of Error terms - Assumption (2)**

A regression model requires independence of error terms. A residual plot can be used to check this assumption.

*How to detect*: plots of residual versus dependent and versus all independent variables are observed. Note evidence of residuals that are becoming larger (i.e., more spread-out). If the residuals seem to increase or decrease in average magnitude with the fitted values, it is an indication that the variance of the residuals is not constant. If the points in the plot lie on a curve around zero, rather than fluctuating randomly, it is an indication that the assumption (the random errors \( \epsilon_i \) have zero mean) is broken. If assumption (ii) is satisfied, the residuals can be expected to vary randomly around zero and the spread of the residuals would be expected to be approximately the same throughout the plot. Figure A.1 addresses different features in the residual plots.
Figure A.1 - Different Features in Residual Plots, Where the Residuals (y-axis) Plotted versus Fitted Values (x-axis). After Draper and Smith (1998)
Figure A.1 (a) shows a residual plot with no systematic pattern. It appears the assumption (ii) is satisfied for the data associated with this residual plot. In Figure A.1 (b), there is a clear curved pattern: Assumption (2) may be broken. In Figure A.1 (c), the random variation of the residuals increases as the fitted values increase. This pattern indicates the variance is not constant. Finally, in Figure A.1 (d), most of the residuals are randomly scattered around 0, but one observation has produced a residual which is much larger than any of the other residuals. The point may be an outlier.

*How to fix:* if the plot of the residuals against dependent variable shows a conical pattern as Figure A.1 (c), a transformation is attempted upon the dependent variable before performing a regression analysis. If it is a trend pattern, as in Figure A.2 (a, c and d), there is a means error in the analysis caused by wrongly omitting a constant term, \( b_0 \) in the model. If the plot takes a curve pattern as in Figure A.1 (b) or Figure A.2 (b), the model may require an extra term or there may be a need for a transformation on \( Y \).

A standardized residual can be plotted instead of a regular residual. However, standardized Residual Plots are used in order to put the residual plot on a meaningful scale. The pattern for a standardized residual plot will be identical to the pattern in the regular residual plot. The only difference is the scale on the y-axis. This scale allows for easy detection of potential outliers. A standardized residual can be calculated as a regular residual divided by a standard error of estimate.

**A.5.2.1 Standard Error of Estimate**

An alternative index of goodness of fit is a standard error of estimate. It is a command measure of the accuracy of the predictions. Usually, there is a desire to know how accurate predictions can be in making predictions of the \( y \) value for any given value of independent variables based on the regression analysis results. The validity of our estimate will depend upon how the errors of prediction will be obtained from our regression analysis, particularly on the average error of prediction. With so many single residuals, it is difficult to look at every one individually and come to a conclusion for all the data. It is desirable to have a single number that reflects all the residuals. If all deviations of observation from the regression are added, zero is obtained. To avoid that hurdle, the error \((y_i - \hat{y}_i)\) is squared to obtain the sum of square of error (SSE).

To obtain the average deviation from the regression line, the square root of the \( SSE \) divided by \( n - p \) is used, where \( p \) is the number of predictors and \( n \) is the sample size. The square root is used because the residuals were squared and \( p \) was subtracted from \( n \) because \( p \) degrees of freedom were lost from using \( p \) independent variables. Thus, it is desirable to have the standard error of estimate as small as possible. However, standard error of estimate has an inverse relationship of \( R^2 \). As \( R^2 \) increases, standard error of estimate decreases, and vice versa. The standard error of estimate is usually computed by the majority of statistical packages.
Figure A.2 - Further Different Features in Residual Plots
A.5.3 Normality - Assumption (3)
To satisfy this assumption, we must answer the question “are the model residuals normally distributed”? The histogram and the normal probability plot can answer this question. The histogram and the normal probability plot are used to check whether or not it is reasonable to assume that the random errors, inherent in the process, have been drawn from a normal distribution. Sometimes the error distribution is “skewed” by the presence of a few large outliers. Figure A.3 illustrates the histogram plot of the residuals. It shows the residual term is almost normally distributed. Figure A.4 demonstrates some shapes of the normal probability plot.

How to detect: the best test for normally distributed errors is a normal probability plot of the residuals. If the distribution is normal, the points on this plot should fall close to the diagonal line. A bow-shaped pattern of deviations from the diagonal, as in Figure A.4 (a and d) indicates that the residuals have excessive skewness (i.e., they are not symmetrically distributed, with too many large errors in the same direction). An S-shaped pattern of deviations indicates that the residuals have excessive kurtosis (i.e., there are either too many or too few large errors in both directions) as in Figure A.4 (b). Figure A.4 (e) shows most points fall close to the diagonal line, concluding that the residual appears to be homoscedastic and more closely follows a normal distribution. In statistics, a sequence or a vector of random variables is homoscedastic if all random variables are in the sequence.

How to fix: violations of normality often arise due to:
- The distributions of the dependent and/or independent variables are themselves significantly non-normal, and/or
- The linearity assumption is violated. In such cases, a nonlinear transformation of variables might cure both problems.
Figure A.3 - Histogram of the Residuals. The Residuals Close Normal Distribution
Figure A.4 - Tests of the Normality. The Histogram and the Normal Probability Plot in (c and e) Show that the Error Terms are Normally Distributed.
A.5.3.1 Outliers, Leverage and Influential Points in Regression

Another use for residual analysis is to see if there are any outlier or leverage values in the data set. An outlier among the residuals is one that is greater than the rest in the absolute value. Leverage basically means that a single data point located well outside the bulk of the data (an outlier), has an over proportional effect on the resulting regression curve. Observations with large leverage values may exert a disproportionate influence upon a model and produce misleading results. However, the problem with the residual distribution is mainly due to one or two very large errors. To obtain a satisfactory model, it is important to take into account these particular points. Automatic rejection of outliers is not always a very wise procedure. Sometimes, the outlier is providing information which other data points cannot provide due the fact that it arises from an unusual combination of circumstances which may be of vital interest and requires further investigation rather than rejection. Such values should be scrutinized closely. Are they genuine, (i.e., not the result of data entry errors)? Are they explainable? Are similar events likely to occur again in the future, and how influential are they in the model-fitting results?

To determine the extent of influence, you can fit the model with and without those observations and compare the coefficients, $p$-values, $R^2$, and residual plots. If the model changes significantly when the influential observation is removed, then first determine whether the observation is a data entry or measurement error. If not, examine the model further to determine if an important term (e.g., interaction term) or variable has been omitted, or has incorrectly specified the model. More data may be required to resolve the issue.

However, the majority of statistical packages have a method with which to see how the data spread and scatter. The box plot, otherwise known as a “box-and-whisker plot” is a graphical representation of data that shows how the data are spread. The plot shows how the data are scattered within those ranges. The advantage of using the box plot is that when it is used for multiple variables, not only does it graphically show the variation between the variables but it also shows the data spread, data variations within the ranges and shows the obviousness of the presence of outliers. Figure A.5 illustrates the components of a box plot.

To determine the presence of outliers, the interquartile range (IQR) must be found first. The IQR measures the vertical distance of the box; it is the difference between the third quartile, $Q_3$ and the first quartile, $Q_1$ values ($IQR = Q_3 - Q_1$). An outlier is defined as any observation away from the closest quartile by more than $1.5 \times IQR$, $(Q_3 + 1.5(Q_3 - Q_1))$. An outlier is considered extreme when it is away from the closest quartile by more than $3 \times IQR$. This is illustrated in Figure A.5.
Figure A.5 - Box Plot

Outlier – an unusually large or small observation. Values beyond the whiskers are outliers.

By default, the top of the box is the third quartile (Q3) – 75% of the data values are less than or equal to this value.

By default, the bottom of the box is the first quartile (Q1) – 25% of the data values are less than or equal to this value.

Median – the middle of the data. Half of the observations are less than or equal to it.

By default, the upper whisker extends to the highest data value within the upper limit.

Upper limit = Q3 + 1.5 (Q3 – Q1)

By default, the lower whisker extends to the lowest value within the lower limit.

