Correction model based ANN modeling approach for the estimation of radon concentrations in Ohio

Pavan Yerrabolu
The University of Toledo
A Thesis

entitled

Correction Model Based ANN Modeling Approach for the Estimation of Radon Concentrations in Ohio

by

Pavan Yerrabolu

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Electrical Engineering

_______________________________________
Dr. Vijay Devabhaktuni, Committee Chair

_______________________________________
Dr. Ashok Kumar, Committee Member

_______________________________________
Dr. Mansoor Alam, Committee Member

_______________________________________
Dr. Patricia R. Komuniecki, Dean
College of Graduate Studies

The University of Toledo
August 2012
An Abstract of

Correction Model Based ANN Modeling Approach for the Estimation of Radon Concentrations in Ohio

by

Pavan Yerrabolu

Submitted to the Graduate Faculty as partial fulfillment of the requirements for the Master of Science Degree in Electrical Engineering

The University of Toledo
August 2012

According to National Cancer Institute, radon is one of the major causes for lung cancer related deaths after smoking in the United States. In order to prevent deaths due to radon inhalation there is a need to determine the level of radon concentration in each locality, e.g., zip-codes. However, factors like inapproachability hinder the process of estimating radon concentration in some places. In such places the radon concentrations could be estimated using several interpolation techniques. In this thesis, a new approach that improves the accuracy of the neural model with the help of sensitivity based correction model for modeling and estimating radon concentrations in Ohio is proposed. The results are compared with commonly used techniques such as kriging, radial basis function (RBF), inverse distance weighting (IDW), global polynomial interpolation (GPI), local polynomial interpolation (LPI) and the recently developed conventional ANN modeling approach. Further, model accuracies of all the above interpolation schemes are evaluated based on the ranked performance measures criteria with emphasis on the extreme-end (peak-end, low-end), and mid-range radon concentrations. The results demonstrate the effectiveness of the proposed approach in estimating the radon concentrations. The prediction accuracy of the proposed approach is found to be improved by 70-80% compared to the other techniques.
For my parents, fiancé, and friends.
Acknowledgements

I would like to express my sincere gratitude to my advisor Dr. Vijay Devabhaktuni and co-advisor Dr. Ashok Kumar for their guidance and encouragement throughout my research. I am grateful to ODH and USEPA for providing the radon data for research purposes to University of Toledo.

I would also like to thank my colleague Lakshman Mareddy for patiently helping me out whenever I was struck in my research. I am thankful to Deepak Bhatt and Arjun Akkala for giving valuable inputs during my research which made my work easier.

Special thanks to all my family especially my elder sister for supporting me morally throughout my research study. Without their support and motivation this research work wouldn’t have been possible.

Above all, I am thankful to Swetha Vanam, my fiancé, for being by my side through my tough times and for inspiring me whenever I lost hope. I am also thankful to my friends in Toledo on whom I could count on whenever I had a problem.
# Contents

Abstract iii

Acknowledgements v

Contents vi

List of Tables ix

List of Figures x

List of Abbreviations xii

1 Introduction 1

1.1 Problem Statement........................................................................................................... 1

1.2 Motivation.........................................................................................................................3

1.3 Organization of Thesis....................................................................................................4

2 Literature Survey 6

2.1 Role of Interpolation in Environmental Studies .........................................................6

2.2 Traditional Interpolation Techniques ...........................................................................8

2.3 Review of Artificial Neural Networks.............................................................................10

3 Artificial Neural Networks Approach 12

3.1 Introduction to ANN.......................................................................................................12

3.2 Structure of MLP3 ..........................................................................................................13

3.3 Data Preparation ..........................................................................................................15

vi
3.4 Methodology ............................................................................................................. 16
3.5 Results and Discussions .......................................................................................... 18

4 Correction Model based ANN Approach ................................................................. 21

4.1 Proposed Methodology .......................................................................................... 21

4.2 Ranked statistical performance measures for IAQ models .................................. 26

4.2.1 Mean of training and validation data ($M_t, M_v$) ........................................... 28
4.2.2 Standard deviation of training and validation data ($SD_t, SD_v$) .................. 28
4.2.3 Least square slope ($m$) and intercept ($b$) regression statistics ..................... 28
4.2.4 Paired peak accuracy ($A_p$) ............................................................................. 29
4.2.5 Geometric mean bias ($MG$) ........................................................................... 29
4.2.6 Geometric mean variance ($VG$) ................................................................. 29
4.2.7 Fractional bias ($FB$) ...................................................................................... 30
4.2.8 Fractional variance ($FV$) ............................................................................. 30
4.2.9 Normalized Mean Square Error ($NMSE$) ....................................................... 30
4.2.10 Factor of Two ($Fa2$) .................................................................................... 31
4.2.11 Coefficient of correlation ($r$) ..................................................................... 31
4.2.12 Spearman correlation coefficient ($\rho$) ......................................................... 31
4.2.13 Model comparison measure ($MCM_2$) ....................................................... 31
4.2.14 Robust highest concentration ($RHC$) ......................................................... 32
4.2.15 Revised Index of Agreement ($IOA_r$) .......................................................... 32
4.2.16 Scatter plots .................................................................................................. 33
4.2.17 Quantile-Quantile ($Q-Q$) plots ................................................................... 33
4.3 Results and Comparison ................................................................. 33

5 Conclusions and Future Work ............................................................ 46

5.1 Conclusions ................................................................................... 46

5.2 Future Work .................................................................................. 47

References ......................................................................................... 48

Appendices .......................................................................................... 59

A Source Code for Correction Model Based ANN modeling approach 59

B Screenshots of the Neuromodeler Software ........................................ 62

B.1 Neuromodeler Software Main Window .............................................. 62

B.2 New Model Creation Window in the Neuromodeler Software.............. 63

B.3 Screenshot of the training window of the Neuromodeler software, while training the MLP3 of the Conventional ANN technique .................................................. 64

B.4 Screenshot of the testing window of the Neuromodeler software, while validating the MLP3 of the Conventional ANN technique .................................................. 65

B.5 Screenshot of the Neuromodeler Software comparison window ............ 66

B.6 Screenshot of interpolated radon concentration surface by ArcGis spatial analyst .......................................................... 67
List of Tables

2.1 A list of available interpolation techniques commonly used in environmental studies. ...7
3.1 Training and Validation Errors of the Conventional ANN Model with the number of hidden neurons fixed at 10. ........................................................................................................19
3.2 Training and Validation Errors of ANN Models trained using Quasi-Newton Algorithm. ........................................................................................................................................20
4.1 Ranked statistical performance measures for IAQ models........................................27
4.2 Training and Validation Errors of candidate correction models. ..............................34
4.3 Comparison of Validation Errors of the Conventional ANN and Correction Model Based ANN Models..............................................................................................................34
4.4 Performance measures computed for traditional interpolation techniques, Conventional ANN and Correction model based ANN model......................................................................37
4.5 Performance measures computed for traditional interpolation techniques, Conventional ANN and Correction model based ANN model (Cont.)........................................38
List of Figures

1-1 Comparison of deaths per year in US due to radon and other causes (Source: http://www.epa.gov/radon/pubs/citguide.html) ................................................................. 2

1-2 Ohio Map showing the GM of radon concentrations(http://aprg.utoledo.edu/radon/). 3

3-1 Structure of MLP3 .............................................................................................................14

3-2 Plot of Sigmoid function........................................................................................................15

3-3 Proposed MLP3 structure for modeling radon concentrations. ..................................16

4-1 Candidate correction model with latitude as the output ...........................................23

4-2 Candidate correction model with longitude as the output ...........................................23

4-3 Flow-chart illustrating the Correction model based ANN correction approach .....26

4-4 Scatter plot of kriging technique ..................................................................................38

4-5 Scatter plot of GPI technique ..........................................................................................39

4-6 Scatter plot of LPI technique.. .........................................................................................39

4-7 Scatter plot of IDW technique .........................................................................................40

4-8 Scatter plot of RBF technique..........................................................................................40

4-9 Scatter plot of Conventional ANN technique .................................................................41

4-10 Scatter plot of Correction model based ANN technique ...........................................41

4-11 Q-Q plot of kriging technique..........................................................................................42

