

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

**A Thesis Submitted for the Degree of PhD at the University of Warwick**

<http://go.warwick.ac.uk/wrap/55721>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

## Library Declaration and Deposit Agreement

### 1. STUDENT DETAILS

Please complete the following:

Full name: MUHD. KHAIKULZAMAN ABDUL KADIR

University ID number: 856844

### 2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (ETHOS) service.

[At present, theses submitted for a Master's degree by Research (MA, MSc, LL.M, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:

#### (a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of ..... months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. YES / ~~NO~~ (Please delete as appropriate)

#### (b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via ETHOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / ~~NO~~ (Please delete as appropriate)

OR My thesis can be made publicly available only after.....[date] (Please give date)  
YES / NO (Please delete as appropriate)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.  
YES / NO (Please delete as appropriate)

OR My thesis cannot be made publicly available online. YES / NO (Please delete as appropriate)

### 3. GRANTING OF NON-EXCLUSIVE RIGHTS

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

### 4. DECLARATIONS

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.
- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.
- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.
- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.
- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

### 5. LEGAL INFRINGEMENTS

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

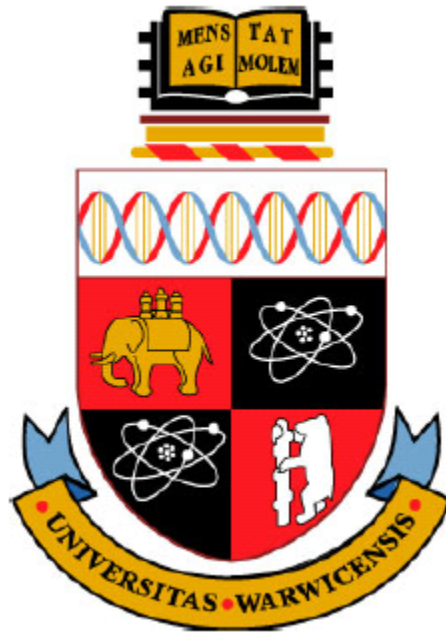
---

*Please sign this agreement and return it to the Graduate School Office when you submit your thesis.*

Student's signature: .....



Date: 11/3/13 .....



# **Food Security Modelling Using Two Stage Hybrid Model and Fuzzy Logic Risk Assessment**

By,

**Muhd Khairulzaman Abdul Kadir**

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Engineering, University of Warwick

March 2013

## **Table of Contents**

List of Figures	8
List of Tables	10
Acknowledgements	11
Declaration	12
List of Author's Publications	13
Abstract	14
Abbreviations	15
<b>Chapter 1: Introduction</b>	<b>16</b>
1.1    Introduction to Food Security	16
1.1.1    Concept and indicator of food security	18
1.2    Food Security Challenges and Modelling	19
1.2.1    Scopes, Challenges and Modelling	20
1.3    Research objectives	22
1.4    Thesis outline	23
Reference	24
<b>Chapter 2: Intelligent system Techniques</b>	<b>31</b>
2.1    Introduction to Intelligent System	31
2.2    Fuzzy Logic	32
2.2.1    Architecture of Fuzzy Logic	33
2.2.1.1    Fuzzification Interface	33

---

2.2.1.2 Defuzzification Interface	35
2.2.2 Purpose of Using the Fuzzy Logic	36
2.3 Artificial Neural Network (ANN)	36
2.3.1 Purpose of Using the ANN	40
2.4 Genetic Algorithm (GA)	40
2.4.1 Purpose of Using the GA	45
2.5 Hybrid Intelligent Systems (HIS)	45
2.5.1 Adaptive Neuro-Fuzzy Inference System (ANFIS)	47
2.5.2 Sensitive Genetic Neural Optimization (SGNO)	49
2.6 Conclusion	51
References	52

**Chapter 3: Two stage modelling using Genetic Algorithm Neural Network (GA-ANN) and Optimized Artificial Neural Network (ANN) 57**

3.1 Overview	57
3.2 Data Pre-processing	59
3.3 Determining GA parameter and fitness function	60
3.3.1 Selection	60
3.3.2 Crossover	61
3.3.3 Mutation	62
3.3.4 Population Vs Generation	63
3.3.5 Fitness function	65
3.4 Determining ANN parameter	66
3.4.1 Hidden layer and number of hidden neuron	66

3.4.2	Activation function	68
3.4.3	Learning function	69
3.5	First stage process	70
3.5.1	Variable Encoding	70
3.5.2	Population evaluation	71
3.6	Second stage process	72
3.6.1	Variable presentation	75
3.7	Remodelling ANN	77
3.8	Benchmarking techniques	77
3.8.1	Principal Component Analysis (PCA)	78
3.8.2	Multi Layer Perceptron - Artificial Neural Network (MLP-ANN)	80
3.8.3	Feature Selection (GA-ANN)	80
3.8.4	Optimized Weight and Threshold Neural Network (OWTNN)	81
3.8.5	Sensitive Genetic Neural Optimization (SGNO)	81
3.8.6	Benchmark Performance	81
3.9	Complexity of TSH	82
3.10	Potential of this modelling in food security area of study	83
3.11	Conclusion	83
	Reference	84

#### **Chapter 4: Farm household output prediction using farm household activities and behaviour** **90**

4.1.	Introduction	90
4.2.	Background	91

---

4.3.	Dataset	92
4.3.1.	ICRISAT	95
4.4.	Data Pre-processing	96
4.5.	First stage process	99
4.6.	Second stage process	101
4.7.	Remodelling ANN	104
4.8.	Benchmarking and discussion	107
4.8.1.	Principal Component Analysis (PCA)	107
4.8.2.	Multi Layer Perceptron - Artificial Neural Network (MLP-ANN)	110
4.8.3.	Feature Selection (GA-ANN)	113
4.8.4.	Optimized Weight and Threshold Neural Network (OWTNN)	115
4.8.5.	Sensitive Genetic Neural Optimization (SGNO)	117
4.8.6.	Summary	123
4.9.	Conclusion	125
	References	125

**Chapter 5: Trends in global output per capita prediction and modelling using  
FAOstat, USDA and World Bank database** **128**

5.1.	Introduction	128
5.2.	Background	129
5.3.	Dataset	130
5.4.	Data Pre-processing	135
5.5.	First stage process	137
5.6.	Second stage process	139



5.7.	Remodelling ANN	141
5.8.	Benchmarking and discussion	144
5.8.1.	Principal Component Analysis (PCA)	144
5.8.2.	Multi Layer Perceptron - Artificial Neural Network (MLP-ANN)	147
5.8.3.	Feature selection (GA-ANN)	149
5.8.4.	Optimized Weight and Threshold Neural Network (OWTNN)	151
5.8.5.	Sensitive Genetic Neural Optimization (SGNO)	153
5.8.6.	Summary	158
5.9.	Conclusion	160
	References	161

**Chapter 6: Food security risk level assessment prediction using Grain China dataset** **164**

6.1.	Introduction	164
6.2.	Background	165
6.3.	Dataset	166
6.4.	Data Pre-processing	168
6.5.	First stage process	169
6.6.	Second stage process	171
6.7.	Remodelling ANN	173
6.8.	Benchmarking and discussion	176
6.8.1.	Principal Component Analysis	176
6.8.2.	Multi Layer Perceptron - Artificial Neural Network (MLP-ANN)	179
6.8.3.	Feature selection (GA-ANN)	181

---

6.8.4. Optimized Weight and Threshold Neural Network (OWTNN)	183
6.8.5. Sensitive Genetic Neural Optimization (SGNO)	185
6.8.6. Summary	190
6.9. Conclusion	192
References	193
<b>Chapter 7: Food security Risk Level Assessment using Fuzzy Logic (FL)</b>	<b>196</b>
7.1. Introduction	196
7.2. Background	197
7.3. Dataset	200
7.4. Data Pre-processing	200
7.5. Fuzzy Logic Modelling	201
7.5.1. Fuzzification Interface	204
7.5.2. Knowledge Base and Decision Making Unit	205
7.5.3. Defuzzification	208
7.6. Risk Level Assessment Model Analysis	208
7.7. Performance Index	214
7.8. Conclusion	219
References	220
<b>Chapter 8: Conclusion and future works</b>	<b>224</b>
8.1. Overview	224
8.2. Research Summary	226

8.2.1. Two Stage Hybrid (TSH) model	226
8.2.2. Application chapter summary and results	228
8.3. Advantages and disadvantages of the TSH model and food security risk level assessment	230
8.4 Future works	231
References	232