Lower limit = Q1 – 1.5 (Q3 – Q1)
A.5.4 Multicollinearity - Assumption (4)
Multicollinearity is a high degree of correlation (linear dependency) among several independent variables. It commonly occurs when a large number of independent variables are incorporated into a regression model. It is because some of them may measure the same concepts or phenomena. If the model fits the data well, the overall $R^2$ value will be high, and the corresponding P value will be low (a great fit is unlikely to be a coincidence). In addition to the overall P value, multiple regression also reports an individual P value for each independent variable. A low P value herein means that this particular independent variable significantly improves the fit of the model. It is calculated by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent variable is omitted. If the fit is much worse when that variable is omitted from the model, the P value will be low, thus stating that the variable has a significant impact upon the model. In some cases, multiple regression results may seem paradoxical. Even though the overall P value is very low, all the individual P values are high. This means that the model fits the data well, even though none of the X variables has a statistically significant impact on predicting Y. How is this possible? When two X variables are highly correlated, they both convey essentially the same information. In this case, neither may contribute significantly to the model after the other is included. But together they contribute significantly. If both variables were removed from the model, the fit would be much worse. So, the overall model fits the data well, but neither X variable makes a significant contribution when it is added to the last model. When this happens, the X variables are “collinear” and the results show “multicollinearity”.

Why is multicollinearity a problem? If the goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem. The predictions will still be accurate, and the overall $R^2$ (or adjusted $R^2$) quantifies how well the model predicts the Y values. If the goal is to understand how the various X variables impact Y, then multicollinearity is a major problem. One problem is that the individual P values can be misleading (a P value can be high, even though the variable is important). The second problem is that the confidence intervals on the regression coefficients will be very wide. The confidence intervals may even include zero, which means there is a non-confidence as to whether an increase in the X value is associated with an increase, or a decrease in Y. Because the confidence intervals are so wide, excluding a subject (or adding a new one) can change the coefficients dramatically and may even change their signs.

Symptoms of multicollinearity may be observed in situations such as
- small changes in the data produce wide swings in the parameter estimates
- coefficients may have very high standard errors and low significance levels even though they are jointly significant and the $R^2$ for the regression is quite high
- coefficients may have the “wrong” sign or an implausible magnitude (Greene 2000)

How to detect: multicollinearity can be detected in the following ways:
1. Examine the correlations and the relations between independent variables to detect a high level of association. High bivariate correlations are easy to spot by running
correlations among the variables. If high bivariate correlations are present, one can be deleted from the two variables. However, this may not always be sufficient.

2. Regression coefficients will change dramatically according to whether other variables are included or excluded from the model. Various approaches can be taken by adding and then removing variables from the regression model. This can be easily accomplished by running a stepwise regression.

3. Examine the Variance Inflation Factor (VIF). VIF measures the impact of multicollinearity among the variables in a regression model. VIF is always greater than or equal to 1. There is no formal VIF value to determine the presence of multicollinearity. Values of VIF that exceed 10 are often regarded as indicating multicollinearity.

In most statistics packages, the results are shown for both the \( R^2 \) value and the Variance Inflation Factor (VIF).

**How to fix:** The best solution is to understand the cause of multicollinearity and remove it. Multicollinearity occurs because two (or more) variables are related (they basically measure the same). If one of the variables does not appear to be logically essential to the model, removing it may reduce or eliminate multicollinearity. Or, perhaps a method can be found to combine the variables. The two most common ways to display the multicollinearity are correlation matrix and stepwise regression.

A.5.4.1 Coefficient of Determination \( R^2 \)

Coefficient of determination, \( R^2 \) is the proportion of the variation in the dependent variable explained by the regression model, and is a measure of the goodness of fit of the model. It is calculated as follows:

\[
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}
\]  

(A.5)

A fundamental equality, \( SST = SSR + SSE \), where \( SST \) stands for the total sum of squared deviations in \( Y \) from its mean, \( SSR \) denotes the sum of squares due to regression, and \( SSE \) represents the sum of squared residuals. Those quantities may be computed from

\[
SST = \sum (y_i - \bar{y})^2
\]

\[
SSR = \sum (\hat{y}_i - \bar{y})^2
\]  

(A.6)

\[
SSE = \sum (y_i - \hat{y}_i)^2
\]
Usually, the values are computed and supplied as part of the regression output by statistical packages named “the ANOVA table”. ANOVA stands for analysis of variance. However, because SSE is less than or equal to SST, the value of \( R^2 \) will always lie between 0.0 and 1.0 (or, in percentage, between 0% and 100%). To achieve \( R^2 \) near 100%, we must minimize the sum of squares of the errors. The equation is often asked, “What is a good value of \( R^2 \)?” That depends upon the area of application. In some applications, variables tend to be more weakly correlated and there considerable noise. Lower values for \( R^2 \) could be expected in these cases. In another area of application, a value of \( R^2 \) of 60% might be considered good. In physics and engineering, where most data comes from closely controlled experiments, a much higher \( R^2 \)'s is expected and a value of 60% would be considered low. Generally, some experience with the particular area is necessary to judge the \( R^2 \)'s well. Table 1.1 shows the range of \( R^2 \) in various PVT correlations to estimate a solution gas-oil ratio, oil formation volume factor and undersaturated oil compressibility.

A.5.4.2 Adjusted \( R^2 \)

However, \( R^2 \) tends to somewhat over-estimate the success of the model when applied to the real world, so an adjusted \( R^2 \) value is calculated which takes into account the number of variables in the model and the number of observations (participants) the model is based upon. This adjusted \( R^2 \) value gives the most useful measure regarding the success of our model. If, for example an adjusted \( R^2 \) value of 85% is given, it can be said that the model has accounted for 85% of the variance in the criterion variable.

Does a large \( R^2 \) mean a good model? \( R^2 \) automatically increases as more variables are added to a multiple regression model, even if the additions are insignificant. The prediction always appears to become superior as more independent variables are added. This is a false indication and does not actually increase confidence in the predictions. When there are similar numbers of variables and data samples, multiple R will always be high. Adjusted R2 is a modified R2 where the addition of independent variables does not automatically increase R. It is an improved measure of the overall fit of a multiple regression model.

\[
adjR^2 = 1 - \left[ (1 - R^2) \frac{n - 1}{n - p - 1} \right] \quad (A.7)
\]

where,

\( n \) = number of samples
\( p \) = number of independent variables

Adjusted \( R^2 \) is sometimes used to compare models having different numbers of predictor variables. In comparing the goodness of fit of models with different numbers of explanatory variables, \( R^2 \) tries to “adjust” for the unequal number of variables in the different models. Usually, statistical packages supplied \( R^2 \) and adj\( R^2 \) as part of the regression output, as shown in Figure A.6.
Figure A.6 - ANOVA Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>9.1773</td>
<td>9.1773</td>
<td>9902.52</td>
<td>0.000</td>
</tr>
<tr>
<td>Residual Error</td>
<td>427</td>
<td>0.3957</td>
<td>0.0009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>339</td>
<td>0.3401</td>
<td>0.0010</td>
<td>1.59</td>
<td>0.005</td>
</tr>
<tr>
<td>Pure Error</td>
<td>88</td>
<td>0.0556</td>
<td>0.0006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>428</td>
<td>9.5730</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **S = 0.0304428**
- **R-Sq = 95.9%**
- **R-Sq(adj) = 95.9%**

**Analysis of Variance**

- Standard error of estimate
- Coefficient of determination, $R^2$
- Adjusted value of $R^2$, the proportion of the variance of response explained by predictors

- F statistics & p-value for hypothesis that all the regression coefficient except $\beta_0$ are zeros
A.5.4.3 Correlation Coefficient among Variables, Cor($X_1, X_2$)

The correlation coefficient is one of the most widely used coefficients to study the relation among variables. In simple linear regression, $R^2$ is equal to the square of the correlation coefficient between the response variable $Y$ and the predictor $X$ or to the square of the correlation coefficient between the response variable $Y$ and the fitted values $Y$. The correlation coefficient between $Y$ and $X$, Cor($Y, X$) is given by

$$
Cor(Y, X) = \frac{\sum (\hat{y}_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (x_i - \bar{x})^2}}
$$

Thus, the correlation coefficient is symmetric, that is $Cor(Y, X) = Cor(X, Y)$. Furthermore, $Cor(Y, X)$ satisfies $0.1 \leq Cor(Y, X) \leq 1.0$. Values near zero mean no linear correlation and values near ± 1 mean a very strong linear correlation. The negative sign means that the two variables are inversely related. That is, as one variable increases the other variable decreases. The properties make the correlation coefficient a useful quality with which to measure both the direction and the strength of the relation between $X_1$, $X_2$ and $X_n$. Figure A.7 supplies different types of correlation coefficients.

