INPUT VARIABLE ANALYSIS AND SELECTION

FOR CORN YIELD PREDICTION USING

ARTIFICIAL NEURAL NETWORK

By

PAWAN RAMESH LAWALE

Bachelor of Engineering in Information Technology

Government College of Engineering

Amravati, Maharashtra, India

2007

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 2015

INPUT VARIABLE ANALYSIS AND SELECTION

FOR CORN YIELD PREDICTION USING

ARTIFICIAL NEURAL NETWORK

Thesis Approved:

Dr. K. M. George

Thesis Adviser

Dr. Blayne Mayfield

Dr. Douglas Heisterkamp

ACKNOWLEDGEMENTS

Pursuing M.S in Computer Science Department of Oklahoma State University was a great experience for me. I take this opportunity to express my gratitude towards my advisor, committee members, collaborators, colleague, relatives, etc. who have provided me the scientific and moral support for the completion of this thesis.

Foremost I’m grateful to my advisor and head of Computer Science Department, Oklahoma State University, Prof. K.M George for his supportive guidance and supervision. This work would not have been possible without his support and encouragement. I convey my deepest gratitude to my committee members Prof. Blayne Mayfield and Prof. Douglas Heisterkamp for their constant encouragement and scientific discussions.

I’m indebted to SST Software for partially supporting me. I would like to explicitly acknowledge Mr. Todd Pugh for helping me in providing data and sorting the variables which were integral part of my research work.

I’m also thankful to lab and administrative staff for helping me throughout in many ways. I would like to thank my batch mate Mr. Naveen Singireddy for his support. Last but not the least, I would like to thank my family member for their care and motivation.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: PAWAN RAMESH LAWALE

Date of Degree: JULY, 2015

Title of Study: INPUT VARIABLE ANALYSIS AND SELECTION FOR CORN YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORK

Major Field: COMPUTER SCIENCE

Abstract: United States of America entered into the advance agricultural practices since 1930. Continuous evolution of agricultural technologies has been witnessed to maximize the production. Efficient production of crop requires effective estimation of yield. Various data-driven models are therefore used to predict the yield. These models rely highly on the retrospective analysis of data. Choosing the right parameters for yield prediction is an essential exercise. This paper evaluates the effective techniques of choosing the right parameters from the data set, finding the correlation between the individual parameters, group of parameters and classifying these parameters which makes an impact on the final yield production. Linear and Nonlinear regression modeling techniques were used for this classification. The selected attributes are then given as an input to an Artificial Neural Network (ANN) to test its prediction capability. Root Mean Squared Error (RMSE), is used as comparison measure. A MatLab program is designed to train and test the models. A dimension reduction approach based on Principal Component Analysis (PCA) gives the best model with minimum RMSE. The crop chosen for our analysis is Corn from the state of Texas. The trained model predicted yield with RMSE of 1206.59 and regression R of 0.63.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Background

Crop yield Prediction is important for agricultural planning and resource distribution decision making [10]. Crop yield is mostly affected by two factors, human decision making and climatic conditions. To make predictions based on these two factors, several variables are involved using which prediction becomes difficult. The most commonly followed procedure to guess the yield is based on the previous year's performance, which is good for guess but is highly susceptible to error and miss judgement. Thus more efficient models were developed, in order to help reduce the error. Out of all models, Crop growth and Data-driven models are the most widely used models. Crop growth model uses site specified experimental data, regional calibration and plot-level observations. This method is robust and considered as very efficient, however it is used for very specific crops and are extremely time consuming [11]. Data-driven models work on the high level information and without the knowledge of actual mechanism which produce the data [12]. Previous studies indicates that Data-driven methodology is more promising than plant growth methodology [13].

Data-driven Modeling

Data-driven models are broadly classified as Statistical and Data-mining models. Statistical models use parametric structures tuned with sum-of-square residue [12]. Most of the data mining models apply machine learning techniques and nonparametric structure [12]. Machine learning helps in finding the nonlinear model for the large set of data which is used to train the model. Part of the data-set (which is most recent) is reserved for validating the model, which is not used in the training process. Most widely used machine learning techniques include Regression trees [14] and Artificial Neural Network. From the past studies and its application into various other sectors, Artificial Neural Network has turned out to be a most promising model, obtaining very good results. However, in order to perform well, it is far more important to select the most appropriate input variables from the large data set.

Prominence of Artificial Neural Network

An extensive and successful application of Artificial Neural Network (ANN) has been observed in the past to broad spectrum of data-intensive problems such as in Finance, Health care, Science, Energy, Agriculture, Water resource [1], etc. However in order to have accuracy in the prediction, the most important step involved in ANN is the identification and selection of right input variables [2]. If any of the influential input variable is not included in the model then the performance of the model will be compromised. If any unwanted input variable is included, computational efficiency will be decreased and calibration becomes difficult [3]. This makes model validation problematic. Our case study focuses on the agricultural data analysis and appropriate input variable selection for ANN model to receive highest possible prediction index.

Significance of the Problem

In this study our primary focus is the selection of the most relevant variables from the given data set for Artificial Neural Network model, so that we can achieve the highest possible prediction indexes i.e. to reduce the Mean Squared Error (MSE) and increase the Regression (R-square). For this experiment we will use data for Corn from the state of Texas. Texas is among the states, that have massive land with varying soil types but still is one of the lowest contributors of overall corn growth in USA. It contributes only 1.83% to the total corn production in USA [15]. Good predictive system could help farmers as an early warning system so that they can take preventive actions in timely manner and increase the productivity. The soil conditions in Texas are relatively different compared to other states. Texas alone has around 1300 different kinds of soil types recognized. Weather conditions in Texas vary widely from arid in the west to humid in the east [16]. Prediction with such a diverse climatic conditions might give out some more correlations between climatic parameter and yield.

SST Software is the primary source of all the data set received for this experiment. SST Software deals largely with the structuring and mining spatial data for the agricultural industry. They are one of the most interested parties in exploring the area of crop yield prediction more accurately. This type of study will give them and similar businesses more insight for future planning.

CHAPTER II

REVIEW OF LITERATURE

A considerable amount of work has been done in input selection techniques and crop yield prediction using various methodologies. Autoregressive state-space models, least-square regression, exponential-linear crop growth algorithm and numerical crop yield model have been used to predict crop yield with moderate success. Some amount of work is also done using Artificial Neural Networks and results from those experiments have shown significant improvement in prediction compared to other methodologies/models.

Artificial Neural Network

Artificial Neural Networks (ANNs) are a computational tool, based on the properties of biological neural system [18]. Interrelationships of the correlated variables or input parameters can be symbolically represented using Nodes in this model (also known as Neurons in biological term) [17]. The nodes are basically classified into three categories: Input Nodes, Hidden Nodes and Output Nodes. The Input Nodes accept the input variables given to the model. Each node represent single input variable to the model. All the input nodes belong to the Input layer of the model. Output Node(s) are the ones, which emits the predicted value after the computation. There can be more than one output node. The layer containing the output nodes is called as Output Layer. The hidden node plays the most important role in computation. A node in the hidden layer is also termed as Perceptron. Perceptrons are the algorithms, which are used to transform input value into a desired output. Along with the input a certain weightage is applied to each perceptron and it is adjusted to reach close to desired output value. Each node in the network is connected to every node in the next layer. Data flow across the layers over the weighted connections [17]. Figure 1 shows the unidirectional Neural network also known as Feed Forward Neural Network. [19]

