Application of artificial neural networks in the quantitative analysis of gas chromatograms

Michael Williams

Follow this and additional works at: https://trace.tennessee.edu/utk_gradthes

Recommended Citation
To the Graduate Council:

I am submitting herewith a thesis written by Michael Williams entitled "Application of artificial neural networks in the quantitative analysis of gas chromatograms." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

M. A. Abidi, Major Professor

We have read this thesis and recommend its acceptance:

P. B. Crilly, M. O. Pace

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)
To the Graduate Council:

I am submitting herewith a thesis written by Michael Williams entitled “Application of Artificial Neural Networks in the Quantitative Analysis of Gas Chromatograms.” I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Electrical Engineering.

M. A. Abidi, Major Professor

We have read this thesis and recommend its acceptance:

Accepted for the Council:

[Signature]

Associate Vice Chancellor
and Dean of The Graduate School
APPLICATION OF ARTIFICIAL NEURAL NETWORKS IN THE QUANTITATIVE ANALYSIS OF GAS CHROMATOGRAMS

A Thesis
Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Michael Williams
May 1996
ACKNOWLEDGMENTS

I wish to express my sincere appreciation to Dr. Mongi Abidi. As my advisor and the chair of my graduate committee, Dr. Abidi provided a tremendous amount of encouragement and time during my two years of graduate school. Special thanks go to Mr. Martin Hunt for his endless guidance, support, and time during this research. Additional thanks go to Mr. Thomas Lewis and Mr. Brian Chase for their assistance along the way. Thanks are due to Dr. P. B. Crilly and Dr. M. O. Pace for their time in reviewing the thesis and being part of my defense committee. I would also like to thank Ms. Katie Jager for her time in reviewing and editing the thesis. I would also like to give credit to the United States Department of Energy (Robotics Technology Development Program within the Office of Science and Technology) for the financial support for this research.
ABSTRACT

This thesis demonstrates the effectiveness of artificial neural networks (ANN) in the quantitative analysis of multicomponent environmental samples. The data used in this research is from simulated soil samples contaminated with environmentally hazardous chemicals. Chromatographic data for three different pure analytes (Aroclors 1242, 1254, and 1260), as well as mixtures of the three, were collected by gas chromatography (GC) as training and testing data for the ANN. The networks were designed so that the concentrations of three different Aroclors in a given soil sample may be determined even when multiple Aroclors are present in one sample. This is significant since traditional linear methods tend not to handle overlapping contributions from multiple components. Due to the nature of a multilayered feedforward neural network, such overlapping contributions can be isolated and nonlinear classification analysis can be employed. Network parameters and architectures were optimized to give maximum performance. This is important because the architecture affects the generalization capabilities of the network; that is, its ability to produce accurate results on patterns outside its training set. Our ANN is as effective, or more effective, in both single and mixture concentration determination as existing methods such as multiple linear regression and principal component regression. Our method also has the advantage of using peak area tables as input data instead of the entire chromatogram.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Statement of the Problem</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Scope of the Current Application</td>
<td>5</td>
</tr>
<tr>
<td>1.3 Review of Related Work</td>
<td>7</td>
</tr>
<tr>
<td>1.4 Synopsis</td>
<td>10</td>
</tr>
<tr>
<td>2. DATA STRUCTURE AND NEURAL NETWORK FEASIBILITY</td>
<td>12</td>
</tr>
<tr>
<td>2.1 Basic Principles of Gas Chromatography</td>
<td>12</td>
</tr>
<tr>
<td>2.2 Features of Chromatographic Data</td>
<td>14</td>
</tr>
<tr>
<td>2.3 Feasibility of a Neural Networks Approach</td>
<td>21</td>
</tr>
<tr>
<td>3. TECHNIQUES EMPLOYED IN NEURAL NETWORK LEARNING</td>
<td>23</td>
</tr>
<tr>
<td>3.1 General Description of Artificial Neural Networks (ANN)</td>
<td>23</td>
</tr>
<tr>
<td>3.2 An Early Neural Network Model: The Perceptron</td>
<td>24</td>
</tr>
<tr>
<td>3.3 Multilayer Neural Networks</td>
<td>26</td>
</tr>
<tr>
<td>3.4 Backpropagation Network (BPN)</td>
<td>29</td>
</tr>
<tr>
<td>3.5 Radial Basis Network (RBN)</td>
<td>34</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>PAGE</td>
</tr>
<tr>
<td>---------</td>
<td>------</td>
</tr>
<tr>
<td>4. ANALYSIS OF CHROMATOGRAPHIC DATA FOR ARTIFICIAL NEURAL NETWORKS</td>
<td>39</td>
</tr>
<tr>
<td>4.1 Data Interpretation</td>
<td>39</td>
</tr>
<tr>
<td>4.2 Determining the Network Input Data</td>
<td>45</td>
</tr>
<tr>
<td>4.3 Data Organization for Neural Network Training</td>
<td>53</td>
</tr>
<tr>
<td>5. RESULTS OF ARTIFICIAL NEURAL NETWORK DESIGN IN AROCLOR CONCENTRATION DETERMINATION</td>
<td>55</td>
</tr>
<tr>
<td>5.1 Introduction</td>
<td>55</td>
</tr>
<tr>
<td>5.2 Optimization of the Backpropagation Network</td>
<td>57</td>
</tr>
<tr>
<td>5.2.1 Selection of the Number of Hidden Nodes</td>
<td>57</td>
</tr>
<tr>
<td>5.2.2 Network with Two Hidden Layers</td>
<td>59</td>
</tr>
<tr>
<td>5.2.3 Utilizing Network Ensembles</td>
<td>62</td>
</tr>
<tr>
<td>5.3 Optimization of the Radial Basis Network</td>
<td>65</td>
</tr>
<tr>
<td>5.3.1 Number of Hidden Nodes Determination</td>
<td>65</td>
</tr>
<tr>
<td>5.3.2 Spread Constant Estimation</td>
<td>66</td>
</tr>
<tr>
<td>6. ARTIFICIAL NEURAL NETWORK PERFORMANCE ANALYSIS</td>
<td>71</td>
</tr>
<tr>
<td>6.1 Introduction</td>
<td>71</td>
</tr>
<tr>
<td>6.2 ANN Performance Across the Developmental Data Set</td>
<td>71</td>
</tr>
<tr>
<td>6.3 Effect of Scaled Test Data on ANN Performance</td>
<td>79</td>
</tr>
<tr>
<td>6.4 Comparison of RBN to Existing Methods</td>
<td>83</td>
</tr>
<tr>
<td>CHAPTER</td>
<td>PAGE</td>
</tr>
<tr>
<td>--------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.5 Calibration of the ANN Using Linear Regression Techniques</td>
<td>89</td>
</tr>
<tr>
<td>6.6 Compatibility of the ANN with Data from Other GC Columns</td>
<td>93</td>
</tr>
<tr>
<td>7. SUMMARY AND CONCLUSIONS</td>
<td>97</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>101</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>106</td>
</tr>
<tr>
<td>VITA</td>
<td>123</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Evaluation of Ensemble Performance.</td>
<td>64</td>
</tr>
<tr>
<td>5.2 Evaluation of Optimal ANN.</td>
<td>70</td>
</tr>
<tr>
<td>6.1 Comparison of RBN Results for Scaled Data and Unscaled Data.</td>
<td>82</td>
</tr>
<tr>
<td>6.2 Calibration Results from Sample 28.</td>
<td>93</td>
</tr>
<tr>
<td>6.3 Evaluation of RBN Performance on Data from the HP db-608 and the HP db-1701.</td>
<td>96</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td></td>
</tr>
<tr>
<td>1.1</td>
<td>Illustration of the use of Artificial Neural Networks (ANN) to determine Aroclor concentrations from a gas chromatogram input.</td>
</tr>
<tr>
<td>2.1</td>
<td>Schematic of a gas chromatograph with a single column, a single detector and a single oven. a-Gas source b-Pressure controller, c-Pressure gauge, d-Gas sampling valve, e-Liquid sampling port, f-Column, g-Detector, h-Detector, θ-Oven.</td>
</tr>
<tr>
<td>2.2</td>
<td>Gas chromatogram featuring 400 ppb each of Aroclors 1242, 1254, and 1260.</td>
</tr>
<tr>
<td>2.3</td>
<td>Gas chromatogram featuring 400 ppb of Aroclor 1242.</td>
</tr>
<tr>
<td>2.4</td>
<td>Gas chromatogram featuring 400 ppb of Aroclor 1254.</td>
</tr>
<tr>
<td>2.5</td>
<td>Gas chromatogram featuring 400 ppb of Aroclor 1260.</td>
</tr>
<tr>
<td>2.6</td>
<td>A close-up of the chromatogram featuring 400 ppb of each of the three Aroclors shows overlapping peaks due to the presence of multiple Aroclors and high concentration contents.</td>
</tr>
<tr>
<td>3.1</td>
<td>Illustration of a perceptron where $x_i$ are the input values, $y$ is the output, and $w_i$ are the weights associated with each input.</td>
</tr>
<tr>
<td>3.2</td>
<td>Single processing node with output $y_j$, bias $\theta_j$, and weighted inputs $w_{ij}$.</td>
</tr>
</tbody>
</table>
FIGURE

3.3 A multilayer network in which $g_i$, $h_i$, and $o_i$ represent unit activation levels of input, hidden, and output units. Weights on connections between the input and hidden layers are denoted by $w_{1ij}$, while weights on connections between the hidden and output layers are denoted by $w_{2jk}$. .................................. 28

3.4 The tan-sigmoid transfer function used in a BPN. ................. 30

3.5 Single linear output radial basis network. ............................ 35

3.6 Transfer function used in a radial basis network. .................. 36

3.7 The effect of the spread constant's value on the output of a radial basis neuron. .................................................. 38

4.1 A typical chromatogram seen in the determination of Aroclor concentrations (10,000 time points). ................................. 40

4.2 Picture indicating the transformation from a raw chromatogram to a peak area graph. ............................................... 41

4.3 Graph of DCBP retention marker areas indicating the variance in peak areas across the data set. ............................... 44

4.4 Portion of plot for each of the standards for Aroclor 1242. The upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb. ........................................ 47
4.5 Portion of plot for each of the standards for Aroclor 1260. The upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.

4.6 Portion of plot for each of the standards for Aroclor 1254 and Aroclor 1242. In each, the upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.

4.7 Portion of plot for each of the standards for Aroclor 1254 and Aroclor 1260. In each, the upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.

5.1 The optimization of the BPN by varying the number of nodes in the hidden layer from 3 to 23.

5.2 Family of curves indicating the optimization of the BPN by varying the number of nodes in each hidden layer.

5.3 A graphical evaluation of BPN performance on each of the three Aroclors with different levels of ensembling. The graphs show the BPN estimation error on each sample in the data set.

5.4 Plot showing the progression toward satisfying the SSQ error while iteratively adding new radial basis nodes. One epoch equals an additional node.
<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5 The optimization of the RBN by varying the spread constant, $\sigma$, from 0.5 to 15.0 in 0.1 increments.</td>
<td>68</td>
</tr>
<tr>
<td>5.6 A precise evaluation of the spread constant about a) $\sigma = 2.0$, b) $\sigma = 2.6$, and c) $\sigma = 4.2$.</td>
<td>69</td>
</tr>
<tr>
<td>6.1 Absolute error analysis for the BPN Aroclor concentration estimation on each sample during the hold-one-out evaluation.</td>
<td>74</td>
</tr>
<tr>
<td>6.2 Absolute error analysis for the RBN Aroclor concentration estimation on each sample during the hold-one-out evaluation.</td>
<td>75</td>
</tr>
<tr>
<td>6.3 An illustration of the contribution of multiple Aroclors to individual peaks. The solid line is from a chromatogram of a sample with 400 ppb for each of Aroclor 1242, 1254, and 1260. The dashed line is from a chromatogram of a sample with 400 ppb of Aroclor 1242 only. The dash-dot line is from a chromatogram of a sample with 400 ppb of Aroclor 1254 only. The dotted line is from a chromatogram of a sample with 400 ppb of Aroclor 1260 only.</td>
<td>77</td>
</tr>
<tr>
<td>6.4 Absolute error analysis for the BPN Aroclor concentration estimation on each sample during the hold-one-out evaluation. Three trials were performed with a different scale factor on the peak area inputs of the test sample.</td>
<td>81</td>
</tr>
</tbody>
</table>
6.5 An RMS error comparison of the RBN versus other existing methods for predicting Aroclor 1242 concentrations. The other methods in the comparison are principal component regression (PCR), multiple linear regression (MLR), and the Target3 (T3) software package. 86

6.6 An RMS error comparison of the RBN versus other existing methods for predicting Aroclor 1254 concentrations. The other methods in the comparison are principal component regression (PCR), multiple linear regression (MLR), and the Target3 (T3) software package. 86

6.7 An RMS error comparison of the RBN versus other existing methods for predicting Aroclor 1260 concentrations. The other methods in the comparison are principal component regression (PCR), multiple linear regression (MLR), and the Target3 (T3) software package. 87

6.8 A plot of the residuals for each of the three Aroclors with error bars showing 95 percent confidence intervals on the residuals. . . . . . . . . . 91

6.9 A hold-one-out evaluation of RBN error performance on data samples run through a different GC column (Hewlett Packard db-1701). 95

A.1 SAMPLE 1: Ar1242: 50 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb . . . 108

A.2 SAMPLE 2: Ar1242: 100 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb . . 108

A.3 SAMPLE 3: Ar1242: 200 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb . . 108

A.4 SAMPLE 4: Ar1242: 400 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb . . 109

A.5 SAMPLE 5: Ar1242: 800 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb . . 109