4-12 Q-Q plot of GPI technique.............................................................................................42
4-13 Q-Q plot of LPI technique ................................................................. 43
4-14 Q-Q plot of IDW technique................................................................. 43
4-15 Q-Q plot of RBF technique ................................................................. 44
4-16 Q-Q plot of Conventional ANN technique ........................................ 44
4-17 Q-Q plot of Correction model based ANN technique .......................... 45
B-1 Screenshot of the main window of the Neuromodeler software .......... 62
B-2 Screenshot of the new neural model creation window in the Neuromodeler software..... 63
B-3 Screenshot of the training window of the Neuromodeler software, while training the MLP3 of the Conventional ANN technique. ................................. 64
B-4 Screenshot of the testing window of the Neuromodeler software, while validating the MLP3 of the Conventional ANN technique. ........................................... 65
B-5 Screenshot of the Neuromodeler software window, showing the validation data (magenta line) and the predicted data (yellow line) ......................................................... 66
B-6 Estimated radon concentrations obtained for Ohio by using four conventional interpolation techniques ................................................................. 67
### List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>AFB</td>
<td>Absolute Fractional Bias</td>
</tr>
<tr>
<td>EPA</td>
<td>Environmental Protection Agency</td>
</tr>
<tr>
<td>Fa2</td>
<td>Factor of Two</td>
</tr>
<tr>
<td>FB</td>
<td>Fractional Bias</td>
</tr>
<tr>
<td>FV</td>
<td>Fractional Variance</td>
</tr>
<tr>
<td>GM</td>
<td>Geometric Mean</td>
</tr>
<tr>
<td>GPI</td>
<td>Global Polynomial Interpolation</td>
</tr>
<tr>
<td>IAQ</td>
<td>Indoor Air Quality</td>
</tr>
<tr>
<td>IDW</td>
<td>Inverse Distance Weighting</td>
</tr>
<tr>
<td>IOA&lt;sub&gt;y&lt;/sub&gt;</td>
<td>Index of Agreement</td>
</tr>
<tr>
<td>KBNN</td>
<td>Knowledge Based Neural Network</td>
</tr>
<tr>
<td>LPI</td>
<td>Local Polynomial Interpolation</td>
</tr>
<tr>
<td>MG</td>
<td>Geometric Mean Bias</td>
</tr>
<tr>
<td>MLP</td>
<td>Multi Layer Perceptron</td>
</tr>
<tr>
<td>MCM&lt;sub&gt;2&lt;/sub&gt;</td>
<td>Model Comparison Measure</td>
</tr>
<tr>
<td>NMSE</td>
<td>Normalized Mean Square Error</td>
</tr>
<tr>
<td>ODH</td>
<td>Ohio Department of Health</td>
</tr>
<tr>
<td>ORIS</td>
<td>Ohio Radon Information System</td>
</tr>
<tr>
<td>Q-Q</td>
<td>Quantile-Quantile</td>
</tr>
<tr>
<td>RBF</td>
<td>Radial Basis Function</td>
</tr>
<tr>
<td>RHC</td>
<td>Robust Highest Concentration</td>
</tr>
<tr>
<td>US</td>
<td>United States</td>
</tr>
<tr>
<td>USEPA</td>
<td>United States Environment Protection Agency</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Problem Statement

Radon is a naturally occurring radioactive, tasteless, and odorless gas which is the decay product of uranium in the soil. Radon enters a house or a building through many possible ways such as openings around floor drains or pipes, or through cracks in the floor or walls and it builds up to dangerous levels in closed spaces. Radon has adverse effects on humans when exposed to higher levels for prolonged durations of time, due to its radioactive nature [1]. In Unites States, radon accounts for about 21,000 lung cancer deaths every year, out of which 2,900 occur among people who have never smoked (EPA’s 2003 Assessment of Risks from Radon in Homes) [2].Fig 1-1 shows the comparison of deaths per year in US due to radon and other causes.
Owing to the serious effects of radon on human health, efforts are being taken by the United States Environmental Protection Agency (USEPA) to identify radon concentrations levels. In Ohio, under the direction of the Ohio Department of Health (ODH) radon concentration levels for more than 200,000 homes/schools have been gathered so far, from the year 1989 [3-7] and the data is being managed using different database management systems [8-12]. The University of Toledo, under the aegis of Ohio Air Quality Development Authority and ODH developed one such database known as Ohio Radon Information Systems (ORIS) [13, 14]. The ORIS also helps in determining the best mitigation system based on pre-mitigation level and post-mitigation level of radon concentrations stored in its database [14]. These mitigation levels and radon concentrations of various zip codes are reported by mitigation testers and contractors allover Ohio. However, radon data is not available for some zip codes owing to factors like inapproachability (refer Fig 1-2). At present out of 1492 zip codes in Ohio, data is available for 1261 zip codes [15]. For the remaining zip codes we need to estimate radon concentrations using interpolation techniques based on the available data.
1.2 Motivation

Artificial Neural Network (ANN)’s have emerged as a powerful tool in signal processing, pattern recognition, and other applications [16-18]. They have the advantage of faster learning capability as they are structurally parallel-distributed systems [19]. ANN based modeling has been proposed in [15] to estimate the radon concentrations in the zip codes where data is not available. This approach proved to be more accurate than the traditional interpolation techniques, such as Kriging, GPI, LPI, RBF, and IDW. The accuracy of the neural models depends on the adequacy of training data. A well-distributed, sufficient, and accurately measured training data is needed to obtain an accurate neural model [20-22]. As mentioned earlier, training data i.e., radon concentration is obtained through measurements at various zip codes by mitigation testers and contractors allover Ohio. Therefore, obtaining
a huge training dataset could be very expensive as more testers and testing equipment is required. However, even though sufficient and uniformly distributed data is provided, ANN’s have the drawback of poor extrapolation capability due to limited generalization of the model [22].

To enhance the generalization ability and extrapolation capability of the neural model, more advanced ANNs known as knowledge-based neural networks (KBNNs) have been developed. KBNN’s incorporate the existing engineering knowledge in the form of another neural model into the neural network structure [15, 18, 21]. Various KBNN models including prior knowledge input method, source difference method and space-mapped neural network were applied to estimate radon concentrations [22]. These KBNN models provided accurate radon estimates compared to ANN’s but at the cost of increased structural complexity.

To overcome the drawbacks of the above techniques, this thesis presents a new correction model based ANN approach which predicts the radon concentrations more accurately than all the previously used techniques.

1.3 Organization of Thesis

The remainder of the thesis is organized as follows. Chapter 2 discusses the role of interpolation in environmental studies and the traditional interpolation techniques commonly used. A review of ANN’s and their recent applications in various fields of science and engineering is provided.

Chapter 3 gives a detail introduction to ANNs, discusses the structure of Multi Layer Perceptrons (MLP), and gives an introduction to radon data preparation. Various performance measures used to assess spatial interpolation techniques are discussed. The
ANN based methodology in estimating radon concentrations is discussed in detail and the results are compared with traditional interpolation techniques.

Chapter 4 discusses the implementation of correction model based ANN approach in estimating radon concentrations, and the comparison of results with ANN based methodology as well as traditional interpolation techniques is provided. Indoor Air Quality (IAQ) performance measures used to select the best model in predicting the radon concentrations are discussed and the selection of best model is described. Finally, Chapter 5 gives the conclusions of the work done and future work.
Chapter 2

Literature Review

2.1 Role of Interpolation in Environmental Studies

Spatially continuous data (or spatially continuous surfaces) play an important role in environmental sciences and management. Environmental analysts usually require spatially continuous data over the region of interest to make effective decisions, and scientists need spatially continuous data across a region to make accurate interpretations [23]. However, such spatially continuous data cannot be acquired manually over the area of interest as the geography of certain areas is inapproachable by humans. Hence, data is collected at discrete locations (i.e. grids/points). In order to generate spatially continuous data, the value of an attribute at unsampled location needs to be estimated. In such a case, spatial interpolation methods are used to estimate the values of an environmental variable at unsampled location using data collected from discrete locations.

Spatial interpolation techniques are based on the concept that points which are close together in space tend to have similar value attributes. This is known as positive spatial autocorrelation. This principle is used in generating spatially continuous data from discrete data points. In the present case of radon concentration, the data provided by the ODH is
used to create radon concentration maps of the entire state of Ohio. This data is used in quantitative research to reduce cancer deaths due to radon gas inhalation. Therefore, it can be concluded that when an interpolation technique output is the basis for other research studies its accuracy becomes a critical factor. There are many performance measures which define the accuracy of an interpolation technique and these are discussed in Chapter 3 and 4 of this document.

Spatial interpolation techniques have been applied in various fields such as environmental studies, mining engineering, soil sciences, civil engineering and probability and statistics [23]. Table 2.1 shows the different interpolation techniques used in environmental studies, their principle of working, advantages, and disadvantages [24].

Table 2.1: Interpolation techniques used in environmental studies.

<table>
<thead>
<tr>
<th>Interpolation Technique</th>
<th>Principle</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPI</td>
<td>Captures global patterns in the data in fitting a polynomial</td>
<td>Interpolates global variations effectively</td>
<td>Exponential increase in prediction errors with increasing complexity in data</td>
</tr>
<tr>
<td>LPI</td>
<td>Similar to GPI, but the curve is fitted to a subset of data</td>
<td>Interpolates local variations effectively</td>
<td>Misses the global variations in data</td>
</tr>
<tr>
<td>Trend Surface Analysis</td>
<td>Separates the data based on regional and local trends</td>
<td>Removes broader variations prior to further analysis</td>
<td>Edge effects caused by spatial autocorrelation</td>
</tr>
<tr>
<td>IDW</td>
<td>Linear combination of data points, inversely weighted as per distance</td>
<td>Simple to use, and works well with any kind of data</td>
<td>Spatial arrangement of samples is not taken into consideration</td>
</tr>
<tr>
<td>Kriging</td>
<td>Similar to the principle of IDW and also includes spatial arrangement of data</td>
<td>Best in linear unbiased spatial prediction; and no edge-effects during polynomial fitting</td>
<td>Requires sophisticated programming</td>
</tr>
<tr>
<td>RBF</td>
<td>Fits a series of data points but the curve is not smooth</td>
<td>Requires few samples</td>
<td>Fails at extrapolation</td>
</tr>
</tbody>
</table>
2.2 Traditional Interpolation Techniques

GPI fits a smooth surface over the sampled data points using least-squares regression. The disadvantage of GPI is that the presence of outliers, i.e., extreme data values in the data leads to poor interpolation. This technique is more useful in determining the overall/global trend of the data and it lacks minor details. Global interpolation is used for:

1. Fitting a smooth surface to the sample points when the surface slowly varies over the area of interest.
2. Eliminating or studying the effects of global trends.