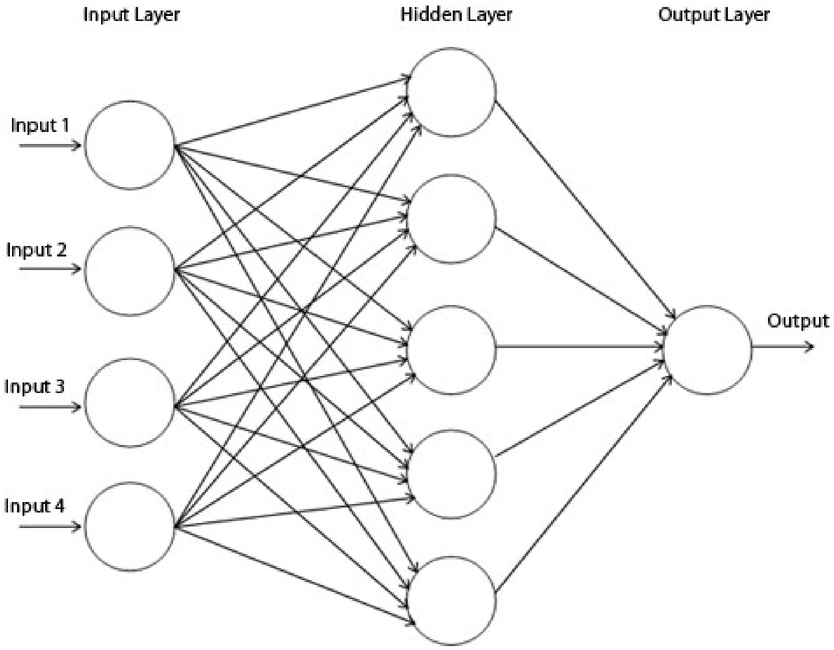

Figure 1. Unidirectional/Feed Forward Neural Network

A node accepts data from the previous layer and calculates a weighted sum of all its inputs, t: [17]

$$t_i = \sum_{j=1}^{n} w_{ij} x_j$$

Where n is the number of inputs, w is the weight of the connection between node i and j, and x is the input from node j. A transfer function is then applied to the weighted value, t_i to calculate the node output, o_i: [17]

$$o_i = f(t_i)$$

The number of nodes in the hidden layer depends on the type of problem one is dealing with, also with the number of input and output nodes. Too many hidden nodes may cause the ANN to over-train, resulting in the poor prediction by memorizing all of the training data [20]. Learning rate decides the magnitude of change required in the weightage during the series of iteration to bring the predicted value within the acceptable range [17]. The error rate is the acceptable error in the network. Errors beyond error rate need to be rectified by adjusting weightages. Once the network converges, an approximate function is developed and utilized for further prediction [21]. After the model is trained it is then tested with separate data set for its accuracy.

Related Experiments

One of the experiments performed to predict the Corn and Soybean yield prediction for Maryland, indicates the significant improvement in the prediction result using ANN than Regression Models. ANN model for corn in Maryland resulted in r^2 and RMSEs of 0.77 and 1036 as compared to 0.42 and 1356 for linear regression, respectively. Similarly, ANN model for soybean in Maryland resulted in r^2 and RMSEs of 0.81 and 214 as compared to 0.46 and 312 for linear regression, respectively [17]. Similar experiment performed for rice yield prediction in the mountainous regions Fujian of China also shows similar trend. The ANN model for rice yield prediction in this region resulted in r^2 and RMSEs of 0.67 and 891 as compared to 0.52 and 1977 for linear regression, respectively [23].

Input selection Techniques

Since the input variables have significant impact on the output variable, it is absolutely necessary to find out a suitable and effective technique for variable selection. Review of current literature shows that several studies use ad-hoc approaches for variable selection [4]. Some of the quantitative approaches used in variable selection in the area of water resource modeling are sensitivity analysis [5], the Gamma test [6], partial mutual information [7], hybrid independent component analysis and input variable selection filter [8], and cross-correlation analysis [9]. One of the experiment [2] performed for input variable selection shows Partial mutual information (PMI) is one of the most promising approach for the selection of the inputs. This experiment was carried out on environment and water resource modeling.

Regression Tree is another technique used to classify and select the appropriate input variables. When data has lot of features and parameters and they interact in a much more complicated and nonlinear way, linear models are inefficient for prediction. Other way of solving this problem is to break the data set into subsets and then further divide the subsets into new subsets. This is called recursive portioning. This process is performed until we get the chunk of data set so tame that they can be fit into a simple model. A tree of node ids are formed for this representation. The top most node is the root node with no incoming edges. All other nodes have exactly one incoming edge [12]. The nodes which do not have outgoing edges are the leaf nodes which has the smallest set of input variables, either grouped with other variables or an individual variable. Each leaf is one class that represents the most appropriate target value. The target value and the class values can be a range of values assigned to that particular node. The widely used algorithms to build regression trees are CART, M5 and M5' [22].

Iterative stepwise selection is another approach of input variable selection [24]. This methodology can be applied to both linear as well as nonlinear models. Whichever model is chosen, each independent variable is first checked for its significance with the dependent variable. The variable with highest significance with respect to the dependent variable is selected and clubbed with the second highest significant variable. These two are then together modeled and checked if the relation between them increases the impact on output significantly. If yes, then the two are selected and move to the selection of other attribute. If no, then the variable is discarded and another variable is checked in combination with it. This processes is repeated until all combinations of variables are checked. This process is effective but long lasting and very much time consuming.

Looking at the application of methodologies and techniques applied so far, we find that each technique has its own benefits and drawbacks. For our experimentation, rather than applying one particular approach, we will club two approaches together. First we will eliminate variables based on the expert's advice then we will classify them and then we will apply the iterative stepwise selection methodology. Classification of variables will help us reduce the number of iterations in next step.

CHAPTER III

METHODOLOGY

From the experiments performed in the past, we inferred that one single method or technique is not sufficient for appropriate selection of the input variables. Some effective techniques take longer time in the selection process and some, which take shorter time, are not so effective. Therefore for this study we will adopt a combination of three techniques. First we will filter out the variables from the whole data set based on expert advice, then eliminate variables which are in non-numeric format and then we eliminate variables which have no relevance with the output variable. In the second step, the variables which are obtained after first step will be classified into groups. Then iterative stepwise technique will be applied for the final selection of variables. These variables will then be given as an input to the neural network in Matlab to find out Root Mean Square Error (RMSE) and Regression (R-square) which will tell us the accuracy in prediction. These steps are elaborated in the following sections.

Data Description

The data for this experiment is provided by SST software. After the discussion with the domain level experts and reading through all the elements, the variables listed in Table 1 are taken into consideration for the experiment. The data provided in Table 1 are on per sub field level. For example, if a single field is divided into 4 sub fields, then the data provided is for all the 4 sub fields uniquely based on the soil type and properties.

Table 1. Input variables and description

Attribute Name	Description
Hectares	Area of land in hectors
Total Mass	Total corn production on this land in Kg.
Minimum Yield	Minimum yield on this land in kg/ha.
Average Yield	Average corn yield per hector in Kg.
Maximum Yield	Max yield produced on that field.
Crop Year	Year in which the corn sowed.
Hybrid Id	Id of the variety of the corn seed.
Hybrid Name	Name of the variety of the corn.
Relative Maturity	Recommended duration of crop maturity.
Available water 0 to 25 CM	Ground water availability between the depth of 0 to 25 CM.
Available water 0 to 100 CM	Ground water availability between the depth of 0 to 100 CM.
Slope	Ground slope on the field.
Sand	Percentage of sand in the soil.
Silt	Percentage of silt in the soil.
Clay	Percentage of clay in the soil.
Organic Matter	Percentage of organic matter in the soil.
CEC7	Percentage of Cat ion Exchange Capacity.
Latitude	Centroid latitude location of the whole field.
Longitude	Centroid longitude location of the whole field.
Component	Percentage content of major soil type.
Rainfall	Daily rainfall from year 2011 to 2014.
Temperature	Daily temperature from year 2011 to 2014.

Elimination of Irrelevant attributes

In this step, first we eliminate attributes which are non-numeric. Next we remove the variables which have identical properties/values but measured in different format/units. For example, yield is measured in Kg as well as bushels. We choose to experiment with more standard measuring unit, that’s why we choose all the yield variables in kg.

For this experiment, we will use the idea of predicting average yield in kg/ha. Therefore of the given yield variables only one is required i.e. “Average yield”. We therefore discard other yield variables which includes “Total Mass”, “Minimum Yield”, and “Maximum Yield”. Crop Year is another attribute which is not relevant in direct prediction, however we will refer to it for segregation of data throughout our experiment. We also exclude the field size i.e. variable "Hectares" since total field are is not of our concern as we are predicting corn yield in kg/ha.

On observing the values of Hybrid Id and Relative Maturity it is found that they are related to each other and are used to recommend duration for a crop to mature. It has no connection to the final corn yield production. All they specify is, for a given hybrid id how many days are recommended for a crop to be on the field. Therefore these two attributes are eliminated from further consideration in the experiment.

Latitude and Longitude variables are again have no impact on the yield prediction, however they will be referred in this experiment for the purpose of identifying the field and its associated records.

Classification of Variables

After the initial elimination of attributes, the variables used for classification are listed in Table 2.

Table 2. Variables for classification

Attributes	Description
Latitude	Centroid latitude location of the whole field.
Longitude	Centroid longitude location of the whole field.
Average yield	Average corn yield per hector in Kg.
Available water 0 to 25 CM	Ground water availability between the depth of 0 to 25 CM.
Available water 0 to 100 CM	Ground water availability between the depth of 0 to 100 CM.
Slope	Ground slope on the field.
Component	Percentage content of major soil type.
Sand	Percentage of sand in the soil.
Silt	Percentage of silt in the soil.
Clay	Percentage of clay in the soil.
Organic Matter	Percentage of organic matter in the soil.
CEC7	Percentage of Cat ion Exchange Capacity.
Rainfall	Average annual rainfall.
Temperature	Average annual temperature.