A.6 SAMPLE 6: Ar1242: 0 ppb, Ar1254: 50 ppb, Ar1260: 0 ppb . . 109
A.7 SAMPLE 7: Arl242: 0 ppb, Arl254: 100 ppb, Arl260: 0 ppb ..... 110
A.8 SAMPLE 8: Arl242: 0 ppb, Arl254: 200 ppb, Arl260: 0 ppb ..... 110
A.9 SAMPLE 9: Arl242: 0 ppb, Arl254: 400 ppb, Arl260: 0 ppb ..... 110
A.10 SAMPLE 10: Arl242: 0 ppb, Arl254: 800 ppb, Arl260: 0 ppb ..... 111
A.11 SAMPLE 11: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 50 ppb ..... 111
A.12 SAMPLE 12: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 100 ppb ..... 111
A.14 SAMPLE 14: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 400 ppb ..... 112
A.15 SAMPLE 15: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 800 ppb ..... 112
A.16 SAMPLE 16: Arl242: 50 ppb, Arl254: 50 ppb, Arl260: 50 ppb ..... 113
A.17 SAMPLE 17: Arl242: 100 ppb, Arl254: 100 ppb, Arl260: 100 ppb ..... 113
A.19 SAMPLE 19: Arl242: 400 ppb, Arl254: 400 ppb, Arl260: 400 ppb ..... 114
A.21 SAMPLE 21: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 800 ppb ..... 114
A.23 SAMPLE 23: Arl242: 800 ppb, Arl254: 0 ppb, Arl260: 0 ppb ..... 115
A.26 SAMPLE 26: Arl242: 50 ppb, Arl254: 0 ppb, Arl260: 100 ppb ..... 116
<table>
<thead>
<tr>
<th>FIGURE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.29 SAMPLE 29: Arl242: 100 ppb, Arl254: 800 ppb, Arl260: 400 ppb</td>
<td>117</td>
</tr>
<tr>
<td>A.30 SAMPLE 30: Arl242: 400 ppb, Arl254: 800 ppb, Arl260: 0 ppb</td>
<td>117</td>
</tr>
<tr>
<td>A.33 SAMPLE 33: Arl242: 50 ppb, Arl254: 800 ppb, Arl260: 0 ppb</td>
<td>118</td>
</tr>
<tr>
<td>A.34 SAMPLE 34: Arl242: 50 ppb, Arl254: 200 ppb, Arl260: 800 ppb</td>
<td>118</td>
</tr>
<tr>
<td>A.35 SAMPLE 35: Arl242: 0 ppb, Arl254: 400 ppb, Arl260: 50 ppb</td>
<td>119</td>
</tr>
<tr>
<td>A.37 SAMPLE 37: Arl242: 200 ppb, Arl254: 0 ppb, Arl260: 0 ppb</td>
<td>120</td>
</tr>
<tr>
<td>A.38 SAMPLE 38: Arl242: 0 ppb, Arl254: 200 ppb, Arl260: 0 ppb</td>
<td>120</td>
</tr>
<tr>
<td>A.40 SAMPLE 40: Arl242: 200 ppb, Arl254: 0 ppb, Arl260: 0 ppb</td>
<td>121</td>
</tr>
<tr>
<td>A.41 SAMPLE 41: Arl242: 0 ppb, Arl254: 200 ppb, Arl260: 0 ppb</td>
<td>121</td>
</tr>
<tr>
<td>A.42 SAMPLE 42: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 200 ppb</td>
<td>121</td>
</tr>
<tr>
<td>A.43 SAMPLE 43: Arl242: 200 ppb, Arl254: 0 ppb, Arl260: 0 ppb</td>
<td>122</td>
</tr>
<tr>
<td>A.44 SAMPLE 44: Arl242: 0 ppb, Arl254: 200 ppb, Arl260: 0 ppb</td>
<td>122</td>
</tr>
<tr>
<td>A.45 SAMPLE 45: Arl242: 0 ppb, Arl254: 0 ppb, Arl260: 200 ppb</td>
<td>122</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Statement of the Problem

The analysis of gas chromatography (GC) has been used in many applications. Of specific interest in the environmental restoration field is composition analysis of chemical samples. Environmental laboratories prefer GC as the primary technique for sample analysis due to reduced complexity, lower equipment costs, and automation of sample preparation. Analyzing chromatograms for Aroclors, petroleum hydrocarbons, and other multi-component materials are among the most difficult analyses to perform. Current data interpretation practice for an unknown Aroclor sample requires a human comparison of the unknown chromatogram with one obtained from a standard sample [9].

Linear regression is the primary method used by chemists for extracting pertinent chemical knowledge from a chromatogram. This technique assumes that the properties of the compounds are linearly related to the properties of the calibrated standards of the compound of interest. This method works well when the GC detector response is linear, the baseline variations are minimal, and only a single
compound with a unique pattern exists. In practice, however, many samples do not meet these ideal conditions and alternate methods must be used to obtain accurate quantization.

Many of the GC samples to be analyzed contain multiple components (Aroclors) and may have background interferences. The result is a complex chromatogram which might have nonlinear data features, making it difficult to quantify the contributing components. The nonlinear features can be attributed to a nonlinear GC detector response and a difficulty in calculating peak areas due to overlapping peaks. A method which can compute an accurate result in the presence of these conditions is necessary in order to analyze these samples.

The purpose of this thesis is to demonstrate the optimal use of artificial neural networks (ANN) as an effective tool for determining Aroclor concentrations in environmental soil samples from chromatogram data, a process which is illustrated in Fig. 1.1. Peak area data are derived from each chromatogram and from this data the ANN is trained with a signature vector chosen from the chromatogram consisting of a set of predetermined peaks. The signature peaks are selected based on such characteristics as isolation from other peaks and the contribution (or lack of) from each of the Aroclors. The ANN is trained to associate the vector patterns with their respective chromatographic profiles and then determine the specific concentration values of samples of unknown Aroclor concentrations.
Figure 1.1: Illustration of the use of Artificial Neural Networks (ANN) to determine Aroclor concentrations from a gas chromatogram input.

Artificial neural networks such as backpropagation networks (BPN) and radial basis networks (RBN) have been shown to be suitable for quantitative analysis of data from sources similar to chromatograms. We have developed and optimized both a BPN and an RBN for the Aroclor concentration estimation task. The primary advantage of using an ANN lies in the structure of the network, which enables both the linear and nonlinear relationships in the data to be represented in the analysis methodology.

A critical issue in network learning is the estimation error. The estimation
error may be affected by two factors. The first factor is the neural network architecture which includes the number of hidden nodes, the number of hidden layers, and the values of the learning parameters. The second factor affecting the estimation error is related to the training set which includes the number of training samples, inaccuracies of the input data, and data preprocessing.

An extensive numerical analysis was performed in this study. This thesis will address the following issues as applied to the Aroclor concentration estimation problem:

1. Preprocessing of training patterns based on the data features.

2. Selection of the number of hidden nodes and hidden layers.

3. Determination of learning parameters.

4. Modification of the network configurations.

5. Analysis of the sensitivity of network estimation to variations in the input pattern.

6. Comparison of results for the RBN and the BPN.

7. Comparison of results between the ANN and existing methods.

8. Demonstration of the ability to calibrate the ANN and calculate a confidence error using linear regression techniques.
9. Description for the adaptability of the ANN in the analysis of other GC data.

1.2 Scope of the Current Application

The ANN, the subject of this thesis, is being developed as part of an expert system for automated analysis of gas chromatograms for Aroclors. Aroclors are commercially produced polychlorinated biphenyl (PCB) mixtures manufactured by chemical manufacturers such as Monsanto. Commercial chromatographic data analysis methods (Target3) have already been incorporated into the expert system as well as other methods such as principal component regression (PCR) and multiple linear regression (MLR). The strategy of the expert system is to use knowledge from all methods (ANN, Target3, PCR, and MLR) and knowledge generated about the sample to create a final determination about the Aroclor concentrations. In this application the concentration of three commercially developed PCBs (Aroclor 1242, Aroclor 1254, and Aroclor 1260) are targeted in soil samples for concentration determination [9].

The U.S. Department of Energy (DOE) has significant amounts of radioactive and hazardous waste stored, buried, and still being produced at many sites around the country. The requirements for sampling and analyzing this waste will
increase sharply as DOE is required to characterize and remediate these sites. Currently, DOE-associated laboratories perform 2 to 3 million chemical, biological, and radiological determinations per year. The goal of the Contaminant Analysis Automation (CAA) program within the DOE is to reduce the cost of the characterization effort by automating the sample analysis procedure [9]. The scope of this thesis is the development of the ANN as a part of the Data Interpretation Module (DIM) in conjunction with CAA.

The DIM is one of several standard laboratory modules (SLM®) incorporated into the CAA project. An SLM® is defined as a logical grouping of laboratory operations which perform a subtask of an analytical protocol. The SLM® paradigm requires the modules to conform to a set of specifications that will allow them to become building blocks of fully automated systems [9]. The DIM has two primary functions: (1) an on-line analysis of data under the laboratory Task Sequence Controller (TSC) and (2) an off-line functionality that provides the chemist with the tools needed to build automated data analysis methods, build calibration functions, and review the analysis process. The focus of this research is to build an effective ANN system that will contribute to both functions.

Currently, data interpretation of an unknown Aroclor sample requires human comparison of the unknown chromatogram with that obtained from a standard sample. Examination of the chromatograms in Chapter 2 illustrates that this is
not a trivial task. Typically, the chemist selects one to five peaks which he chooses as characteristic of the individual Aroclors. Linear regression techniques in commercially developed software such as Target3 are then used to develop calibration functions from a series of standard samples. Once manual identification of the Aroclor component in the unknown chromatogram is complete, the calibration functions applicable to that Aroclor are used to determine the concentration [10]. Using the current data interpretation process, errors in Aroclor concentration estimation typically range from 20 to 600 percent [9].

1.3 Review of Related Work

Most pattern recognition applications make excellent candidates for an ANN-based system. ANN have been proven effective in a variety of pattern recognition tasks, including handwritten character recognition [3], medical diagnosis [31], and the classification of sonar signals and radar waveforms [42].

Analysis of chromatograms using pattern recognition techniques has been effective in determining the composition of complex chemical and environmental samples. Specifically, multivariate analysis techniques have been used to chromatographically classify such substances as fuel samples [25], essential oils [33], and whiskeys [43].
ANN are not only effective in pattern classification, but also in signal component quantization. Automated identification of signal data is an application for which an ANN-based system is well suited. ANN-based systems have been used successfully to classify spectra from various modalities, including infrared spectroscopy [40], mass spectrometry [38], and Raman spectroscopy [26]. A paper by Majcen et al. [28] compares the use of ANNs to multivariate linear regression in the quantitative prediction of “total color difference” from complex oxide concentration measurements. In this study the quantitative results of the ANN compare favorably to the results of the multivariate linear regression.

Most of the current research in multivariate GC analysis has been in the use of modified linear methods — particularly PCR. Geiser et al. [13] use PCR to correlate, visualize, and verify parameter relationships in free-fatty-acid model systems. Singh [39] discusses the use of PCR as an approach for estimating multivariate decision and detection limits for GC data. Lavine et al. [25] use PCR to establish relationships between chemical constitution and environmental variables in jet fuels. Hida et al. [20] experiment with PCR and cluster analysis for the quantitative analysis of sodium dodecylbenzene sulfonate and polyoxyethylene lauryl ether mixtures. They report their results from either method as only being “in fair agreement with theoretical values.” [20]. All of these PCR applications effectively quantize chemical components one at a time.
The GC data of chemical mixtures is nonlinear. This presents a problem when using linear methods for the quantitative analysis of the GC mixture data. A method for quantitative analysis of the nonlinear data would be more appropriate. For this reason, research into the use of ANNs for multivariate GC analysis is growing in popularity. Gemperline et al. [14] compare the use of ANN and PCR in detecting and modeling nonlinear regions of spectral response in multivariate, multicomponent spectroscopic assays. They established that ANN can be used to develop nonlinear calibration models that perform better than PCR when a nonlinear response is present. Francelin et al. [12] use ANNs for the classification of vegetable oils using GC data. While the data contains mixtures of the oils, the ANNs developed only recognize the presence of the oils and do not quantitize them.

The BPN has been central to most of the current study in the area of interest such as mass spectrometry and Raman spectroscopy. Research for Aroclor concentration estimation, however, is not being limited to BPN only. Another multilayered feedforward network, the RBN, exhibits features which are found to be quite appropriate for the analysis of chromatographic data. Although a relatively new concept in ANN research, the RBN has already been shown to be at least as effective as the BPN in some applications such as handwritten character recognition [32] and ceramic strength and density prediction [6]. The ceramics
application is similar to our application in that the RBN was not used as a mere classifier but instead was used to predict a ceramic's strength and density based on input parameters such as milling time, sintering time, and sintering gas pressure.

In addition to the current manual analysis of chromatograms, several more sophisticated algorithms have been implemented for Aroclor concentration determination as part of the CAA program. Target3 is used as part of the manual analysis. It is an integrated two-dimensional and three-dimensional gas chromatography (GC), liquid chromatography, and mass spectrometry data processing, data review, and data reporting software package [9]. Target3 was developed to estimate a chemical's concentration in a sample where only one analyte is present. It is not intended to solve the nonlinear problem where more than one analyte is present in a sample. Other techniques such as principal component regression (PCR) [21] and multiple linear regression [37] have been used with more success but do not handle nonlinearities in the data very well. Recently, the interest in applying ANN in this area has been steadily increasing [27].