Recently, GPI has been employed in the field of image rectification [25] and climate characterization [26].

Unlike GPI, the LPI technique fits many polynomials for a subset of the data, each of which is optimized for a neighborhood. Although it is more flexible than GPI, LPI is not an exact interpolator like IDW. This technique captures local variations in data effectively and misses the global trend, which is a disadvantage of this technique. Therefore, this technique is best suited when the data is in the form of clusters with no significant global trends. Recent applications of LPI include hydrologic modeling of daily precipitations [27], modeling seasonal rainfall erosivity [28], estimating air temperatures [29], and mapping the spatial distribution of soil contaminants [30].

IDW technique weights the points nearer to the prediction location more than those that are far away, hence the name inverse distance weighting. In other words, as the distance of a surrounding point to the predicted location approaches zero, its relative weight in determining the value at predicted location becomes one. An advantage of IDW is that it can
efficiently estimate extreme changes in data such as: cliffs, fault lines etc. This technique works best when the training data is evenly distributed and is moderately dense. A disadvantage of IDW is that the presence of data clusters and/or outliers leads to ineffective interpolation. Another drawback is that this technique cannot estimate above maximum or below minimum values. Recently, IDW has been applied in the field of environmental sciences to map soil nitrogen, phosphorous and potassium [31], to map soil properties in Iran [32], to determine spatial variability of methane flux [33] and the variation in marine temperature and salinity [34].

Interpolation using radial basis functions is conceptually similar to exactly fitting a fairly stiff rubber sheet that bends and folds to the sample data points. This is achieved mathematically by approximating the relation between input and output data by using a linear combination of different radial basis functions. The five important radial basis function used by RBF are thin-plate spline, spline with tension, completely regularized spline, multiquadric function, inverse multiquadric function. The RBF’s create a rubber like membrane that is fits into each of the measured data points minimizing the area of curvature [35]. This technique is considered as a form of ANN where the radial basis functions functionality is similar to the activation function in a neural network.

Disadvantage of this technique is that the radial basis functions work effectively only when the data is smooth i.e., there should not be significant changes in the data values [36]. Most recently, a predictive control based on RBF is being used to control the supercritical main stream temperature of a power plant [37]. Other recent applications of RBF include in the field of photovoltaic’s [38], electric power components [39], health sciences [40] and antennas [41].
Kriging provides estimates for accuracy in predictions apart from predicting the values at unknown locations. It is similar to IDW, but the weights are determined based on spatial autocorrelation (function of distance) between data points. Recently kriging has been used to estimate the daily particulate matter PM 2.5 concentration [42], to develop Carbon: Nitrogen ratio surface [43], in remote sensing image analysis and geostatistics[44].

2.3 Review of Artificial Neural Networks

ANN’s due to their faster learning capability and efficient generalization of the model are being used in computer aided modeling applications. When the data pattern is complex, which is practically the case in environmental data, the traditional interpolation techniques fail to derive the input-output relationship efficiently. In such cases, ANN’s have been applied and the prediction accuracy has been found to be much better than the traditional interpolation techniques. The application of ANN’s in the field of environmental studies is dated back to early 1900's. The remaining section discusses the most recent applications of ANNs and its growing importance.

Recently, ANN’s were used to predict the mean monthly maximum and minimum temperatures to lake-basin scale in an arid region in India. These predictions were used to downscale the surface temperature for the Pichola lake region in Rajasthan State in India [45]. In Iran, ANN’s has been used to forecast evapotranspiration, which is the sum of evaporation and plant transpiration from earth’s surface [46]. The data provided a better understanding of the loss of water through soil and future groundwater level could be estimated. ANN’s are also being used to determine the amount of suspended sediment concentration, an important factor in the design and management of water resource projects [47].
In industrial countries air pollution has become a major problem and is being considered as a hindrance to socio-economic development. ANN’s have been proved to be effective in estimating the air quality by learning the air pollutants concentrations from the previous years [48]. For example, recently in Northern Italy, ANN’s were used to estimate the secondary pollutants concentrations namely VOC, NOx, NH3, primary PM and SO2 [49].

Recent applications of ANN in the field of material sciences include parametric modeling and optimization of LASOX cutting of mild steel [50], modeling adhesive wear resistance of Al-Si-Mg-/SiC PM compacts fabricated by hot pressing process [51], investigating boiling point heat transfer characteristics in TiO2 nanofluids [52] and studying the synthesis of TiO2 nanoparticles in different thermal conditions and modeling its photocatalytic activity [53].

Recently, ANN’s were used to model and optimize the production of biogas [54]. They are being applied for quantitative analysis of radioactive sources [55]. Most recent applications of ANN’s include in the field of metallurgy [56], polymer sciences [57], mineral processing [58], environmental research and risk assessment [59], mechatronics and automation [60].

The use of ANN based modeling techniques to predict atmospheric pollutant concentrations are increasing. Lately, [15, 22, 61-64] have used ANN’s to predict particulate matter concentrations and compared the prediction accuracy with other traditional interpolation techniques. Recently, the use of additional knowledge input to aid the learning process of ANN’s in predicting particulate concentrations has been demonstrated in [22]. From, [15] it was concluded that the ANN based modeling approach provided much better prediction accuracy compared to traditional interpolation techniques.
Chapter 3

Artificial Neural Network Approach

In this chapter we discuss a conventional ANN approach using quasi-newton algorithm to predict radon concentrations in the state of Ohio. A simple MLP3 network is used to develop the artificial neural network model. This chapter illustrates the procedure of data preparation, methodology of training and validation of the ANN model and comparison of the proposed methodology with traditional interpolation techniques. At first, the introduction to ANN and the structure of the most commonly used ANN known as Multi layer perceptrons (MLP) is discussed.

3.1 Introduction to ANN

ANN is a mathematical or computational model inspired by biological neural networks. ANN’s consists of information processing units known as neurons which are organized in layers. Generally there are three different types of neurons: input neurons, output neurons and hidden neurons. The input neurons are present in the input layer and receive information from outside the neural network, whereas the output neurons present in the output layer send processed information to an external recipient. Hidden neurons are the
ones that receive information from other neurons and their output serves as input to other neurons. They are present in the intermediate layers between input and output layers known as hidden layers. Generally ANN has an input layer, an output layer and one or more hidden layers. The neurons in these layers are interconnected through links, each of which has a weight assigned to it. For an artificial neuron, the inputs are multiplied by the weights assigned to it and then computed by a mathematical function known as activation function which determines the output of the neuron. By adjusting the weights of the artificial neuron we can obtain the desired output from a specified input. The process of adjusting the weights is known as training or learning. Different training algorithms are used to adjust the weights of the artificial neurons. ANNs combine artificial neurons in specific order to process information.

Of late, the most commonly used ANN’s are the multilayer feed-forward networks known as multilayer perceptrons or MLP networks. MLP networks are characterized by multiple hidden computational layers [65, 66] and multiple hidden neurons in each layer. The universal approximation theory states that MLP networks with a single hidden layer consisting of finite number of hidden neurons are universal approximators. Therefore, the structure of ANN used in this research is MLP3 neural network. In the next section the structure of MLP3 used in this research is discussed.

3.2 Structure of MLP3

As the name indicates a MLP3 neural network has three layers: input layer containing the input neurons, the hidden layer containing the hidden neurons and the output layer containing the output neurons. The structure of a typical MLP3 is shown in the next page.
The training algorithm used in this research to train the weights of MLP3 is Quasi Newton. The activation function for the input neurons is the relay function whereas for the output neurons it is a linear function of the weighted sum of inputs given to the output neurons. The hidden layer neurons are activated by sigmoid function because it is simple to compute \[ 67 \]. The sigmoid function is given by

\[
Z(\gamma) = \frac{1}{1 + e^{-\gamma}},
\]

(3.1)

where \( \gamma \) is the weighted sum of inputs to the hidden neuron and \( Z(\gamma) \) is the output of the hidden neuron. The plot of sigmoid function is shown in the next page.
3.3 Data Preparation

As mentioned in Chapter 1, radon concentration levels gathered from more than 200,000 homes/schools across Ohio are being stored in a database known as ORIS which is developed and maintained by the University Of Toledo. The data is provided by university researchers, government agencies, and commercial testing companies and has been published in a series of technical papers and reports published by

The radon data reported consists of the following information: the zip code of the tested building, type of room where the test was conducted, type of radon detection device used, and the season of the year when the test was performed. Out of the above information, during training of ANN or while using ArcGis spatial analyst software, the input data consists of only the latitude and longitude of a zip code and corresponding geometric mean of the radon concentration at that zip code. The procedure to obtain the geometric mean of the radon concentration at a particular zip code is explained in [24]. The radon data used in
this research was obtained during the year 2010 and had been reformatted by Arjun Akkala [24].