## List of Figures

Figure 1.1:	General block diagram of food chain	19
Figure 2.1:	Architecture of FL	21
Figure 2.2:	Basic single artificial layer neural network	24
Figure 2.3:	ANN activation function	35
Figure 2.4:	Selection process	38
Figure 2.5:	Crossover example	39
Figure 2.6:	Mutation example	40
Figure 2.7:	Basic GA flowchart	41
Figure 2.8:	Original data and new offspring data for generation (i+1)	42
Figure 2.9:	Architecture of a 2 input ANFIS	45
Figure 2.10:	Basic SGNO system	48
Figure 3.1:	General structure of the two stage GA model	55
Figure 3.2:	Stochastic selection function	58
Figure 3.3:	Example of crossover scattered	59
Figure 3.4:	Example of population size in one generation	61
Figure 3.5:	Tangent Sigmoid activation function	65
Figure 3.6:	First stage process flow	68
Figure 3.7:	Second stage process flow	70
Figure 3.8:	Interconnection of weights and thresholds for ANN	72
Figure 4.1:	First stage performance via GA generation	93
Figure 4.2:	Performance of second stage via number of generation	95
Figure 4.3:	Regression of TSH model	97
Figure 4.4:	Performance of ANN based on MSE vs Epochs	97
Figure 4.5:	Overall regression for each part – training, validation and testing	98
Figure 4.6:	Variance in Principal component	100
Figure 4.7:	Regression for PCA using ANN	101
Figure 4.8:	MSE between PCA and TSH model	102
Figure 4.9:	Regression for MLP-ANN	103
Figure 4.10:	MSE performance of ANN models using TSH against MLP-ANN with original data	104
Figure 4.11:	Regression for feature selection using ANN	106
Figure 4.12:	MSE performance of ANN models using TSH against feature selection GA-ANN	107
Figure 4.13:	Regression for OWTNN using ANN	108
Figure 4.14:	MSE performance of ANN models using TSH against OWTNN	109
Figure 4.15:	Performance of SGNO chromosomes via number of generation	110
Figure 4.16:	Mean for each of the feature variables	112
Figure 4.17:	ANN performance based on SGNO	114
Figure 4.18:	MSE performance of ANN models using TSH against SGNO	114
Figure 4.19:	Benchmarking on overall ANN regression performance	116
Figure 4.20:	Benchmarking on MSE performance of ANN	116
Figure 5.1:	First stage performance via GA generation	131
Figure 5.2:	Performance of second stage via number of generation	132
Figure 5.3:	Regression of TSH model	134
Figure 5.4:	Performance of ANN based on MSE vs Epochs	135
Figure 5.5:	Overall regression for each part – training, validation and testing	135
Figure 5.6:	Variance in Principal component	137
Figure 5.7:	Regression for PCA using ANN	138

Figure 5.8:	MSE between PCA and TSH model	138
Figure 5.9:	Regression for original ANN	140
Figure 5.10:	MSE for original ANN and TSH	140
Figure 5.11:	Regression for feature selection using ANN	142
Figure 5.12:	MSE for feature selection and TSH using ANN	143
Figure 5.13:	Regression for OWTNN using ANN	144
Figure 5.14:	MSE for OWTNN and TSH using ANN	145
Figure 5.15:	Performance of SGNO chromosomes via number of generation	146
Figure 5.16:	Mean for each of the feature variables	147
Figure 5.17:	ANN performance based on SGNO for 9 input selections	149
Figure 5.18:	MSE performance of ANN models using TSH against SGNO	150
Figure 5.19:	Benchmarking on overall ANN regression performance	151
Figure 5.20:	Benchmarking on MSE performance	151
Figure 6.1:	First stage performance via GA generation	163
Figure 6.2:	Performance of second stage via number of generation	163
Figure 6.3:	Regression of TSH model	166
Figure 6.4:	Performance of ANN based on MSE vs Epochs	166
Figure 6.5:	Overall regression for each part – training, validation and testing	167
Figure 6.6:	Variance in Principal component	169
Figure 6.7:	Regression for PCA using ANN	170
Figure 6.8:	MSE between PCA and TSH GA-ANN model	171
Figure 6.9:	Regression for original ANN	172
Figure 6.10:	MSE for original ANN and TSH	173
Figure 6.11:	Regression for feature selection using ANN	174
Figure 6.12:	MSE for feature selection and TSH using ANN	175
Figure 6.13:	Regression for OWTNN using ANN	176
Figure 6.14:	MSE for OWTNN and TSH using ANN	177
Figure 6.15:	Performance of SGNO chromosomes via number of generation	178
Figure 6.16:	Mean for each of the feature variables	179
Figure 6.17:	ANN performance based on SGNO for 6 input selections	181
Figure 6.18:	MSE performance of ANN models using TSH against SGNO	182
Figure 6.19:	Benchmarking on overall ANN regression performance	183
Figure 6.20:	Benchmarking on MSE performance	184
Figure 7.1:	Food security risk assessment model	195
Figure 7.2:	Rule list showing the connection between the inputs and the output	199
Figure 7.3:	Control surface diagram	202
Figure 7.4:	Plots for each input membership function and its range	202
Figure 7.5:	Plots for the output membership function and its range	204
Figure 7.7:	Food security risk level for year 1988 – 2008	205
Figure 7.8:	UK performance index for risk level via producer price	206
Figure 7.9:	Australia performance index for risk level via producer price	207
Figure 7.10:	Germany performance index for risk level via producer price	207
Figure 7.11:	China performance index for risk level via producer price	208
Figure 7.12:	India performance index for risk level via producer price	208

## List of Tables

Table 1.1:	Food definition on broader term	17
Table 4.1:	Features variables on farm household	88
Table 4.2:	Statistics of all feature variables	89
Table 4.3:	Ranking selection for each of the features	111
Table 5.1:	Features categories for trends in global output per capita	124
Table 5.2:	List of basic statistic for input variables	128
Table 5.3:	Ranking selection for each of the features	148
Table 6.1:	Features categories for trends in global output per capita	159
Table 6.2:	List of basic statistic for input variables	160
Table 6.3:	Ranking selection for each of the features	190
Table 7.1:	Basic statistic of the dataset for each country	194
Table 7.2:	The grade of fuzzy inputs	197
Table 7.3:	The grade of fuzzy output	197
Table 7.4:	Summary of input and output for the year 1988	204
Table 7.5:	List of natural disasters from year 1988 – 2008	210

## **Acknowledgements**

I would like to thank my supervisor, Prof. Evor L. Hines for the great supports, supervision, ideas and endless guidance throughout the entire research and the structure of my thesis.

I would also like to thank my wife Norashima Mohd Albakri and my children for supporting, motivating and for bearing all the hardship right from the start of my research and through my writing process. A special thanks to my parents in helping in many other ways although they live far away.

Not forgotten are Dr. Mark Leeson, Dr. Fu Zhang, Dr. Xu Qin Li, Dr Jian Hua in the School of Engineering, Dr. Rosemary Collier, Dr. Elizabeth Dowler, Prof. Wyn Grant, Dr. Keith Richards, Dr. Richards Napier in School of Social Sciences, University of Warwick and Dr. Arjunan Subramanian in the University of Glasgow for helping me in find my way during the research process, thank you all. Finally, I would like to express my sincere gratitude to Universiti Kuala Lumpur (UniKL) and Majlis Amanah Rakyat (MARA) for giving a chance for me to study abroad and financial support.

## **Declaration**

This thesis is presented in accordance with the regulations for the degree of doctor of philosophy. The work described in this thesis is entirely original and my own, except where otherwise indicated.



## **List of Publications**

### **Journal paper**

1 – Muhd Khairulzaman Abdul Kadir, E Hines, Kefaya Qaddoum, Rosemary Collier, Elizabeth Dowler, Wyn Grant, Mark Leeson, Daciana Iliescu, Arjunan Subramanian, Keith Richards, Yasmin Merali & Richard Napier, 2013, Food Security Risk Level Assessment: A Fuzzy Logic based Approach, Applied Artificial Intelligence.

2 - Zulhilmy Sahwee, Halil Hussain, Muhd Khairulzaman Abdul Kadir, Mohamad Fiteri Razali, Mohamad Zikri Zainol, Shahliza Azreen Sarmin, 2012, Automatic Monitoring of Photovoltaic Cells Performance on Solar Aircraft, Recent Advances in Aerospace Technology, Applied Mechanics and Materials Vol. 225, Trans Tech Publications, Switzerland, pp. 356-360.