1.4 Synopsis

This thesis is organized in seven chapters. Chapter 1 introduces the problem related to the present application and reviews pertinent literature. The principles
of GC and the features of chromatographic data are described in Chapter 2 along with a feasibility description for a neural network approach. Chapter 3 contains a general description of ANN and an in-depth theoretical and mathematical analysis of the BPN and RBN algorithms used in this research. Data interpretation and preprocessing issues are discussed in Chapter 4. The method of data organization is discussed in Chapter 4 as well. Extensive numerical studies related to optimizing the BPN and RBN architectures are described in Chapter 5. The optimal ANN structure for Aroclor concentration estimation is discussed in this chapter. Chapter 6 presents an analysis of a variety of ANN performance issues. First, a numerical study is performed to investigate the ANN performance across the developmental data set. Some of the conclusions from this study are supported by an analysis of the effects of scaled test data on ANN performance. The results of the optimal ANN architecture are compared with the results of existing methods for the Aroclor concentration estimation task. A discussion on the ability to calibrate the ANN using linear regression techniques is presented. Chapter 6 concludes with a numerical analysis of a compatibility study of the ANN with data from other GC columns. Chapter 7 is an outline of the conclusions from this research.
CHAPTER 2

DATA STRUCTURE AND NEURAL NETWORK

FEASIBILITY

2.1 Basic Principles of Gas Chromatography

Gas chromatography (GC) is a powerful separation technique for gas and vapor mixtures. The combination of separation and on-line detection permits accurate quantitative analysis of complex mixtures, including traces of compounds down to parts per billion. GC is important in quality control and process monitoring in the chemical and drug industry, in clinical analysis, and in environmental pollution investigations [16].

Organic compounds are separated due to differences in the partitioning behavior between the mobile gas phase and the stationary phase in a GC column. The system's design is made so that the rate constant of the kinetics of mass transfers between the two phases is maximized. Fast exchanges between the two phases is critical to achieve a high separation efficiency. Using gas as the mobile phase in chromatography was first suggested as an analytical method by Martin and Synge in 1941 [30].
Modern columns are swept by a constant flow-rate stream of the mobile phase (a gas) that carries the sample to be analyzed. The stream is forced under pressure through a column and an on-line detector. This requires a source of carrier gas and a flow-rate controller. The flow meter is a convenient but not necessary auxiliary device. The column must be heated at constant temperature, or alternatively, the column temperature may be programmed following a reproducible method. For this method, an oven with a temperature controller is needed. Devices permitting the injection of a sample of convenient size into the gas stream and on-line detectors are provided. The signal of this detector is handled by electronic circuits and the resulting chromatogram is recorded and displayed. A schematic of a chromatography system is shown in Fig. 2.1. A more detailed description of the GC process is given by Guiochon and Guillemin [16].

For those (humans or machines) who know how to read the results of a chromatogram, it tells which compounds are present in the analyzed sample, and their concentrations. The chromatogram also gives information on the thermodynamics and kinetics of the molecular interactions used as a basis for the retention mechanisms involved. GC is a rather inexpensive and fast separation method that uses small samples and is applicable to any substance which has a vapor pressure exceeding 1 cm mercury at a temperature at which a stationary phase can be found.
2.2 Features of Chromatographic Data

The utilized output of a GC system is the chromatogram. The chromatogram contains information about the presence of chemical compounds (congeners) in each sample. The presence of a peak at a particular time on the chromatogram indicates the presence of a particular congener. The height of a particular peak indicates the quantity of a particular congener. Some congeners are unique to one Aroclor, but most are shared by two or more Aroclors.
Chemical samples containing various mixtures of Aroclors 1242, 1254, and 1260 were used in the basic study of GC. Figure 2.2 shows a chromatogram of a mixture of the three Aroclors in a 1:1:1 ratio at a 400 ppb concentration of each. The two large peaks on either end of the chromatogram are retention markers. The chemicals used for the retention markers are specifically chosen so that all pertinent information in the chromatogram will lie between the times at which the GC detects the retention markers. For our data set the retention markers occur at the approximate times of 7 minutes and 31 minutes. The retention times are determined by the set-up of individual GC systems.

It is important to note that specific peaks in the chromatogram have a relationship to specific chemical components. For example, the occurrence of peaks located between the approximate range of 9 to 18 minutes in Fig. 2.3 is characteristic of Aroclor 1242. Similarly, the occurrence of peaks located between the approximate range of 13 to 25 minutes in Fig. 2.4 is characteristic of Aroclor 1254 and the peaks located between the range of 16 to 30 minutes in Fig. 2.5 are characteristic of Aroclor 1260. A comparison between Fig. 2.2 and Figs. 2.3–2.5 gives some insight on the overlapping that occurs between the Aroclors.

Obviously, the relationship between the chemical components and time range may not always be one-to-one. Many peaks are shared by more than one Aroclor. Because many peaks are shared, the peaks are sometimes quite large when one
Figure 2.2: Gas chromatogram featuring 400 ppb each of Aroclors 1242, 1254, and 1260.
Figure 2.3: Gas chromatogram featuring 400 ppb of Aroclor 1242.
Figure 2.4: Gas chromatogram featuring 400 ppb of Aroclor 1254.
Figure 2.5: Gas chromatogram featuring 400 ppb of Aroclor 1260.
or more of the Aroclors is present, especially at higher concentrations. When the peaks are large and/or are in close proximity there are often overlaps. Figure 2.6 is a close-up view of a region of the chromatogram in Fig. 2.2. Here, the concentration of all three Aroclors is high (400 ppb each) which contributes to a large amount of overlapping between peaks. This overlap contributes to a nonlinearity in the relationship between chromatographic data and sample composition, thus reducing the sensitivity of linear composition techniques.

![Raw Chromatogram Data](image)

Figure 2.6: A close-up of the chromatogram featuring 400 ppb of each of the three Aroclors shows overlapping peaks due to the presence of multiple Aroclors and high concentration contents.

The overlap of portions of some peaks is not only a result of high concentrations of Aroclors, but is also the result of the GC system and the natural close
proximity of some of the congeners present. The relative distance between retention times is determined by parameters set by the chemist such as temperature and column choice. Longer retention times yield more resolution in the chromatogram. Overlapping peaks cause peak area estimation problems because of the lack of a true baseline.

2.3 Feasibility of a Neural Networks Approach

Artificial neural networks (ANN) have been successfully applied in many areas during the past few years. In this research, a neural network's methodology is applied to gas chromatogram data to produce quantitative estimation of chemical components. Multilayer feedforward networks were chosen to estimate the concentration of chemical components of environmental samples from chromatogram data. The feasibility of using multilayer feedforward networks for this application was studied based on the following observations:

1. ANN are very effective in relating quantities for which a physical or empirical model is not fully described.

2. The given data set requires a supervised learning model.

3. Multilayer feedforward neural networks are the most suitable network for generating nonlinear relationships for a given problem.
4. Pattern-mapping is the most suitable problem to model using multilayer feedforward networks.

5. Preliminary tests using the backpropagation network and radial basis network showed encouraging estimation results.

A supervised learning algorithm uses a set of selected values taken from the chromatographic data as input to the ANN. The output is a vector of concentrations from chemical components. First, the ANN will be presented a sequence of inputs and a set of expected outputs in order to learn the functional relationship between the input and output vectors. In this research, the training data pair is a sequence of GC peak areas as inputs and concentrations of the three Aroclors as outputs. Once the network is trained, it will estimate a sample's concentration by presenting peak area data of an unknown GC sample. In order to build a good model that can interpolate concentrations of future unknown samples, the training data must completely span the domain of interest. The GC data set used came from Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee. Although the data set is relatively small (45 samples), it amply covers our range of interest.
CHAPTER 3

TECHNIQUES EMPLOYED IN NEURAL NETWORK LEARNING

3.1 General Description of Artificial Neural Networks (ANN)

Knight [23] said "In our quest to build intelligent machines, we have one naturally occurring model: the human brain. It follows that one could achieve artificial intelligence by simulating the functioning of the human brain on a computer." The idea for building an intelligent machine out of artificial neurons is not a new one. Inspired by neurophysiologists such as Donald Hebb, work in the field of neural networks began in the 1940s [29]. By the early 1970s, however, research virtually came to a halt when the networks under study were shown to be weak computationally [23]. In the past decade, there has been a resurgence of interest in neural networks. There are several reasons for this, including the development of faster digital computers on which to simulate larger networks, interest in building massively parallel computers, and most importantly, the discovery of powerful new network learning algorithms.

For the most part, these neural network architectures are not meant to du-
plicate the operation of the human brain, but rather to receive inspiration from known facts about how the brain works. They can be characterized by several features: large numbers of simple neuron–like processing elements; large numbers of weighted connections between the elements in which the weights on the connections encode the network’s knowledge; highly parallel, distributed control; and emphasis on learning internal representations automatically. The weighted connections are modeled after synapses found in the human brain. A brain neuron receives signals from other neurons through synapses. These synapses regulate how much of each incoming signal passes into the neuron. The signals are added together inside the neuron and when enough signal energy is present, cause the neuron to fire a signal out to the other neurons. ANN work in much the same way.

3.2 An Early Neural Network Model: The Perceptron

The perceptron, an invention of Frank Rosenblatt, was one of the earliest neural network models [23]. A perceptron models a neuron by taking a weighted sum of its inputs and transmits as output a “1” if the sum is greater than some adjustable threshold value. Otherwise it transmits a “0”. The inputs \((x_1, x_2 \cdots x_n)\) and connection weights \((w_1, w_2 \cdots w_n)\) are typically real values, both positive and negative. Learning is the process of modifying the values of the weights and the threshold. If the presence of some feature \(x_i\) tends to cause the perceptron to
fire, the weight $w_i$ will be positive; if the feature $x_i$ inhibits the perceptron, the weight $w_i$ will be negative. The perceptron itself consists of the weights, the summation processor, and the adjustable threshold processor. A simple perceptron network, therefore, consists of only two layers: an input layer and the perceptron layer which is shown in Figure 3.1. The output of the network is simply the binary output of the perceptron.

Figure 3.1: Illustration of a perceptron where $x_i$ are the input values, $y$ is the output, and $w_i$ are the weights associated with each input.

While the Perceptron Convergence Theorem [23] guarantees correct classification of linearly separable data, most problems don’t provide such nice data. Here, an approach with two separate perceptron stages is used. The output of the first
perceptron stage becomes the input to the next stage. Using this idea, a multilayer perceptron with the ability to solve nonlinear problems can be constructed. This concept, however, introduces a serious learning problem—the Convergence Theorem does not extend to multilayer perceptrons. The perceptron learning algorithm can correctly adjust weights between inputs and outputs, but it cannot adjust weights between perceptron layers [23].

Despite the identification of this important research problem, actual research in perceptron learning came to a halt in the 1970s [23]. The field saw little interest until the 1980s, when several learning procedures for multilayer perceptrons were proposed.

### 3.3 Multilayer Neural Networks

The ability to train multilayer networks is an important step in building intelligent machines out of neuron-like components. The following is a discussion of a subclass of multilayer networks, namely fully connected, layered, feedforward networks. Nodal transfer function activation flows from the input layer through a hidden layer, then to the output layer. A typical processing node is shown in Fig. 3.2 and has a functional description given by

\[
y_j = f \left( \sum_{i=1}^{N} w_{ij} x_i + \theta_j \right),
\]
where $x_i$ is one of the $N$ inputs to processing node $j$, $w_{ij}$ is the connection weight between node $i$ and node $j$, $\theta_j$ is the bias for node $j$, and $y_j$ is the output from node $j$.

![Figure 3.2: Single processing node with output $y_j$, bias $\theta_j$, and weighted inputs $w_{ij}$](image.png)

Each neuron in one layer is connected in the forward direction to every nodal unit in the next layer. The network knowledge is encoded in the weights on connections between units. Typically, several similar units are combined in order to form a useful ANN topology as shown in Fig. 3.3. This network has three layers, although it is possible and sometimes useful to have more. Each layer of a network can be expressed in matrix form as

$$y = f(w \cdot x + \theta)$$
where $y$ is a column vector equal in length to the number of nodes ($M$) in the present layer, $x$ is a column vector equal in length to the number of inputs from the previous layer, $w$ is a $N \times M$ matrix of connection weights, and $\theta$ is a column vector of length $M$.

The existence of hidden units allows the network to develop complex feature detectors, or internal representations. Most importantly, the behavior of these

Figure 3.3: A multilayer network in which $g_i$, $h_i$, and $o_i$ represent unit activation levels of input, hidden, and output units. Weights on connections between the input and hidden layers are denoted by $w_{1ij}$, while weights on connections between the hidden and output layers are denoted by $w_{2jk}$.
hidden units is automatically learned, not preprogrammed. When using a neural network to solve a problem, the network should learn the input/output relationships it is trained on, and be able to generalize an output based on inputs it has not seen yet. According to Maren et al. [29], with the proper network topology and transfer functions, an ANN can approximate any function well.

3.4 Backpropagation Network (BPN)

One way to train a multilayer feedforward network is to employ the backpropagation training algorithm. The backpropagation training algorithm was first proposed by Rumelhart et al. [36] and is now the most widely used algorithm for supervised learning in multilayer neural networks. The goal of such an algorithm is to teach the network to associate specific input patterns (e.g., chromatogram peak area data) with their corresponding output patterns or values (e.g., Aroclor concentrations) by adjusting the connection weights in order to minimize the error between target output and actual output of the network. A gradient descent algorithm is generally used to perform the optimization [19]. The nodes in a BPN require a monotonically increasing, differentiable activation function. A sigmoidal function is used in this research. This function transmits a continuous output between 0 and 1.
Each node in a BPN sums up its weighted inputs, but unlike the basic perceptron, it produces a real value between 0 and 1 as output, based on the sigmoid function in Fig. 3.4. The key attribute of this transfer function (and many others used in modern ANN) is the nonlinearity in part of the input range. Using this transfer function, if $W$ is a matrix of weights and $x$ is the vector output from the previous layer, the equation for a node's output, $o_j$, is given by

$$o_j = \frac{1}{1 + e^{-\sum_{i=1}^{a} W_{ij} x_j}},$$

where $a$ is equal to the number of nodes on the previous layer, $i$ ranges from 1 to $a$, and $j$ ranges from 1 to the number of nodes on the current layer.