3.4 Methodology

In the case of radon modeling, where the model inputs, latitude and longitude are represented by $\mathbf{x}$, and the corresponding model output by $y$, such that

$$y = f(\mathbf{x}),$$

\hspace{1cm} (3.2)

where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ represents the input vector, $n$ is the total number of inputs, which is two for radon modeling as latitude and longitude are the inputs to the neural network and $f$ represents the functional relationship between $\mathbf{x}$ and $y$. Since there is only one output \textit{i.e.}, radon concentration, $y$ is a scalar. In conventional ANN modeling approach a MLP3 neural network (refer Fig 3-3) is trained using the Quasi-Newton algorithm to obtain the ANN model $f_{\text{ann}}$ that approximates $f$.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3-3.png}
\caption{Proposed MLP3 structure for modeling radon concentrations.}
\end{figure}
To train the MLP3 neural network, the available dataset containing latitude, longitude and radon concentration is randomly divided into training and validation datasets represented by

\[
\{(x_i, y_i), i = 1... N_t\}, \tag{3.3}
\]

\[
\{(x_i, y_i), i = 1... N_v\}, \tag{3.4}
\]

where \(y_i\) is the desired output for the corresponding input vector \(x\) and \(N_t, N_v\) represent the number of training and validation samples respectively. Initially, 90\% of the given data is selected as training data. The ANN is trained iteratively using the training data which involves the adjustment of weights in the neural network. In such a case, eq. (3.2) can also be represented as

\[ y = f_{\text{ann}}(x, w), \tag{3.5} \]

where \(w\) is the ANN weight vector. Once the ANN is trained using the Quasi-Newton algorithm the final step is the validation of the model which is done by using the validation data. The validation of the model is based on certain quality measures. The quality of the developed neural network is assessed using the quality measures \(i.e.,\) user defined error \(E_{\text{user}},\) average validation error \(E_{\text{avg}}\) and worst-case error \(E_{\text{worst}}.\) The \(E_{\text{user}}\) is user defined and is set as 5\%. \(E_{\text{avg}}\) and \(E_{\text{worst}}\) are expressed as

\[
E_{\text{avg}} = \sqrt{\frac{1}{N_v} \sum_{i=1}^{N_v} \left( \frac{d_i - f_{\text{ann}}(x_i, w)}{d_i} \right)^2} * 100, \tag{3.6}
\]

and

\[
E_{\text{worst}} = \max_i \left( \frac{d_i - f_{\text{ann}}(x_i, w)}{d_i} \right) * 100, \tag{3.7}
\]
where \( f_{ann}(x, w) \) is the output of ANN and \( d_i \) is the desired output corresponding to the input vector \( x_i \), and \( N_v \) represents the number of validation samples. The above procedure is repeated by varying the percentages of training and validation data and in each case the average validation error \( E_{avg} \) and the worst-case error \( E_{worst} \) are calculated. Then, the percentage of training data for which the neural network model gives least validation average error is identified.

Now, with the percentage of training data fixed, the number of hidden neurons is varied in steps of 10. In each step, the neural network model is trained and validated as described earlier and the average validation error \( E_{avg} \) and the worst-case error \( E_{worst} \) are calculated. The neural network model which gives the least average validation error, \( E_{avg} \), is identified as the best model for the conventional ANN approach. Finally, the performance of the conventional ANN approach is assessed based on whether the condition \( E_{avg} \leq E_{user} \) is satisfied.

### 3.5 Results and Discussion

The given radon concentration data is randomly divided into training and validation data. Initially, 90% of the given data is selected as the training data and the remaining 10% as the validation data. The number of hidden neurons in the conventional ANN model is fixed at 10. The model is then trained and validated using the training and validation data respectively. The percentage of training and validation data is varied and the average validation error \( E_{avg} \) and \( E_{worst} \) shown in eq. (3.6) and eq. (3.7) are calculated in each case. Table 3.1 shows the change in the validation error with the change in percentage of training and validation data. From Table 3.1 it is clear that the MLP3 network performs best when
the training data is 90% of the given radon concentration as it gives least validation error in that particular case.

Table 3.1: Training and Validation Errors of the Conventional ANN Model with the number of hidden neurons fixed at 10.

<table>
<thead>
<tr>
<th>% Training Data</th>
<th>% Validation Data</th>
<th>Training Error (%)</th>
<th>Validation Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>10</td>
<td>10.81</td>
<td>12.13</td>
</tr>
<tr>
<td>80</td>
<td>20</td>
<td>11.06</td>
<td>13.26</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>11.69</td>
<td>12.66</td>
</tr>
<tr>
<td>60</td>
<td>40</td>
<td>12.14</td>
<td>12.85</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>12.41</td>
<td>12.36</td>
</tr>
</tbody>
</table>

With the training and validation data fixed at 90% and 10% of the total dataset the number of hidden neurons in the conventional ANN model is varied. The average validation error $E_{\text{avg}}$ and $E_{\text{worst}}$ shown in eq. (3.6) and eq. (3.7) are calculated in each case. The neural model which gives the least average validation error, $E_{\text{avg}}$, is identified as the best model for the conventional ANN approach. Table 3.2 shows the training and validation error with varying number of hidden neurons. From Table 3.2, it is clear that the neural model which consists of 10 hidden neurons gives the least validation error. Therefore, the number of hidden neurons in the conventional ANN model is fixed at 10 and the $E_{\text{avg}}$ of this model is 12.13%.
Table 3.2: Training and Validation Errors of ANN Models trained using Quasi-Newton Algorithm.

<table>
<thead>
<tr>
<th>No. of Hidden Neurons</th>
<th>Training Error (%)</th>
<th>Validation Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$E_{avg}$ (%)</td>
</tr>
<tr>
<td>10</td>
<td>10.81</td>
<td>12.13</td>
</tr>
<tr>
<td>20</td>
<td>10.59</td>
<td>12.89</td>
</tr>
<tr>
<td>30</td>
<td>10.71</td>
<td>12.91</td>
</tr>
<tr>
<td>40</td>
<td>10.63</td>
<td>12.74</td>
</tr>
<tr>
<td>50</td>
<td>10.66</td>
<td>12.60</td>
</tr>
<tr>
<td>60</td>
<td>10.54</td>
<td>12.86</td>
</tr>
<tr>
<td>70</td>
<td>10.69</td>
<td>12.60</td>
</tr>
<tr>
<td>80</td>
<td>10.46</td>
<td>12.74</td>
</tr>
<tr>
<td>90</td>
<td>10.21</td>
<td>13.15</td>
</tr>
</tbody>
</table>

From Table 3.2 it is clear that the condition $E_{avg} \leq E_{worst}$ is not achieved by the conventional ANN approach. Owing to the limited generalization capability, the conventional ANN modeling approach is not suitable for extrapolation. To overcome these drawbacks, in the next chapter a new correction model based ANN approach is proposed.
Chapter 4

Correction Model Based ANN Approach

Due to the inadequacy of training data the conventional ANN modeling accuracy is limited and the condition $E_{\text{avg}} \leq E_{\text{user}}$ is not achieved. Owing to the limited generalization capability, the conventional ANN modeling approach is not suitable for extrapolation. In such a case, the next best alternative can be the KBNN’s.

Various KBNN models like prior knowledge input method, source difference method and space-mapped neural networks were applied to estimate radon concentrations [22]. However, using uranium as the knowledge input in addition to latitude and longitude tends to increase the structural complexity of the ANN. To overcome these issues, the correction model based ANN approach is proposed.

4.1 Proposed Methodology

In this approach, apart from the original less-accurate neural model $f_{\text{ann}}$ a set of neural models $f_{\text{ann},j}$ referred to as candidate correction models are defined [68] to improve the accuracy of $f_{\text{ann}}$ such that condition $E_{\text{avg}} \leq E_{\text{user}}$ is achieved. Let us first define the structure of
The structure of $f_{\text{ann}}$ is problem dependent whereas, that of $f_{\text{ann},j}$ is not known initially. Let the desired neural model be

$$y = f_{\text{ann}}(x_1, x_2, x_3, \ldots, x_n, w),$$  \hspace{1cm} (4.1)

and the corresponding set of candidate correction models be

$$x_{1,c} = f_{\text{ann},1}(y, x_2, \ldots, x_n, w_1),$$
$$x_{2,c} = f_{\text{ann},2}(x_1, y, x_3, \ldots, x_n, w_2),$$
$$\vdots$$
$$\vdots$$
$$x_{t,c} = f_{\text{ann},t}(x_1, x_2, \ldots, x_{n-1}, y, w_t),$$  \hspace{1cm} (4.2)

where $x_1, x_2, x_3, \ldots, x_t$ are the scalar inputs, $n$ is the total number of inputs, $f_{\text{ann},j}$, $x_{i,c}$ and $w_i$ represents the $i$th candidate correction model, $1 \leq i \leq t$, its corresponding output, and its corresponding weight vector respectively. From eq. (4.2), it is clear that the number of candidate correction models is equal to the total number of inputs in the original less accurate neural model. In the case of radon modeling, as mentioned earlier two inputs are fed to the neural model i.e., latitude and longitude, so we have two candidate correction models (refer Fig 4-1 and Fig 4-2). The eq. (4.2) can be rewritten as

$$x_{1,c} = f_{\text{ann},1}(y, x_2, w_1),$$  \hspace{1cm} (4.3)
$$x_{2,c} = f_{\text{ann},2}(x_1, y, w_2),$$  \hspace{1cm} (4.4)

where $x_{1,c}$ is the output of the first candidate correction model i.e., latitude, $x_{2,c}$ is the output of the second candidate correction model i.e., longitude, $y$ is the radon concentration, $x_1, x_2$ are the latitude and longitude inputs, and $w_1, w_2$ are the ANN weight vectors of the first and second candidate correction models respectively.
The first step is to determine the best correction model out of the two candidate correction models. The error criterion employed in selecting the correction model is the quality measure, $E$, mentioned in eq. (3.6). First, one of the candidate correction models, $f_{ann,1}$ is considered. As shown in Fig 4-1, the inputs to the first candidate correction model are radon concentration and longitude with latitude being the output. Accordingly, the dataset is
modified by swapping the latitude and radon concentration columns. The modified dataset is randomly divided into training data and validation data. Using the training data, the \( f_{\text{ann},1} \) is trained by adjusting \( \mathbf{w}_i \) iteratively using Neuromodeler [69] software. This model gives an output \( x_{i,c} \) (which closely approximates \( x_i \)). Then, quality measure \( E_i \) of the \( f_{\text{ann},1} \) is evaluated. This process is repeated for the \( f_{\text{ann},2} \) and \( E_2 \) is calculated. Out of the two candidate correction models the one with the least validation error is considered as the correction model \( f_{\text{ann},j} \) according to

\[
j = \arg \min_i \left( E_i \right),
\]

where \( i = 1, 2 \).