### **Conference papers**

1 – Muhd Khairulzaman Abdul Kadir, Evor Hines, Saharul Arof, Daciana Iliescu, Mark Leeson, Elizabeth Dowler, Rosemary Collier, Richard Napier, Qaddoum Kefaya and Reza Ghaffari, 2011 “Grain Security Risk Level Prediction Using ANFIS”, 3<sup>rd</sup> International Conference on Computational Intelligence, Modelling and Simulation, Malaysia, pp. 103-107.

2 - Muhd Khairulzaman Abdul Kadir, Evor Hines, Saharul Arof, Daciana Iliescu, Mark Leeson, Elizabeth Dowler, Rosemary Collier, Richard Napier, Arjunan Subramanian, 2012, “Neural Network for Farm Household Output Prediction.” International Conference on Statistics In Science, Business And Engineering. Langkawi, Malaysia.

3 - Kefaya Qaddoum, Evor Hines, Daciana Illiescu and Muhd Khairulzaman Abdul Kadir, 2011, “Self-Organizing Maps and Principal Component Analysis for Tomato Yield Datasets”, Middle Eastern Simulation and Modelling Conference, Amman, Jordan.

## **Abstract**

Food security has become a key issue worldwide in recent years. According to the Department for Environment Food and Rural Affairs (DEFRA) UK, the key components of food security are food availability, global resource sustainability, access, food chain resilience, household food security, safety and confidence of public towards food system. Each of these components has its own indicators which need to be monitored. Only a few studies have been made towards analysing food security and most of these studies are based on conventional data analysis methods such as the use of statistical techniques. In handling food security datasets such as crops yield, production, economy growth, household behaviour and others, where most of the data is imprecise, non-linear and uncertain in nature, it is better to handle the data using intelligent system (IS) techniques such as fuzzy logic, neural networks, genetic algorithm and hybrid systems, rather than conventional techniques. Therefore this thesis focuses on the modelling of food security using IS techniques, and a newly developed hybrid intelligent technique called a 2-stage hybrid (TSH) model, which is capable of making accurate predictions. This technique is evaluated by considering three applications of food security research areas which relate to each of the indicators in the DEFRA key food security components. In addition, another food security model was developed, called a food security risk assessment model. This can be used in assessing the level of risk for food security.

The TSH model is constructed by using two key techniques; the Genetic Algorithm (GA) module and the Artificial Neural Network (ANN) module, where these modules combine the global and local search, by optimizing the inputs of ANN in the first stage process and optimizing of weight and threshold of ANN, which is then used to remodel the ANN resulting in better prediction. In evaluating the performance of the TSH prediction model, a total of three datasets have been used, which relate to the food security area studied. These datasets involve the prediction of farm household output, prediction of cereal growth per capita as the food availability main indicators in food security component, and grain security assessment prediction. The TSH prediction model is benchmarked against five other techniques. Each of these five techniques uses an ANN as the prediction model. The models used are: Principal Component Analysis (PCA), Multi-layered Perceptron-Artificial Neural Network (MLP-ANN), feature selection (FS) of GA-ANN, Optimized Weight and Threshold (OWTNN) and Sensitive Genetic Neural Optimization (SGNO). Each of the application datasets considered is used to show the capability of the TSH model in making effective predictions, and shows that the general performance of the model is better than the other benchmarked techniques. The research in this thesis can be considered as a stepping-stone towards developing other tools in food security modelling, in order to aid the safety of food security.

## Abbreviations

ANN	Artificial Neural Network
ANFIS	Adaptive Neural Fuzzy Inference System
CGIR	Consultative Group on International Agriculture Research
DEFRA	Department for Environment Food and Rural Affairs
FAO	Food and Agriculture Organization of United Nations
FL	Fuzzy Logic
FS	Feature Selection
GA	Genetic Algorithm
HIS	Hybrid Intelligent System
ICRISAT	International Crops Research Institute for the Semi-Arid Tropics
IDSS	intelligent decision support system
IS	Intelligent System
LM	Lavernberg Marquadt
MF	Membership function
MLP	multi layer perceptron
MSE	mean square error
NN	neural network
OWTNN	optimize weight and threshold neural network
PCA	principal component analysis
RMSE	root mean square error
SAT	semi arid tropics
SGNO	sensitive genetic neural optimization
TSH	two stage hybrid
VLS	Village level study

# Chapter 1: Introduction

## 1.1 Introduction to Food Security

Food security has become an important issue, and is being discussed and studied globally. In 1996 at the World Food Summit, it was agreed that food security can be described as “a situation when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life”. This definition was issued again in 2006 in the policy brief of the *Food and Agriculture Organization* (FAO) (FAO, 2006).

The UK Department of Environment, Food and Rural Affairs (DEFRA) are one organization that provides guidelines in the UK for food and agriculture. The DEFRA definition and concept of food security is stated in a food chain analysis group paper (DEFRA, 2010, DEFRA, 2009), and follows the same outline as the definition by the FAO.

Food is an absolute necessity for life, not only for short periods, but for the long-term survival of life. Consequently, the safety or security of food is an important consideration for all living beings, especially humans.

A study by Maxx Dilley and Boudreau seeks to improve the practice of vulnerability assessment for food security purposes by addressing long-standing issues that have hampered the development of both theory and practical methods of maintaining food security (Dilley and Boudreau, 2001). However the study focuses mainly on the vulnerability concept being used in comparison with risks to food security. Another paper gives a broader definition of food security, which translates the definition of food security as described by FAO, and is shown in table 1.1 (Khanya-aicdd, 2006).

Table 1.1: Food definition on broader term (Khanya-aicdd, 2006)

Everyone has	Equity; all people
At all times	Stability of food (availability, access and utilization) in short term or long term. Protection against the risk of food security
Access to	Right food, enough food, affordable food and land rights for own food production
And control over	Power of decision relating to food production, distribution, consumption, etc.
Sufficient quantities	Enough food for daily requirement and sufficient stock for both the household level and community level.
Of good quality food	Variety of quality food (nutritious, safe, culturally appropriate)
For an active and healthy life	Proper consumption and a good biological utilization of food, resulting in an adequate nutritional status of people.

### **1.1.1 Concept and indicator of food security**

Each country generally has different concepts of food security. This concept depends on the dominant food in a country, and is also closely linked to the economic and social health of a nation, community and individual household (AusAID, 2004, Initiative, 2009). For example, in Laos, food security is defined as “to assure enough food and foodstuffs for every person at any time, both in material and economic aspects, with increasing demand on nutritional quality, hygiene and balance so as to improve health and enable normal development and efficient work”(NAPP, 2000). Other research states that, since the dominant food item for Laos is rice, a sufficient quantity of rice year round is the key to achieving food security. Even in the lowlands, where rice production is in surplus, the food supply has not yet been secured. The food regime at the household level is very poor in terms of protein, fat and micronutrients. Many experts suggest that diversification of household food production can ensure a better balanced diet, especially for protein intake (Khemmarath, 2005). In this case, food security concepts can also involve cultural, environmental and political aspects (Khemmarath, 2005). In another example, in Sub Saharan Africa, food security is described as the function of food production at the macro level and income at the household level (Jones, 2008).

In the DEFRA analysis, food security concepts are based on different levels such as individual or household food security, regional food security, national food security and global food security (DEFRA, 2010). The main focus for this report is on the national food security, but it also refers to other level. The key themes for food security in DEFRA paper discussion are around food availability, universal access to the food, affordability, nutrition and quality, safety, resilience and confidence from all groups of people in UK (DEFRA, 2010).

In other words, food security is very important in ensuring that people can maintain good nutrition and continue to live in a good health. In studying the flow required to ensure food security, it is necessary to start from the beginning of the food production chain, through to the end product. This flow is called a food chain. Figure 1.1 shows a general food chain.