Figure 3.4: The tan-sigmoid transfer function used in a BPN.
During the training process, a BPN typically starts out with a random set of weights. The network adjusts its weights each time it is trained with an input/output data pair. Each pair requires a forward pass and a backward pass. The forward pass begins with the presentation of an input vector and a target output vector. The nodal calculations (activations) are propagated from the input layer to the hidden layer using the sigmoid activation function of Fig. 3.4 modeled by the equation given above. Activations are then propagated from the units in the hidden layer to the units in the output layer using the same function and equation.

During the backward pass, the network’s actual output \((o_j)\) is compared to the target output \((y_j)\) and error estimates are computed for the output units denoted as \(\delta_2_j\):

\[
\delta_2_j = o_j(1 - o_j)(y_j - o_j),
\]

for all \(j = 1, \ldots, c\), where \(c\) is equal to the number of nodes on the output layer. Similarly, the errors for the units in the hidden layer denoted by \(\delta_1_j\) are calculated.

\[
\delta_1_j = h_j(1 - h_j) \sum_{i=1}^{c} \delta_2_i \cdot w_{2ji},
\]

for all \(j = 1, \ldots, b\), where \(b\) is equal to the number of nodes on the hidden layer, and \(h_j\) are the outputs from the hidden layer. These error formulas are related to the derivative of the activation (sigmoid) function \([23]\).
Once the error estimates have been established between layers, the weights between the layers may be modified. The traditional method uses a gradient descent algorithm to update weight values. For instance, the weights between the hidden layer and output layer may be adjusted based on the error estimate $\delta_2_j$,

$$\Delta w_{2ij} = \eta \cdot \delta_2_j \cdot h_i,$$

for all $i = 1, \ldots, b$, and $j = 1 \ldots c$, where $\eta$ is the learning rate. The learning rate, $\eta$, is a scale factor for determining how far to move in the direction of the gradient. The basic idea here is that each hidden unit tries to minimize the errors of output units to which it connects. Errors are then propagated back to the connections stemming from the input units and the weights between the input layer and the hidden layer are adjusted based on the error estimate $\delta_1_j$.

$$\Delta w_{1ij} = \eta \cdot \delta_1_j \cdot x_i,$$

where $i = 1, \ldots, a$, and $j = 1, \ldots, b$.

A drawback of the gradient descent method is the slow rate at which the weights converge to a minimum error. The Levenberg-Marquardt (LM) method has been used to improve the speed in training the network used in this research [2]. The LM method uses an interpolation between two approaches based on the maximum neighborhood in which the truncated Taylor series gives an adequate representation of the nonlinear model [2].
Instead of using the $\Delta w$ equations from the gradient descent method, the following expressions are used in the LM method for modifying the weights between layers:

$$\Delta w_{2jk} = (J^T \cdot J + \mu \cdot h_j)^{-1} J^T \cdot \delta_{2k},$$

$$\Delta w_{1ij} = (J^T \cdot J + \mu \cdot x_i)^{-1} J^T \cdot \delta_{1j},$$

where $J$ is the Jacobian matrix of derivatives of each error to each weight, $\mu$ is a scaler, $k$ ranges from 1 to the number of nodes in the output layer, $j$ ranges from 1 to the number of nodes in the hidden layer, and $i$ ranges from 1 to the number of nodes in the input layer. The LM method begins the minimization process with a large value for $\mu$, thus approximating the gradient descent method. As the algorithm approaches a solution, $\mu$ decreases. When the value for $\mu$ is small the LM method functions in a Gauss-Newton optimization mode [7], where a search direction is obtained at each iteration. The search direction is a solution of the linear least squares problem [15].

The backpropagation algorithm usually updates its weights incrementally after seeing each input/output pair. After it has seen all the training pairs and adjusted its weights accordingly, we say that one epoch (or iteration) has been completed. We repeat this process for as many epochs as are desired or are necessary to meet a predefined network sum of squares (SSQ) error.

Several issues need to be considered when using the backpropagation algo-
rithm. For example, the optimal number of hidden layers and the corresponding number of nodes in each layer should be decided. These and other means of optimization such as ensembling are studied in Chapter 5.

3.5 Radial Basis Network (RBN)

One disadvantage to using most feedforward layered neural networks is the high degree of nonlinearity in the parameters. Learning must be based on nonlinear optimization techniques (i.e. backpropagation), and the parameter estimate may become trapped at a local minimum of the selected optimization criterion during the learning procedure. An alternative to such neural networks is to use the radial basis function (RBF) as a transfer function.

As pointed out in Chen et al. [5], the strong connection between the RBF and neural networks makes it reasonable to believe that a radial basis network (RBN) can offer approximation capabilities similar to other feedforward, layered neural networks, provided that the hidden layer of the RBN is fixed appropriately. This belief is strongly supported by the theoretical results from the RBF method as a multidimensional interpolation technique [35].

Figure 3.5 shows a general RBN with n inputs and one linear output. The
network performs a mapping \( f: \mathbb{R}^n \rightarrow \mathbb{R} \) with the equation

\[
f(x) = \lambda_0 + \sum_{i=1}^{m} \lambda_i \phi(||x - c_i||),
\]

where \( x \in \mathbb{R}^n \) is the input vector, \( c_i \ (0 \leq i \leq m) \) are the RBF centers, \( \phi(\cdot) \) is a function from \( \mathbb{R}^n \rightarrow \mathbb{R} \), \( || \cdot || \) denotes the Euclidean norm, \( \lambda_i \ (0 \leq i \leq m) \) are the weights of the output node, and \( m \) is the number of centers.

Figure 3.5: Single linear output radial basis network.

Studies have shown that the choice of the nonlinear function \( \phi(\cdot) \) is not crucial
to the network’s overall performance [5]. One of the more common functions used for \( \phi(\cdot) \) is the Gaussian function shown in Fig. 3.6 which is implemented with the formula

\[
\phi(||x - c_i||) = \exp\left(-\frac{||x - c_i||^2}{\sigma_i^2}\right),
\]

where \( \sigma_i \) is the spread constant of the \( i \)-th node. The Gaussian function has a maximum value of 1 when \( ||x - c_i|| \) is 0, and drops off to 0 as \( ||x - c_i|| \) approaches infinity.

An RBN can be regarded as a special network which never requires more than one hidden layer because it allows the different regions of the input space to be represented satisfactorily by different hidden layer nodes. An RBN is actually parameter-linear because all RBF centers and nonlinearities in the hidden layer

Figure 3.6: Transfer function used in a radial basis network.
are fixed. The hidden layer, therefore, performs a fixed nonlinear transformation with no adjustable parameters and maps the input space onto a new space. The output layer is simply a linear combiner applied to this new space, so only the weights of this linear combiner need adjusting. These weights, therefore, can be determined using the linear least squares method.

The performance of an RBN is greatly determined by the chosen centers in each of the RBFs. In practice, the centers are often chosen to be a subset of the training data. The fixed centers should suitably sample the input domain. In this research a systematic approach to the center selection process is taken. The orthogonal least squares method [15] is used to select a suitable set of centers from a larger set of candidates. At each step of the process, the increment toward the success of the desired output is maximized. Additionally, the problems of network oversizing and ill-conditioning that occur in random center selection processes are automatically eliminated. This method provides an efficient learning algorithm for fitting adequate RBFs and is covered in more detail along with experimental support in Chapter 5.

The net input for neurons in most multilayer feedforward networks is the weighted outputs from the neurons of the previous layer. Each radial basis neuron, however, receives as net input the vector distance between the connection weight vector, $c_i$, and the input vector, $x$. As the distance between the weight
vector and the input vector decreases, the output increases. In this way, a radial basis neuron acts as a detector which outputs a value of 1 whenever the input is identical to its weight vector.

The spread constant, $\sigma_i$, determines the width of an area in the input space to which each neuron responds. It is important that the RBF of the hidden layer overlap to allow good generalization. However, the RBF should not be so spread out that the radial basis neurons return outputs near 1 for a large number of the input vectors used. Figure 3.7 demonstrates the output of a single radial basis neuron with its spread constant arbitrarily set to 2.5. Each neuron responds with a 0.5 or more to any input vectors within a vector distance of 2.5 from their weight vector. Determining the value for the spread constant of an RBN is a heuristic process that is described with experimental support in Chapter 5.

Figure 3.7: The effect of the spread constant's value on the output of a radial basis neuron.
CHAPTER 4

ANALYSIS OF CHROMATOGRAPHIC DATA FOR ARTIFICIAL NEURAL NETWORKS

4.1 Data Interpretation

The Oak Ridge National Laboratory (ORNL) supplied us a training set of forty-five chromatograms. Fifteen of these chromatograms represent standards of each of the three Aroclors at five different concentrations (50 ppb, 100 ppb, 200 ppb, 400 ppb, and 800 ppb). Another 21 of the samples consist of various mixtures and concentrations of the Aroclors. The other nine samples represent reproducibility tests of individual Aroclors. The complete collection of chromatograms is included in the Appendix.

The graph in Fig. 4.1 is a typical chromatogram used to determine Aroclor concentrations. While the data in the entire chromatogram is important, it is difficult to train a neural network on the more than 10,000 time points that were collected in the 35-minute chromatogram. A training vector with a dimensionality of anywhere close to 10,000 is quite unreasonable for any neural network. We wanted to reduce the dimensionality of the input vector for use in the neural
network. By reducing the input dimensionality with careful preprocessing to select good features or combinations of features, the available data becomes more practical and manageable [11].

![Raw Chromatogram Data](image)

Figure 4.1: A typical chromatogram seen in the determination of Aroclor concentrations (10,000 time points).

We are not necessarily interested in each data point in the chromatogram, but rather in the area under each of the peaks which make up the chromatogram. The area value under each peak is precalculated using peak area integration software such as the commercially available Target3 or HP ChemStation packages. Such software yields a table containing time and area data that is anywhere from 50 to 100 entries long, depending on the number of peaks the software interprets are
present in the chromatogram. The number of peaks is determined by the number of Aroclors present and the concentration of each. Figure 4.2 is an illustration of the conversion of a peak in the chromatogram to a peak area value. In this figure the area under the highlighted peak in the first graph (the continuous chromatogram) has been calculated and is shown in the second graph (peak area plot) as the highlighted peak area.

![Figure 4.2: Picture indicating the transformation from a raw chromatogram to a peak area graph.](image)

Each gas chromatography (GC) system has different set-up parameters. These parameters affect the actual time at which each chemical compound’s (congener) peak should occur in the chromatogram. Because these parameters vary from
one GC system to another, the time at which each congener's peak occurs in the chromatogram will vary. Fortunately, the times relative to the end markers within a chromatogram are mostly uniform regardless of what the GC set-up is.

The end markers are two artificially added congeners, tetrachlorometaxylene (TCMX) and decachlorobiphenyl (DCBP), that signify the beginning and end of the GC process. These chemicals are usually injected at very high concentrations so that the beginning and end of the process are easily detected in the chromatogram. One can see from the peak area graph of Fig. 4.2 that the peaks labeled TCMX and DCBP are greater than any of the other peak area values.

Because each congener's retention time is uniform relative to the end markers from one GC system to another, we may normalize all retention times in the peak area plot based on a percentage of time between the two end retention markers for each individual sample. The normalization yields new retention times with values between 0 and 1 with the equation

\[ nrt_i = \frac{rt_i - rt_{TCMX}}{rt_{DCBP} - rt_{TCMX}} \]

for all \( i \) ranging from 1 to the number of peaks in the peak area table, where \( rt_i \) are the actual retention times in minutes, \( rt_{TCMX} \) is the time at which the beginning TCMX retention marker occurs, and \( rt_{DCBP} \) is the time at which the end DCBP retention marker occurs. This normalization allows the identification of particular congener peaks without concern for the particular GC system in which a sample was evaluated.

42
When training a neural network, it is desirable to have training data that are consistent across the entire set. Ideally, this data set consists of a wide variety of samples spanning the anticipated input range with as few variables as possible contributing to variations in the data. As with most chemical processes, gas chromatography has a certain amount of variability in its output. Of specific interest in the data set used in this research is the variance in the calculated areas of the end markers.

Each sample in a set of contrived data has the same amount of chemical injected for each of its end markers (TCMX and DCBP). However, due to the various processes the samples are exposed to during GC analysis, the calculated area values for the end markers aren’t equal across the set. If the end marker area in a sample is off by some factor, then all the area values in the sample are off by the same factor. For this reason, it is desirable to establish the fractional margin by which each of the data samples is off. In order to establish this margin, a plot of each sample’s end marker area is made. Figure 4.3 shows a plot of the DCBP end marker areas for each sample in the data set. By calculating the mean area value of these DCBP markers, a good approximation of the actual amount of DCBP injected is established.

\[ \text{mean} = \frac{\sum_{j=1}^{n} (ar_{dcbp})_j}{n}, \]

where \( n \) is the number of samples in the data set, and \((ar_{dcbp})_j\) is the area under
Figure 4.3: Graph of DCBP retention marker areas indicating the variance in peak areas across the data set.

the DCBP peak for sample number \( j \). Each peak area in a particular sample may be normalized \((nar_j)\) with respect to the ratio between the mean and that sample’s DCBP area, \((ar_{dcbp})_j\), using the formula

\[
    nar_j = ar_j * \frac{\text{mean}}{(ar_{dcbp})_j},
\]

for all \( j \) ranging from 1 to the number of peak areas in that particular sample, where \( ar_j \) represents the individual peak area values in the sample. By normalizing the peak area values in each of the samples, we essentially eliminate one variable that would otherwise contribute to variations in the data.
The importance of scaling input data to a reasonable range, say [0,1], is well accepted in neural network study. Although such a scaling could be accomplished with the bias and connection weights between the input layer and hidden layer, research has shown that training is easier when scaling is performed explicitly [11]. To accomplish this, the entire table of normalized peak areas within a sample, excluding the two end markers, is scanned for the largest value. The value of the largest area is designated as $area_{\text{largest}}$ and the rest of the peak areas are scaled ($sar_j$) using the formula

$$sar_j = \frac{nar_j}{area_{\text{largest}}}$$

for all $j$ ranging from 1 to the number of peak areas in the data table for that particular sample, and $nar_j$ are the individual area values after normalization.