A.6 Stepwise Regression (Variables Selection):

There are different ways that the relative contribution of each predictor variable can be assessed. As each variable is entered into the model, its contribution is assessed. If adding the variable does not significantly increase the predictive power of the model then the variable is dropped. In “statistical” methods, the order in which the predictor variables are entered into or taken out of the model is determined according to the strength of their correlation with the criterion variable. Actually there are two versions of this method, called “stepwise selection” via “forward” and “backward” selection. Stepwise regression is the most sophisticated of the statistical methods. Each variable is entered in sequence and its value assessed. If adding the variable contributes to the model then it is retained, but all other variables in the model are then re-tested to see if they are still contributing to the success of the model. If they no longer contribute significantly they are removed. Thus, this method should ensure that the smallest possible set of predictor variables is included in the model.
Figure A.7 - Different Features in Correlation Coefficient
In forward selection, the variables enter into the model one at a time in an order determined by the strength of their correlation with the criterion variable. The effect of adding each is assessed as it is entered, and the variables that do not significantly add to the success of the model are excluded. However, in backward selection, all the predictor variables are entered into the model. The weakest predictor variable is then removed and the regression re-calculated. If this significantly weakens the model, then the predictor variable is re-entered – otherwise it is deleted. This procedure is then repeated until only useful predictor variables remain in the model.

It is emphasized that the regression assumptions should be checked before drawing statistical conclusions from the analysis because the validity of the statistical procedures hinges on the validity of the assumptions. As explained previously, most of the assumptions were detected graphically. This manner of testing requires some experience. Is there a way to detect some of the assumptions quantitatively?

As explained earlier, that regression analysis is used to produce an equation that will predict a dependent variable using one or more independent variables (Equation 1.1). However, when running the regression, the statistical packages provide very useful information. Beside the model coefficients (estimated parameters), it presents the standard deviation of each estimated parameters, T-statistics and P-value. This is illustrated in Figure A.8. First of all, the purpose is to discover whether the coefficients on the independent variables are really different from zero, and if the independent variables are having a genuine effect on the dependent variable, or if any apparent differences from zero are just due to random chance. The null (default) hypothesis is always that each independent variable is having absolutely no effect (has a coefficient of zero) and a reason to reject this theory is sought.

A.7 T-statistics and P-value
The T-statistic is the coefficient divided by its standard error. The standard error is an estimate of the standard deviation of the coefficient; the amount it varies across cases. It can be thought of as a measure of the precision with which the regression coefficient is measured. If a coefficient is large compared to its standard error, then it is probably different from zero. How large is large? The regression software compares the T-statistic on the variable with values in the Student's T-distribution to determine the P-value, which is the number to be observed. The Student's T-distribution, (Figure A.8) describes how the mean of a sample with a certain number of observations is expected to behave.

A P-value of 5% or less (significant level 95%) is the generally accepted point at which to reject the null hypothesis. With a P-value of 5% (0.05), there is only a 5% chance that the estimated parameters would have come up in a random distribution, and thus, it can be said that with a 95% probability of being correct that the variable is having some effect, assuming the model is specified correctly. The test is carried out by comparing this observed value with the appropriate critical value obtained from statistical tables. Critical value can be read directly from the statistical tables at \( n-p \) and 0.95/2, where \( n \) denotes an observation number. Note, the number was divided by 2 because it has a two-sided alternative hypothesis. Also, \( n-p \) was used as a degree of freedom because \( p \) is the
estimated parameters (including the constant, $b_0$). The P-value is the sum of the two shaded areas under the curve in Figure A.8.

Assume regression analyses are conducted on $Y$ where $X_i$ is one of the explanatory variables. By observing the P-values for each estimated coefficient, if anyone is less than 0.05, the null hypothesis is rejected. Here, the default test is to determine if the estimated coefficient for $X_i$ is different from zero. Therefore, the null hypothesis states that the coefficient equals zero while the alternate hypothesis states that it is not equal to zero. Generally, if any estimated coefficients of the independent variables (including the constant) have a P-value greater than 0.05, it is removed from the model and the analysis is re-run.

Furthermore, while model completeness is one of the most important aspects of model adequacy, this type of test does not address other important aspects of model quality. In statistical jargon, this type of test for model adequacy is usually called a "lack-of-fit" test.

The most common strategy used to test for model adequacy is to compare the amount of random variation in the residuals from the data used to fit the model with an estimate of the random variation in the process using data that are independent of the model. If these two estimates of the random variation are similar, that indicates that no significant terms are likely to be missing from the model. If the model-dependent estimate of the random variation is larger than the model-independent estimate, then the model miss-specified, (a model that provides an incorrect description of the data).

A.8 Analysis of Variance (ANOVA Table)

A convenient way to present the results of a regression analysis is in an analysis of variance table, or, as it is usually known, an ANOVA table (from ANalysis Of VAriance). The table contains various summary statistics of the data. The entries in the table enable the testing of some important hypotheses in regression, which could not be tested previously. For example, testing for overall regression or testing the significance of sets of independent variables.

The ANOVA table contains the degrees of freedom, the sum of squares, the mean squares, an F-statistic and a corresponding P-value. For a multiple linear regression model with $p$ independent variables, the model has $p$ degrees of freedom (i.e. the number of independent variables). Since the total degrees of freedom is unchanged at $n-1$, it follows that the degrees of freedom for the error is $n-p-1$. A summary of the first table is given by

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>$p$</td>
<td>$SSR$</td>
<td>$SSR/p$</td>
<td>$SSR/p$</td>
<td>$SSE/(n-p-1)$</td>
</tr>
<tr>
<td>Error</td>
<td>$n-p-1$</td>
<td>$SSE$</td>
<td>$SSE/(n-p-1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>$n-1$</td>
<td>$TSS$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

335
Figure A.8 - Probability Density Function Graph of a T-distribution. The P-value is the sum of the two shaded areas under the curve.
In a multiple linear regression model, the hypothesis of overall regression is the hypothesis that the independent variables (as a whole) provide some useful information about the response variable. Thus, the hypothesis of overall regression is accepted, if the null hypothesis that none of the explanatory variables influence the response is rejected. That is, H₀: \( b_1 = b_2 = \ldots = b_p = 0 \) against the alternative hypothesis H₁: \( b_j \neq 0 \) for at least one \( j = 1, 2, \ldots, p \). If H₀ is false, that is, if the dependent variable depends upon at least one of the independent variables, then model \( \hat{Y} = b_0 + b_1 x_1 + \ldots + b_p x_p \) should explain a substantial amount of the variation in the data, relative to the residual variation. Thus, if H₀ is false, the model variation (SSR) should be large relative to the residual variation (SSE). Then, reject H₀, if the value of the ratio \( \text{SSR}/p \) / \( \text{SSE}/(n-p-1) \) is large. The F-value in the ANOVA table is exactly this ratio. However, the hypothesis can be rejected if the P-value is less than 0.05 (5% significance level) as explained earlier.

Occasionally, it is of interest to test the significance of a set of \( p \) variables. This model will be referred to as the full model (or Model #1). Now, suppose there is reason to believe that a model with only \( m \) (with \( m < p \)) independent variables will fit the data just as well (Model #2). Model #1 can be tested as to whether it describes the variation in the data as well as Model #2 by testing the significance of the \( p - m \) extraneous variables. That is, by testing the null hypothesis H₀: \( b_{m+1} = \ldots = b_p = 0 \) against the alternative hypothesis H₁: \( b_j \neq 0 \) for at least one \( j = m + 1, \ldots, k \).