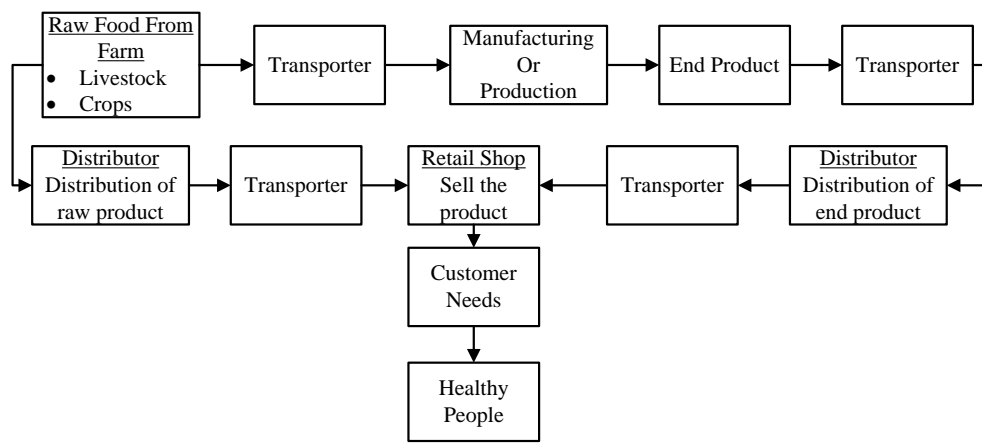


Figure 1.1: General block diagram of food chain

Each component in the food chain represents the flow of the raw food product until it becomes the end product for consumption. These components need to be monitored and assessed in terms of quality and quantity, in order to maintain food security.

## 1.2 Food Security Challenges and Modelling

There are a very few studies which involve the modelling of food security as (Men et al., 2009, Jianling and Yong, 2010a, Jianling and Yong, 2010b, Kadir et al., 2011), either in terms of risk management or the assessment of any effect related to

food security, such as: food availability, food access, food chain resilience, safety and public confidence. Food security behaviour also needs to be predicted to ensure that the level of risk of food security is at an acceptable level.

### **1.2.1 Scopes, Challenges and Modelling**

In the first step to ensure food security is at a safe level, it is suggested to have a model which can be used to predict and to monitor the behaviour of food security. Not many studies in the analysis of food security have been made, such as studies at the household level (Kirkpatrick, 2008) and (Cordell, 2010). In Kirkpatrick, the studies examine the adequacy of nutrition and factors relating to the level of household food security. In Cordell, consideration of the availability of Phosphorus forms a main part of the study, towards ensuring the maintenance of food security at a safe level. In addition to these two studies, most of the other studies which have been performed are based on conventional data analysis (statistical techniques). Another example of a food security study is also made by using a conventional techniques is the study of China grain warning model, which use Analytical Hierarchical Process (AHP) as China's grain security warning indicator as in (Men et al., 2009).

In handling food security datasets such as crops yield data, production data, and economic growth data, mostly consisting of multiple non-linear and uncertain data, this is best handled using the Intelligent System (IS) technique.

Generally, IS are used with Soft Computing and the Computational Intelligence techniques, which differ from conventional techniques. IS can be defined as a system which learn and acts accordingly to the environment until a pre-defined



objective is achieved. IS are more suited to applications which involve data imprecision, partial truth and approximated reasoning (Venugopal K.R., 2009, Mitra S., 2002, Fritz, 1997). FL, ANN and GA are the fundamentals techniques of IS (Mitra S., 2002).

Generally, IS techniques have been used and studied mainly in the area of food production, manufacturing, crop production or food technology; for example Ahmend et al. used FL to enhance crop control (M. Ahmend, 1999), Odetunji et al. studied the gari fermentation plant control system using FL (Odetunji and Kehinde, 2005) and Davidson et al. developed a fuzzy risk assessment tool for microbial hazards in food system (Davidson et al., 2006). There are further studies involving the above area by (Ding et al., 2007, Dohnal et al., 1993, El-Sebakhy et al., 2007, Inglis et al., 1997, Perrot et al., 2006, Xiaojun et al., 2008, Xiaping. Fu, 2007, Xie et al., 1998).

Research on the monitoring the risk level of food security is rare; only a few pieces of research are available; for example, grain security modelling in China has been studied using the IS technique. These studies show the importance of grain as a main food source where it can be used as the main indicator of food security risk level in China (Jianling and Yong, 2010a, Jianling and Yong, 2010b, Xiaojun et al., 2008, Xiaoping et al., 2009, Yong and Jianling, 2010).

In term of prediction, a study had been done performed by Muhd Khairulzaman using the same dataset as the grain China security risk level. The prediction is based on three main indices; consumptives index, productive index and disaster index in predicting the risk level of food security (Kadir et al., 2011). Another study by him which also relates to food security using intelligent systems is based on

farm household behaviour or activities, to predict the farm household crop output (Muhd Khairulzaman Abdul Kadir, 2012).

Based on the above discussion, any changes to the indicators (main indicators or sub-indicators) as in (DEFRA, 2010), key components, or themes of food security which can impact food access, food availability and food risk management need to be considered. These are becoming the three most challenging areas that need to be considered in modelling food security using IS techniques.

### **1.3 Research Objective**

The aims of this research is to establish a systematic relationship between the food security and the effect of the food security itself either by using a clustering method for unsupervised learning, or through supervised learning for the study involving the prediction model. The study can also gives a basic overview of the impact on overall food security key components as described in earlier of this chapter, in terms of food production, food quality and food access for all people.

In referring to the aims, the primary objectives of this study is to develop an effective general purpose prediction model using Hybrid Intelligent System (HIS) techniques, known as Two-Stage Hybrid (TSH) model. The developed model can also perform automatic feature selection or input selection, while maintaining the amount of information from the original dataset which gives the best prediction performance. In the determination of the relationship between the features, regression plots will be used.

This developed model will be applied to the food security research area in conjunction with the aims. The possibility also exists for applying the TSH to related areas of study of food security, such as a farm household output prediction model, a food growth per capita prediction model and a grain security warning prediction model. In addition, each of these studies will be compared with several IS techniques and a conventional technique to test the performance of TSH.

The final objective of this thesis is to develop a risk assessment model by using Fuzzy Logic (FL). This study will examine the risk level of food security, which can be used to determine the overall level prevailing food security risk by monitoring various risk elements related to food security, as described in DEFRA assessment report indicated in the previous section. This part of the study will also explore possible outcomes by using a combination of strategies and sensitivities to local and global conditions to account for any uncertainties.

#### **1.4 Thesis Outline**

The current chapter present the overview of the food security concept and the general applications which are involved in modelling it. This chapter also describes the general research objectives and overall thesis structure.

Chapter 2 reviews all of the IS techniques used in the research that are relevant to the development of the proposed prediction model, including the ANN technique and GA technique. Additionally, a Sensitive Genetic Neural Optimization (SGNO) is discussed, and will be used as one of the benchmark techniques in terms of its prediction performance.

Chapter 3 describes details of the proposed TSH model and its construction. This model aims to become one of the prediction models. Options for the implementation of the benchmark techniques are also reviewed and explained.

Chapter 4 demonstrates the application of the farm-household output prediction model by using the TSH model, which will be based on the farm household activities and behaviour. The performance of the proposed model will be benchmarked with other techniques.

Chapter 5 illustrates another application by using the TSH model in predicting the food growth per capita. The performance is also discussed and benchmarked by using other techniques.

Chapter 6 presents another related food security prediction model, where the TSH model is used to predict the grain security level based on the three categories – production indices, consumption indices and disaster indices.

Chapter 7 demonstrates a newly developed food security risk level assessment model by using a FL as the unsupervised learning method. In this chapter, three major components are assumed to have a major impact on food security risk level.

Chapter 8 summarises and concludes all of the discoveries presented in the previous chapters. It also suggests future work which could be undertaken in this research area, and possible improvement to the model itself.

---

**References**

AUSAID. 2004. *Food Security*

*Strategy* [Online]. Australian Government. Available:  
[http://www.usaid.gov.au/publications/pdf/food\\_security\\_strategy04.pdf](http://www.usaid.gov.au/publications/pdf/food_security_strategy04.pdf)  
[Accessed 16/5/11 2011].

CORDELL, D. 2010. *The Story of Phosphorus: Sustainability Implication of Global Phosphorus Scarcity for Food Security*. PhD, Linköping University.

DAVIDSON, V. J., RYKS, J. & FAZIL, A. 2006. Fuzzy risk assessment tool for microbial hazards in food systems. *Fuzzy Sets and Systems*, 157, 1201-1210.

DEFRA 2009. UK Food Security Assessment: Out Approach.: Department for Environment and Rural Affairs.

DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In*: DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). Department for Environment, Food and Rural Affairs.

DILLEY, M. & BOUDREAU, T. E. 2001. Coming to terms with vulnerability: a critique of the food security definition. *Food Policy*, 26, 229-247.