4.2 Determining the Network Input Data

It is not necessary to use all peak areas for input to the neural network. Instead, an analysis of the chromatogram is performed to select only a few of the peaks and use them to represent the entire chromatogram as input. Specifically, six peak areas are sufficient for satisfactory network performance as long as the peaks are chosen so that the data range is fully represented. In other words, some peaks should at least partially represent Aroclor 1242; some should represent Aroclor 1254; and some should represent Aroclor 1260. The six peaks are chosen so
that each of the three Aroclors have at least two peaks which contain information about them.

Certain regions of the chromatogram contain information for each of the Aroclors. For instance, the data set used in this research has most of the peaks making up Aroclor 1242 in the time region between 9 and 18 minutes. Figure 4.4 is a portion of a plot of the Aroclor 1242 standards. Each line in the plot represents a specified concentration of Aroclor 1242. Two peaks were chosen from this time interval (approximately 10.8 minutes and 12.3 minutes) as inputs for the neural network. They were chosen because the contribution to their areas is almost exclusively from Aroclor 1242 and because of their relative isolation from nearby peaks. When a peak is isolated from the other peaks, the interference from other peaks on peak area calculation is eliminated. This is because the baseline is more accurate. This is evident in the figure where the peaks at 10.8 minutes and 12.3 minutes have a nice isolation from neighboring peaks, thus a true baseline on either side is maintained.

Most of the peaks with contributions from Aroclor 1260 occur in the approximate time range of 16 to 30 minutes. Figure 4.5 is a portion of a plot of the Aroclor 1260 standards where each line in the plot represents a specified concentration. Two peaks were chosen from this time interval (approximately 25.7 minutes and 29.8 minutes) as inputs for the neural network. Although there is
Figure 4.4: Portion of plot for each of the standards for Aroclor 1242. The upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.
Figure 4.5: Portion of plot for each of the standards for Aroclor 1260. The upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.
a small contribution from Aroclor 1254 to the area under these peaks, the isolation of the two peaks make them good input peaks. Even though there are some peaks in the region with no contribution from other Aroclors, their isolations are generally inferior to those of the two peaks selected. The importance of isolation can be seen in Fig. 4.5 between the 26 minute and 26.5 minute markers. The two peaks located in this range have an overlap between them, so the baseline has been raised. A raised baseline can have adverse affects on peak area calculations.

The peaks to which Aroclor 1254 contribute are in the time range of 13 to 25 minutes. There are no peaks, however, to which Aroclor 1254 exclusively contributes. This is because the Aroclor 1254 region is located among the region of the other two Aroclors in the chromatogram. The neural network should, however, be able to determine the concentration of Aroclors despite the presence of multiple Aroclors. For the network to accomplish this it should be trained on data that contains combinations of Aroclors so that it can “learn” what impact the combinations have on the input data. The desire, then, is to choose one input peak with contributions from both Aroclor 1254 and Aroclor 1242, and another input peak with contributions from both Aroclor 1254 and Aroclor 1260.

Figure 4.6 shows the region from which the input peak shared by Aroclor 1254 and Aroclor 1242 was selected. A notable problem with the peaks in the Aroclor 1254 time range is that most of the peaks with any substantial Aroclor 1254 con-
Figure 4.6: Portion of plot for each of the standards for a) Aroclor 1254 and b) Aroclor 1242. In each, the upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.
tribution have poor isolation. A comparison, however, of Fig. 4.6a and Fig. 4.6b shows that the peak at approximately 17.1 minutes has some of the best isolation properties in the region and has a good contribution from both Aroclor 1254 and Aroclor 1242. For these reasons this peak was chosen as an input peak.

Figure 4.7 shows the region in which the input peak shared by Aroclor 1254 and Aroclor 1260 was selected. A comparison of Fig. 4.7a and Fig. 4.7b shows that the peak at approximately 21.1 minutes has some of the best isolation properties in the region and has a good contribution from both Aroclor 1254 and Aroclor 1242. For these reasons it was selected as an input peak.

Once the chromatograms have been studied and the six input peaks selected, the neural network can be used. Six peak areas are taken from the peak area tables based on the selected input peaks. Because the retention time percentages may vary slightly from one sample to another, a small range is used to scan the tables for each of the calculated percentage retention times. The six specified peak areas are used as the network input vector whether the network is trained on a sample from the data set or the network is used to evaluate an unknown sample.
Figure 4.7: Portion of plot for each of the standards for a) Aroclor 1254 and b) Aroclor 1260. In each, the upper solid line is 800 ppb, the dash-dot line is 400 ppb, the dotted line is 200 ppb, the lower solid line is 100 ppb, and the dashed line is 50 ppb.
4.3 Data Organization for Neural Network Training

Most artificial neural networks (ANN) attempt to synthesize modules that transduce inputs into desired outputs from a set of correct input/output pairs [34]. To show how this can be achieved an analogy is used in which the problem of learning from examples can be viewed as a problem of approximating a multivariate function. An input/output mapping is modeled with a two-dimensional surface reconstruction from sparse data. Learning simply means collecting samples with input coordinates \((x_i, y_i)\) and the corresponding output values at those locations (the heights of the surface \(d_i\)). A generalization is made by estimating \(d\) at locations \((x, y)\) where there are no examples. This requires interpolating the surface between the data points. The mapping of \((x, y)\) to \(d\) is assumed to be smooth: small changes in the inputs cause a small change in the output. The effectiveness of estimating by interpolation increases as the amount of training data increases.

In training ANN, a situation often arises in which enough data exists to adequately train the network, but not enough to hold out for a valid test set. The 45 samples available for this research, do not represent enough data to form separate training and test sets that would provide a valid evaluation of the network. To overcome the lack of data, a method of training and testing known as "hold-one-out" is used in this research. For example, suppose there are \(n\) samples available (\(n = 45\,\text{here}\)). The network is trained \(n\) different times using \(n - 1\) of the samples
each time for training. The one remaining sample is held out as a solitary test set. On each of these \( n \) trials, the training set is used for testing the network during training. Then the error of the network on each of the held-out examples is calculated. After doing this \( n \) times, the square root of the mean of the squared errors (RMS) made on each of the three Aroclor concentration estimates of the held-out samples is calculated. The RMS for each Aroclor is calculated as

\[
RMS = \sqrt{\frac{1}{n} \sum_{k=1}^{n} \epsilon_k^2},
\]

where \( \epsilon_k \) are the errors by which the concentration estimates are off. This method produces an estimate of the overall network performance that would be achieved if more data were available.
CHAPTER 5

RESULTS OF ARTIFICIAL NEURAL NETWORK DESIGN IN AROCLOR CONCENTRATION DETERMINATION

5.1 Introduction

One critical aspect in artificial neural network (ANN) design is choosing an appropriate network structure for a given application. For layered network architectures, network design involves the number of layers in the network, the number of nodes per layer, and the choice of network parameters. Many researchers agree that the quality of a solution found by an ANN depends strongly on the network’s size and parameter choices [1]. In general, a network’s structure affects network complexity and learning time. It also affects the generalization capabilities of the network—its ability to produce accurate results on samples outside its training set. An undersized network lacks the representational capacity to accurately describe nonlinear complex relations. Conversely, an oversized network may fit noise in the training set, yielding poor prediction capability on independent data samples.
Training an ANN typically requires many runs with different choices of structures and parameter settings for each [41]. In such a trial-and-error process, two nested optimizations are performed. The outer loop of optimization finds the optimal network size and parameter settings. Given a choice of size and parameters, the inner loop of optimization trains the neural network. Training attempts to optimize the connection weights between layers to yield the best performance of the neural network over the training data.

In this research, two forms of multilayer neural networks, the backpropagation network (BPN) and the radial basis network (RBN), are analyzed for their ability to determine Aroclor concentrations from gas chromatography (GC) data samples. In Chapter 4, it was established that the input vector for each neural network should consist of six peak area values. It has also been established that the output of each neural network consists of the concentration of each of the three Aroclors (Aroclor 1242, 1254, and 1260). Bebis and Georgiopoulos [1] state that the number of nodes in the input and output layers of a neural network can be determined by the dimensionality of the problem. Subsequently, the number of input nodes has been set to six and the number of output nodes has been set to three.
5.2 Optimization of the Backpropagation Network

The outer loop of optimization in BPNs involves determining the number of hidden layers and the number of nodes in each. The inner loop of optimization involves training the network. This includes the adjustment of the learning scalar, $\mu$, during the Levenberg-Marquardt training algorithm discussed in Chapter 3. The optimization takes place in a MATLAB® environment where $\mu$ is automatically adjusted during training. As long as the error gets smaller, $\mu$ is made bigger. If the error increases, $\mu$ is made smaller. The following results are from the experimentation performed to optimize the BPN for the determination of Aroclor concentrations from the GC peak area values.

5.2.1 Selection of the Number of Hidden Nodes

In the BPN structure, the hidden layer is a critical part in the network's learning phase. Determining the right number of hidden nodes is the most challenging task in designing a BPN structure. With too few hidden nodes, the network is unable to adequately create complex function estimations. Too many hidden nodes, however, and the training may become computationally costly, making it more difficult for the trained network to create a generalized mapping. According to Kung and Hwang [24], when there is some regularity embedded in the patterns, as in GC peak areas, the number of hidden nodes will be dictated by the number
of regularity features rather than by the number of training patterns.

The optimal number of nodes in the hidden layer was selected through experimental analysis. The three-layer BPN were trained with an increasing number (3–23) of hidden nodes. To minimize the effects of initial random connection weights and resulting local minima, the network training procedure for each iteration was repeated eight times with different initial conditions. The concentration for each Aroclor was obtained by averaging these eight estimation values at each iteration. Figure 5.1 is a graph of the total RMS error for each iteration through the experiment. Each iteration represents the number of nodes in the hidden layer, ranging from three to 23. The total RMS error for each iteration is a composite value obtained from all three Aroclor estimation errors. After the hold-one-out training is completed for an iteration, the RMS value of the three Aroclor RMS errors is computed. This is the value plotted in Fig. 5.1. This graph reveals that the network performs better with a relatively small number of nodes in the hidden layer. This result is theoretically supported by Bebis and Georgiopoulos [1] who state that smaller, less complex networks exhibit better generalization properties especially when the training data set is limited. From the graph in Fig. 5.1 it is evident that a BPN with seven nodes in the hidden layer is the optimal network structure for the task described in this research. This seven hidden node BPN has a total RMS error of only 20.57 ppb over the 45 sample data set.
Figure 5.1: The optimization of the BPN by varying the number of nodes in the hidden layer from 3 to 23.

5.2.2 Network with Two Hidden Layers

The number of hidden layers is another important consideration when optimizing a BPN. One must decide how many hidden layers are required for the given problem. Although Hecht-Nielsen [18] has proved that a network with only one hidden layer can approximate any arbitrary nonlinear function, sometimes it is more efficient to incorporate a second hidden layer. No strict proof exists that a network with two hidden layers is better than one with a single hidden layer or vice versa. We can conclude, however, that a BPN with a maximum of two hidden layers can approximate any relationship between input and output.
An experiment was performed to determine whether there is an advantage to using two hidden layers over one in a BPN for the task described in this research. Instead of using just one outer loop of optimization, an additional nested loop was used so that two hidden layers with varying numbers of nodes could be implemented. The four-layer networks were trained with increasing numbers (3–8) of nodes on each hidden layer. With four-layer networks a smaller range of nodes in each layer is used because the larger network architecture generally requires fewer nodes on each hidden layer than the smaller three-layer network [29]. As with the three-layer BPN experiment, an ensemble of eight networks was used with each iteration.

Figure 5.2 is a family of curves showing the results of the four-layer BPN optimization experiment. The graph in Fig. 5.2(a) shows the total network RMS error where the number of nodes in the first hidden layer is fixed at three and the number in the second hidden layer increases from three to seven across the x axis. Similarly, the graphs in Figs. 5.2(b)—5.2(f) show the nodes in the first hidden layer fixed from four to eight, and increases from three to seven in the second hidden layer. This family of curves shows that the optimal four-layer BPN contains four nodes in the first hidden layer and five in the second hidden layer. With a total RMS error of just 20.03 ppb this four-layer, 4 × 5 hidden node architecture is the optimal BPN structure for this research.
Figure 5.2: Family of curves indicating the optimization of the BPN by varying the number of nodes in each hidden layer.
5.2.3 Utilizing Network Ensembles

Even after optimizing a BPN architecture, one can still improve the estimation error. To reduce this error, a tool from fault tolerant computing [8], called ensembling, is used. Ensembles are desirable due to the basic fact that the selection of weights, \( w \), is an optimization problem with many local minima [17]. All global optimization methods yield "optimal" parameters which can vary greatly from one run of the algorithm to the next due to the many local minima present. As a result, each network in the ensemble makes generalization errors on different subsets of the input space. It has been shown by Hansen et al. [17] that the collective decision produced by an ensemble yields a smaller network error than an individual network across a data test set in many BPN applications.