Why is the H₀ tested? If the full model explains the variation in the data better than the reduced model (Model #2), it must be because the \( p - m \) extraneous variables provide more useful information regarding the response variable. Hence, H₀ should be rejected and the extraneous variables in the model should be retained. If, on the other hand, the full model (Model #1) and the reduced model (Model #2) explain (almost) the same amount of the variation, the simpler, reduced, model may as well be used as the full one. Thus, we do not reject H₀. More specifically, we reject H₀ if the difference between the model variation for the full model and the model variation for the reduced model is large, relative to the residual variation under the full model.

A.9 Comparing Models
When fitting the data with regression, the main objective is to discriminate between different models. Model building in regression takes many forms, but the test will be whether the data are more consistent with one possible model relative to another and to see if including one extra variable improves the model significantly. In multiple regressions, two models can be compared where one model has an additional term by using the F-test. In practice, this means that the second model has the greater number of parameters. When comparing two nested models, the full model will almost always have higher \( R^2 \) and a lower sum of squares than the reduced model. It is not enough to compare \( R^2 \) or sum of squares. A statistical approach must be used to decide which model should be accepted.

If the models are nested, two distinct approaches can be used to compare models. The first method is based on statistical hypothesis and ANOVA. It is based on analysis of the difference between the sum of squares of the two models. It also takes into account the number of parameters of each model. The ANOVA calculations compute an F-ratio and
P-value. If the P-value is lower, it can be concluded that there are sufficient data to convince and to reject the reduced model (the null hypothesis) in favour of the alternative hypothesis if the full model is correct. If the P-value is small, there are two possibilities:

- The full model is correct
- The reduced model is correct, but random scatter led the full model to be better.

The P-value is able to tell how frequently this could be expected to happen. So, which model is correct? Which model fits better? The full model is correct but this would be expected to be by chance and thus, that is the wrong question to ask!! Which model should be accepted? This is answered by statistical hypothesis testing. This method is only valid when comparing nested models. It cannot be used for non-nested models.

The second method for comparing models is not based on hypothesis testing at all. After fitting a number of different models to a given data set, they can be compared by:

- Error measures in the estimation period
- Error measures in the validation period

To compare the models and before running the regression analysis, the data set must be divided into two sets. One set for fitting the regression model and the other set (about 10% of the whole data) to validate and compare the models. Average relative error, absolute average relative error, minimum and maximum of both averages give a good advice. See Appendix C for more details.
A.10 References

A.10.1 Regression Analysis

A.10.2 Regression Analysis in Seismic

A.10.3 Regression Analysis in Petrophysical


### A.10.4 Regression analysis in Natural Gas Engineering


### A.10.5 Regression analysis in Relative Permeability


A.10.6 Regression Analysis in Reservoir Recovery Factor

Arps, J.J. and Roberts, T.G. 1955. The Effect of Relative Permeability Ratio, the Oil Gravity, and the Solution Gas - Oil Ratio on the Primary Recovery from Depletion Type Reservoir," Trans., AIME, Petroleum Development and Technology, 204,120.

A.10.7 Regression Analysis in Drilling Engineering


Yidan Li et al., 1995. Correlation between filter cake structure and filtration properties of model drilling fluids. International symposium on oilfield chemistry, San Antonio TX, 14-17.
A.10.8 Regression analysis in Flow in Porous Media
APPENDIX B

ARTIFICIAL NEURAL NETWORKS

B.1 Introduction
In order to find relationship between the input and output data derived from experimental work, a more powerful method than the traditional ones are necessary. For some time, ANN methods have been used in many diverse fields as a function approximation tools, producing results comparable with (or better than) regression models and other statistical methods, see for example Negnevitsky (2005). However, one of their key advantages is their ability to easily model complex, non-linear systems. ANN is an especially efficient algorithm to approximate any function with finite number of discontinuities by learning the relationships between input and output vectors (bozorgmehry et al., 2005; Haganet et al., 1990). These algorithms can learn from the experiments, and also are fault tolerant in the sense that they are able to handle noisy and incomplete data. ANN methods are a powerful tool and widely used technique for supervised neural networks learning in many application areas, ranging from pattern recognition, and function approximation, to speech recognition and robot control. Thus, ANNs are able to deal with non-linear problems, and once trained can perform prediction and generalization rapidly (Sozen et al., 2004). They have been used to solve complex problems that are difficult to be solved if not impossible by the traditional methods and provide more accurate and robust solutions for problems. Examples include; regression, classification, control, optimization, pattern recognition, and so on. Especially it is desired to have the minimum difference between the predicted and actual outputs (Richon and Laugier, 2003).

B.2 Backpropagation Neural Network
In recent years, ANN methods have become increasing popular for prediction and forecasting in a number of disciplines, including petroleum and gas engineering. The application of ANNs has blossomed since the introduction of the backpropagation training algorithm, BP for feedforward ANNs. BP neural network may thus be considered as a fairly new tool in the field of production and forecasting in petroleum and gas engineering.

The BP algorithm consists of two paths; forward path and backward bath. The backward bath is described for the backpropagation training algorithm. Forward bath contain creating a feedforward network, initializing weight, simulation and training the network. The network weights biases are updated in backward path. Once the network weights have been initialized, the network is ready for training. The training process requires a set of proper inputs and targets as outputs. During training, the weights are interactively adjusted to minimize the network performance function (mean square error, MSE). MSE is the default performance function for forward networks which is the average squared error between the network outputs and the target output. The basic procedure for training the network is embodied in the following description:
i. Apply an input vector from training dataset to the network and calculate the corresponding output value,

ii. Compare the actual output with the computed output value and determine the difference (error).

iii. Determine in which direction (increase or decrease, + or −) to change each weight in order to reduce the error,

iv. Determine the amount by which to change each weight,

v. Apply the corrections to the weights, and

vi. Repeat items a through e with all the training vectors until the error for all vectors in the training set is reduced to an acceptable value.

Whoever, BP network is a layered, feedforward network that is fully interconnected by layers. Thus, there are no feedback connections and no connections that bypass one layer to go directly to a later layer. A basic multilayer arrangement of a typical interconnected neural network is shown in Figure B.1. It consists of three layers of processors; an input layer, output layer and one hidden layer between them. Each layer has a number of neurons or nodes. However, the Figure consists of an input layer with four neurons, one hidden layer with five neurons, and an output layer with single neuron. Although only one hidden layer is used in the Figure B.1, more than one hidden layer is permissible.

B.3 Mathematic Description of Backpropagation Algorithm

To begin, let’s assume a set of \( P \) vector-pairs, \((x_1, y_1), (x_2, y_2), \ldots, (x_p, y_p)\), which are examples of a set of a function mapping \( g : x \in R^N, y \in R^M \). The three layer network must be trained so that it will find an appropriate set of weights. To achieve that goal, Freeman and Skapura (1999) presented all of the relevant equations for the BP network in the order in which they would be used during training as follows:

1. Apply the \( N \)-dimensional input vector, \( x_p = (x_{p1}, x_{p2}, \ldots, x_{PN})^t \) to the input layer of the network, where the ‘\( t \)’ superscript means transpose. The input units distribute the values to the hidden-layer units.

2. Calculate the net-input values to the hidden layer units. The net-input to the \( j \)th hidden unit is

\[
\text{net}_j^h = \sum_{i=1}^{N} w_{ji}^h x_i + \theta_j^h \quad \text{(B.1)}
\]

The weight on the connection from the \( j \)th node to the \( i \)th node is denoted \( w_{ji} \), where \( w_{ji}^h \) is the weight on the connection from the \( i \)th input units, the “\( h \)” superscript refer to quantities on the hidden layer and \( \theta_j^h \) is the bias term.
Figure B.1- Schematic of Typical Multilayer Neural Network Model. It consists of an input layer with four neurons ($x_{p1}$, $x_{p2}$, $x_{p3}$, and $x_{p4}$), one hidden layer with five neurons ($hn_1$, $hn_2$, $hn_3$, $hn_4$, and $hn_5$) and an output layer with single neuron, $y$. 
3. Calculate the outputs from the hidden layer. The function used for this calculation is referred to as the activation function as illustrated in Figure B.2. The activation of this node is equal to the net-input; then, the output of this node is

\[ i_{pj} = f^h_j(\text{net}_p^h) \]  

(B.2)

where \( f^h_j \) is the input activation function. The activation function can be linear so that we have a linear network, or nonlinear. Activation functions for the hidden units are needed to introduce nonlinearity into the network. Without nonlinearity, hidden units would not make nets more powerful than just plain perceptrons (which do not have any hidden units, just input and output units). However, it is the nonlinearity that makes multilayer networks so powerful.