DING, Z.-H., LI, J.-T. & FENG, B. Year. Radio Frequency Identification in Food Supervision. *In*: Advanced Communication Technology, The 9th International Conference on, 12-14 Feb. 2007 2007. 542-545.

DOHNAL, M., VYSTRCIL, J., DOHNALOVA, J., MARECEK, K., KVAPILIK, M. & BURES, P. 1993. Fuzzy food engineering. *Journal of Food Engineering*, 19, 171-201.

EL-SEBAKHY, E. A., RAHARJA, I., ADEM, S. & KHAERUZZAMAN, Y. Year. Neuro-Fuzzy Systems Modeling Tools for Bacterial Growth. *In*: Computer

- Systems and Applications, 2007. AICCSA '07. IEEE/ACS International Conference on, 13-16 May 2007 2007. 374-380.
- FAO 2006. Policy Brief : Food Security. *In: ECONOMICS*, A. A. D. (ed.). FAO's Agriculture and Development Economics Division (ESA) with support from the FAO Netherlands Partnership Programme (FNPP) and the EC-FAO Food Security Programme.
- FRITZ, W. 1997. *Intelligent Systems and their Societies* [Online]. Available: <http://www.intelligent-systems.com.ar/intsys/index.htm> [Accessed 5/3/13 2013].
- INGLIS, I. R., FORKMAN, B. & LAZARUS, J. 1997. Free food or earned food? A review and fuzzy model of contrafreeloading. *Animal Behaviour*, 53, 1171-1191.
- INITIATIVE, L. A. F. S. 2009. "L'Aquila" Joint Statement on Global Food Security. L'Aquila Food Security Initiative.
- JIANLING, X. & YONG, D. Year. Food safety risk analysis based on generalized fuzzy numbers. *In: Advanced Management Science (ICAMS)*, 2010 IEEE International Conference on, 9-11 July 2010 2010a. 699-702.
- JIANLING, X. & YONG, D. Year. Linguistic ranking model and its application in food management. *In: Computer Design and Applications (ICCD)*, 2010 International Conference on, 25-27 June 2010 2010b. V5-208-V5-212.
- JONES, M. P. 2008. *Achieving Food Security And Economic Growth in Sub-Saharan Africa: Key Institutional Levers*. [Online]. Africa FARA. Available: <http://www.fara-africa.org/library/browse/ACHIEVING FOOD SECURITY AND ECONOMIC GROWTH IN SUB-SAHARAN AFRICA> 2 .pdf [Accessed August 2010].

- KADIR, M. K. A., HINES, E. L., AROF, S., ILLIESCU, D., LEESON, M., DOWLER, E., COLLIER, R., NAPIER, R., KEFAYA, Q. & GHAFARI, R. Year. Grain Security Risk Level Prediction Using ANFIS. *In: Computational Intelligence, Modelling and Simulation (CIMSIM), 2011 Third International Conference on, 20-22 Sept. 2011*. 103-107.
- KHANYA-AICDD. 2006. *Food security concept paper* [Online]. Available: [http://www.khanya-aicdd.org/siteworkspace/files/foodsecurityconceptpaperfinal06\\_09\\_07.pdf](http://www.khanya-aicdd.org/siteworkspace/files/foodsecurityconceptpaperfinal06_09_07.pdf) [Accessed 13/05/11 2011].
- KHEMMARATH, S. 2005. *Key Concepts of Food Security* [Online]. NAFRI. Available: [http://www.nafri.org.la/document/sourcebook/Sourcebook\\_eng/Volume1/13\\_conceptsfoodsec\\_sitha.pdf](http://www.nafri.org.la/document/sourcebook/Sourcebook_eng/Volume1/13_conceptsfoodsec_sitha.pdf) [Accessed 16/5/2011 2011].
- KIRKPATRICK, S. 2008. *Household Food Insecurity in Canada: Examination of Nutrition Implications and Factors Associated with Vulnerability*. PhD, University of Toronto.
- M. AHMEND, E. D., ET AL. 1999. A general purpose fuzzy engine for crop control. *Computational Intelligence*, 1625, 473-481.
- MEN, K., WEI, B., TANG, S. & JIANG, L. Year. China's Grain Security warning based on the integration of AHP-GRA. *In: Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on, 10-12 Nov. 2009*. 655-659.
- MITRA S., P. S. K., MITRA P. 2002. Data Mining in Soft Computing Framework: A Survey. *IEEE Transactions on Neural Network*, 13, 3 - 14.

- MUHD KHAIRULZAMAN ABDUL KADIR, E. L. H., SAHARUL AROF, DACIANA ILIESCU, MARK LEESON, ELIZABETH DOWLER, ROSEMARY COLLIER, RICHARD NAPIER, ARJUNAN SUBRAMANIAN 2012. Neural Network for Farm Household Output Prediction. *International Conference on Statistics In Science, Business And Engineering*. Langkawi, Malaysia.
- NAPP 2000. Lao PDR Food Security Strategy In the Period 2001-2010. Vientiane: National Institute of Agriculture Planning and Projection.
- ODETUNJI, O. A. & KEHINDE, O. O. 2005. Computer simulation of fuzzy control system for gari fermentation plant. *Journal of Food Engineering*, 68, 197-207.
- PERROT, N., IOANNOU, I., ALLAIS, I., CURT, C., HOSSENLOPP, J. & TRYSTRAM, G. 2006. Fuzzy concepts applied to food product quality control: A review. *Fuzzy Sets and Systems*, 157, 1145-1154.
- VENUGOPAL K.R., S. K. G., PATNAIK L.M. 2009. *Soft Computing for Data Mining Applications*, New York, Springer-Verlag.
- XIAOJUN, W., DONG, L. & XIANLIANG, S. Year. A fuzzy enabled model for aggregative food safety risk assessment in food supply chains. *In: Service Operations and Logistics, and Informatics*, 2008. IEEE/SOLI 2008. IEEE International Conference on, 12-15 Oct. 2008 2008. 2898-2903.
- XIAOPING, W., YU, F. & JIASHENG, W. Year. Information systems security risk assessment on improved fuzzy AHP. *In: Computing, Communication, Control, and Management*, 2009. CCCM 2009. ISECS International Colloquium on, 8-9 Aug. 2009 2009. 365-369.



- XIAPING. FU, Y. Y., ET AL. 2007. Principal Components-Artificial Neural Networks for Predicting SSC and Firmness of Fruits based on Near Infrared Spectroscopy. *ASABE Annual meeting Paper*.
- XIE, G., XIONG, R. & CHURCH, I. 1998. Comparison of Kinetics, Neural Network and Fuzzy Logic in Modelling Texture Changes of Dry Peas in Long Time Cooking. *Lebensmittel-Wissenschaft und-Technologie*, 31, 639-647.
- YONG, D. & JIANLING, X. Year. Fuzzy evidential warning of grain security. *In: Advanced Management Science (ICAMS), 2010 IEEE International Conference on, 9-11 July 2010 2010. 703-706.*

## **Chapter 2: Intelligent System Techniques**

### **2.1 Introduction to Intelligent System**

Artificially Intelligent systems have been used in studying the behaviour of food-related areas, for example in the study of crop control by Moataz Ahmend (M. Ahmend, 1999). This study relied on the feeding of remote sensing data, together with an evolvable model of crop behaviour, into an intelligent decision support system (IDSS), based on the backward chaining of a fuzzy reasoning engine. Another example is in data fusion for a food transportation system by using neural networks in (Jabbari et al., 2008) and (Farkas et al., 2000); the objective of this study is was to achieve final moisture distribution in the material bed of a grain processing system, with minimum consumption of energy in the shortest possible drying time. There are a lot of other studies relating to food involving artificial intelligence, which, use various IS technique such as fuzzy logic (Odetunji and Kehinde, 2005), neural networks (Fu Xiaping, 2007), neural-fuzzy methods (El-Sebakhy et al., 2007) and genetic algorithms (Chemin and Honda, 2006). Each of these techniques has their own principles and concepts, depending on the application.

As previously stated, most of the existing studies have been in the food-related area with IS as the technique employed, which is only a sub-area of the study of food security. Very few studies have looked at food security in its entirety. However, one example of paper which did consider this area is by Nathalie et al (Perrot et al., 1999) and Oscar Castillo et al. (Jones, 2006). The paper itself concentrates on ensuring the quality of food by using a fuzzy set in considering wet-milling for maize, and for manufacturing in the food industry, both using fuzzy logic techniques. This study can be considered as a sub area of food security; as explained in the previous chapter in the food security themes describe by (DEFRA, 2010), the main indicators for assessing food security are: food availability, every people access to food, affordability, nutrition and quality, safety, resilience and confidence by all people.