As per Monisha Kaul, Robert L. Hill, Charles Walthall [17] water availability is crucial factor for corn growth. Sadras and Calvino (2001) [25] determined that 90% of soybean and 76% of corn yield variation were linked to water deficits. Rainfall was deemed to be primarily responsible for yield variability within a region [1]. Soil water holding capacity and land capability class are important factors in the yield prediction. Environmental factors, such as climatic information, in addition to multiple soil properties related to crop rooting depth and water availability, are significant factors for crop yield models [26] [27]. Based on these conclusions, we can broadly infer that there are three important factors which are responsible for crop growth and those are Land characteristics, Soil Properties, and Weather.

Now this new set of variables can further be classified into these three categories. By going through each variable and asking the question “Which class it belongs to?” we can segregate our input variables as below:

1. Soil Properties:
    - i. Component
    - ii. Sand
    - iii. Silt
    - iv. Clay
    - v. Organic Matter
    - vi. CEC7
2. Weather:
    - i. Rainfall
    - ii. Temperature
3. Land Characteristics:
    - i. Available water 0 to 100 CM
    - ii. Available water 0 to 25 CM
    - iii. Slope

Stepwise Iteration

In this step each independent variable is first checked for its significance with the dependent variable. The variable with highest Regression (R-square) and lowest RMSE, with respect to the dependent variable, is selected and clubbed with the second highest R-square and lowest RMSE variable. These two are then together modeled and checked if the relation between them increases the impact on output significantly. If yes, then the two are selected and then moved to the selection of other attributes. If no, then the variable is discarded and another variable is check in combination with it. This processes is repeated until we have checked combinations with all the variables. We perform this experiment with nonlinear regression using Neural Network Toolbox in Matlab.

We apply this process on each class separately. Once we have a subset from each class those will be our final input variables to be given as an input to an Artificial Neural Network.

Principal Component Analysis

Principal Component Analysis (PCA) is a well-established dimension reduction technique commonly used to eliminate unwanted or redundant input variables keeping the structure of the data intact. PCA reduces the dimensionality of a data set having large number of variables [28]. The reduced set of data is called as a principal components. Principal Components are ordered so that the first variable represents most of the variations in the original variables [28]. Below are the steps to compute PC [28]:

i. Let X denotes a set of raw scores of size N. Then the average of X is defined as avg(X)= $(\sum X)/N$

ii. The deviation from mean is defined as $X_d = x - \text{avg}(X)$ where $x \epsilon X$.

iii. The correlation between two variables X and Y is given by

$$r_{xy} = \frac{\sum X_d Y_d}{\sqrt{\sum X_d^2}\sqrt{\sum Y_d^2}}$$

iv. The correlation matrix of n variables is defined as an n x n symmetric matrix R whose diagonal elements are 1 and the (i, j) th entry is the correlation between ith and jth variable. This Correlation matrix is of the form r_{xy}

$$\begin{bmatrix} 1 & r_{xy} \\ r_{yx} & 1 \end{bmatrix}$$

v. If A is a n x n symmetric matrix, then the real number λ is called eigenvalue of A if and only if there is a non-zero vector V in R^n for which AV = λV. Any such vector is called *eigenvector* associated with the *eigenvalue* λ.

vi. The first PC is the eigenvector associated with the largest eigenvalue.

To calculate the eigenvalue and eigenvector we use matlab function princomp(I) which gives us *eigenvectors* for each *eigenvalue*. The first PC gives the order of the most significant variables based on vector values. Let n be the total number of initial variables, "w_{1i}" be the weightage of the ith variable in 1st PC, "w_{2i}" be the weightage of the ith variable in 2nd PC and so on and "v_i" be the ith variable. Then the first dimension V_1 can then be calculated as

$V_1 = w_{11}*v_1 + w_{12}*v_2 + w_{13}*v_3 +$ …………. + $w_{1n}*v_n$ ……………………………............. (1)

Similarly second dimension V_2 can be calculate as

$V_2 = w_{21}*v_1 + w_{22}*v_2 + w_{23}*v_3 +$ …………. + $w_{2n}*v_n$ ……………………………………… (2)

We consider dimensions which shows input variables variance up to 98-99%. In this experiment we considered up to dimension V10, the reason for which is elaborated in chapter IV (FINDINGS). These reduced dimensions are then given as an input to Neural Network to verify its accuracy and to compare it with the Stepwise Iteration method. The variance is defined as $Variance_i = \frac{\sum_1^i \lambda_j}{\sum_1^n \lambda_j}$

which is interpreted as the explanation of variance provided by the first *i* eigenvectors. Usually it is represented as percentages and 99% is considered an acceptable number. This is the cutoff we used for feature eigenvector selection.

Input Normalization

All inputs to ANN training and PCA analysis are normalized using MatLab function ‘mapminmax’ to the interval [0, 1]. The formula implemented by the function is: y = (ymax-ymin)*(x-xmin)/(xmax-xmin) + ymin; where y represents the normalized value, x the input value. The max and min represent the highest and lowest values among x and y. The mapminmax function allows us to specify the range of y as an argument to it. Furthermore, no assumptions on the distribution of data is necessary and examination of our dataset did not produce any outliers.

Output Renormalization

Output is always renormalized by the MatLab Neural network tool box itself. The function used for renormalization is also mapminmax. Following example code based on MatLab documentation illustrate the process (http://www.mathworks.com/help/nnet/ref/mapminmax.html):

```
[y, PS] = mapminmax(x, ymin, ymax);
x_again = mapminmax('reverse', y, PS);
```

PS is a structure that keeps the transformation information such as the max and min of x.

Neural Network Analysis

Neural Network Toolbox in Matlab is used for both input selection iterative method and final yield prediction. We use ANN feed forward back propagation model as implemented in MatLab. All inputs to the ANN are normalized. The final step of the prediction de-normalizes the data to the standard form.

Description of ANN

The general architecture of the ANN is shown in Figure 2. MatLab documentation recommends the default as better when choices are available. So for the work reported in this thesis, we always considered the default given by the Neural Net Toolbox. Following are the specifics:

1) The total number of neurons in the hidden layer are ten. (This choice was made since it is the default in MatLab and used by research reported in the literature). Input layer depends on the number of input variables and the out layer consists of only one node.
2) There are three transfer functions provided by Neural Network toolbox which are, TANSIG (tangent sigmoid transfer function), PURELIN (linear transfer function) and LOGSIG (log-sigmoid transfer function). The most widely used of them all is TANSIG. So for our experiment we have selected this function.
3) For every analysis, the model is trained 5 times with the same data set and the best results obtained of the 5 trainings will be taken into consideration. For example in the case of the iterative input variable selection, when the second variable is selected, we run the model five times each with different combinations for the second variable. We compare the best fit in each case and select the one with the best fit. MatLab computes two fitness measures, namely R and RMSE. We chose RMSE as the fitness measure in each case. This choice can be justified by the fact that MatLab uses MSE as the training stopping condition. Lower the RMSE, the better the fit. However we display the R value also as appropriate.
4) Levenberg-Marquardt algorithm is chosen for training as it is considered to be the best among all the training algorithms supported by MatLab. The MatLab function used is ‘trainlm’.

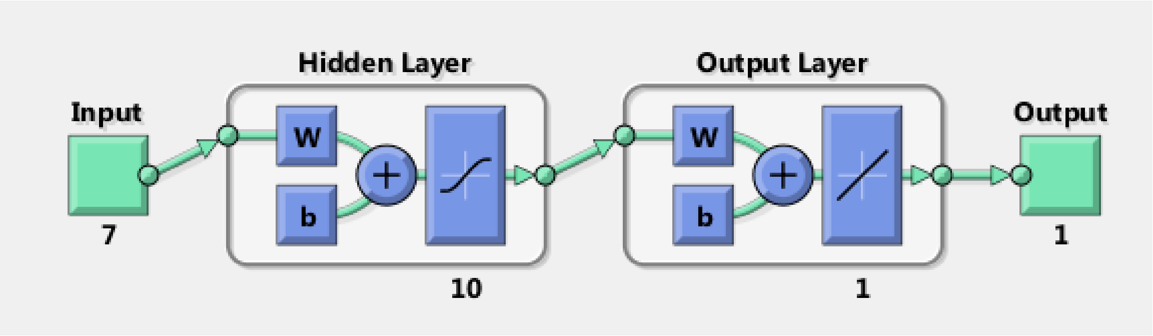

Figure 2. Feed-forward back propagation AAN model using NN toolbox in MatLab [29]

Training -validation-test Partition

For optimal training, MatLab automatically partitions input data randomly into three sets – training, validation, and test. However as we are using ANN model for variable selection, we wanted to make sure the same indices of the random partition be used throughout the experiment. Therefore we use random partitioning only for the very first execution of the model. After this execution, we retrieve the random indices generated by MatLab function 'dividerand'. These indices are accessible through MatLab variables *tr.trainInd, tr.valInd* and *tr.testInd* for training, validation and testing indices respectively. Once we obtain these indices, we then set the partitioning function to 'divideind' which require us to set the division parameters *net.divideParam.trainInd*, *net.divideParam.valInd* and *net.divideParam.testInd*. These parameters are set to the values of indices obtained from the first execution of model. This ensures that the same portioning is maintained throughout the experiment.