4. Move to the output layer. Calculate the net-input values to each unit:

\[ \text{net}_{pk} = \sum_{i=1}^{L} w_{ki} i_{pj} + \theta_k \]  

(B.3)

The “\( o \)” superscript refers to quantities of the output layer.

5. Calculate the outputs, \( o_{pk} \):

\[ o_{pk} = f^o_k(\text{net}_{pk}) \]  

(B.4)

where \( f^o_j \) is the output activation function.

6. Calculate the error terms, \( \delta_{pk}^o \) for the output units:

\[ \delta_{pj}^h = (y_{pk} - o_{pk}) f^o_k(\text{net}_{pk}^o) \]  

(B.5)

where \( y_{pk} \) are the desired output values and \( o_{pk} \) is the model output values from Equation (3.4).

7. Calculate the error terms, \( \delta_{pj}^h \) for the hidden units:

\[ \delta_{pj}^h = f_j^h(\text{net}_{pj}^h) \sum_k \delta_{pk}^o w_{kj} \]  

(B.6)

Note that, \( \delta_{pj}^h \) is calculated before the connection weights to the output-layer unit have been updated.
8. Update weights on the output layer:

\[ w_{ij}^{(o)}(t+1) = w_{ij}^{(o)}(t) + \eta \delta_{pk}^{(o)} i_j \] (B.7)

where \( \eta \) is the learning rate parameter.

9. Update weights on the hidden layer:

\[ w_{ij}^{(h)}(t+1) = w_{ij}^{(h)}(t) + \eta \delta_{pk}^{(h)} x_i \] (B.8)

Every weight update on the hidden layer depends on all the error terms on the output layer, \( \delta_{pk}^{(o)} \). This result is where the notion of BP arises. The known errors on the output layer are propagated back to the hidden layer to determine the appropriate weight changes on that layer. The order of the weight updates on an individual layer is not important. Be sure to calculate the error term, \( E_p \)

\[ E_p = \frac{1}{2} \sum_{k=1}^{M} \delta_{pk}^{2} \] (B.9)

since \( E_p \) is the measure of how well the network is learning. When the error is acceptably small for each of the training-vector pairs, training can be discontinued. Thus, our goal is to minimize the network performance function (Mean Squares Errors, MSE) Eq. (3.10).

\[ \text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (E_p)^2 \] (B.10)

In a neural network model the function \( y = g(x) \) is modeled as a nonlinear function, and the neural network model can be rewrite as

\[ Y_i = f(X_i, \alpha_1, \ldots, \alpha_{m-1}, \beta) + \varepsilon_i \] (B.11)

Where, \( \beta, \alpha_1, \ldots, \alpha_{m-1} \) are unknown parameter vectors, \( X_i \) is a vector of known constants and \( \varepsilon_i \) are residuals (error terms).

**B.4 Design of Neural Networks Architecture**

Selection of suitable artificial neural network architecture is probably the hardest part of the problem and critical to obtaining a useful artificial neural network. It is analogous to selecting the form and independent variables of a regression equation as explained in Appendix A. One of the advantages of the artificial neural network approach is that it is easy to automate searching for an optimum architecture, either by using a simple exhaustive search or by using some sort of heuristic search methodology. Alyuda’s
NeuroIntelligence software (Alyuda 2006) was used for the artificial neural networks calculations. This proved to be a very powerful tool for carrying out the artificial neural network design and analysis. Usually, trial-and-error is required to achieve appropriate network architecture. In this chapter, a multiple feedforward backpropagation network was used to build the following six models:

1) Estimating bubble point pressure, $P_b$,
2) Estimating bubble point oil FVF, $B_{ob}$,
3) Estimating bubble point solution GOR, $R_{sob}$
4) Estimating stock-tank vent GOR, $R_{ST}$,
5) Predicting reservoir oil production prediction, and
6) Average oil reservoir pressure profile

One of the most important problems that neural network designers face today is choosing an appropriate network size for a given application. Network size involves in the case of layered neural network architectures,

- the number of input variables
- the data preparation
- the type of neural network architecture
- the number of hidden layers
- the number of hidden units per hidden layer
- the type of combination, transfer or error function
- the type of optimization technique

**B.4.1 Number of Input Variables**
If you are very knowledgeable in the function you are fitting, you may be able to select some input variables that you believe are predictive towards the output. Certainly before building the network there is no way of telling if one input will be better than another but you can obviously reduce the amount of variables you will use by doing a step ways technique to filter out those variables which are evidently and heavily correlated. The forward and backward step ways techniques allow you to eliminate those variables which might be redundant within the network and therefore only increase complexity without increasing the quality of the results. Once you have a set of variables which are not correlated it is time to test how much predictive power they actually have against your desired output. After well-selection network inputs, the next step is;

**B.4.2 Data Pre-processing**
Understanding the structure of the data in preparing the data for analysis is extremely important to a well-designed neural network design. In general, before building a model the balk dataset must partition randomly into four sets, namely training set, validation dataset, testing dataset, and deployment dataset.

1. **Training Dataset**
The training dataset is used to train and build a network model. In other words, this dataset is used to obtain the network weights.
Figure B.2- A Neural Net Perceptron
ii. Validation Dataset
However, during the training process the accuracy of the model is computed with a dataset that was not used in the training process which is validation dataset. The task of this dataset is to check the performance of the network model. In order to avoid overtraining which result poor network model performance, it is necessary to use an early stopping technique as illustrated in Figure B.3. Training can be stopped when the performance of the model on the validation dataset gives minimum error. The training set error will naturally decrease with increasing numbers of epochs of training. However, the error on the unseen validation dataset will start off decreasing as the under-fitting is reduced, but then it will eventually begin to increase again as over-fitting occurs. Indeed, the solution to get the best generalization and lowest error, the validation dataset is used for early stopping (Bishop 1995 and Amari et al. 1997).

i. Testing Dataset
The testing dataset is used to fine-tune models. This dataset is used neither in training nor in validation; it is used to pick the best model architecture. Typically, several network models are tried out with various architectures (as will be explained in detail in Section B.5) the accuracy of each one on the testing dataset is tested to choose the competing architecture. The lowest error on the testing dataset is used for selecting the appropriate model and to indentify the optimum network architecture.

The training and testing dataset error will naturally decrease with increasing numbers of training epochs after the number of epochs of training the testing error begin to increase as over-fitting takes place. In this iteration, the network should track.

ii. Deployment Dataset
The concept of deployment refers to the application of a model for prediction to new data. Building a model is generally not the end of the project. The knowledge gained will need to be organized and presented in a way that we can use it in the future. Thus, this portion of data is used to examine the model for prediction.

B.4.3 Type of Neural Network Architecture
There are a number of different ways to classify network architecture. A primary differentiator is single-layer network versus multilayer network. Single-layer networks only consist of an input and output layer, where multilayer networks have one or more hidden layer, as the name ‘hidden’ derives from the fact that it is not connected to the outside world. The networks without any feedback loop are called feedforward networks where is only one way flow from input units to output units. However, the feedforward Multilayer Perceptron (MLP) is one of the most widely used ANNs among other network models. In this chapter, a multiple feedforward backpropagation network was used to build all the six models:
Figure B.3- Illustration of the Early-stopping Rule. The trained network is saved at the point defined by the early-stopping criterion, and training continues exclusively to check whether the error will fall again. This will ensure that the increase in the validation error is not a temporary event.
B.4.4 Number of Hidden Layers and Hidden Nodes per Hidden Layer

A major problem in designing a neural network is establishing the optimal number of layers and hidden nodes per hidden layer. The hidden layer is responsible for internal representation of the data and the information transformations input and output layers. A search of the available literature suggested many rules of thumb for determining the number of neurons in the hidden layer. It became apparent that there is no general rule by which a suitable number of hidden layer neurons can be chosen and that the optimum number is very specific to the problem being solved (Krose and Van der Smaget, 1996) and (Heaton, 2005). Usually, trial-and-error is required to achieve a better network model architecture. It is arbitrary to the selection of the various neural network architectures and configuration settings. Hence, it is advised that building a well-designed neural network model takes patience and involves a large amount of time spent in network training and training a series of networks with single and multiple hidden layers with different number of nodes in each hidden layer.