From a statistical point of view, ensemble averaging is desirable because the collective decision from a set of networks shows less error compared with the decision made by any one of the individual networks. This argument is supported by an experiment presented here in which the previously determined optimal BPN is used for the Aroclor concentration estimation task with a varying number of networks per ensemble. Figure 5.3 is a series of graphs demonstrating the performance of the various sizes of network ensembles for each of the three Aroclors. The graphs indicate the estimation error on each of the samples during the hold-
Figure 5.3: A graphical evaluation of BPN performance on each of the three Aroclors with different levels of ensembling. The graphs show the BPN estimation error on each sample in the data set.
Table 5.1: Evaluation of Ensemble Performance.

<table>
<thead>
<tr>
<th></th>
<th>Ar. 1242 RMS</th>
<th>Ar. 1254 RMS</th>
<th>Ar. 1260 RMS</th>
<th>Total RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>single network</td>
<td>31.6 ppb</td>
<td>35.8 ppb</td>
<td>47.7 ppb</td>
<td>39.0 ppb</td>
</tr>
<tr>
<td>ensemble of 5</td>
<td>18.6 ppb</td>
<td>25.3 ppb</td>
<td>23.0 ppb</td>
<td>22.5 ppb</td>
</tr>
<tr>
<td>ensemble of 8</td>
<td>22.8 ppb</td>
<td>23.9 ppb</td>
<td>24.1 ppb</td>
<td>23.6 ppb</td>
</tr>
<tr>
<td>ensemble of 10</td>
<td>19.2 ppb</td>
<td>24.8 ppb</td>
<td>24.5 ppb</td>
<td>23.0 ppb</td>
</tr>
</tbody>
</table>

one–out training procedure. The experiment was performed first with only one network and then with ensembles of five, eight, and ten.

The results of the total RMS error analysis for each Aroclor are shown in Table 5.1. For each of the four experimental ensemble runs, the table shows the RMS results of the hold–one–out analysis for each of the three Aroclors, along with a total RMS error. This value is simply the RMS value of the three Aroclor RMS values. From Table 5.1, it is evident that an ensemble of networks dramatically improves the BPN performance. Additionally, it is clear that an ensemble of just five networks is enough for optimal performance. Adding more networks to the ensemble is unnecessary and a waste of training time.
5.3 Optimization of the Radial Basis Network

The outer loop of optimization for an RBN involves estimating the optimal spread constant. The inner loop of optimization involves training the network, which attempts to optimize the connection weights to yield the best performance of the neural network over the training set. This includes determining the number of nodes in the hidden radial basis layer. The optimization takes place in a MATLAB® environment in which the number of hidden nodes is automatically adjusted during training. The following are results from experimentation performed to optimize the RBN for determining Aroclor concentration from GC peak area values.

5.3.1 Number of Hidden Nodes Determination

Similar to the BPN, the determination of the optimum number of nodes in the hidden layer of the RBN is essential. With the RBN, however, a different approach is taken in determining this number. A constructive approach is taken, whereby the process begins with a small one hidden node network and nodes are gradually added as needed until the sum-squared (SSQ) error goal is met [1]. This approach begins by using the first of the \( n \) input vectors, \( x_1 \), to determine the center vector of the first node, \( c_1 \). The center vector is determined by setting the dot product between \( x_1 \) and \( c_1 \) equal to 0 and solving for the resulting orthogonal vector, \( c_1 \).
The process continues by iteratively using each of the other \( n - 1 \) input vectors, \( x_i \), for the center of the second node. At each iteration, the SSQ error is calculated and the input vector which lowers the network error the most is used to create the second radial basis node. Each node added to the radial basis layer is fully connected to both the input and output layers. If the SSQ error of the new network structure does not meet the specified error goal then the process continues until enough nodes are created to satisfactorily meet the SSQ error goal.

The graph in Fig. 5.4 is an example of the iterative process of adding radial basis nodes to the hidden layer of the RBN. In this example, all of the 45 data samples were used to train the network, which required nine nodes in the radial basis layer. Typically, the number of nodes varied from nine to twelve throughout the training process in this research, depending on which of the samples was held out of the training set.

### 5.3.2 Spread Constant Estimation

To find the optimal network structure for an RBN, the optimal value for the spread constant must be estimated. The spread constant determines the width of an area in the input space to which each neuron responds. It is important that the spread constant be large enough that the radial basis neurons respond to overlapping regions of the input space, but not so large that all the neurons respond in
Figure 5.4: Plot showing the progression toward satisfying the SSQ error while iteratively adding new radial basis nodes. One epoch equals an additional node.

essentially the same manner. If the spread constant is set too small, the network might very well meet the error goal, but since the radial basis functions do not overlap, the function the network forms does not generalize well. With this lack of generalization capabilities, the network will not give meaningful outputs for inputs not in the training set. If, on the other hand, the spread constant is too large, the radial basis neurons will output large values (near 1.0) for a large number of the input vectors used to design the network. Then no matter what the input is, most all of the outputs of the hidden radial basis layer will be 1. In summary, one should choose a spread constant larger than the distance between adjacent input vectors, but smaller than the distance across the whole input space, in order to get good generalization [7].
In this experiment, a heuristic approach to estimating the optimal spread constant is taken. Starting with a spread constant of 0.5, this value is incremented by 0.1 until a final value of 15.0 is reached. For each spread constant value, the network is fully evaluated on each sample in the data set using hold-one-out as the method for training and testing. For each iteration through the hold-one-out process, the RMS error for each Aroclor is tabulated and the total RMS is calculated and plotted. Figure 5.5 is the plot of total network RMS error for the RBN with the spread constant set within the range of 0.5 to 15.0.
To optimize the spread constant value even further, the three lowest RMS values are chosen so that the corresponding spread constants (2.0, 2.6, and 4.2) may be evaluated more precisely. An evaluation about these three spread constants is performed with increments of 0.02. Figure 5.6 is a family of plots containing the results of this experiment. From this figure, it is evident that a spread constant of \( \sigma = 2.02 \) is optimal for the RBN in this research.

Table 5.2 shows a comparison between the optimal three layer BPN, the optimal four layer BPN, and the optimal RBN from this research. With a total RMS
Table 5.2: Evaluation of Optimal ANN.

<table>
<thead>
<tr>
<th></th>
<th>Ar. 1242 RMS</th>
<th>Ar. 1254 RMS</th>
<th>Ar. 1260 RMS</th>
<th>Total RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-layer BPN</td>
<td>17.8 ppb</td>
<td>24.7 ppb</td>
<td>18.6 ppb</td>
<td>20.6 ppb</td>
</tr>
<tr>
<td>4-layer BPN</td>
<td>18.2 ppb</td>
<td>23.9 ppb</td>
<td>17.3 ppb</td>
<td>20.0 ppb</td>
</tr>
<tr>
<td>RBN</td>
<td>17.2 ppb</td>
<td>20.9 ppb</td>
<td>14.9 ppb</td>
<td>17.8 ppb</td>
</tr>
</tbody>
</table>

error of only 17.8 ppb, the RBN whose nodal spread constants are $\sigma_i = 2.02$ is the optimal neural network for determining the Aroclor concentrations from gas chromatograms in this research.
CHAPTER 6

ARTIFICIAL NEURAL NETWORK PERFORMANCE ANALYSIS

6.1 Introduction

The purpose of this research is to demonstrate the accuracy of artificial neural networks (ANN) as a tool for determining Aroclor concentrations in environmental soil samples using data from gas chromatograms. This chapter presents an in-depth analysis of the results from the ANN and then compares the ANN results to the results of other existing methods for Aroclor concentration estimation. This chapter also demonstrates the ability to calibrate the ANN and calculate a confidence error using linear regression techniques. Finally, a description of the adaptability of the ANN in analyzing other GC data is presented.

6.2 ANN Performance Across the Developmental Data Set

After optimizing each of the two ANN used in this analysis, it is desirable to evaluate the performance of each network on the individual data samples in the
data set. The optimal radial basis network (RBN) and the optimal backpropagation network (BPN) are both examined for their accuracy in estimating the concentration of the three Aroclors for each of the 45 samples in the data set. It is important to recall that with the hold-one-out process used in evaluating each ANN, the estimations produced by the network for each sample result from training the network on the other 44 samples and testing on the one sample held out. In other words, the results of the network estimation are for a sample that the network never "learned" during the training process.

The data set used in this research consists of 45 samples. The samples adequately cover the full range of input that the ANN might see in a sample of unknown concentration. Some of these samples represent standards where only one of the Aroclors is present in some concentration. Some of the samples are reproduced standards samples. Some of the samples represent extreme cases with combinations of very high and very low concentrations of the three Aroclors. Still, other samples represent more consistent cases of data one could expect to see occurring in soil samples naturally, where the concentrations are reasonably low. The samples which represent the more consistent data should yield network estimations which are indicative of the network’s true accuracy in determining the Aroclor concentrations of unknown samples. Similarly, those samples which represent the more extreme cases should yield network results which are more representative of the maximum errors expected in determining Aroclor concentra-
While optimizing each of the two ANN used in this research, the results from estimating Aroclor concentrations of each of the 45 samples are recorded. Figure 6.1 is a graphical representation of the performance of the optimal BPN on each of the individual samples during the hold-one-out process. The figure shows the absolute error in determining the concentration of each of the three Aroclors for each sample that is held out as the test sample.

Figure 6.2 is a graphical representation of the performance of the optimal RBN on each of the individual samples during the hold-one-out process. The plots show the absolute error in determining the concentration of each of the three Aroclors for each sample that is held out as the test sample. It was established through experimentation that the RBN performs best for the task in this research. The RBN, therefore, is the neural network implemented by the Data Interpretation Module (DIM) team as part of the Department of Energy (DOE) funded Contaminant Analysis Automation (CAA) project. Although Fig. 6.1 is nice for comparing the performance of the BPN with the performance of the RBN from Fig. 6.2, the results of the RBN only, will be used when comparing the results of this research with other existing methods.

In Fig. 6.2, some important results of the RBN performance should be noted.
Figure 6.1: Absolute error analysis for the BPN Aroclor concentration estimation on each sample during the hold-one-out evaluation.
Figure 6.2: Absolute error analysis for the RBN Aroclor concentration estimation on each sample during the hold-one-out evaluation.
The first 15 samples are a subset of standards samples. Each of these samples contains only a single Aroclor with known concentrations between 50 ppb and 800 ppb. The network is quite accurate in determining Aroclor concentrations in these samples, especially when a particular Aroclor is not present in the sample. This is important because incorrectly reporting an Aroclor's presence could result in the unnecessary decontamination of an area. The highest errors reported in the standards subset generally occur when one of the Aroclors is present in high concentrations. Even though the absolute errors for some of these higher concentration samples are a little high, the percentage error is actually quite low. For example, in Sample 5, the concentration of Aroclor 1242 is reported as 830 ppb although its actual concentration is 800 ppb. The margin of error is actually only 3.8 percent. Additionally, when Aroclors are present at such high concentrations, the absolute reported concentration is not as critical as it is for low concentrations.

The true accuracy of the ANN is determined by how well the network determines Aroclor concentrations when more than one is present in a given sample. When more than one Aroclor is present in a sample, each peak in the chromatogram may have contributions from multiple Aroclors. Figure 6.3 is an illustration of the effects that the presence of multiple Aroclors has on individual peaks in a chromatogram. The solid line represents an excerpt from a chromatogram that contains 400 ppb each of Aroclor 1242, 1254, and 1260. The dashed line represents an excerpt from a chromatogram that contains 400 ppb of Aroclor 1242.
Figure 6.3: An illustration of the contribution of multiple Aroclors to individual peaks. The solid line is from a chromatogram of a sample with 400 ppb for each of Aroclor 1242, 1254, and 1260. The dashed line is from a chromatogram of a sample with 400 ppb of Aroclor 1242 only. The dash-dot line is from a chromatogram of a sample with 400 ppb of Aroclor 1254 only. The dotted line is from a chromatogram of a sample with 400 ppb of Aroclor 1260 only. Similarly, the dash-dot line and the dotted line represent excerpts from chromatograms containing 400 ppb of Aroclor 1254 and Aroclor 1260, respectively. This illustration shows that more than one Aroclor may have a significant contribution to an individual peak and that these contributions are not linear from peak to peak. Because ANN tend to handle nonlinearities in the input data better than traditional linear regression techniques, the ANN is appropriate for determining the concentrations of individual Aroclors in gas chromatograms when
multiple Aroclors are present.

Samples 16–36 contain mixtures of the three Aroclors at varying concentrations from 0 to 800 ppb. It has been noted already that some of these samples contain extreme combinations of the Aroclors. These extreme samples have one important feature in common: they have a large amount of one or more Aroclors and a small amount of the other Aroclors. When this feature is present in a chromatogram, the peak area contributions from an Aroclor with a high concentration might cause a misproportional interpretation of peak area concentration for each of the Aroclors. For instance, the results for Sample 28 can be explained with this logic. The known concentrations for this sample are 200 ppb of Aroclor 1242, 50 ppb of Aroclor 1254, and 800 ppb of Aroclor 1260. Due to the high concentration of Aroclor 1260 and relatively low concentrations of Aroclors 1242 and 1254, a study of Fig. 6.2 shows that the estimated concentration for Aroclor 1260 is low (754.9 ppb) and that the estimated concentration for Aroclors 1242 and 1254 are high (217.3 ppb and 57.4 ppb respectively). These estimates are actually quite accurate considering the complexity of the data sample. The key result is the accuracy in which the RBN estimates the concentration of Aroclor 1254. Since it is known to be present at a low concentration (50 ppb), the concentration at which the RBN estimates its presence determines whether or not the Aroclor is present at a concentration high enough to warrant soil decontamination.
Other samples also contain data with extreme combinations of Aroclors (i.e. Samples 20, 24, 27, 29, 31, 32, and 34). These samples are present in the mixtures subset so that the RBN may be evaluated for some of the most extreme combinations that may be encountered. The results in Fig. 6.2 from these samples should represent some of the highest absolute error values the RBN will yield once the ANN system is implemented by the DOE for data interpretation.