The training set is used to compute the gradient and updating the network weights. The validation set is used to compute the error during training. Both training error and validation error should decrease. If validation error begins to increase it indicates over fitting. Testing is done after training using the test data set. The test set error is used for model comparison. MatLab gives R and RMSE statistics for all three cases when training and test are complete. During the training 70% of the data will be used to train network, 15% of the data will be used to validate and 15% will be used to test the network.

Training-test Strategy

The total data available to us are for the four years 2011, 2012, 2013, and 2014. As the yield data for one year does not have variations, data from several years is necessary to build the model. We used ANN for final input variable selection process. During the selection phase we trained the model for all the four years of data i.e. 2011, 2012, 2013 and 2014. Based on the RMSE values generated during the procedure, final set of variables were selected

Once the final set of input variables are selected, we use the data for the first three years 2011, 2012, and 2013 for model building. Fourth year data (2014 data) are used as user test case to compare against the model fit measures. We did not use data from all four years for the model building because then no data will be left for conducting user testing and we will have to solely rely on the model fit supplied by MatLab. However, we also have the model with all four years data available as a by-product of the variable selection process. That is, when we determine the final set of variables, we have trained the model with all four years of data.

Accuracy Measurements

MatLab provides two model fit metrics as default. They are R and RMSE. RMSE is the square root of average squared difference between the predicted values and the expected values. If the value of RMSE is low then better is the prediction by the model. Zero value means no errors. If x_i is a predicted value and y_i is the actual expected value, and if there are N observations then RMSE is calculated as:

$$RMSE = \sqrt{\frac{\sum(y_i - x_i)^2}{N}}$$

R compares the predicted value to the expected value. We have not used R in our choices. As mentioned earlier, we used RMSE.

CHAPTER IV

FINDINGS

After performing attribute elimination step as described in the Chapter 3, Methodology, the variable used for further experiments are listed in Table 2. Those elements were then classified as per their properties into 3 classes i.e. Soil Properties, Weather and Land Characteristics. These classified elements were then processed thorough the Stepwise iteration methodology to determine the final set of input variables for model training. The crucial observations of this methodology are listed in Tables 3-7.

Table 3 shows the statics on the availability of data we have available for this experiment. During the variable selection process we have used ANN to find out RMSE value, based on which the best performing variables are selected. (Fertilizer data was not available for the experiments). We implemented MatLab program to generate all results. After training we get three RMSE values for training, validation and test. We use the test RMSE for all the comparisons. Other measures are for information only.

Table 3. Texas corn yield number of observations

Year	**Number of observations**
2011	837
2012	1723
2013	2790
2014	1055
Total	**6405**

In this process we used all years' data i.e. 2011, 2012, 2013 and 2014. However after the variables are selected in each method (i.e. Stepwise iteration and PCA), ANN model will be trained for three years of data i.e. 2011, 2012 and 2013 using the selected input variables. Data of year 2014 will then be used for testing and prediction. For building (training) the final ANN model, we use 5350 observations (years 2011, 2012, 2013). This model is then tested using the 2014 data with total observation of 1055. In the next section results of variable selection using iterative method is given. Results from the best of 5 runs will be chosen in all cases.

Results from Stepwise Iteration Method

Table 4 shows results of variable selection from the Soil Properties class and is divided into 5 sections. First section finds R and RMSE for each variable with respect to Yield. R values are given for information only. As mentioned previously, it is not used in variable selection. The different sections of Table 4 show the combinations of 2, 3, 4, 5 and 6 variables respectively of the Soil Properties class which has six elements.

Table 4. Observations of Stepwise iteration on Soil Properties vs Yield.

Input Variables	R	RMSE
Component	0.30	1792.45
Clay	0.56	1540.24
Silt	0.57	1516.78
Sand	***0.58***	***1501.5***
Organic Matter	0.47	1638.35
CEC7	0.47	1641.82
Sand, Silt	0.60	1488.33
Sand, Clay	0.53	1542.52
Sand, Organic Matter	***0.58***	***1449.59***
Sand, CEC7	0.58	1509.68
Sand, Component	0.57	1487.47
Sand, Organic Matter, Component	0.60	1436.29
Sand, Organic Matter, Silt	***0.63***	***1422.61***
Sand, Organic Matter, CEC7	0.63	1429.08
Sand, Organic Matter, Clay	0.62	1432.39

Sand, Organic Matter, Silt, CEC7	0.63	1435.1
Sand, Organic Matter, Silt, Clay	***0.62***	***1426.95***
Sand, Organic Matter, Silt, Component	0.61	1459.31
Sand, Organic Matter, Silt, Clay, CEC7	0.62	1435.97
Sand, Organic Matter, Silt, Clay, Component	0.61	1465.17
Sand, OM, Silt, Clay,CEC7, Component	***0.62***	***1456.86***

The highlighted row shows the most significant variable or combination of variables in a section. In Table 4 variable which has lowest RMSE value is selected as the most significant variable(s) in that iteration. This variable is then clubbed with the variable which have second lowest RMSE value and so on. We have selected RMSE as a deciding factor as our ultimate goal is to minimize the error so that we can have the most accurate prediction.

Table 5 is for the variables belonging to the Weather class. There are only two variables in this class. Highlighted combination of variable in Table 5 shows the most influenced variables in this class with respect to yield.

Table 5. Observations of Stepwise iteration on Weather vs Yield.

Input Variable	**R**	**RMSE**
Rainfall	0.34	1743.52
Temperature	0.67	1372.1
Rainfall, Temperature	***0.68***	***1339.21***

Table 6 is for the variables that belongs to Land Characteristics class which has 3 elements. This table is again divided into 3 section, based on the number of variables combined. Highlighted variable is the most influenced variable of this class

Table 6. Observations of Stepwise iteration on Land Characteristics vs Yield

Input Variable	R	RMSE
Water at 100 cm	***0.35***	***1704.38***
Water at 25 cm	0.19	1800.25
Slope	0.12	1842.42
Water100, Water25	0.28	1726.6
Water100, Slope	0.27	1771.21
Water at 100 cm, Water at 25 cm and Slope	0.28	1748.19

We then choose most influential set of attributes from each category and then combine them together to process the next iteration. Table 7 shows this observation:

Table 7. Observations of clubbed subsets vs Yield.

Input Variable	R	RMSE
{Rainfall, Temperature},{Sand, Silt, OM}	0.68	1337.71
{Rainfall, Temperature}, {Water at 100 cm}	0.66	1367.5
{Rainfall, Temperature},{Sand, Silt, OM},{Water at 100 cm}	***0.70***	***1334.8***

From Table 7 we say that, variables which contributes most to the corn yield prediction are Rainfall, Temperature, percentage Sand in soil, percentage Silt in soil, Organic Matter and Water at 100 cm.

The Neural Network model was then trained for data set of 2011, 2012 and 2013 for the above six variables. This trained model was then tested for the same 6 variables but for the data set of year 2014.

Table 8. Result of Neural Network Model after training and testing for 6 variables

Crop Year	R			RMSE			Observations
	Training	**Validation**	**Test**	**Training**	**Validation**	**Test**	
2011-2013	0.71	0.72	0.70	1347.14	1143.57	1393.04	5350
2014	-	-	0.63	-	-	1281.03	1055

Figure 3 shows the Regression plot for Training, Validation, testing and overall for the data of year 2011 - 2013. Figure 4 shows the regression plot of testing of the data of year 2014. Comparing the results of the test of 2014 data to the average yield, we see that the error is 12% (1281/9862). In the case of the RMSE for the training of 2011-2013 data, the result is 14% (1393/9610). Based on this we make two observations:

1) We do not have a benchmark to compare and so we cannot say the goodness of the model.
2) Fertilizer is an important factor for yield. Availability of fertilizer data could have given a better error.