B.4.5 Activation Function

Transfer function at the active nodes of the feedforward is required to be fixed before training the model in order to introduce nonlinearity into the network. In practice, the data set presented to a neural network must be pre-processed before training. During pre-processing, the data is compressed to fit into the active range of the activation function. This scaled dataset is then used for training purposes. To interpret the results obtained from the neural network, the outputs must be rescaled to the original range. However, the accuracy obtained by the neural network refers to this rescaled data set. Therefore, in order to yield the proper training neural network, the input/output data must first be scaled into the input/output interval of the neural network by using a suitable activation function. These activation functions scale the output of the neural network into proper ranges. However, the output from the activation function is either between 0 and 1, or between −1 and 1, depending on which activation function is used. In computational networks, the activation function of a node defines the output of that node given an input or set of inputs. Most neural networks pass the output of their layers through activation functions as illustrated above in Figure (B.2). Mathematically this can be represented as:

\[
y = f \left( \sum_{i=0}^{n} w_j x_j + b \right)
\]  

(B.11)

where \( w_i \) represents the weight vector, \( x_i \) is the input vector \((i = 1, 2, ..., n)\), \( b \) is the bias, \( f \) is the activation function, and \( y \) is the output.

Many types of activation functions have been proposed and tried, but the surviving, most used are the logistic and hyperbolic tangent activation function.

- Logistic activation function defined for any variable \( s \) as:

\[
f \ (s) = \frac{1}{1 + e^{-s}}
\]

(B.12)
It is an S-shaped (sigmoid) curve and their outputs take values in the ranges of [0,1].

- Hyperbolic tangent activation function defined for any variable $s$ as:

$$f(s) = \frac{2}{1 + e^{-2s}} - 1$$  \hspace{1cm} (B.13)

It a sigmoid curve, like the logistic function, except that output lies in the range $[-1, +1]$. Often performs better than the logistic function because of its symmetry.

Again, the activation function plays a critical role in NN modeling. However, the selection of the correct activation function to the NN design is not that critical. Since the hyperbolic tangent function produces both positive and negative target values, therefore, it tend to yield faster network training than logistic function, (Matignon, 2005).

### B.4.6 Type of Optimization Technique

As illustrated above, the objective of training the network is to minimize the global error by adjusting weights. Several algorithms have been proposed to speed-up the convergence of the backpropagation learning algorithm. Despite the wide range of available learning algorithms to train the network model, no single learning algorithm exists which works well on all learning problems. The most popular methods are:

1. **Quasi-Newton**
   Quasi-Newton is an enhancement of the BP training algorithm. Since, back propagation adjusts the network weights after each case, Quasi-Newton works out the average gradient of the error surface across all cases before updating the weights once at the end of the epoch. There is also no need to select learning or momentum rates for Quasi-Newton, and thus, it can be much easier to use than back propagation. However, Quasi-Newton is the most popular algorithm in nonlinear optimization, with a reputation for fast convergence (StatSoft, 2011).

2. **Levenberg-Marquardt**
   A Levenberg-Marquardt training algorithm can only be used on networks with a single output unit.

3. **Conjugate Gradient Descent**
   As with the Quasi-Newton, the Conjugate Gradient Descent algorithm is also a batch updating algorithm. It firstly determines the direction of the steepest descent then projects a straight line in that direction and locates a minimum gradient value along the line (StatSoft, 2011).
iv. **Quick Propagation**

Quick propagation works by making the assumption that the error surface is locally quadratic, with the axes of the hyper-ellipsoid error surface aligned with the weight. If this is true, then the minimum of the error surface can be found after only a couple of epochs. Based on this assumption, quick propagation needs to select a learning rate and algorithm coefficient (training algorithm’s parameter), (StatSoft, 2011).

In fact, the selection of a training algorithm is based upon the characteristics of the problem itself. Essentially, training a neural network model means selecting one model from the set of allowed models that minimize the cost function (minimum square error, Equation B.10). As a result, the selection of the neural network architecture will come down to a trial-and-error procedure. If the model architecture and learning algorithm are selected appropriately, the resulting ANN can be extremely robust.

Finally, there are concerns regarding the possibility of converging to a local minimum in weight space. A typical neural network might have a couple of hundred weights whose values must be found to produce an optimal solution. Since the output of a neural network, as a function of the inputs, is often highly nonlinear, this makes the optimization process complex. If an error was plotted as a function of the weights, a rough surface with many local minima would likely be observed, as shown in Figure B.4a. A final topic concerns the possibility of converging to a local minimum in weight space. Figure B.4b illustrates the idea. It shows a cross-section of a hypothetical error surface in weight space. The point, $z_{\text{min}}$, is called a “global minimum”. Notice, however, that there are other minimum points called “a local minimum”. A search for the global minimum might accidentally find one of these local minima instead of the global minimum.

Once a network settles on a minimum, whether local or global, learning stops. If a local minimum is reached, the error at the network outputs may still be unacceptably high. Fortunately, this problem does not appear to cause much difficulty in practice. If a network stops learning before reaching an acceptable solution, a change in the number of hidden nodes or in the learning parameters will often fix the problem or simply start over with a different set of initial weights (retrain the network), (Freeman and Skapura, 1991). Actually, the retraining approach was adopted in the modeling process in this chapter. However, when a network reaches an acceptable solution, there is no guarantee that it has reached the global minimum rather than a local one. If the solution is acceptable from an error standpoint, it does not matter whether the minimum is global or local, or whether the training was halted at some point before a true minimum was reached. Thus, the algorithm does not always step in the direction of the global minimum of the error surface. As a result, to avoid local minima, simply try a number of ANN models with different optimization algorithms with random starting points and use the one with the best value. In this way, randomly jumps in the hope that the location is getting closer to the global minimum.
Figure B.4- (a) Hypothetical Surface in Weight Space with Many Local Minimums. (b) Cross-section of a Hypothetical Error Surface in Weight Space.
B.5 **Selections of the Best Architecture**

However, the development of any ANN model is based on a thorough understanding of the research problem. Designing an ANN model, however, can be summarized in the following eight-step procedure:

*Step 1:* Data collection,

*Step 2:* Variable selection (number of inputs and outputs, if applicable). Generally, Steps 1 and 2 would be considered simultaneously prior to developing a model.

*Step 3:* Data pre-processing (normalizing, log transformation, standardization).

*Step 4:* Selection of training, test and validation sets.

*Step 5:* Neural network training parameters and configuration values (number of input neurons, percentage iterations for constant initial learning rate, initial weights, learning rate increment, and radius decrement).

*Step 6:* Neural network training (presentation of records and the number of iterations).

*Step 7:* Evaluation criteria (quantitative and graphical analysis), and

*Step 8:* Model deployment.

The development of ANN models starts with finding the best network structure that will represent the complicated relationship between variables. Possible ANN architectures, starting from a low number of layers and nodes to a higher number of layers and nodes, should be tested to select the best architecture. Successful architectures are those that converge to the target error or reach the possible MSE and exhibit stable performance to new data not included in the training. Topically, the best ANN architecture is the one that simultaneously fulfils all the following criteria:

- Lowest training error
- Lowest test error
- Lowest difference between test and training error
- Lowest difference between test and validation error,
- Highest $R^2$, and
- Simplest structure. Generally, a simple model that has approximately the same performance as a more complex model should be preferred.

These criteria, however, can be adjusted to a case of ANN for prediction.