Some of the samples in the mixtures subset are Aroclor mixtures that are more consistent. For example, Sample 26 is known to have 50 ppb of Aroclor 1242, 0 ppb of Aroclor 1254, and 100 ppb of Aroclor 1260. From Fig. 6.2, the estimation results of the RBN for this particular sample are 57.1 ppb of Aroclor 1242, 0.0 ppb of Aroclor 1254, and 105.0 ppb of Aroclor 1260. It is evident from this example and other similar samples that the network is quite accurate in estimating Aroclor concentrations for the more consistent data samples.

6.3 Effect of Scaled Test Data on ANN Performance

This experiment is performed to determine the effect that placing an intentional scale factor on the peak area inputs has on RBN performance. At each iteration through the hold–one–out procedure, the network is trained as usual with 44 unscaled samples, but for the one held–out sample the peak area values
are multiplied by a scale factor. If the input data were linear, then one could expect the input scale factor to influence the output by the same linear scale factor. The estimation error for each sample would remain the same. Figure 6.4, however, indicates that this is not the case. The figure compares the estimation results for the three scaling trials on each sample during hold-one-out. The dashed line represents the absolute errors from the trial whose input scale factor is 0.8, while the dash-dot line represents the absolute errors from the trial whose scale factor is 1.2. The solid line represents the absolute errors from the trial with no input scale factor from Fig. 6.2.

A careful evaluation of the graphs in Fig. 6.4 shows that the 0.8 scale factor actually increases the accuracy of the RBN for Aroclor concentration estimations on many samples. Conversely, the 1.2 scale factor decreases the accuracy of the RBN for almost every sample. Table 6.1 shows a comparison between the 0.8 scaled input test data, the 1.2 scaled input test data, and the input test data with no scale factor. The RBN is actually more accurate in total concentration estimation for the 0.8 scaled input data (total RMS 17.0 ppb) than it is on the data when no scale factor is applied (total RMS 17.8 ppb). Conversely, the RBN is considerably less accurate in total concentration estimation when the input data is scaled by a factor of 1.2 (total RMS 34.6). Reducing the values of the input peak areas with a 0.8 scale factor should theoretically scale the estimated output by 0.8, therefore not changing the estimation error. Since the peak area input
Figure 6.4: Absolute error analysis for the BPN Aroclor concentration estimation on each sample during the hold-one-out evaluation. Three trials were performed with a different scale factor on the peak area inputs of the test sample.
Table 6.1: Comparison of RBN Results for Scaled Data and Unscaled Data.

<table>
<thead>
<tr>
<th>Scale Factor</th>
<th>Ar. 1242 Error (ppb)</th>
<th>Ar. 1254 Error (ppb)</th>
<th>Ar. 1260 Error (ppb)</th>
<th>Total RMS (ppb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No scale</td>
<td>17.2 ppb</td>
<td>20.9 ppb</td>
<td>14.9 ppb</td>
<td>17.8 ppb</td>
</tr>
<tr>
<td>Scale X 0.8</td>
<td>15.1 ppb</td>
<td>18.9 ppb</td>
<td>16.8 ppb</td>
<td>17.0 ppb</td>
</tr>
<tr>
<td>Scale X 1.2</td>
<td>34.7 ppb</td>
<td>30.0 ppb</td>
<td>38.8 ppb</td>
<td>34.6 ppb</td>
</tr>
</tbody>
</table>

Values are relatively smaller, less overlap among peaks is “observed” by the RBN. The network, therefore, can estimate the “true” Aroclor concentrations more effectively. Similarly, raising the values of the peak area inputs with a 1.2 scale factor should theoretically scale the estimated output by 1.2, thus not changing the estimation error. Since the peak area input values are relatively larger, more overlap is simulated and the network has more difficulty in estimating the correct Aroclor concentrations. The negative effect is more pronounced for the samples with Aroclor concentrations of 800 ppb because some of the peak area input values are outside the range of values on which the network was trained.
6.4 Comparison of RBN to Existing Methods

The accuracy of the RBN developed in the present research should be compared to the results of the three other methods developed as part of the CAA data interpretation module project elsewhere. Target3 is a commercially developed software package developed by ThruPut Systems Incorporated and implemented at the Los Alamos National Laboratories (LANL). Principal component regression (PCR) is a multivariate analysis implemented at the Oak Ridge National Laboratories (ORNL) to use the MATLAB® signal processing toolkit. The MATLAB® computation engine is run as a UNIX® background task. Multiple linear regression (MLR), a method in which a series of linear equations is used to represent data, is another method used at LANL.

Target3 is written for the UNIX® operating system and designed to serve as a single-point data processing and data management software for use in high-volume environmental and pharmaceutical laboratories [9]. The software can directly read raw GC data files. It follows an object-oriented paradigm to process the data. The data in the fifteen standards samples is examined to extract five major peaks unique to each Aroclor. In the case of Aroclor 1254, peaks have to be selected with minor contributions from Aroclors 1242 and 1260 because no distinct peaks are present. Calibration equations for each of the fifteen peaks are created with a linear fit of the peak areas to the Aroclor concentration. Chromatograms
of unknown mixtures are analyzed by finding the peaks nearest in retention time to the standard peaks. The areas of these peaks are used to calculate Aroclor concentrations with the calibration equations. The Aroclor concentration is reported as the average concentration calculated with each peak in the standard set.

PCR calibration modeling is a technique that utilizes all of the data produced by modern analytical instruments such as GC while simultaneously reducing the mathematical dimensionality of the problem to a few orthogonal features. The PCR models are built by performing a principal component analysis on the standards data set and identifying the resulting eigenvalues that express most of the variance in the standards. MLR techniques are then used to construct the calibration functions using the eigenvalues as the basis set. Advantages of this approach, besides a reduction of the problem's dimensionality, are the removal of irrelevant information, such as noise and correlated responses, and the ability to apply statistical tools to assess the validity of the calculated concentrations [9]. An in–depth discussion of PCR is given by Jolliffe [21].

MLR is a method used to establish a quantitative relationship between a group of predictor variables and a response. The response is each point in the gas chromatogram of a sample of unknown Aroclor concentrations. The predictor variables, then, are the points in a chromatogram of a known Aroclor concentration. The predictor chromatograms are chosen through an iterative process in which the
chromatogram of unknown Aroclor make-up is matched to the chromatograms from the standards data set. Once the closest matches are established, a linear regression between each point in the unknown chromatogram and the standard chromatogram is performed.

The Aroclor concentration determination results of the RBN created in this research are compared to the results of the other three methods (Target3, PCR, and MLR) as applied to the GC data. Figure 6.5 is a comparison of overall RMS errors from estimating Aroclor 1242 concentration. Similarly, Fig. 6.6 and Fig. 6.7 are error comparisons from estimating the concentrations of Aroclors 1254 and 1260, respectively. Each figure is broken into an analysis for both the reproducibility subset and the mixtures subset. The reproducibility subset consists of nine samples which contain only one Aroclor each. These samples are especially valuable for testing the PCR, MLR, and Target3 methods for situations in which only a single Aroclor is present.

Some important observations can be made from the results presented in Figs. 6.5–6.7. The most forthright observation is that the commercially developed software package, Target3, is not a valid method for the Aroclor concentration determination task. Not only is the Target3 software ineffective in determining Aroclor concentrations of the mixture subset (RMS errors of 848.8 ppb for Aroclor 1242, 355.8 ppb for Aroclor 1254, and 1128.8 ppb for Aroclor 1260), but it is
Figure 6.5: An RMS error comparison of the RBN versus other existing methods for predicting Aroclor 1242 concentrations. The other methods in the comparison are principal component regression (PCR), multiple linear regression (MLR), and the Target3 (T3) software package.

Figure 6.6: An RMS error comparison of the RBN versus other existing methods for predicting Aroclor 1254 concentrations. The other methods in the comparison are principal component regression (PCR), multiple linear regression (MLR), and the Target3 (T3) software package.
Figure 6.7: An RMS error comparison of the RBN versus other existing methods for predicting Aroclor 1260 concentrations. The other methods in the comparison are principal component regression (PCR), multiple linear regression (MLR), and the Target3 (T3) software package.

...also ineffective in determining concentrations of two of the Aroclors of the reproducibility subset (RMS errors of 87.6 ppb Aroclor 1254, and 237.4 ppb Aroclor 1260). Another immediate observation is that a comparison of the three valid methods (PCR, MLR, and RBN) in estimating the Aroclor concentrations of the reproducibility subset shows that while PCR is good, both the MLR and RBN techniques are nearly perfect.

The accuracy of each of the three methods can best be evaluated by noting how well each method estimates concentrations of each Aroclor in the mixtures subset. An evaluation of the Aroclor 1242 concentration estimates in the mixtures subset
reveals a high RMS error (93.3 ppb) for the MLR method, while the RMS errors for the PCR (27.8 ppb) and RBN (23.8 ppb) methods are quite low. Similarly, an evaluation of Aroclor 1260 estimates reveals a considerably higher RMS error for the PCR (30.7 ppb) method than the RMS for the MLR and RBN methods. An evaluation of the RMS errors from Aroclor 1254 concentration estimation shows that both the PCR (46.1 ppb) and MLR (45.9 ppb) methods have considerably higher performance errors than those of the RBN method (28.9 ppb). It is evident from these results that even though all three methods (PCR, MLR, and RBN) are effective in estimating the proper Aroclor concentrations, the RBN created in this research has a better error performance and is the most consistent method of the three.

Another important issue to consider when deciding which of the methods are appropriate is the manner in which they use the given data. Because of the massive size of the chromatogram data files, they must be purged from the DIM system on a regular basis. The PCR and MLR methods are disadvantaged by having to rely on the entire chromatogram for its data set because they are subject to being stranded without their original data once the chromatogram is destroyed. Conversely, the peak area data files used by the RBN are relatively small in size and easily maintained without using large amounts of memory for its storage, making the RBN more data-efficient than the other two methods.
All three methods (PCR, MLR, and RBN) have been shown to be valid for the estimation of Aroclor concentrations in soil samples. For this reason, the DIM team has proposed that all three methods be used for the Aroclor concentration estimation task simultaneously. A data fusion process shall be implemented so that the results of all three methods may be used to optimize the Aroclor concentration determination process.

6.5 Calibration of the ANN Using Linear Regression Techniques

The gas chromatography (GC) process has parameters that may change periodically, therefore, the ANN must be able to adjust to changing trends in the chromatograms. It is undesirable and practically impossible to constantly retrain a new ANN on a regular basis. To accommodate for the subtle changes that might occur in the output characteristics of a chromatogram, the ANN must be calibrated on a regular basis [37].

Linear regression [4] is used to calibrate the ANN in this research. Linear regression's purpose is to establish a quantitative relationship between a predictor variable, $x$, and a response, $y$. Such a relationship is useful for understanding which predictors have the most effect, knowing the direction of the effect (i.e., increasing $x$ increases/decreases $y$), and using the model to predict future values.
of the response when only the predictors are currently known.

The process of linear regression begins with the establishment of a $2 \times 1$ vector of parameters, $\beta$, for each Aroclor. This is accomplished for each Aroclor by first letting $y$ be a $5 \times 1$ vector of observations based on the five standards samples. These values are the concentration outputs of the ANN for a particular Aroclor and are typically a value between 0 and 1. A corresponding $5 \times 2$ matrix of regressors, $x$, is developed based on known Aroclor concentrations in parts per billion (ppb) for each standards sample. Using the linear model

$$y = x\beta + \epsilon,$$

where $\epsilon$ is a $5 \times 1$ vector of random disturbances, we establish the regression parameters, $\beta$ [22].

Using the linear regression function built into the MATLAB® statistics toolbox [22], the regression parameters, $\beta$, are calculated along with a confidence interval for $\beta$, of confidence $\alpha$. Additionally, the residuals, $r$, along with a confidence interval for each residual of confidence $100(1 - \alpha)$ are calculated. The residuals and their confidence intervals are valuable for determining how well the linear regression establishes the calibration parameters. Figure 6.8 is a series of plots showing the residuals and their confidence intervals for each of the three Aroclors. If any of the residual interval bars do not pass through the zero line, then there is an outlier present for that particular sample. An outlier occurs when
Figure 6.8: A plot of the residuals for each of the three Aroclors with error bars showing 95 percent confidence intervals on the residuals.
an estimated Aroclor concentration value does not fall within the specified error range for a given confidence interval. For instance, when estimating the concentration of Aroclor 1254 in Sample 10 [Fig. 6.8b], the ANN was off by \(-7.2\) ppb, but did not fall within the \(\pm 6.3\) confidence interval established at the 95\% confidence level.