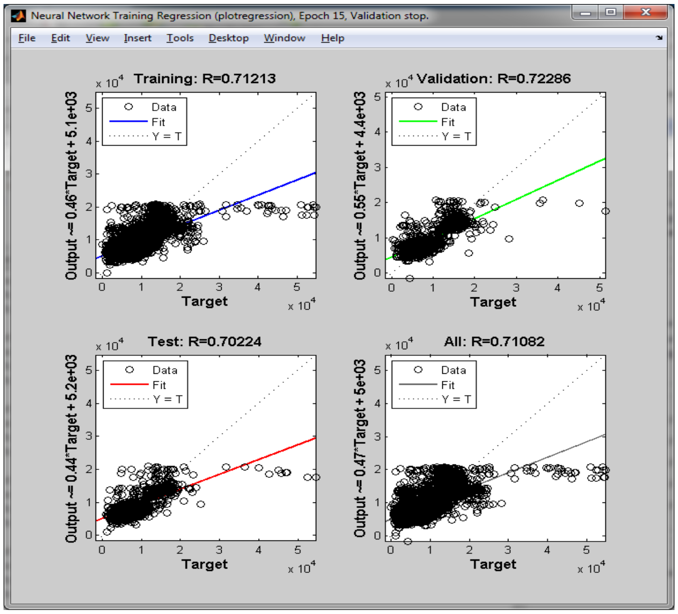

Figure 3. Regression plot for model building for the data-set of 2011 – 2013 (Stepwise Iteration)

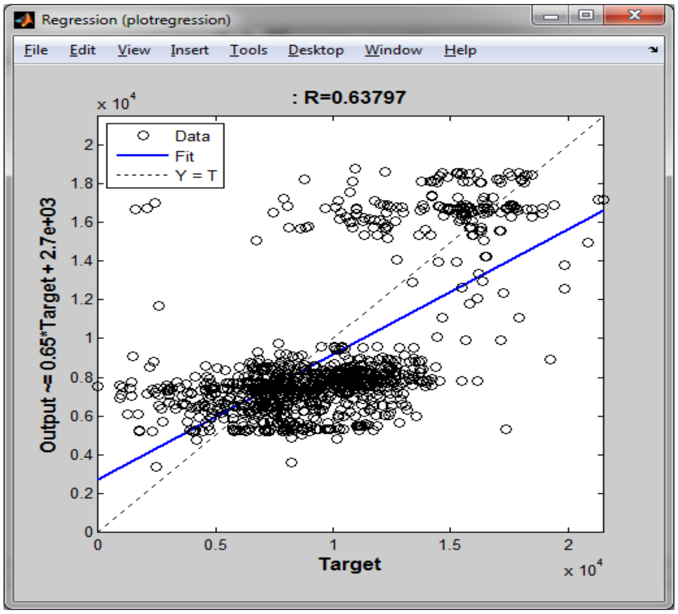
Regression (plotregression)
File Edit View Insert Tools Desktop Window Help
: R=0.63797
x 10^4
Data
Fit
Y = T
Output ~= 0.65*Target + 2.7e+03
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
0.5
1
1.5
2
Target
x 10^4

Figure 4. Regression plot for the testing of data for year 2014 (Stepwise Iteration)

Input Selection Based on Principal Component Analysis

First step in this analysis is to normalize the data as described in the methodology section. We performed PCA using matlab function princomp(I). This function gives us two results, which are eigenvector and eigenvalues. Table 9 and Table 10 shows the eigenvector and eigenvalues for our input variables.

Table 9. Eigenvectors

variables	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**	**PC11**
Water 25	0.107987	0.306188	0.245047647	-0.21349	0.426987	0.06633	0.035609	0.071353	0.108198	0.764314	6.69E-09
Water 100	0.048729	0.309328	0.170959863	-0.24388	0.626754	0.046745	-0.06959	-0.04496	0.154837	-0.62241	-9.73E-09
Slope	0.007701	-0.07866	0.003161656	0.102454	-0.14674	-0.09736	-0.00964	0.11467	0.969017	0.001019	-8.71E-09

Comp	0.087475	0.330538	-0.31580282	0.726881	0.173043	0.428072	-0.20112	0.008844	0.016446	0.031909	2.51E-09
Sand	-0.4415	-0.30332	-0.22988691	0.036025	0.333612	-0.02406	0.046103	-0.21258	0.049446	0.094067	0.698892
Silt	0.098922	0.456708	0.441204354	0.035529	-0.43293	0.032421	-0.09697	-0.33167	0.007131	-0.05495	0.524169
clay	0.527531	-0.05632	-0.14508175	-0.09001	-0.01281	-0.00037	0.038235	0.662578	-0.0787	-0.07591	0.486619
OM	0.2506	-0.01191	-0.0971127	-0.01354	-0.00881	0.324123	0.861015	-0.2708	0.070684	-0.05133	-5.23E-09
CEC7	0.640062	-0.22985	-0.24234654	-0.05282	0.119999	-0.22942	-0.30358	-0.55596	0.040442	0.077783	1.35E-09
Rain	0.149371	-0.45435	0.672414208	0.49747	0.225513	-0.07634	0.082821	0.044607	-0.0708	-0.03308	-2.24E-09
Temperature	0.021636	-0.36327	0.148732584	-0.30819	-0.09573	0.79721	-0.31621	-0.03229	0.068716	0.001017	-4.57E-09

Table 10. Eigenvalues for Principal Components

Principal Component	**Eigenvalue**
PC1	0.170283522804821
PC2	0.0579925481848019
PC3	0.0342306870813013
PC4	0.0265192908198251
PC5	0.0216235091272856
PC6	0.0142817981274225
PC7	0.00962676979229931
PC8	0.00670140331125785
PC9	0.00388434090994049
PC10	0.00331893234867045
PC11	3.01549930098553e-17

We then compute the variance vector for these eigenvalues using equation (3) in METHEDOLOGY section. Table 11 list the variance for the inclusion of each PC.

Table 11. Variance of PC in dimensions

Principal Component	Variance
PC1	0.488670588594875
PC2	0.655094516105855
PC3	0.753327919599625
PC4	0.829431568623238
PC5	0.891485563975618
PC6	0.932470707941192
PC7	0.960097099403990
PC8	0.979328429873219
PC9	0.990475504631236
PC10	1.00000000000000
PC11	1

From Table 11 we observe that inclusion of up to 9 PCs will contribute to the variance of approximately 99%. Therefore, PC1 through PC9 will contribute to the total variance of approximately 99%. There using equations (1) and (2) we calculate 9 dimensional data for prediction i.e. V1 – V9 corresponding to PC1 – PC9.

Input Selection Based on Variance

From principal component analysis we concluded that we will use 9 dimensions for prediction i.e. V1 – V9. Table 12 shows the Regression (R) and RMSE for measured to predict yield. Each variable is first measured individually, and then combined to another one by one. All these values are measured against output variable i.e. “Corn Yield”.

Table 12. Regression and RMSE for reduced dimensions against Corn Yield

Variable	R	RMSE
V1	0.50	1595.85
V1, V2	0.62	1454.23
V1, V2, V3	0.62	1452.87
V1,V2, V3, V4	0.63	1435.2
V1, V2, V3, V4, V5	0.64	1416.25

V1,V2, V3, V4, V5, V6	0.68	1358.15
V1,V2, V3, V4, V5, V6, V7	0.68	1349.47
V1,V2, V3, V4, V5, V6, V7, V8	0.69	1340.49
V1,V2, V3, V4, V5, V6, V7, V8, V9	***0.69***	***1338.09***

From Table 12 we see that by reducing the dimensions we can achieve the maximum R of 0.69 and minimum RMSE of 1338.09.

The Neural Network model was then trained for data set of 2011, 2012 and 2013 with V1 – V9 as input. This trained model was then tested for the same 9 variables but for the data set of year 2014. Results are given in Table 13. The RMSE for the model and the test are noticeably higher than the stepwise iteration method.

Table 13. Neural Network results for the dimension reduction technique

Crop Year	**R**			**RMSE**			**Observations**
	Training	**Validation**	**Test**	**Training**	**Validation**	**Test**	
2011 - 2013	0.70	0.73	0.70	1357.81	1118.7	1374.61	5350
2014	-	-	0.63	-	-	1294.05	1055

Figure 5 shows the Regression plot for training, validation, testing and overall for the data of years 2011 - 2013. Figure 6 shows the regression plot of testing of the data of year 2014.