B.6 **Model Verification**

Model verification and deployment is the final stage in the modeling process and is the most important step in the modeling development. After a satisfactory model has been identified as the best model in the previous stages, both qualitative and graphical analyses
are used to evaluate the model. Quantitative analysis is determined in terms of correlation coefficient ($R^2$), standard deviation (SD), average, maximum and minimum absolute error (AE), average, and maximum and minimum absolute relative error (ARE). One common distinction between the absolute error and the relative error is the absolute error is the magnitude of the difference between the exact value and the approximation. The relative error is the absolute error divided by the magnitude of the exact value. The percent error is the relative error expressed in terms of per 100. Average absolute error is a quantity used to measure how close forecasts or predictions are to the actual values. Standard deviation shows how much variation or dispersion there is from the average. Correlation coefficient $R^2$ provides a measure of how well future outcomes are likely to be predicted by the model.

Graphical analysis is determined in terms of scatter diagrams that compare the predicted ANN values versus the actual observation values. From this plot, it can be observed as to how the two comparable datasets agree with each other. In this case, an identity line ($y=x$ line, or a 1:1 line) is often drawn as a reference. The more the two datasets agree, the more the scatters tend to concentrate in the vicinity of the identity line. This indicates an excellent agreement between the actual data and network model predicted data. If the two datasets are numerically identical, the scatters fall on the identity line exactly.
APPENDIX C

LEAST SQUARES SUPPORT VECTOR MACHINE, LS-SVM

C.1 Introduction

The Support Vector Machine (SVM) can be used for both regression and calcification problems, as with ANN. A SVM is derived from statistical learning theory and was found in 1995 by Vapnic et al. SVM is one of the effective algorithms in machine learning and data mining with the advantages of simple model structure selection, fast processing speed, and high learning precision. It is widely used in handling regression and classification problems based on statistical learning theory. The method adheres to the principle of structural risk minimization, seeking to minimize an upper bound of the generalization error, rather than minimize the training error. Suykens et al. (2002) proposed the LS-SVM method, in which the least squares loss function was modified and introduced to the SVM. LS-SVM is based on equality constraints and a sum square error cost function as opposed to earlier approaches that utilize inequality constraints and solve complex convex optimization problems. The LS-SVM reformulation simplifies the problem and solution by adopting a linear system rather than quadratic problems that are difficult to compute. This principle is illustrated in Figure C.1. Consider a nonlinear relationship in the input space (original space, Figure C.1, left panel). Then, the inputs ($x$) are mapped into a high dimensional space by means of $\Phi$ (Figure C.1, right panel). In this space, a linear model is fitted. There are, however, only two parameters to be tuned, $\gamma$ and $\sigma^2$, which are less than that for standard SVM (three parameters). When adopting LS-SVM with the RBF kernel function, the parameter combination ($\gamma$, $\sigma^2$) should be established, where $\gamma$ denotes the regularization parameter, namely, the trade-off between minimizing the training error and minimizing the complexity of the LS-SVM model; $\sigma^2$ denotes the RBF kernel parameter and represents the change in the RBF kernel function (width of kernel) and it is a positive real constant.

SVM and LS-SVM is a kernel-base algorithm. A kernel is a function that transfers the input data to a high-dimensional space where the problem is solved. Kernel function can be linear or nonlinear. The model parameters of an LS-SVM are given by the solution of a system of linear equations. The generalization performance of the LS-SVM is however, strongly dependent upon the model selection process, in this case the careful selection of an appropriate kernel function and the good values of the model parameters (Suykens et al., 2002).
Figure C.1 - Illustration of the Key Ingredient of LS-SVM: Transformation $\Phi$ of the input data to the feature space.
C.2 LS-SVM for Function Approximation

The following paragraph is a brief introduction to the LS-SVM regression model. The main objective of function approximation is to approximate a function \( f(x) \) from a given training set of \( N \) samples \( \{x_i, y_i\}_{i=1}^N \) in which \( x_i \in \mathbb{R}^N \) (ND vector space) is the input data and \( y_i \in \mathbb{R} \) (1D vector space) is the corresponding output value. The least square method learns in a linear relation between a set of indicators \( f_i(x), i = 1 \ldots N \), and the output is

\[
\hat{y}(x) = b_1 f_1(x) + b_2 f_2(x) + \ldots + b_n f_n(x) \tag{C.1}
\]

In this equation, the \( b \)'s are the \( n \) parameters that receive a value during the training by minimizing the summed squared approximation error over all examples. The indicators can be non-linear functions of the input vectors. Therefore, the training data \( X \) and the output \( Y \) can be defined as

\[
X = \begin{bmatrix}
f_1(x_1) & \ldots & f_n(x_1) \\
\vdots & \ddots & \vdots \\
f_1(x_N) & \ldots & f_n(x_N)
\end{bmatrix}, \quad Y = \begin{bmatrix}
y_1 \\
\vdots \\
y_N
\end{bmatrix} \tag{C.2}
\]

The goal of function approximation is to train the underlying relation between the inputs and the output values. Then, the LS-SVM model is constructed in this high dimensional feature space and takes the form

\[
\hat{y}(x) = w^T \varphi(x) + b \tag{C.3}
\]

where the nonlinear mapping \( \varphi(.) \) maps the input data into a higher dimensional feature space. \( b \) is the bias and \( w \) is a weight vector having the same dimension as the feature space. In LS-SVM, the weight vector \( (w) \) is that will give the smallest summed squared approximation error over all samples is located. After training, a function approximation implements a mapping \( x \rightarrow \hat{y} \) that can be evaluated for any \( x \). The performance of function approximation is measured by means of the approximation error \( e_i \), which is defined as:

\[
e_i = y_i - \hat{y}(x_i) \tag{C.4}
\]

Thus, nonlinear function is estimated in its original space because the linear function estimation is in the feature space. Then, the optimization problem of LS-SVM is formulated wherein the regularization parameter \( (\gamma) \) is used to avoid overfitting. As a result, the regularization parameter needs to be tuned while in the training process.

\[
y(x) = \sum_{i=1}^n \alpha_i k(x, x_i) + b \tag{C.5}
\]

where \( k(x, x_i) \) is the kernel function, \( x_i \) is the input vector, \( \alpha_i \) is Lagrange multiplier called the “support value”, \( b \) is the bias term. However, \( \alpha_i \) and \( b \) are the results from the optimization. The details of the LS-SVM algorithm could be found in Vapnok, (1995), Gue et al., (2006) and Chen et al., (2007).
There are several common kernel functions such as linear, polynomial, radial basis function (RBF) kernel, and Multilayer Perceptron (MLP). Theoretically speaking, the linear kernel type is the simplest and most efficient kernel to perform similarity calculation. RBF kernel is a nonlinear function and a more compact supported kernel, and could reduce the computational complexity of the training procedure while giving good performance under general smoothness assumptions (Wu et al., 2008). Moreover, in comparison with some other feasible kernel functions, the RBF is a more compactly supported kernel function and able to improve generalization performance of the LSSVM (Brabanter et al., 2002 and Shi and Nabil, 2007). As a result, RBF kernel (Equation C.6) was recommended as the kernel function for the LS-SVM regression models in this thesis.

\[ K(x_1, x_2) = \exp\left( -\|x_1 - x_2\|^2 / 2\sigma^2 \right) \]  

(C.6)

where, \( \sigma^2 \) is a positive real constant denoting the kernel width parameter. As a result of kernel regression, the nonlinear relationship in the data space is transferred to a linear relationship in the feature space by means of kernels.

Consequently, from the training LS-SVM problem, it can be observed that there are two free parameters. Regularization parameter (\( \gamma \)) and kernel width parameter (\( \sigma^2 \)), may affect the LS-SVM generalization performance. The regularization parameter \( \gamma \) determined the trade-off between minimizing the training error and minimizing model complexity. The parameter \( \sigma^2 \) was the bandwidth and implicitly defined the nonlinear mapping from the input space to a high dimensional feature space. The small value of \( \sigma^2 \) can eventually lead to overfitting. On the contrary, a large value can make the model more parsimonious and not accurate (Espinoza et al., 2003).