Once the calibration parameters and confidence intervals have been established, the ANN output for each Aroclor is passed through its respective linear calibration "layer" on the output of the neural network. This layer not only adjusts the network results to the calibrated GC parameters, but also scales the ANN outputs from a decimal number between 0 and 1 to a concentration in ppb. The actual concentration for each Aroclor, \(y\), is calculated with the linear formulas

\[
y_{42} = x_{42} \cdot (\beta_{42})_2 + (\beta_{42})_1 ,
\]

\[
y_{54} = x_{54} \cdot (\beta_{54})_2 + (\beta_{54})_1 ,
\]

\[
y_{60} = x_{60} \cdot (\beta_{60})_2 + (\beta_{60})_1 ,
\]

where \(x\) is the actual output from the ANN and the \(\beta\) values are the scale factor and offset factor from the regression parameters established during ANN calibration. A confidence error is established from the confidence intervals produced with the regression parameters. The ANN outputs the confidence error along with each Aroclor concentration estimate. This is demonstrated in Table 6.2 for one of the samples from the data set.
Table 6.2: Calibration Results from Sample 28.

<table>
<thead>
<tr>
<th></th>
<th>Aroclor 1242 (ppb)</th>
<th>Aroclor 1254 (ppb)</th>
<th>Aroclor 1260 (ppb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Known</td>
<td>200</td>
<td>50</td>
<td>800</td>
</tr>
<tr>
<td>Estimated</td>
<td>217.3</td>
<td>57.4</td>
<td>754.9</td>
</tr>
<tr>
<td>Calibration</td>
<td>218.2 ± 2.5</td>
<td>57.3 ± 0.9</td>
<td>755.4 ± 9.9</td>
</tr>
</tbody>
</table>

6.6 Compatibility of the ANN with Data from Other GC Columns

There are a variety of GC columns used in the chemical industry today. The ANN developed in this research has been developed with versatility so GC data may be interpreted no matter what kind of column is used or what the GC parameters are. This is made possible by the manner in which peak areas are extracted and the manner in which peak area retention times are normalized. The process of adapting the ANN for use with different GC columns is outlined as follows:

1. Analyze the chromatograms from the standards subset to select peaks that will represent the entire chromatogram as input as discussed in Chapter 4. Note the approximate retention times at which these peaks occur.

2. Scan the peak area files to determine the actual retention times, \( r_t \), by matching the estimated times from step 1 with those in the peak area tables.
3. Determine the end marker retention times, $rt_{DCBP}$, $rt_{TCMX}$, from the peak area tables.

4. Normalize the chosen peak retention times based on a percentage of time between the two end retention time markers for each sample:

$$nrt_i = \frac{r_{t_i} - r_{t_{TCMX}}}{r_{t_{DCBP}} - r_{t_{TCMX}}}$$

for all $i$ ranging from 1 to the number of peaks chosen as ANN input.

5. Modify the software that extracts selected peak areas from the peak area data files so that the correct peak areas, $nrt_i$, are extracted for ANN input.

The peak area data files used in this research came from chromatograms generated by a Hewlett Packard (HP) db-608 column at the Oak Ridge National Laboratories. An experiment, however, was performed to demonstrate the versatility of the RBN using peak area files calculated from chromatograms generated by a different GC column: an HP db-1701. The adaptation steps above were used to preprocess the data for use with the RBN. Figure 6.9 shows the absolute error results for each sample during the hold-one-out process for the data run on the HP db-1701. The absolute error for each sample is plotted for each of the three Aroclors. The total RMS errors for each Aroclor is presented in Table 6.3 along with the same errors from data samples run on the HP db-608. Although the RMS errors are not quite as low for the HP db-1701, it is evident that the RBN
Figure 6.9: A hold-one-out evaluation of RBN error performance on data samples run through a different GC column (Hewlett Packard db-1701).
Table 6.3: Evaluation of RBN Performance on Data from the HP db-608 and the HP db-1701.

<table>
<thead>
<tr>
<th></th>
<th>Ar. 1242 error</th>
<th>Ar. 1254 error</th>
<th>Ar. 1260 error</th>
<th>Total RMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>HP db-608</td>
<td>17.2 ppb</td>
<td>20.9 ppb</td>
<td>14.9 ppb</td>
<td>17.8 ppb</td>
</tr>
<tr>
<td>HP db-1701</td>
<td>21.9 ppb</td>
<td>31.2 ppb</td>
<td>28.3 ppb</td>
<td>27.4 ppb</td>
</tr>
</tbody>
</table>

is successful in Aroclor concentration estimation even when data from a different column is analyzed.

Today, there are nearly one million gas chromatographs operated around the world [16]. The main reasons for this overwhelming presence are the ease in which a process development can be carried out, the high degree of precision of the analyses that can be achieved, and the ease with which this technique can be automated and computerized. The number of published separation applications using GC technology is extremely large and it would be impossible to review all the applications [16]. It has been demonstrated, however, that the neural network technology developed in this research is not limited to one specific GC system. In a broader sense, the technology developed here can be used to identify and quantitize any analyte or combination of analytes present in any gas chromatogram given that the ANN is properly trained.
CHAPTER 7

SUMMARY AND CONCLUSIONS

The Department of Energy (DOE) is in the process of identifying for clean-up land that has been contaminated with environmentally hazardous chemicals. To identify the contaminants themselves, a multivariate quantitation analysis must be performed on chromatographic data obtained from soil samples. Because multiple contaminants may be present in a soil sample, the chromatographic data may be nonlinear. It is necessary then, to develop a method of analysis that can determine contaminant concentrations despite the nonlinearity of the data.

Artificial neural networks (ANNs) were identified as a pertinent method for determining the concentrations of multiple contaminants (Aroclors) because of their ability to handle nonlinear data. The ANNs were optimized to the task of identifying three Aroclors that might be present in the soil sample. The ANNs were trained with backpropagation and radial basis function algorithms to relate the data in a gas chromatogram to generalized Aroclor concentrations. Several experiments were conducted to determine optimal network architectures. Having established the optimal network architecture, additional experiments were conducted to compare the effectiveness of the ANN with other existing methods. Ex-
periments were also performed to demonstrate the ability to statistically calibrate the network and to demonstrate the versatility of the ANN for identifying contaminant concentrations from chromatographic data of different gas chromatography systems.

Several conclusions have been established about the research presented in this thesis. These conclusions are made mainly in three areas: data analysis, ANN design, and ANN performance analysis. The conclusions are summarized as follows:

1. The gas chromatograms are best represented using peak area values as input to the ANN. Additionally, the peak retention times should be normalized with respect to a set time such as that of a marker peak, and the peak area values should be scaled between 0 and 1 with the highest peak area in the chromatogram receiving a value of 1.

2. Using every peak is unnecessary in training the ANN. Instead, an analysis of the chromatograms should be performed to choose a few select peaks to represent the entire chromatogram. The selection should be based on such features as isolation from other peaks and the contribution (or lack of) from each of the Aroclors.

3. Both of the experimental ANN used in the present research, the backpropagation network (BPN) and the radial basis network (RBN), benefit highly from optimizing their respective architectures.
4. A heuristic approach was used to determine the optimal number of hidden layers and the number of nodes in each hidden layer for the BPN in the present research. It was determined that a network with two hidden layers (four and five nodes, respectively) is the optimal BPN architecture.

5. The Aroclor concentration estimation errors in a BPN can be reduced by using the results from an ensemble of networks.

6. A heuristic approach was taken in estimating the optimal radial basis function spread constant for the RBN in the present research. It was determined that a network whose radial basis neurons have a spread constant of 2.02 is the optimal RBN architecture.

7. A comparative analysis between the optimal BPN and the optimal RBN revealed that the RBN is the most effective ANN for the present research.

8. The use of ANN to estimate the concentration of chemical components in GC provides a high degree of accuracy.

9. The RBN compares positively with existing methods (multiple linear regression and principal component regression) used for Aroclor concentration estimation. It has been shown that the RBN is as good as, or better than the existing methods and may be used in conjunction with the other methods through data fusion to optimize the Aroclor concentration estimation task.

10. The RBN may be calibrated using linear regression techniques to accommo-
date changes in the GC system. In addition to calibration, linear regression may be used to provide a confidence error for the output.

11. The ANN method can be easily extended to process other analytes from GC data.
BIBLIOGRAPHY


APPENDIX
APPENDIX

GAS CHROMATOGRAPHY DATA USED IN THIS RESEARCH
Figure A.1: SAMPLE 1: Ar1242: 50 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb

Figure A.2: SAMPLE 2: Ar1242: 100 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb

Figure A.3: SAMPLE 3: Ar1242: 200 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb
Figure A.4: SAMPLE 4: Arl242: 400 ppb, Arl254: 0 ppb, Arl260: 0 ppb

Figure A.5: SAMPLE 5: Arl242: 800 ppb, Arl254: 0 ppb, Arl260: 0 ppb

Figure A.6: SAMPLE 6: Arl242: 0 ppb, Arl254: 50 ppb, Arl260: 0 ppb
Figure A.7: SAMPLE 7: Ar\textsubscript{1242}: 0 ppb, Ar\textsubscript{1254}: 100 ppb, Ar\textsubscript{1260}: 0 ppb

Figure A.8: SAMPLE 8: Ar\textsubscript{1242}: 0 ppb, Ar\textsubscript{1254}: 200 ppb, Ar\textsubscript{1260}: 0 ppb

Figure A.9: SAMPLE 9: Ar\textsubscript{1242}: 0 ppb, Ar\textsubscript{1254}: 400 ppb, Ar\textsubscript{1260}: 0 ppb
Figure A.10: SAMPLE 10: Ar1242: 0 ppb, Ar1254: 800 ppb, Ar1260: 0 ppb

Figure A.11: SAMPLE 11: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 50 ppb

Figure A.12: SAMPLE 12: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 100 ppb
Figure A.13: SAMPLE 13: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 200 ppb

Figure A.14: SAMPLE 14: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 400 ppb

Figure A.15: SAMPLE 15: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 800 ppb
Figure A.16: SAMPLE 16: Ar1242: 50 ppb, Ar1254: 50 ppb, Ar1260: 50 ppb

Figure A.17: SAMPLE 17: Ar1242: 100 ppb, Ar1254: 100 ppb, Ar1260: 100 ppb

Figure A.18: SAMPLE 18: Ar1242: 200 ppb, Ar1254: 200 ppb, Ar1260: 200 ppb
Figure A.19: *SAMPLE 19: Ar1242: 400 ppb, Ar1254: 400 ppb, Ar1260: 400 ppb*

Figure A.20: *SAMPLE 20: Ar1242: 800 ppb, Ar1254: 800 ppb, Ar1260: 800 ppb*

Figure A.21: *SAMPLE 21: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 800 ppb*
Figure A.22: SAMPLE 22: Ar1242: 0 ppb, Ar1254: 800 ppb, Ar1260: 0 ppb

Figure A.23: SAMPLE 23: Ar1242: 800 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb

Figure A.24: SAMPLE 24: Ar1242: 800 ppb, Ar1254: 800 ppb, Ar1260: 50 ppb
Figure A.25: SAMPLE 25: Ar1242: 0 ppb, Ar1254: 50 ppb, Ar1260: 100 ppb

Figure A.26: SAMPLE 26: Ar1242: 50 ppb, Ar1254: 0 ppb, Ar1260: 100 ppb

Figure A.27: SAMPLE 27: Ar1242: 200 ppb, Ar1254: 800 ppb, Ar1260: 800 ppb
Figure A.28: SAMPLE 28: Ar1242: 200 ppb, Ar1254: 50 ppb, Ar1260: 800 ppb

Figure A.29: SAMPLE 29: Ar1242: 100 ppb, Ar1254: 800 ppb, Ar1260: 400 ppb

Figure A.30: SAMPLE 30: Ar1242: 400 ppb, Ar1254: 800 ppb, Ar1260: 0 ppb
Figure A.31: SAMPLE 31: Ar1242: 800 ppb, Ar1254: 100 ppb, Ar1260: 200 ppb

Figure A.32: SAMPLE 32: Ar1242: 800 ppb, Ar1254: 50 ppb, Ar1260: 800 ppb

Figure A.33: SAMPLE 33: Ar1242: 50 ppb, Ar1254: 800 ppb, Ar1260: 0 ppb
Figure A.34: SAMPLE 34: Ar1242: 50 ppb, Ar1254: 200 ppb, Ar1260: 800 ppb

Figure A.35: SAMPLE 35: Ar1242: 0 ppb, Ar1254: 400 ppb, Ar1260: 50 ppb

Figure A.36: SAMPLE 36: Ar1242: 800 ppb, Ar1254: 200 ppb, Ar1260: 200 ppb
Figure A.37: SAMPLE 37: Ar1242: 200 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb

Figure A.38: SAMPLE 38: Ar1242: 0 ppb, Ar1254: 200 ppb, Ar1260: 0 ppb

Figure A.39: SAMPLE 39: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 200 ppb
Figure A.40: SAMPLE 40: Ar1242: 200 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb

Figure A.41: SAMPLE 41: Ar1242: 0 ppb, Ar1254: 200 ppb, Ar1260: 0 ppb

Figure A.42: SAMPLE 42: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 200 ppb
Figure A.43: SAMPLE 43: Ar1242: 200 ppb, Ar1254: 0 ppb, Ar1260: 0 ppb

Figure A.44: SAMPLE 44: Ar1242: 0 ppb, Ar1254: 200 ppb, Ar1260: 0 ppb

Figure A.45: SAMPLE 45: Ar1242: 0 ppb, Ar1254: 0 ppb, Ar1260: 200 ppb
VITA

Michael Williams was born in Kingsport, Tennessee, on February 4, 1970. He graduated from Kingsport’s Sullivan South High School in 1988. He enrolled in the University of Tennessee in Knoxville and he received a Bachelor of Science degree in Electrical Engineering in December 1993. In 1994, he entered the University of Tennessee graduate school as a Graduate Teaching Assistant in the Department of Electrical Engineering. Later that year he became a Graduate Research Assistant and began his research towards a Master of Science degree in Electrical Engineering. Williams received his master’s degree in EE in May 1996. He now works for Honeywell Space Systems in Clearwater, Florida, as a Systems Engineer.