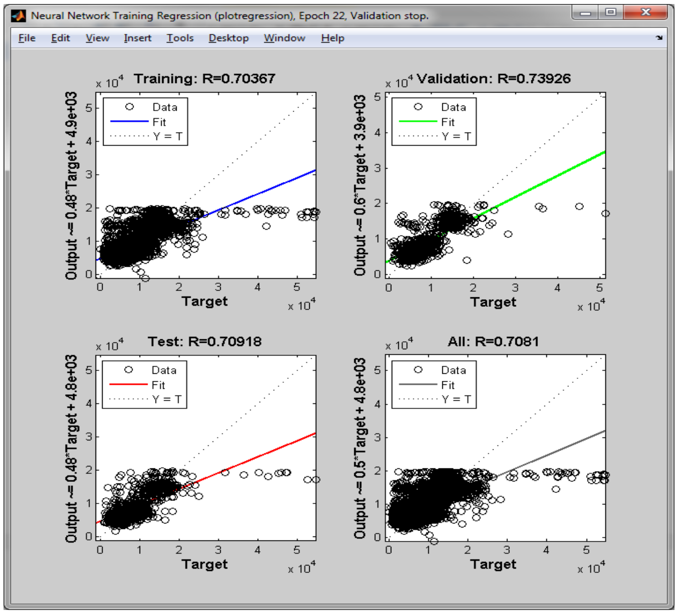

Figure 5. Regression plot for model building for the data-set of 2011 – 2013 (Dimension reduction)

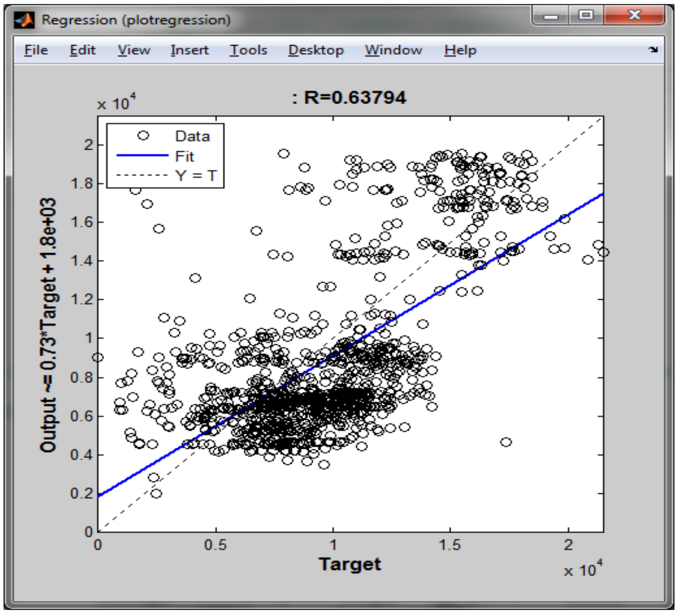

Figure 6. Regression plot for the testing of data for year 2014 (Dimension Reduction)

Neural Network Analysis Based on First Principal Component

First PC corresponds to the highest eigenvalue that explains the highest variance in data. In this section we propose an approach to input variable selection for ANN training based only on the first PC. Each variable can be associated to a component of the first PC. We propose to use the magnitude of the first eigenvector components for variable selection. Our intuition is that the higher the absolute value of a component corresponding to variable in PC-1, higher is the significance of that variable. However, we are unable find mathematical justification. So we follow an empirical approach, i.e. compare against other methods previously used. In this method we test if we select variables in the order of decreasing magnitude of the components of the first PC one at a time combining it with the one in the next higher order till what level we can achieve the accuracy in prediction. Our objective is to determine if it can be used as a feasible technique for the selection of input variables for prediction. If it is feasible, then it will save on computation time as opposed to the iterative process. Table 14 shows the order of significance for all the input variables.

Table 14. Order of significance of input variables as per PCA

Variable	Absolute Weightage
CEC7	0.640061931
Clay	0.527531058
Sand	0.441496262
OM	0.250600268
Rain	0.149371188
Water at 25 cm	0.107986892
Silt	0.098921569
Component	0.087474999
Water at 100 cm	0.048729382
Temperature	0.021636484
Slope	0.007700631

From Table 14, ‘CEC7’ is the most significant variable and ‘Slope’ is the least significant variable. We then find out the relation between these variables with the output variable as per this order and grouping these variables as per the order in the table. If the addition of a variable doesn’t make much difference in the model fit from the previous group we simply ignore that variable and move on to other variable in the order. For example, in Table 15 row 8, RMSE increases from the previous combination. So, we ignore ‘Component’ form the selection and it is not used as an input variable. Similarly we ignore ‘Water at 100 cm’ as its addition increases RMSE. Table 15 shows these relations.

Table 15. Relation between input variables and Corn Yield.

Variable	R	RMSE
CEC7	0.50	1607.72
CEC7, Clay	0.55	1537.13
CEC7, Clay, Sand	0.58	1441.42
CEC7, Clay, Sand, OM	0.63	1419.78
CEC7, Clay, Sand, OM, Rainfall	0.65	1402.14
CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm	0.66	1396.2
CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm, Silt	0.67	1385.53
CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm, Silt, Component	0.65	1424.13
CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm, Silt, Water at 100 cm	0.65	1439.4
CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm, Silt, Temperature	***0.70***	***1320.77***
CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm, Silt, Temperature, Slope	0.68	1352.92

Comparing the results from Table 7 and Table 15, we find that in Stepwise iteration we get the best results as R and RMSE of 0.70 and 1334.80, respectively. However as per Table 15, selection of different set of variables based on First PC, gives a better result than what we obtained in Stepwise iteration method i.e. R and RMSE of 0.70 and 1320.77, respectively. The R and RMSE for the variance based input selection are 0.69 and 1338.09 respectively. Thus, for our data-set, reducing dimensions using variance wasn’t helpful in getting better results. However the input variables selected using the magnitude of the components of the first PC was helpful in finding a better set of input variables which gave us better results than Stepwise iteration.

As stated earlier, we used this method to test our proposal empirically. Our analysis supports our proposal to uses the components of the first PC for input selection. Although this method has shown good results compare to stepwise iterative process in terms of RMSE, further confirmation via testing is required for validation.

The Neural Network model was then trained for data set of 2011, 2012 and 2013 for the above eight variables (as in the case of stepwise iteration). This trained model was then tested for the same eight variables but for the data set of year 2014. Results are shown in Table 16. We observe that the percentage errors for training and testing are 14% (1372/9610) and 12% (1206/9862) compared to actual average yield for the corresponding periods. This is comparable to the stepwise iteration approach.

Table 16. Result of Neural Network Model, First PC based input selection.

Crop Year	**R**			**RMSE**			**Observations**
	Training	**Validation**	**Test**	**Training**	**Validation**	**Test**	
2011 - 2013	0.71	0.74	0.71	1342.4	1139.66	1372.57	5350
2014	-	-	0.63	-	-	1206.59	1055

Figure 7 shows the Regression plot for Training, Validation, testing and overall for the data of year 2011 - 2013. Figure 8 shows the regression plot of testing of the data of year 2014.

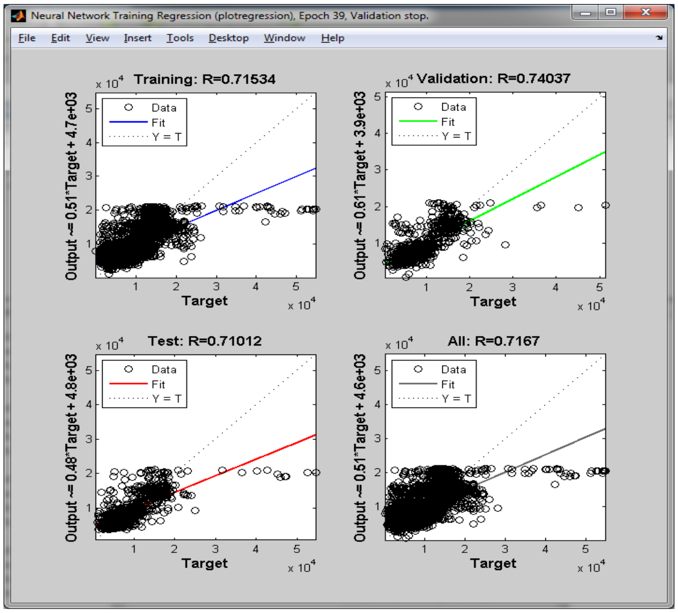

Figure 7. Regression plot for model building for the data-set of 2011 – 2013 (PCA)

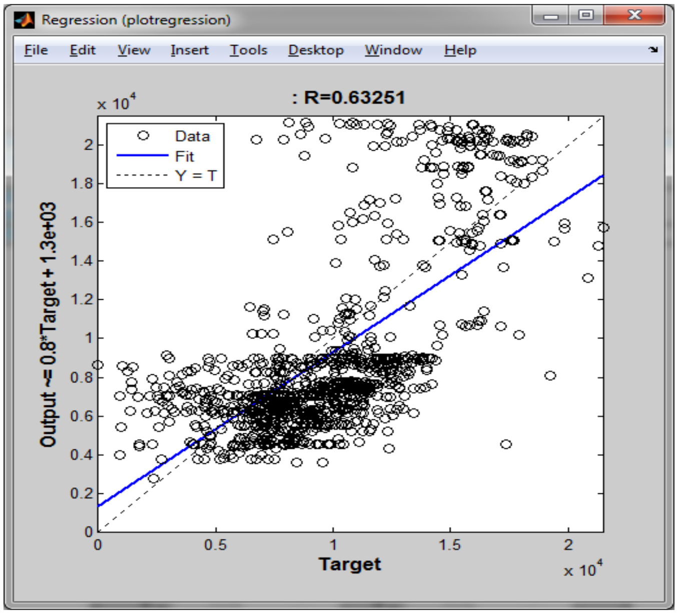

Figure 8. Regression plot for the testing of data for year 2014 (PCA)