In this chapter, it will be emphasized all of the IS techniques such as Fuzzy Logic (FL), Artificial Neural Network (ANN), Genetic Algorithms (GA) and other hybrid IS methods that used in the modelling of food security, either in predicting it or clustering it to get the feature from the parameter being used. At the same time, along with the usage of the existing IS techniques, a new method for the modelling of food security is developed, based on these techniques.

## **2.2 Fuzzy Logic (FL)**

Lofti Zadeh (1965) is attributed as being the key contributor to the modern era of FL and its applications. This study was introduced to demonstrate the vagueness of linguistics and to describe the expression of human 'knowledge' in a natural way (Haslum, et al, 2007). Most of the applications that involved FL were based on its reasoning process and expression in terms of understandable to both operators and experts (Perrot, et al, 2006).

### 2.2.1 Architecture of Fuzzy Logic

The basic architecture of FL is based on the concept of a ‘crisp’ input and ‘crisp’ output. Crisp means the actual data or parameter being used, is described either in quantitative or qualitative parameters. Between the crisp inputs and crisp output, all of the process is based on ‘fuzzy’ parameters which are converted at the beginning of the process. The full architecture of FL is shown in Figure 2.1.

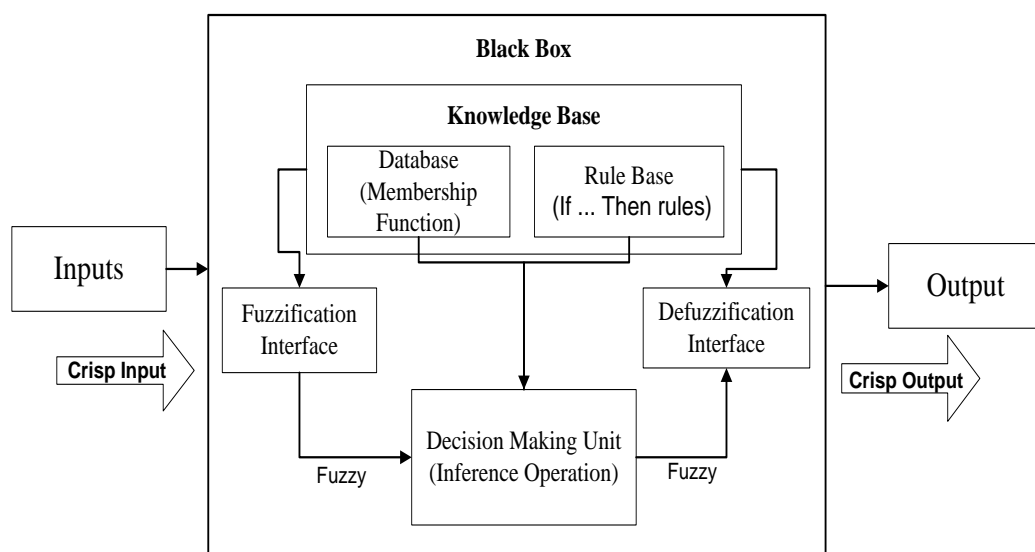


Figure 2.1: Architecture of FL

#### 2.2.1.1 Fuzzification Interface

The first step in the process is known as the ‘fuzzification processes’. It involves rule evaluation and aggregation, which is done mostly in the Knowledge Base block. There are two popular techniques which are used in this process; the Mamdani method and the Sugeno method. The advantage of the Mamdani method is in capturing expert knowledge in its entirety, whereas the Sugeno method only uses the singleton rule output, which only works well with linear techniques (Negnevitsky, 2005, J.-S.R. Jang, 1997).

A very important part of this process is the way in which the fuzzy sets are determined. Fuzzy sets consist of an element which can belong partly to two different sets, each of which have different memberships (Negnevitsky, 2005). In classical sets, each set has a definite answer. In deciding this, the set will be based on the inputs and outputs of the system being modelled; for example the study by (Huey-Ming, 1996), used 11 grades of risk and 11 grades of importance, from definitely unimportant to definitely important, as its fuzzy sets function.

In the design and implementation of a FL system, there is the option of using the three most popular membership functions (MFs) which are known as the triangular, Gaussian and trapezoidal functions (Negnevitsky, 2005). Each of these MFs has its own characteristics in describing the fuzzy sets, and will have a different performance rate and accuracy. For example, using triangular and trapezoidal functions means that the performance rate will be very fast, however the level of accuracy will be lower than would be the case with either of the other membership functions; this is known as the ‘normal speed versus complexity scenario’ (Xie et al., 1998). This process and implementation is performed in the Database block.

The next step is to determine the rules relationship for each of the inputs and the output, which is done in Rule Base block. This is where the inference system is used, specifically based on ‘If...then...rules’ or Bayesian rules (Negnevitsky, 2005) as below:-

*IF x is  $A_i$  and IF y is  $B_i$  then z is  $C_i$*

The above rule shows typical conditions of the input function of x and y with z as the output function, the fuzzy states of  $A_i$ ,  $B_i$  and  $C_i$  of i-th as the Rule Base condition.

The design of these rules usually depends on the inputs and each of the membership functions, for example;

*“Assume given a 3 inputs and a 3 membership functions for each input where the numbers of rules that can be generated based on these condition are  $3^3 = 27$  rules.”*

As also described in (Negnevitsky, 2005, J.-S.R. Jang, 1997), rule evaluation in the fuzzy inference system model is based on either an ‘AND’ function or an ‘OR’ function, in terms of the fuzzy operation as algebraic product or algebraic sum, which is used to compare the inputs. The level of the truth value of the antecedent is then determined, and the consequent membership function is either clipped (correlation minimum) or scaled (correlation product) (Negnevitsky, 2005). This fuzzy operator will compare each of the inputs, and this operation takes place between the Fuzzification Interface and Knowledge Base, and will be processed in the Decision Making Unit as shown in Figure 2.1.

#### **2.2.1.2 Defuzzification Interface**

The final process is to convert all of the fuzzy values to the crisp values. This is done by defuzzification or aggregation of the rule output. Therefore, the fuzzy values that have been declared will be used to evaluate the rules, but, the final output should be a crisp value as described in (A.S. Sodiya, 2007). In order to perform the defuzzification, a number of different approaches can be used, such as; the centre of gravity or centroid, the maximum membership principle, the weighted average method, the mean-max membership method, the centre of sum method, the centre of largest area method, or the first (or last) maxima method (J.-S.R. Jang, 1997,

Negnevitsky, 2005, Ross, 2007). The most widely used defuzzification technique is the centroid method as shown in equation (2.1), which is also discussed in later a chapter when considering clustering methods using FL.

$$z^* = \frac{\int \mu_c(z) \cdot z \, dz}{\int \mu_c(z) \, dz} \quad (2.1)$$

### 2.2.3 Purpose of Using the FL

In analyzing a certain dataset, some features of the data need to be known before the output can be predicted. In one of the studies in this thesis, some data do not have any targets, which mean an unsupervised learning method needs to be used, and FL is one such unsupervised learning method in IS. In addition, FL offers advantages in converting any vagueness in linguistics, and it has a capability to explain terms based on human understanding of knowledge. FL also can be used to represent any qualitative parameter or quantitative parameter, which can help to model a food security risk level assessment. In modelling this, a lot of assumptions are made based on experience of the relationship between each of the input parameter and previous work related to this study, which will be explain in Chapter 7.

### 2.3 Artificial Neural Network (ANN)

The Artificial Neural Network (ANN) is a well established technique, describe by (Eduard Llobet, 1999, Gardner et al., 1992, J W Gardner, 1990, Negnevitsky, 2005). Generally, ANNs are based on a single layer network consisting of 3 basic layers; the input layer, the hidden layer and the output layer. This type of network is called as the ‘perceptron’ as shown in Figure 2.2.