As the final step, once the correction model \( f_{\text{ann},j} \) is determined, it is paired with the less-accurate original neural model \( f_{\text{ann}} \). A root-finding method is iteratively implemented to improve the accuracy of neural model \( f_{\text{ann}} \) until the condition \( E_{\text{avg}} \leq E_{\text{user}} \) is achieved.

In this thesis we define a new root finding technique which has the advantages of Newton’s method and the steepest-descent method [70]. In this technique, the sensitivity information, \( i.e., \) partial derivatives of the correction model is used to determine the step-size and update direction. The chain rule of calculus is applied to the correction model \( f_{\text{ann},j} \) to get the partial derivatives of the correction model \( f_{\text{ann},j} \) with respect to output \( y \)

\[
\frac{\partial x_{j,c}}{\partial y} = \sum_{k=1}^{r} \frac{\partial x_{j,c}}{\partial z_{k}} \frac{\partial z_{k}}{\partial \gamma_{k}} \frac{\partial \gamma_{k}}{\partial y},
\]

and

\[
z_k = \frac{1}{1 + e^{-\gamma_k}}.
\]

In eq. (4.6) and eq. (4.7), \( \gamma_k \) is the weighted sum of all inputs to \( f_{\text{ann},j} \), \( z_k \) is the sigmoid activation function and \( r \) is the number of hidden neurons in \( f_{\text{ann},j} \).
The output \( y \) from \( f_{ann} \) serves as the initial solution or starting point to the root finding algorithm based on “\( f_{ann} \) and \( f_{ann,j} \) pair” and from their output \( y \) is iteratively improved until the condition \( E_{avg} \leq E_{user} \) is achieved. The derivative information from the eq. (4.6) and eq. (4.7) play an important role in determining both step-size and update direction which result in much accurately predicted output \( y \) compared to the conventional ANN approach. The pseudo code and the flow chart (refer to Fig 4-3) for the proposed approach are shown below.

**Pseudo Code of the Correction model based Approach**

**Step 1:** Feed \((x_1, x_2, x_3, \ldots, x_n)\), to \( f_{ann} \) to which gives an output \( y \);

**Step 2:** Feed \((x_1, x_2, x_3, \ldots, x_{j-1}, y, x_{j+1}, \ldots, x_n)\) to the candidate correction model \( f_{ann,j} \) leading to \( x_j, c \), an approximation of \( x_j \);

\[
E_{obj} = \frac{x_j - x_{j,c}}{|x_j|} * 100
\]

**Step 3:** Calculate the objective function,

**Step 4:** If \( E_{obj} << E_{user} \), then return \( y \) and STOP, else, calculate residue \( R = x_j - x_{j,c} \) and sensitivity of the correction model (eq. (4.6));

**Step 5:** Calculate \( \delta \) which is the ratio of residue and sensitivity. Update \( y = y - \alpha * \delta \) (To accelerate convergence, \( \delta \) is multiplied by a constant \( \alpha > 0 \));

**Step 6:** Go to **Step 2** and updated \( y \);

**Step 7:** Return \( y \) and STOP.
4.2 Ranked statistical performance measures for IAQ models

In this section a brief introduction of different performance measures used to assess the performance of different interpolation schemes is discussed. The rank (or weightage) of the performance measures change depending on whether the assessed air quality models
estimate extreme-end (peak-end, low-end), and mid-range contaminant concentrations. For example while assessing the performance of the air quality models when estimating extreme-end (peak-end, low-end) contaminant concentrations the rank of the performance parameter $FV$ is four whereas the same parameter is ranked sixth when the assessed air quality models estimate mid-range contaminant concentrations. Since our study is focused on the indoor radon concentrations in the Ohio homes, we will follow the performance measures suggested for IAQ models discussed in [71]. Table 4.1 shows the ranked performance measures for IAQ models along with the acceptable limits for each measure based on extensive literature review [71].

Table 4.1: Ranked statistical performance measures for IAQ models.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Peak-End Estimates</th>
<th>Low-End Estimates</th>
<th>Mid-Range Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$0.9 \leq \rho \leq 1.0$</td>
<td>$0.9 \leq \rho \leq 1.0$</td>
<td>$0.9 \leq \tau \leq 1.0$</td>
</tr>
<tr>
<td>2</td>
<td>$0.75 \leq m \leq 1.25$</td>
<td>$0.75 \leq m \leq 1.25$</td>
<td>$0.75 \leq m \leq 1.25$</td>
</tr>
<tr>
<td>3</td>
<td>$-25 \leq (b/C0) % \leq 25$</td>
<td>$-25 \leq (b/C0) % \leq 25$</td>
<td>$-25 \leq (b/C0) % \leq 25$</td>
</tr>
<tr>
<td>4</td>
<td>$-0.5 \leq FV \leq 0.5$</td>
<td>$-0.5 \leq FV \leq 0.5$</td>
<td>$0 \leq \text{NMSE} \leq 0.25$</td>
</tr>
<tr>
<td>5</td>
<td>$0.8 \leq Fa2 \leq 1.2$</td>
<td>$0.8 \leq Fa2 \leq 1.2$</td>
<td>$-0.25 \leq FB \leq 0.25$</td>
</tr>
<tr>
<td>6</td>
<td>$0.8 \leq RHC \leq 1.2$</td>
<td>$0 \leq MCM_2 \leq 1.2$</td>
<td>$-0.5 \leq FV \leq 0.5$</td>
</tr>
<tr>
<td>7</td>
<td>$0 \leq MCM_2 \leq 1.2$</td>
<td>$0.8 \leq MG \leq 1.2$</td>
<td>Bootstrap CI over $r$, NMSE, FB</td>
</tr>
<tr>
<td>8</td>
<td>$0.8 \leq MG \leq 1.2$</td>
<td>$0.8 \leq VG \leq 1.2$</td>
<td>$0.8 \leq Fa2 \leq 1.2$</td>
</tr>
<tr>
<td>9</td>
<td>$0.8 \leq VG \leq 1.2$</td>
<td>Bootstrap CI over $MG$, $VG$</td>
<td>$0 \leq MCM_2 \leq 1.2$</td>
</tr>
<tr>
<td>10</td>
<td>Bootstrap CI over $MG$, $VG$</td>
<td>Scatter plots</td>
<td>$0.8 \leq MG \leq 1.2$</td>
</tr>
<tr>
<td>11</td>
<td>Scatter plots</td>
<td>QQ plots</td>
<td>$0.8 \leq VG \leq 1.2$</td>
</tr>
<tr>
<td>12</td>
<td>QQ plots</td>
<td>-</td>
<td>Bootstrap CI over $MG$, $VG$</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>$0.8 \leq IOA_r \leq 1.0$</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-</td>
<td>Scatter plots</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
<td>QQ plots</td>
</tr>
<tr>
<td>16</td>
<td>$0.8 \leq IOA_r \leq 1.0$</td>
<td>$0.8 \leq IOA_r \leq 1.0$</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>$-15 \leq A_{ip} \leq 15$</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The following representation is used in the performance measures equations: \( y_k \) is the output of the correction model based ANN; \( d_k \) is the corresponding desired output of the \( k^{th} \) sample in the validation data, and \( N_v \) represents the number of validation samples.