Farmer's Approach for Prediction

In this section we use the simple average yield as a baseline to compare against the predicted output from the ANN. We call this Farmer's approach. We define the Farmer's prediction as the average of the yield for the previous years. There are in total 57 common subfields across all the years. The data used for this analysis is the data for these 57 fields. As per Farmer's approach the estimate for coming year is calculated by simply taking the average of previous year for every subfield. We have data for years 2011 – 2014. To estimate yield of 2014 by farmer's approach we use below formula for each subfield.

$$Estimate_{farmers} = \frac{\text{Yield in 2011} + \text{Yield in 2012} + \text{Yield in 2013}}{3}$$

We then trained the ANN model with the input variables selected by the First PC approach as it gave good results. The variables selected are ***CEC7, Clay, Sand, OM, Rainfall, Water at 25 cm, Silt, and Temperature***. For training we use data from year 2011, 2012 and 2013 for the 57 subfields mentioned earlier. 2014 data is used for testing. Table 17 compares the estimates of farmer's approach with predictions made by ANN after training.

Table 17. Comparison of Farmer's estimate and ANN prediction.

Subfield	**Yield for years**				**Farmers Method for 2014**		**ANN Method for 2014**	
	2011	**2012**	**2013**	**2014**	**Estimate**	**Error**	**Prediction**	**Error**
Field 1	4018.32	7186.06	9106.16	10248.33	6770.18	3478.15	8031.12	2217.21
Field 2	3738.14	6488.67	8400.25	10918.20	6209.02	4709.18	8341.33	2576.87
Field 3	3253.67	7325.67	9122.42	9024.08	6567.25	2456.83	8148.55	875.53
Field 4	3297.33	8923.51	9055.38	10665.79	7092.07	3573.72	8375.63	2290.16
Field 5	2987.98	8627.64	9102.25	10463.42	6905.96	3557.46	12187.93	1724.51
Field 6	2824.68	9528.84	7510.31	9303.53	6621.28	2682.25	7952.31	1351.22
Field 7	4007.68	8073.95	8058.91	9102.64	6713.51	2389.13	9491.36	388.72
Field 8	3669.60	7175.36	9551.34	8393.25	6798.77	1594.48	8601.66	208.41
Field 9	3174.52	6257.72	8446.87	8776.18	5959.70	2816.48	9991.87	1215.69
Field 10	2478.15	6080.12	7010.83	7263.44	5189.70	2073.74	9270.83	2007.39
Field 11	3684.78	8363.63	9314.32	9145.81	7120.91	2024.90	7746.42	1399.39
Field 12	***3276.37***	***6403.47***	***7900.28***	***6098.31***	***5860.04***	***238.27***	***11601.61***	***5503.30***

Field 13	3336.49	7036.81	8749.59	8253.45	6374.30	1879.15	6617.52	1635.93
Field 14	3793.95	8357.22	8095.99	8706.86	6749.05	1957.81	9920.22	1213.36
Field 15	3990.28	8097.85	9241.15	8450.28	7109.76	1340.52	7915.60	534.68
Field 16	4248.02	8091.69	8018.74	10538.36	6786.15	3752.21	7941.01	2597.35
Field 17	4661.99	7180.65	8425.99	10570.85	6756.21	3814.64	8945.83	1625.02
Field 18	4028.83	7608.60	9513.31	10381.14	7050.25	3330.89	8083.16	2297.98
Field 19	4128.30	9010.74	9142.55	9266.35	7427.20	1839.15	9054.37	211.98
Field 20	6754.61	7703.94	9168.04	10253.26	7875.53	2377.73	10600.59	347.33
Field 21	***6915.98***	***8652.49***	***9218.41***	***13004.45***	***8262.29***	***4742.16***	***8115.21***	***4889.24***
Field 22	6908.01	8514.13	8999.67	13401.46	8140.60	5260.86	12482.81	918.65
Field 23	7226.00	8089.49	8989.99	11212.06	8101.83	3110.23	8706.44	2505.62
Field 24	***7075.96***	***8258.65***	***9952.18***	***12447.33***	***8428.93***	***4018.40***	***8115.21***	***4332.12***
Field 25	8600.85	8207.67	10222.61	13072.30	9010.38	4061.92	10600.59	2471.71
Field 26	7062.38	8635.10	9707.98	12793.62	8468.49	4325.13	12482.81	310.81
Field 27	6678.68	8351.84	11400.12	11399.81	8810.21	2589.60	10877.04	522.77
Field 28	***5404.87***	***9080.24***	***9592.78***	***10400.47***	***8025.96***	***2374.51***	***14066.58***	***3666.11***

Field 29	5218.36	8143.92	9674.58	11250.68	7678.95	3571.73	13457.03	2206.35
Field 30	***5214.01***	***8502.95***	***9697.58***	***12010.51***	***7804.85***	***4205.66***	***4700.63***	***7309.88***
Field 31	4265.03	8019.64	9907.20	12725.25	7397.29	5327.96	9608.81	3116.44
Field 32	5193.82	6288.13	9316.29	10885.93	6932.75	3953.18	12690.18	1804.25
Field 33	***5525.46***	***8934.60***	***9218.42***	***11067.38***	***7892.83***	***3174.55***	***6464.65***	***4602.73***
Field 34	4925.82	8488.43	9146.40	10506.32	7520.22	2986.10	10909.81	403.49
Field 35	1741.31	4198.66	8269.84	9953.87	4736.60	5217.27	13518.42	3564.55
Field 36	***4216.03***	***8088.69***	***7915.81***	***8836.54***	***6740.18***	***2096.36***	***13602.24***	***4765.70***
Field 37	***4686.68***	***5918.59***	***8686.77***	***10070.50***	***6430.68***	***3639.82***	***14541.11***	***4470.61***
Field 38	***3439.28***	***7682.25***	***7209.86***	***6355.32***	***6110.46***	***244.86***	***12982.33***	***6627.01***
Field 39	5941.07	8190.45	10177.29	11851.84	8102.94	3748.90	14541.11	2689.27
Field 40	6080.15	8462.87	14109.75	12632.12	9550.92	3081.20	13602.24	970.12
Field 41	6036.38	6777.94	14616.94	11723.32	9143.75	2579.57	12982.33	1259.01
Field 42	5779.50	8453.11	9433.49	9301.74	7888.70	1413.04	8175.11	1126.63
Field 43	***4880.51***	***7466.70***	***8209.71***	***6896.76***	***6852.31***	***44.45***	***12982.33***	***6085.57***
Field 44	***4696.79***	***7959.00***	***8826.30***	***8631.01***	***7160.70***	***1470.31***	***11094.63***	***2463.62***

Field 45	***5733.89***	***6880.07***	***7239.45***	***8755.24***	***6617.80***	***2137.44***	***13378.91***	***4623.67***
Field 46	***4818.41***	***9064.82***	***7300.56***	***10039.86***	***7061.26***	***2978.60***	***14075.61***	***4035.75***
Field 47	***3172.32***	***6538.75***	***4992.59***	***7603.01***	***4901.22***	***2701.79***	***12695.78***	***5092.77***
Field 48	2612.12	5109.56	5249.56	10700.18	4323.75	6376.43	12706.67	2006.49
Field 49	5041.21	8019.96	7388.91	8794.20	6816.69	1977.51	10486.83	1692.63
Field 50	6552.10	7084.73	7430.23	13436.64	7022.35	6414.29	10960.21	2476.43
Field 51	5666.37	11343.27	10497.53	10026.38	9169.06	857.32	10565.51	539.13
Field 52	6112.75	4924.80	6643.43	11347.17	5893.66	5453.51	12166.03	818.86
Field 53	5779.01	8817.63	7676.16	12550.47	7424.27	5126.20	10960.21	1590.26
Field 54	***5499.71***	***9204.86***	***7919.31***	***11183.58***	***7541.29***	***3642.29***	***3830.60***	***7352.98***
Field 55	5378.37	9275.33	7396.34	11463.47	7350.01	4113.46	9943.96	1519.51
Field 56	***6020.99***	***9381.96***	***6971.78***	***7845.89***	***7458.24***	***387.65***	***6410.34***	***1435.55***
Field 57	***6161.45***	***8646.52***	***11143.59***	***13396.20***	***8650.52***	***4745.68***	***3281.17***	***10115.03***

From Table 17 we can see that the subfields highlighted are the ones for which Farmer's approach is better than ANN approach. In total there are 17 such subfields. So out of 57 sub fields we have got better results from ANN for 40 fields. Thus 70% (40/57) of the time we achieve better results using ANN approach overs Farmer's approach.