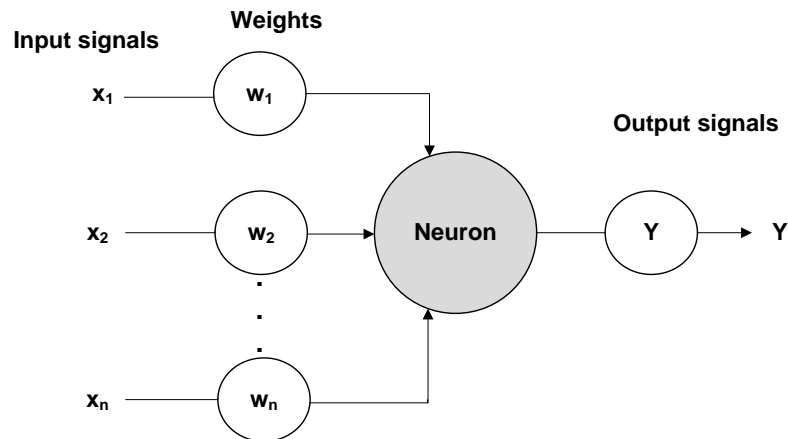


Figure 2.2: Basic single artificial layer neural network (Negnevitsky, 2005)

This ANN architecture use ‘back propagation feed forward’ network, where the input pattern is applied to the network and forward propagated through the network using the initial node connection weights (S.A. Shearer, 2000). The back-propagation algorithm consists of many techniques, each of which has different ways of updating the bias and the weight of the network (H. Demuth, 2004).

In the back propagation training algorithm, the first step is to initialize all of the weights and thresholds (if any) to a random value. After that, the network must activate its back-propagation process by using the activation function. Then, the weight for each neuron will be updated accordingly, based on the number of neurons for each layer. Finally, all of the above processes will be repeated until the sum square error is less than 0.001 (Negnevitsky, 2005).

Rather than using several hidden layers, the ANN is usually constructed based on one hidden layer, because additional hidden layers will increase the processing time on some applications. A significant number of researchers have proved that the use of one hidden layer is sufficient to approximate any function which contains a continuous mapping from one space to another in solving real life problems (Foster et al., 1999, Zhang, 2011).



In deciding on the number of neurons, as in (Weigend, 1994, Geman, 1992, Tetko, 1995), there are a number of ways to determine the required quantity of neuron in the hidden layer; usually the optimal number of neurons will be determined only after several training processes have been completed or based on a ‘rule of thumb’.

In an ANN, a lot of activation functions are tested, but only a few can be applied in real applications. These include the step function, sign function, sigmoid or hyperbolic function and linear function (Negnevitsky, 2005, J.-S.R. Jang, 1997). These activation functions are showed in Figure 2.3.

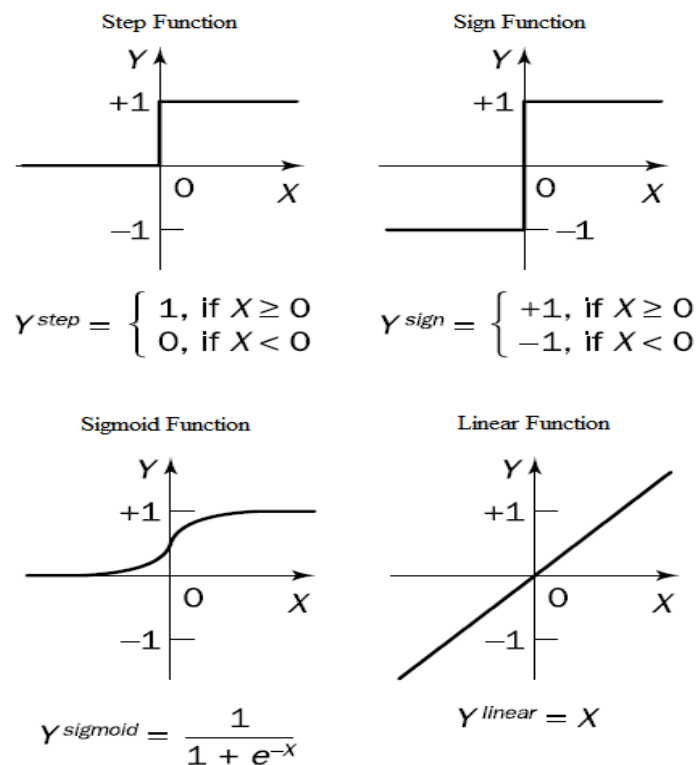


Figure 2.3: ANN activation function (J.-S.R. Jang, 1997, Negnevitsky, 2005)

In a back-propagation multilayer network, the sigmoid function and linear function are usually used as the activation function. It provides a smooth and non-zero derivative with respect to input signals. Sometimes, this function is known as a ‘squashing function’, but normally it is called as a sigmoidal function (Gardner et al.,

1992, J W Gardner, 1990, J.-S.R. Jang, 1997). Referring to (Negnevitsky, 2005), a hyperbolic tangent or tangent sigmoid can accelerate the learning procedure for the models. This can greatly improve the efficiency of the system and result in a faster processing speed for the prediction system.

In training the data for the ANN, Levenberg-Marquadt (LM) technique is one of the techniques most widely used as the training algorithm, because of its fast performance, however it can impact on memory usage if the network is too complex. There are many types of learning technique as stated in (H. Demuth, 2004). According to (B. Safa, 2004), the best training method in a multilayer network is the ‘steepest descent’ method; however in some models, the variable learning rate algorithm is slower and does not give an accurate results.

The LM algorithm is designed to give intermediate optimization between the Gauss-Newton (GN) method and gradient descent algorithm. It also addresses all of the limitations of both techniques (Kermani et al., 2005, Lourakis, 2005). The use of this technique is usually desirable because of it offers the fastest convergence and also it gives accurate training to the model. It is also able to obtain lower mean square errors for most models, but if the number of weights in the network increases, the accuracy of the model will decrease.

There is one major problem in the ANN architecture, which occurs when the error of the training set becomes a small value in the system - this is called the ‘over-fitting’ or generalization problem. However, when new data is presented to the network, the error will be large in value. In order to prevent this problem, the hidden neuron must be determined, and this is dependent on the number of input parameters in the network. Alternatively, the problem can be prevented by improving the

generalization, by setting the early stopping values in the validation dataset during the training process (H. Demuth, 2004).

### **2.3.1 Purpose of Using ANN**

The decision on which technique will be used will be based on the number of datasets, the accuracy of certain techniques and how the technique relates to each of the input parameter with the target. In this thesis, the ANN is used as the main technique because its reliability and its performance with a number of datasets can be analyzed. The ANN also offers a very good local search method for finding a solution, and can be considered as one of the best prediction models.

The consideration of one technique in isolation cannot, however, result in any conclusions about best overall technique. Therefore the ANN will be compared with other hybrid IS techniques in later a chapter, related to food security modelling.

## **2.4 Genetic Algorithm (GA)**

Genetic Algorithm (GA) is based on natural genetics, which consist of a number of chromosomes in a population of individuals. These genetic structures, called ‘genotypes’ can be used to select the best solution to a problem based on their ‘fitness’ (Michalewicz, 1992, Holland, 1992).

In IS, GAs are used as optimization and search algorithm which use the same concepts as natural genetics, mimicking evolution in nature (Goldberg, 1989, Randy L. Haupt, 2004, Negnevitsky, 2005, Z. Didekova, 2009). The idea in developing GA comes from John Holland; he and his team tried to explain the adaptive process of natural systems, and at the same time developed artificial software that retains the important mechanisms of natural and artificial systems (Goldberg, 1989).