### 4.2.1 Mean of training and validation data \((M_t, M_v)\)

The mean statistics of training and validation data are expressed as

\[
M_t = \frac{1}{N_v} \sum_{k=1}^{N_v} d_k, \tag{4.8}
\]

and

\[
M_v = \frac{1}{N_v} \sum_{k=1}^{N_v} y_k. \tag{4.9}
\]

### 4.2.2 Standard deviation of training and validation data \((SD_t, SD_v)\)

\( SD_t \) and \( SD_v \) statistics measure the variability from the mean statistics in training and validation data, respectively. They are expressed as

\[
SD_t = \sqrt{\frac{1}{N_v} \sum_{k=1}^{N_v} |d_k - M_t|}, \tag{4.10}
\]

And

\[
SD_v = \sqrt{\frac{1}{N_v} \sum_{k=1}^{N_v} |y_k - M_v|}. \tag{4.11}
\]

### 4.2.3 Least square slope \((m)\) and intercept \((b)\) regression statistics

The least square slope ‘\(m\)’ and intercept ‘\(b\)’ for a line ‘\(y = b + mx\)’ is expressed as
\[ m = \left[ \frac{\sum_{k=1}^{N_v} d_k y_k - \left( \sum_{k=1}^{N_v} d_k \right) \left( \sum_{k=1}^{N_v} y_k \right)}{\sum_{k=1}^{N_v} d_k^2 - \left( \sum_{k=1}^{N_v} d_k \right)^2} \right], \] (4.12)

and

\[ b = \frac{\left[ \sum_{k=1}^{N_v} y_k - m \left( \sum_{k=1}^{N_v} d_k \right) \right]}{N_v}. \] (4.13)

### 4.2.4 Paired peak accuracy \((A_p)\)

Paired peak accuracy \((A_p)\) measure serve as a supplementary test of model performance and is given by

\[ A_p = \frac{y_r - d_r}{d_r} \times 100\%, \] (4.14)

where \(d_r\) is the peak value of the validation data and \(y_r\) is the corresponding output from the correction model based ANN.

### 4.2.5 Geometric mean bias \((MG)\)

Geometric mean bias \((MG)\) is defined as the exponential function of the difference between means of the logarithmic function of observed values and logarithmic function of predicted values and is expressed as

\[ MG = \exp \left[ \ln d_r - \ln y_r \right]. \] (4.15)

### 4.2.6 Geometric mean variance \((VG)\)

Geometric mean variance \((VG)\) is defined as the exponential function of mean of square of the difference between logarithmic function of observed values and logarithmic function of observed values and is expressed as
\[ VG = \exp\left(\ln d_k - \ln y_k\right)^2 \]. \hfill (4.16)

**4.2.7 Fractional bias (FB)**

Fractional bias (FB) is defined as the mean error that measures the residual of the observed and predicted concentrations. Typically, FB varies between -2 and +2 with an ideal value of zero. It is expressed as

\[ FB = 2 \left( \frac{d_k - y_k}{d_k + y_k} \right). \hfill (4.17) \]

**4.2.8 Fractional variance (FV)**

The fractional variance (FV) is defined as the normalization of the mean bias of the variances of the observed and predicted values. The ideal value of FV is zero.

\[ FV = 2 \left( \frac{\sigma_{d_k} - \sigma_{y_k}}{\sigma_{d_k} + \sigma_{y_k}} \right), \hfill (4.18) \]

where \( \sigma_{d_i} \) is the variance of the observed values \( i.e., \) the input and \( \sigma_j \) is the variance of the corresponding outputs from correction model ANN.

**4.2.9 Normalized Mean Square Error (NMSE)**

Normalized mean square error (NMSE) emphasizes the scatter of the whole data and estimates the total deviation between observed and predicted values. To ensure that the NMSE is not biased towards over prediction or under prediction normalization is done by the denominator. Smaller values of NMSE indicate better model performance, the ideal value being zero.

\[ NMSE = \frac{(d_k - y_k)^2}{d_k \cdot y_k}. \hfill (4.19) \]
4.2.10 Factor of Two ($Fa2$)

Factor of two ($Fa2$) is defined as the percentage of predictions of the observed values within a factor of two. The ideal value of $Fa2$ is 1 (or 100%).

\[ Fa2 = \text{Fraction of data satisfying } 0.5 \leq \frac{y_k}{d_k} \leq 2.0. \]  \hspace{1cm} (4.20)

4.2.11 Coefficient of correlation ($r$)

Coefficient of correlation ($r$) describes the proportional change of the means of two quantities. The ideal value of $r$ is 1.0 which implies good agreement between the observed and predicted values.

\[ r = \frac{\left( d_k - \bar{d}_k \right) \left( y_k - \bar{y}_k \right)}{\sigma_k \cdot \sigma_{y_k}}. \]  \hspace{1cm} (4.21)

4.2.12 Spearman correlation coefficient ($\rho$)

Spearman correlation coefficient ($\rho$) is defined as the correlation coefficient ($r$) between the ranked variables. It's ideal value is 1.0 and is expressed as

\[ \rho = -\frac{\sum_{k=1}^{N} \left( d_k - \bar{d}_k \right) \left( y_k - \bar{y}_k \right)}{\sqrt{\sum_{k=1}^{N} \left( d_k - \bar{d}_k \right)^2 \sum_{k=1}^{N} \left( y_k - \bar{y}_k \right)^2}}. \]  \hspace{1cm} (4.22)

4.2.13 Model comparison measure ($MCM_2$)

Model comparison measure ($MCM_2$) is expressed as

\[ MCM_2 = \text{Avg}(AFB) + 2 \times \text{Std}(AFB), \]  \hspace{1cm} (4.23)
\[ AFB = 2 \left( \frac{\bar{d}_k - \bar{y}_k}{\bar{d}_k + \bar{y}_k} \right), \quad (4.24) \]

where \( \text{Avg} \) (AFB) and \( \text{Std} \) (AFB) are the average and standard deviation of the absolute fractional bias (AFB) which is calculated from eq. (4.24). The smaller value of \( MCM_2 \) indicates that the model on average compares best with observations.

### 4.2.14 Robust highest concentration (RHC)

Robust highest concentration (RHC) performance measure is used to study the peak concentrations that are highly variable. It is expressed as

\[ RHC = C(R) + \Theta \times \ln \left( \frac{3R-1}{2} \right), \quad (4.25) \]

Where \( C(R) \) is the \( R \)th largest value; \( \Theta \) is the average of the \( R-1 \) largest values; and \( R \) is the minimum of the number of values used to distinguish the “upper end” of the distribution and number of values exceeding a threshold value. \( R \) value is “26” for frequency distributions involving a year’s worth of values (averaging times of 24 hr or less).

### 4.2.15 Revised Index of Agreement (IOA)

Revised Index of Agreement (IOA) indicates the sum of the magnitudes of the differences between the predicted and observed deviations about the observed mean of the sum of the magnitudes of the ideal model and observed deviations about the observed mean. \( IOA_i \) ranges from -1 to +1 and is expressed as

\[ IOA_i = 1 - \frac{\sum |y_r - y_o|}{2 \sum |y_o - M|}, \quad (4.26) \]

where \( M_o \) is the mean of the observed values.
4.2.16 Scatter plots

Scatter plots are used to determine the bias and error in the observed and predicted values. They are better suited for correlation measures.

4.2.17 Quantile-Quantile (Q-Q) plots

Quantile-Quantile (Q-Q) plots are cumulative frequency distributions of observed and modeled values over their entire ranges. They provide a visual characterization of the scattering of predicted and estimated values with respect to the 1:1 line. Over the 1:1 line signifies a model over prediction and under the 1:1 line signifies a model under prediction.

Quality of the proposed correction based ANN model against the existing techniques is evaluated based on these performance measures and is discussed in next section.

4.3 Results and Comparison

For the correction modeling approach, out of the two candidate correction models the model with the least validation error is selected for the implementation of root finding algorithm along with the less-accurate original ANN model. Table 4.2 shows the training and validation errors of both the candidate correction models and it is clear that the first correction model with latitude as output has the least validation error and is preferred over the latter. Since the correction model and the original less accurate conventional ANN model form a mutually supportive pair while implementing the algorithm, it is logical to keep the structure of both the models same. For this purpose, initially the number of hidden neurons in the correction model is fixed at 10.
Table 4.2: Training and Validation Errors of candidate correction models.

<table>
<thead>
<tr>
<th>Model(s)</th>
<th>Training Error (%)</th>
<th>Validation Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$E_{\text{avg}}$ (%)</td>
</tr>
<tr>
<td>Candidate Correction Model 1</td>
<td>15.88</td>
<td>15.79</td>
</tr>
<tr>
<td>Candidate Correction Model 2</td>
<td>21.19</td>
<td>23.15</td>
</tr>
</tbody>
</table>

As the correction model is determined, the proposed pseudo code is applied to the mutually supportive pair of the conventional ANN model and the correction model, each with 10 hidden neurons and the radon output is estimated. The validation error $E_{\text{avg}}$ shown in eq. (3.6) is calculated. It is observed that the validation error $E$ for the correction model ANN approach is well below the $E_{\text{user}}$ of 5% compared to the conventional ANN approach as shown in Table 4.3.

Table 4.3: Comparison of Validation Errors of the Conventional ANN and Correction Model Based ANN Models.

<table>
<thead>
<tr>
<th>Model(s)</th>
<th>Validation Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$E_{\text{avg}}$ (%)</td>
</tr>
<tr>
<td>Conventional ANN</td>
<td>12.13</td>
</tr>
<tr>
<td>Correction model based ANN</td>
<td>3.40</td>
</tr>
</tbody>
</table>

Further, the performance of the proposed method is compared with the traditional interpolation techniques and the conventional ANN method based on various model performance measures discussed in Section 4.2. ArcGIS geostatistical analyst is used for estimating radon data utilizing traditional interpolation techniques such as kriging, RBF, IDW, GPI, and LPI. The same set of training and validation data that were used to model
radon concentrations in the ANN based approaches are used for modeling radon concentrations from the interpolation techniques.