C.3 LS-SVM Regression Modeling Procedure

The LS-SVM modeling procedure can be summarized as follows (Deng and Yeh, 2010):

- Divide the entire dataset into training, validation and testing datasets. The training and validation datasets are used to build the LS-SVM model and the testing dataset is used to verify the model’s performance.
- Initialize the hyper parameters \( \gamma \), \( \sigma^2 \) using the training dataset.
- Optimize the posterior probabilities of the model parameters with respect to different levels in Bayesian inference.
- Use the grid search algorithm with the cross-validation method to perform fine-tuning of the optimal parameter combination (\( \gamma_{opt}, \sigma^2_{opt} \)), as obtained in Step 3. The present research applied a ten-fold cross-validation method, dividing the training into ten aliquot parts. The grid training data comprises nine aliquot parts and the other part is the grid validation dataset. Train the LS-SVM model with the grid training data and the optimized parameter combination (\( \gamma_{opt}, \sigma^2_{opt} \)). Test the
LS-SVM model with the validation data and iterate the process ten times. The average training error is collected and calculated afterwards. Insert the new parameter combination and repeat the process until the stop criteria are approached. This process can successfully obtain the optimal parameter combination \((\gamma, \sigma^2)\) with the minimized error.

- Adopt the tuned parameter combination \((\gamma, \sigma^2)\) to build the LS-SVM prediction model. Substitute the testing dataset into the LS-SVM model and the prediction data can now be obtained successfully.

- Finally, use performance criteria to calculate the prediction performance.

The above steps are also illustrated in the flowchart presented in Figure C.2.

C.4 Leave-one-out Cross-validation for LS-SVM

Cross-validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments - one used to train a model and the other used to validate the model. There are two possible goals in cross-validation:

- To estimate performance of the learned model from available data using one algorithm. In other words, to gauge the generalizability of an algorithm.

- To compare the performance of two or more different algorithms and find the best algorithm for the available data, or alternatively, to compare the performance of two or more variants of a parameterized model.

In order to avoid the possibility that the LS-SVM model over fits the data it learns from, it is important to measure its performance when approximating data which have not been used during the training process. One of the most common methods of obtaining a model with good generalization capabilities consists in minimizing not just the training error but the cross-validation error. The \(l\)-fold cross-validation error of a model is obtained by dividing the available data in \(l\) sub-sets and, alternately using one of the sub-sets as test data and the rest as training data. Therefore, a total of \(l\) models are trained and cross-validated.

L-fold cross validation is an abstract pattern used for the estimation of the classification accuracy of learning algorithms. It is used to determine the accuracy with which a learning algorithm is able to predict inferences using unused sample data during the training phase. However, 1-iteration of the L-fold cross-validation is performed in the following way:
Figure C.2 – LS-SVM Modeling Procedure
• First, a random permutation of the validation dataset is generated and partitioned into L folds of approximately equal size. Of the L subsets, a single subset is retained as the validation data to estimate the performance of a predictive model, and the remaining L – 1 subsets, together, are used as training data.

• Then a model is trained on the training set and its accuracy is evaluated on the L-fold. Model training and evaluation is repeated L times, with each of the L subsets used exactly once as a testing set. Figure C.2 shows 1-iteration of a 3-fold cross validation algorithmic pattern using 30 samples of data.

C.5 Grid-search Technique
Before using the grid search, we need to identify the option range of parameters, grid spacing and the objective function of the grid points, etc. must be identified. The main steps are as follows:

• initial value assignments of $\gamma$ and $\sigma^2$ were given and their search ranges were determined

• according to the initial value assignments, the position of the first grid point of the cross validation was determined

• the ranges of parameters were respectively divided into $n$ equal spaces, and the n-cross validation was used as the objective function of the grid calculation to calculate all the grid points, and finally,

• the contour line was drawn based on the results of the cross validation. Under the conditions of the contour line, the parameters $\gamma$ and $\sigma^2$ were sought. If the obtained $\gamma$ and $\sigma^2$ do not meet the accuracy requirements, the range of the parameters should be narrowed down and the above steps repeated until they meet the accuracy requirements. Sequentially, the optimal parameter values were obtained.

C.6 Statistical Measures for the Evaluation of Prediction Performance
The performance of the model is estimated based on the cross validation. As explained previously, all the different estimates of the performance are combined. The assumption is made that the input data is distributed independent and identically over the input space. For each combination of $\gamma$ and $\sigma^2$ parameters, MSE was calculated and the optimum parameters were selected when smaller MSE was produced. In addition, the error term should be analyzed by means of Absolute Error (AE), Relative Error (RE) and regression coefficient ($R^2$).
Figure C.2 – One-iteration of a 3-fold Cross Validation Algorithmic Pattern
C.7 References
Wu, Di et al., 2008. Application of image texture for the sorting of tea categories using multi-spectral imaging technique and support vector machine. JFE, vol. 88
APPENDIX D

ERROR ANALYSIS

D.1 Introduction
Statistical error analysis is used to analyse errors \( (e_i) \) in order to check the performance and the accuracy of the developed models. The difference (error) occurs due to randomness or because the estimator does not account for information that could produce a more accurate estimate. In trying to ascertain whether the error measures in the training process are reliable, it should be considered whether the model under consideration is likely to have over fitted the data. The overall data must split into training, validation and testing subsets. The error measures in the training, validation and testing datasets are very important. Indeed, the model's performance in the testing dataset is the best guide to its ability to predict the future. However, it should be expect that the errors made in predicting the future could be larger than those made in fitting the past.

For the purpose of communicating model results to others, it is usually best to report the Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Absolute (AE) and Absolute Relative Percent Error (ARE), Error Standard Deviation (SD), multiple regression coefficients \( (R^2) \) and minimum, maximum and average values of both relative and absolute relative error.

The error term \( (e_i) \) can be defined as the deviation of the calculated value from the true value.

\[
e_i = y_i - \hat{y}_i
\]

where, \( y_i \) is an observation and \( \hat{y}_i \) is a corresponding fitted value obtained by use of the fitted regression model.

D.2 Mean Squared Error (MSE)
MSE is a measure of how close a fitted line is to data points. In other words, it is a way to quantify the difference between the modeled and actual. It measures the average of the squares of the errors. MSE can refer to the mean value of the squared deviations of the predictions from the true values, over an out-of-sample test space, generated by a model estimated over a particular sample space. MSE has the same units of measurement as the square of the quantity being estimated.

\[
MSR = \frac{e^2}{n}
\]

where \( n \) is the number of observations.
D.2 Root Mean Square Error (RMSE)
RMSE is frequently used to measure the differences between values predicted by a model and the values actually observed. It also known as a quadratic mean. The RMSE can be compared to an observed variation in errors of a typical model.

\[
RMSE = (\frac{e^2}{n})^{0.5}
\]  
(D.3)

D.3 Relative Error (RE)
This is an indication of the relative deviation in percent from the true values and is given by;

\[
RE = \frac{e_i}{y_i} \times 100
\]  
(D.4)

D.4 Absolute Relative Error (ARE)
It indicates the relative absolute deviation in percent from the true values, and is defined by;

\[
ARE = |RE|
\]  
(D.5)

Maximum, minimum and average of RE and ARE
After the RE and ARE, for each data point, are calculated, the minimum, maximum, and average values are scanned to know the error range of each model. Lower values imply a better model.

D.5 Standard Deviation (SD)
Standard deviation is a measure of dispersion, defined by

\[
SD = \sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2}
\]  
(D.6)

where, \( \bar{x} \) is the average value of the \( n \) observations of \( x_i \). A low SD indicates that the data points tend to be very close to the average. Whereas high SD indicates that the data are spread out over a large range of values.

D.6 Correlation Coefficient (R²)
R² is a statistic that will give some information regarding the goodness of fit of a model. In other words, it represents the degree of success in reducing the SD by regression analysis. An R² of 100% indicates that the regression model perfectly fits the data.

\[
R^2 = \left(1 - \frac{\sum(y_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}\right) \times 100
\]  
(D.7)

where \( \bar{y} \) is the average of \( y_i \).