CHAPTER V

CONCLUSION

This paper describes three approaches of the input variable selection for the prediction of corn yield using Artificial Neural Network. The first methodology uses the Stepwise iteration method which uses ANN to incrementally add to the input variable collection starting with one until all combinations are tested. The second and third methodology use Principal Component Analysis (PCA), which is used to reduce the dimensionality of data-sets. The second method uses variance of variables for input variable selection. The third method uses the magnitude of the components of the first principal component for final variable selections. The results are summarized in Table 18.

Table 18. Summary of three methods

Method	**RMSE for model training Data of 2011 - 2013**	**RMSE for model test Data of 2014**
Stepwise Iteration	1393.04	1281.03
Selection based on variance	1374.61	1294.05
Selection based on first Principal Component.	1372.77	1206.59

RMSE for all the three models are close, however out of all three, selection based on first principal component techniques performs better in prediction than other techniques.

We analyzed corn yield data from the state of Texas. Fertilizer data was not used in the analysis.

If fertilizer data is not available, from the results we obtained, we can conclude that the 8 variables which together give the best corn yield prediction in the state of Texas are:

i. Rainfall
ii. Temperature
iii. Percentage of Sand in soil
iv. Percentage of Organic Matter in soil
v. Percentage of Clay in soil
vi. Percentage of CEC7 in soil
vii. Ground Water level up to 25 cm
viii. Percentage of Silt in soil.

Future Scope

The highest Regression value achieved for R is 0.63 and lowest RMSE is **1206.59**. We expect these numbers to improve if more information is available in the data-set. For example, one of the most influential factor of crop growth is the use of fertilizer. However our data-set lacks this information. Having this information could lead us to more accurate result. Also we have trained this model for data of 2011 – 2013 and testing it for the data of 2014. However, prediction may go wrong if there is a significant change in the weather conditions. This problem can be further addressed using domain adaptation technique [30]. This can be used to further enhance the prediction capability of the model. Also, further empirical or mathematical validation of the significant variable selection based on the magnitude of the components of the first PC is proposed as future work.

REFERENCES

1. Abrahart R.J., Anctil F., Coulibaly P., Dawson C.W., Mount N.J., See L.M., Shamseldin A.Y., Solomatine D.P., Toth E., Wilby R.L.
   Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting, Prog. Phys. Geogr., 36 (4) (2012), pp. 480–513.

2. Xuyuan Li, Holger R. Maier, Aaron C. Zecchin
   Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models, Volume 65, March 2015, Pages 15–29

3. Bowden G.J., Dandy G.C., Maier H.R.
   Input determination for neural network models in water resources applications. Part 1– background and methodology, J. Hydrol., 301 (1–4) (2005), pp. 75–92

4. Maier H.R., Jain A., Dandy G.C., Sudheer K.P.
   Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions, Environ. Model. Softw., 25 (8) (2010), pp. 891–909

5. Dandy G.C., Maier H.R.
   Determining inputs for neural network models of multivariate time series. Computer-Aided Civ. Infrastruct. Eng., 12 (5) (1997)

6. Agalbjörn S., Končar N., Jones A.
   A note on the gamma test
   Neural Comput. Appl., 5 (3) (1997)

7. Noori R., Karbassi A.R., Moghaddamnia A., Han D., Zokaei-Ashtiani M.H., Farokhinia A., GhafariGousheh M.
   Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. J. Hydrol., 401 (2011), pp. 177–189.

8. Trappenberg T., Ouyang J., Back A.
   Input variable selection: mutual information and linear mixing measures
   IEEE Trans. Knowledge Data Eng., 18 (1) (2006)

9. Chua L.H.C., Wong T.S.W.
Improving event-based rainfall-runoff modeling using a combined artificial neural network-kinematic wave approach, J. Hydrol., 390 (1–2) (2010)

10. Raorane A. A., Kulkarni R. V.
"Data mining: an effective tool for yield estimation in the agricultural sector," International Journal of Emerging Trends of Technology in Computer Science, vol. 1, no. 2, pp. 75–79, 2012.

11. Drummond S. T., Sudduth K. A., Joshi A., Birrell S. J., and Kitchen N. R.
"Statistical and neural methods for site-specific yield prediction," Transactions of the American Society of Agricultural Engineers, vol. 46, no. 1, pp. 5–14, 2003.

12. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, and Ojeda-Bustamante Waldo
Attribute Selection Impact on Linear and Nonlinear Regression Models for Crop Yield Prediction. The Scientific World Journal Volume 2014 (2014), Article ID 509429, 10 pages.

13. Irmak A., Jones J. W., Batchelor W. D., Irmak S., Boote K. J., and Paz J. O.
"Artificial neural network model as a data analysis tool in precision farming," Transactions of the ASABE, vol. 49, no. 6, pp. 2027–2037, 2006.

14. Roel A., Plant R. E.
"Factors underlying yield variability in two California rice fields," Agronomy Journal, vol. 96, no. 5, pp. 1481–1494, 2004.

15. Natural Resources Conservation Service, U.S. Department of Agriculture.

16. National Weather Service Forecast Office, Houston/Galveston, Texas (2004-12-25) .

17. Kaul, Monisha a, Hill, Robert L. b, Walthall, Charles a
Artificial neural networks for corn and soybean yield prediction
   a. Hydrology and Remote Sensing Laboratory, USDA-ARS, 10300 Baltimore Blvd., BARC-W, BLD007, RM 104, Beltsville, MD, 20705, USA
   b. Natural Resource Sciences and Landscape Architecture, University of Maryland, College Park, MD, 20742, USA.

18. URL: http://en.wikibooks.org/wiki/Artificial_Neural_Networks#ANN_Models.

19. URL: http://www.codeproject.com/Articles/175777/Financial-predictor-via-neural-network

20. Lawrence, J., 1994
Introduction to Neural Networks. California Scientific Software Press, Nevada City, CA.

21. Schmueli, D., 1998.
Application of neural networks in transportation planning, Progress in Planning 50, 143–201.

22. Witten, I. H. and Frank, E.
Data Mining, Practical Machine Learning Tools and Techniques, Morgan Kaufmann, 2nd edition, 2005.

23. JI, B. 1, 2, Sun, Y. 2, Yang, S. 1 and WAN, J. 1 2007.
1 National Key Laboratory for Crop Genetics and the Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China
2 College of Crop Science, Fujian Agriculture and Forestry University, Fuzhou 350002, China

24. Mundry, R. and Nunn, C. L.
Stepwise model fitting and statistical inference: turning noise into signal pollution
The American Naturalist, vol. 173, no. 1, pp. 119–123, 2009.

25. Sadras, V.O., Calvino, P.A., 2001.
Quantification of grain yield response to soil depth in soybean, maize, sunflower, and wheat. Agronomy Journal 93, 577–583.

26. Gbadegesin, A., 1987.
Soil rating for crop production in the Savanna Belt of South-Western Nigeria. Agricultural Systems 23, 27–42.

27. Whisler, F.D., Acock, B., Baker, D.N., Fye, R.E., Hodges, H.F., Lambert, J.R., Lemmon, H.E., McKinion, J.M., Reddy, V.R., 1986.
Crop simulation models in agronomic systems. Advances in Agronomy 40, 141–208

28. George, K. M.
An Application of PCA to Rank Problem Parts, the Second International Conference on Computational Science, Engineering and Information Technology, Coimbatore, India 26-28. 2012

29. Documentation of Matlab.

30. Jing Jiang, A literature survey on Domain Adaptation of Statistical classifiers, 2008.

VITA

Pawan Ramesh Lawale

Candidate for the Degree of

Master of Science

Thesis: INPUT VARIABLE ANALYSIS AND SELECTION FOR CORN YIELD PREDICTION USING ARTIFICIAL NEURAL NETWORK

Major Field: Computer Science

Biographical:

Education:

Completed the requirements for the Master of Science in Computer Science at Oklahoma State University, Stillwater, Oklahoma in July, 2015.

Completed the requirements for the Bachelor of Engineering in Information Technology at Government College of Engineering, Amravati, Maharashtra, India in 2007.

Experience:

Worked with Tata Consultancy Services as an IT Analyst at Pune, India
From Feb, 2008 – Mar, 2013