Table 4.4 and Table 4.5 present a summary of the performance measures of the traditional interpolation techniques; the conventional and proposed ANN based models for evaluation purposes. BOOT v2.0 is used to compute the performance measures of $M, SD, r, NMSE, FB, Fa2, MG, VG$ and $FV$. The remaining performance measures of $b/C_0, m, IOA_r, A_p, RHC, MCM_2, \rho$ can be computed using Microsoft® Excel and the equations presented in Section 4.2.

The ranked performance measures discussed in Section 4.2 are used to determine the best model in estimating the extreme-end (peak-end, low-end), and mid-range radon concentrations. This is achieved by eliminating all the models whose performance measures are not within the limit mentioned in Table 4.1. Elimination of the models is a step by step procedure where in each step a different performance measure is used to determine which models are to be considered for further evaluation. Initially, all the models are evaluated based on rank one performance measure and those models that do not satisfy the rank one performance measure criterion are eliminated. In the next step, evaluation of the remaining models is continued based on the rank two performance measure and so on. This evaluation procedure is continued until one is left with only one model which is considered the best compared to other models in predicting radon concentrations.

From Table 4.4, one can observe that only the correction model based ANN was within the acceptance criterion of $0.9 \leq \rho \leq 1.0$ (ranked one for extreme-end estimates) and $0.9 \leq r \leq 1.0$ (ranked one for mid-range estimates). Apart from $\rho$ and $r$, the traditional interpolation techniques and the conventional ANN model failed to satisfy the $0.75 \leq m \leq 1.25$ criterion (ranked two for the extreme-end and mid-range estimates) and $-25 \leq b/C_0 (%) \leq 25$ criterion.
(ranked three for the extreme-end and mid-range estimates). The correction model based ANN satisfied the rank two and rank three criterions as well. Therefore, at this step, the traditional interpolation techniques and the conventional ANN model should not be considered for further evaluation. Alternatively, incorporating more variables such as gravel deposits and the presence of glacial sand that are correlated to radon concentrations helps in reconsidering these models [36]. However, even after reconsideration only the correction model based ANN meet the criterion of \(-0.5 \leq FV \leq 0.5\) (ranked four for the extreme-end estimates) while the remaining interpolation techniques were out of range. At this point, we can conclude that only the correction model based ANN is considered best compared to the remaining models in estimating the extreme-end values of radon concentrations.

Let us continue the evaluation of the remaining models after reconsideration in estimating the mid-range values. From Table 4.4 and Table 4.5, it is clear that kriging, LPI, conventional ANN approach and the correction model based ANN approach meet the criteria of \(0 \leq \text{NMSE} \leq 0.25\) (ranked four) and \(-0.25 \leq FB \leq 0.25\) (ranked five). At this stage, all the above interpolation techniques are considered for further evaluation. However, only the correction model based ANN satisfied the \(-0.5 \leq FV \leq 0.5\) criterion (ranked six) and the remaining interpolation techniques are eliminated. Eventually, even in the case of estimating the mid-range values of radon concentration the correction model based ANN proved to be the best model.

As mentioned in Chapter 2, the training data pattern greatly affects the accuracy of traditional interpolation schemes. Since the training data in this case is randomly selected from the total radon dataset no traditional interpolation technique works universally best which leads to poor interpolation and eventually these techniques fail to satisfy the above criterion for the ranked performance measures. From Table 4.4 and Table 4.5, it is clear that
the performance measures of conventional ANN model is better than all the traditional interpolation techniques because its performance does not entirely depend on the training data pattern and the network weights are adjusted accordingly during training. However, the model accuracy of the conventional ANN model is not enough to satisfy the ranked performance parameters criterion. Table 4.4 and Table 4.5 shows that the correction model based ANN model satisfies the criterion for all the ranked performance measures. Therefore, it can be concluded that the proposed approach is flexible and it interpolates data accurately for any randomly generated dataset even with a lesser degree of correlation between them.

Table 4.4: Performance measures computed for traditional interpolation techniques, Conventional ANN and Correction model based ANN model.

<table>
<thead>
<tr>
<th>Model(s)</th>
<th>M</th>
<th>SD</th>
<th>$E_{avg}$ (%)</th>
<th>NMSE</th>
<th>r</th>
<th>g</th>
<th>m</th>
<th>(b/C_0) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>4.22</td>
<td>2.05</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Kriging</td>
<td>3.71</td>
<td>1.03</td>
<td>13.96</td>
<td>0.25</td>
<td>0.369</td>
<td>0.417</td>
<td>0.186</td>
<td>69.35</td>
</tr>
<tr>
<td>GPI</td>
<td>3.38</td>
<td>0.21</td>
<td>18.98</td>
<td>0.34</td>
<td>0.151</td>
<td>0.166</td>
<td>0.015</td>
<td>78.48</td>
</tr>
<tr>
<td>LPI</td>
<td>3.69</td>
<td>1.00</td>
<td>15.96</td>
<td>0.25</td>
<td>0.382</td>
<td>0.441</td>
<td>0.185</td>
<td>68.96</td>
</tr>
<tr>
<td>IDW</td>
<td>3.76</td>
<td>1.29</td>
<td>17.98</td>
<td>0.30</td>
<td>0.261</td>
<td>0.337</td>
<td>0.164</td>
<td>72.77</td>
</tr>
<tr>
<td>RBF</td>
<td>3.72</td>
<td>1.15</td>
<td>15.52</td>
<td>0.29</td>
<td>0.274</td>
<td>0.335</td>
<td>0.153</td>
<td>72.93</td>
</tr>
<tr>
<td>Conventional ANN</td>
<td>3.85</td>
<td>1.05</td>
<td>12.13</td>
<td>0.18</td>
<td>0.583</td>
<td>0.597</td>
<td>0.297</td>
<td>61.44</td>
</tr>
<tr>
<td>Correction model based ANN</td>
<td>4.51</td>
<td>1.83</td>
<td>3.40</td>
<td>0.03</td>
<td>0.947</td>
<td>0.909</td>
<td>0.845</td>
<td>22.33</td>
</tr>
</tbody>
</table>
Table 4.5: Performance measures computed for traditional interpolation techniques, Conventional ANN and Correction model based ANN model (Cont.).

<table>
<thead>
<tr>
<th>Model(s)</th>
<th>FB</th>
<th>FV</th>
<th>MCM₂₄</th>
<th>F₂a</th>
<th>IOA₁₃</th>
<th>Ap</th>
<th>VG</th>
<th>MG</th>
<th>RHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
<td>1.000</td>
<td>0.000</td>
<td>1.00</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Kriging</td>
<td>0.128</td>
<td>0.661</td>
<td>0.846</td>
<td>0.69</td>
<td>0.124</td>
<td>-64.89</td>
<td>1.19</td>
<td>1.076</td>
<td>0.369</td>
</tr>
<tr>
<td>GPI</td>
<td>0.222</td>
<td>1.636</td>
<td>0.952</td>
<td>0.62</td>
<td>0.043</td>
<td>-70.86</td>
<td>1.21</td>
<td>1.137</td>
<td>0.151</td>
</tr>
<tr>
<td>LPI</td>
<td>0.133</td>
<td>0.693</td>
<td>0.874</td>
<td>0.58</td>
<td>0.124</td>
<td>-66.21</td>
<td>1.18</td>
<td>1.079</td>
<td>0.382</td>
</tr>
<tr>
<td>IDW</td>
<td>0.114</td>
<td>0.553</td>
<td>0.967</td>
<td>0.67</td>
<td>0.042</td>
<td>-64.45</td>
<td>1.26</td>
<td>1.084</td>
<td>0.261</td>
</tr>
<tr>
<td>RBF</td>
<td>0.125</td>
<td>0.563</td>
<td>0.952</td>
<td>0.71</td>
<td>0.079</td>
<td>-65.18</td>
<td>1.24</td>
<td>1.084</td>
<td>0.274</td>
</tr>
<tr>
<td>Conventional ANN</td>
<td>0.092</td>
<td>0.649</td>
<td>0.786</td>
<td>0.78</td>
<td>0.203</td>
<td>-49.46</td>
<td>1.14</td>
<td>1.038</td>
<td>0.583</td>
</tr>
<tr>
<td>Correction model based ANN</td>
<td>-0.067</td>
<td>0.113</td>
<td>0.549</td>
<td>1.00</td>
<td>0.728</td>
<td>0.42</td>
<td>1.04</td>
<td>0.909</td>
<td>0.947</td>
</tr>
</tbody>
</table>

To complete the performance analysis of all the models, the graphical tools i.e., the scatter plots and $Q-Q$ plots for all the interpolation schemes are shown in Fig 4-4 to Fig 4-10 and Fig 4-11 to 4-17 respectively. The interpolation scheme that has the observed and predicted value points plotted close to the 1:1 line (identity line) for both the scatter plots and the $Q-Q$ plots is considered as the best model.

![Scatter plot of kriging technique.](image)

Fig 4-4: Scatter plot of kriging technique.
Fig 4-5: Scatter plot of GPI technique.

Fig 4-6: Scatter plot of LPI technique.
Fig 4-7: Scatter plot of IDW technique.

Fig 4-8: Scatter plot of RBF technique.
Fig 4-9: Scatter plot of Conventional ANN model.

Fig 4-10: Scatter plot of Correction Model Based ANN model.
Fig 4-11: $Q-Q$ plot of kriging technique.

Fig 4-12: $Q-Q$ plot of GPI technique.
Fig 4-13: $Q-Q$ plot of LPI technique.

Fig 4-14: $Q-Q$ plot of IDW technique.
Fig 4-15: $Q-Q$ plot of RBF technique.

Fig 4-16: $Q-Q$ plot of Conventional ANN technique.
Fig 4-17: Q-Q plot of Correction model based ANN technique.

Considering the closeness of predicted values to the given values of each performance measure and graphical representations, one can infer that the correction model based ANN approach proposed in this paper is the best model for modeling and estimating the extreme-end (peak-end, low-end), and mid-range radon concentrations.
Chapter 5

Conclusions and Future Work

5.1 Conclusions

This thesis demonstrates the use of ANNs for modeling radon concentrations in Ohio. The available radon concentration data was reformatted to suit ANN training by converting the zip code information into latitude and longitude data. From Table 3.2, the best neural model, i.e. ANN model with 10 hidden neurons, trained using quasi-Newton is chosen as the best MLP3 approach for the purpose of estimating radon concentrations in zip codes with no available data. From Table 3.1, it can be seen that more training data results in a lesser validation error. Based on this observation, it is recommended that 90% of the available data be used for training to accurately predict the radon concentrations in the remaining zip codes with no available data. However, from Table 3.2 it is clear that the condition $E_{\text{avg}} \leq E_{\text{user}}$ is not achieved by the conventional ANN approach.

Therefore, in the second phase of the thesis, a new correction based ANN modeling approach for estimating radon concentrations is proposed. Simple MLP3 networks are used to develop both the original less-accurate model and the correction model. The proposed approach is compared with the conventional ANN modeling approach and traditional interpolation techniques. From Table 4.4 and Table 4.5, it can be confirmed that the
proposed correction model based ANN modeling approach delivers the best model compared to other models in predicting both the extreme-end (peak-end, low-end), and mid-range radon concentrations. The proposed approach also eliminates the need for additional knowledge input \textit{i.e.}, uranium which tends to increase the structural complexity as in the case of KBNNs.

5.2 Future Work

Exploring further advanced techniques based on ANN modeling could lead to better prediction results. Furthermore, several other techniques exist in literature that have not yet been explored for application to this problem, \textit{e.g.} Support Vector Machines, which is being used extensively in the field of image processing for reconstruction of images. Alternatively, a different root finding method can be applied in the correction based ANN approach. This requires mathematical knowledge in assessing the pros and cons of different root finding methods necessitating additional research.
References


34. Li, W., Liu, W.L., Zhang, Y., Yin, W., and Liu, Y., “On the comparison of spatial interpolation methods of marine temperature and salinity based on Arcgis software: a


69. Zhang, Q.J., *NeuroModeler*, Department of Electronics, Carleton University, 1125 Colonel By Drive, Ottawa K1S5B6, Ontario, Canada, 1999.


Appendix A

Source Code for Correction Model Based ANN Modeling approach

\begin{verbatim}
m = 1; c = 0;
m=1;
for k=1:110
  fid1 = fopen('data.dat');
inp = fscanf(fid1, '%g', 330);
inp1 = inp(m);
inp2 = inp(m+1);
outp = inp(m+2);
inpv = [inp1 inp2];
fclose(fid1);

  ybad = orig(inpv);
  fid2 = fopen('trad.txt','a');
  fprintf(fid2, '%f
', ybad);
  fclose(fid2);

  inp1c = ybad;
  inp2c = inp2;
inpc = [inp1c inp2c];
outpc = correction1(inpc);
  fid3 = fopen('mod.txt','a');
  fprintf(fid3, '%f
', outpc);
  fclose(fid3);

  R = outpc-inp1;
  fid4 = fopen('error.txt','a');
  fprintf(fid4, '%f
', R);
  fclose(fid4);
\end{verbatim}
$x(1) = -1.0 + (2.0) \cdot \text{inp1c}(1.0) / ((12.24989796) - (1.0))$;

$x(2) = -1.0 + (2.0) \cdot \text{inp2c}(-84.79165) / ((-80.53175) - (-84.79165))$;

$z(1) = 1.0 / (1.0 + \exp(-1.0 \cdot (2.8966 + x(1) \cdot 1.84256 + x(2) \cdot 4.24076)))$;

$z(2) = 1.0 / (1.0 + \exp(-1.0 \cdot (2.55345 + x(1) \cdot -7.48095 + x(2) \cdot 18.9876)))$;

$z(3) = 1.0 / (1.0 + \exp(-1.0 \cdot (10.3315 + x(1) \cdot 8.56266 + x(2) \cdot -17.4902)))$;

$z(4) = 1.0 / (1.0 + \exp(-1.0 \cdot (3.9155 + x(1) \cdot -10.1916 + x(2) \cdot 25.4966)))$;

$z(5) = 1.0 / (1.0 + \exp(-1.0 \cdot (4.87492 + x(1) \cdot 3.81473 + x(2) \cdot 6.86083)))$;

$z(6) = 1.0 / (1.0 + \exp(-1.0 \cdot (7.82337 + x(1) \cdot 8.1067 + x(2) \cdot -12.013)))$;

$z(7) = 1.0 / (1.0 + \exp(-1.0 \cdot (-2.16329 + x(1) \cdot 12.3752 + x(2) \cdot -12.6631)))$;

$z(8) = 1.0 / (1.0 + \exp(-1.0 \cdot (5.63282 + x(1) \cdot -0.774374 + x(2) \cdot 21.7944)))$;

$z(9) = 1.0 / (1.0 + \exp(-1.0 \cdot (0.0256262 + x(1) \cdot -5.97796 + x(2) \cdot 18.1607)))$;

$z(10) = 1.0 / (1.0 + \exp(-1.0 \cdot (1.42298 + x(1) \cdot 0.396632 + x(2) \cdot -0.010451)))$;

$y(1) = 3.17523 + z(1) \cdot (1.12634 + z(2) \cdot (-1.54783) + z(3) \cdot (-0.715785) + z(4) \cdot (1.50854) + z(5) \cdot (-0.514922) + z(6) \cdot (0.877342) + z(7) \cdot (0.2171) + z(8) \cdot (-0.300982) + z(9) \cdot (0.269883) + z(10) \cdot (-4.0892))$;

$output(1) = 38.438617 + (y(1) - (0.0)) \cdot ((41.934916) - (38.438617)) / ((1.0) - (0.0))$;

$\text{bw} = [1.84256 \quad -7.48095 \quad 8.56266 \quad -10.1916 \quad 3.17523]$;

$\text{dw} = [1.12634 \quad -1.54783 \quad -0.715785 \quad 1.50854 \quad -0.514922 \quad 0.877342 \quad 0.2171 \quad -0.300982 \quad 0.269883 \quad -4.0892]$;

$\text{ew} = 3.17523$;

$sensitivity1 = 0.0$;

for $i = 1:10$

$sensitivity1 = sensitivity1 + \text{dw}(i) \cdot z(i) \cdot (1 - z(i)) \cdot \text{bw}(i)$;

end

$sensitivity1$;

$sensitivity2 = sensitivity1 \cdot ((12.24989796) - (0.1) / 2.0)$;

$sensitivity = sensitivity2 \cdot ((41.934916) - (38.438617))$;

$J = sensitivity$;

$R$;

alpha = 0.1;

Euser = 0.1; iter = 0; ymod = ybad;

Eobj = ((inp1 - outpc) / inp1) * 100;

while(abs(Eobj) > Euser)

deltay = abs(-R/J);

if ybad < outp

ybad = ymod + (alpha * deltay);

elseif ybad > outp

ybad = ymod - (alpha * deltay);

end

ymod = ybad;

if ybad > outp

end
break;
end
inp = [ybad inp2];
outpc2 = correction1(inp);
Eobj = ((inp1-outpc2)/inp1)*100;
R = outpc2 - inp1;
iter = iter + 1;
if (iter > 1000)
    c = c + 1;
    break;
end
end
ybad;
fid5 = fopen('revmod.txt','a');
fprintf(fid5, '%fn', ybad);
fclose(fid5);
m = m + 3;
k = k + 1;
end
Appendix B

Screenshots of the *Neuromodeler* Software

B.1 *Neuromodeler* Software Main Window

![Neuromodeler Main Window](image)

*Fig B-1*: Screenshot of the main window of the *Neuromodeler* software.
B.2 New Model Creation Window in the *Neuromodeler* Software.

Fig B-2: Screenshot of the new neural model creation window in the *Neuromodeler* software.
B.3 Training Window of the *Neuromodeler* Software.

Fig B-3: Screenshot of the training window of the *Neuromodeler* software, while training the MLP3 of the Conventional ANN technique.
B.4 Testing Window of the *Neuromodeler* Software.

Fig B-4: Screenshot of the testing window of the *Neuromodeler* software, while validating the MLP3 of the Conventional ANN technique.
B.5 Screenshot of the *Neuromodeler* Software comparison window.

Fig B-5: Screenshot of the *Neuromodeler* software window, showing the validation data (magenta line) and the predicted data (yellow line).
B.6 Screenshot of interpolated radon concentration surface by ArcGis spatial analyst.

Fig B-6: Estimated radon concentrations obtained for Ohio by using four traditional interpolation techniques.