In the GA, there are 3 main processes involved; selection, crossover and mutation. The selection process is used to select which pair of chromosomes contributes the most in term of giving the best results or solutions to the problem. This is based on the determination of the fitness function as in equation (2.3) (Michalewicz, 1992, Negnevitsky, 2005). A commonly used selection function is based on a roulette wheel with slots sized according to the fitness values as in equation (2.4) (Michalewicz, 1992). In (Negnevitsky, 2005) a method is described to find the maximum value of a function  $15x-x^2$ . Figure 2.4 shows the selection process of this function.

$$F = \sum_{i=1}^{pop\ size} eval(v_i) \quad (2.3)$$

$$r < v_i - v_1 \text{ (chromosome 1)} \quad (2.4)$$

*otherwise*

$$v_i (2 \ll i \ll population\ size)$$

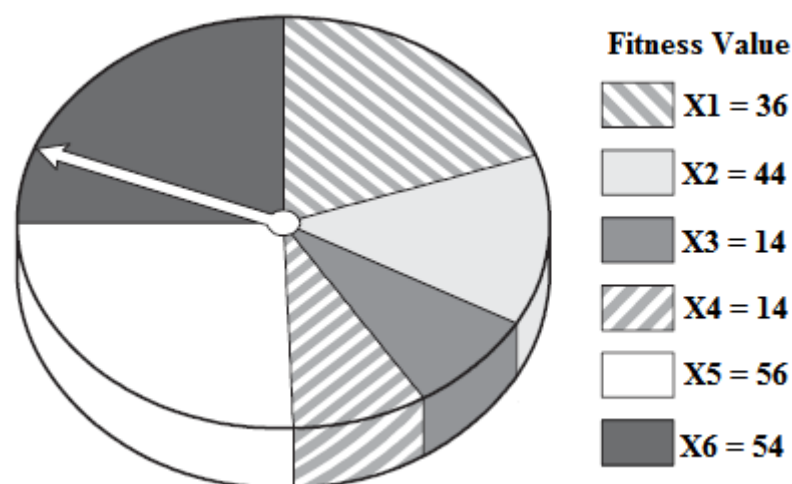


Figure 2.4: Selection process (Negnevitsky, 2005)

In this selection process, the division of the area in the pie chart will depend on the fitness value of each chromosome's population. In other words, if the fitness value is high, the area of the division in the chart will be reflected by this. With the roulette selection process, the wheel will spin randomly six times due to there being six populations, where the greater of the area in the chart, the higher the chance that a particular chromosome's population will be selected.

The crossover process involves the use of the crossover probability functions as shown in Figure 2.5, which shows the same case problem as in Figure 2.4. This function will compare a pair of chromosomes from the selection output to get a new offspring based on the probability parameter. It can be either a 1-point crossover or multi-point crossover such as X6, X2, X1 and X5 in Figure 2.5, depending on the problem being solved. If no crossover occurs, as shown in the population of X2 and X5 cloning will not take place (Negnevitsky, 2005). In the Matlab GA and direct search toolbox, a variety of crossover functions are given that can be used with any model of an application (Mathwork, 2004).

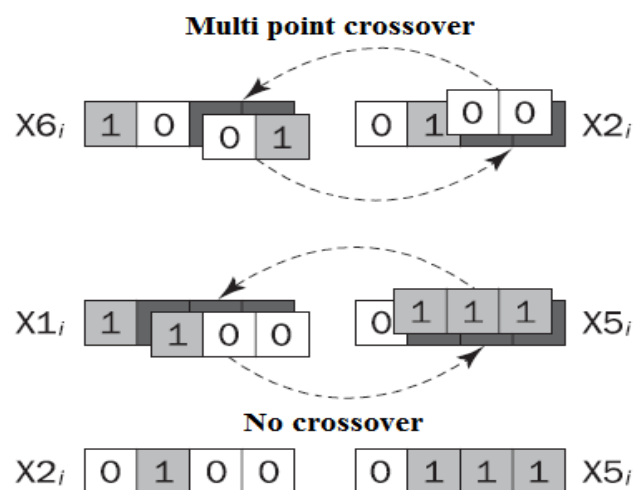


Figure 2.5: Crossover example (Negnevitsky, 2005)

The last stage of process is mutation, which generally comes after the crossover but sometimes before crossover. Mutation is performed on a randomly selected gene in a pair of chromosomes. Usually mutation can be beneficial solution to the problem, but sometimes it can be harmful to the results. However the process is useful in guaranteeing that the search algorithm is not trapped in a local optimum, by keeping the GA from converging too fast before sampling the entire search region (Holland, 1992, Negnevitsky, 2005, Randy L. Haupt, 2004). Therefore the mutation probability is usually given a small value, typically in the range 0.001 to 0.01, but this also depends on the problem being handled (Michalewicz, 1992, Negnevitsky, 2005). An example of mutation is shown in Figure 2.6 which considers the same problem as in the crossover process (Negnevitsky, 2005, Mathwork, 2004). The Figure shows that only X1 and X2 are being mutated at the second gene and the third gene.

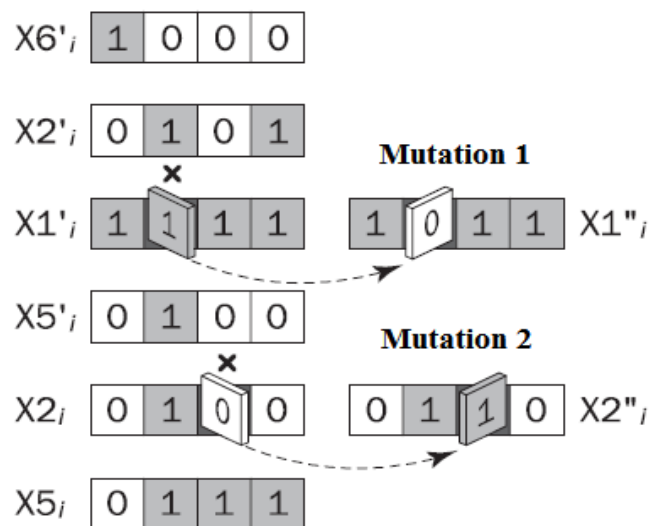


Figure 2.6: Mutation example (Negnevitsky, 2005)

The overall basic GA process flowchart is shown in Figure 2.7. For each main process – selection, crossover or mutation, a new offspring will be given to the output,

and this new offspring usually has the best fitness value for the solution to the problem, for example in the problem shown in Figure 2.8 (Negnevitsky, 2005). This shows the new offspring for the next generation process and the starting point for data before selection process in Figure 2.3. The process will run a few times until it achieves its target or reaches the stopping criterion which has been set. Every time the GA processes start, it will randomly pick the chromosomes and determine their fitness values based on the fitness function. The process will run until the population size and the number of generations ends. The dataset can be encoded either based binary or decimal numbers, whichever is most beneficial to the system.

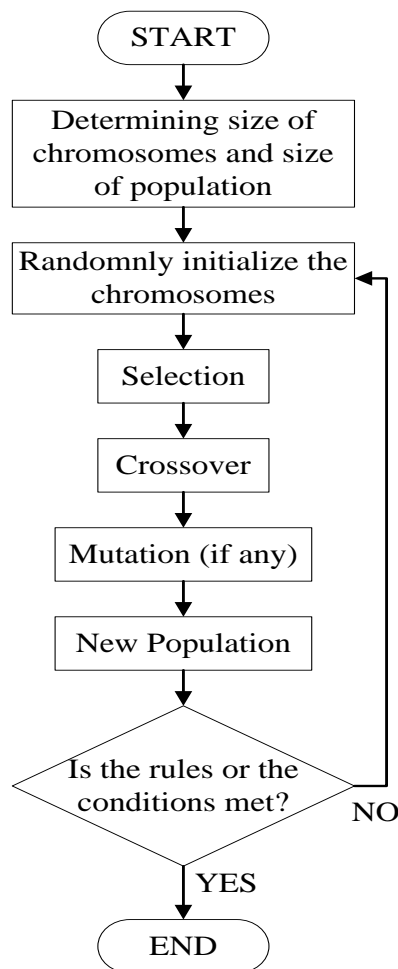


Figure 2.7: Basic GA flowchart


<b>Generation i</b>			<b>Generation (i + 1)</b>		
$X1_i$	1   1   0   0	$f = 36$	$X1_{i+1}$	1   0   0   0	$f = 56$
$X2_i$	0   1   0   0	$f = 44$	$X2_{i+1}$	0   1   0   1	$f = 50$
$X3_i$	0   0   0   1	$f = 14$	$X3_{i+1}$	1   0   1   1	$f = 44$
$X4_i$	1   1   1   0	$f = 14$	$X4_{i+1}$	0   1   0   0	$f = 44$
$X5_i$	0   1   1   1	$f = 56$	$X5_{i+1}$	0   1   1   0	$f = 54$
$X6_i$	1   0   0   1	$f = 54$	$X6_{i+1}$	0   1   1   1	$f = 56$

Figure 2.8: Original data and new offspring data for generation (i+1) (Negnevitsky, 2005)

### 2.4.1 Purpose of Using GA

In terms of optimizing and feature selection, the GA is one of the best optimization and search technique, as it has the advantages of being able to avoid local minima and it also can search all possible solution in all every region simultaneously within the data itself (Muhd Khairulzaman Abdul Kadir, 2012). Therefore the GA is used to optimize the network architecture and to select best feature for a certain application to achieve better prediction, which also can extend the decision making process of the study cases in this thesis.

## 2.5 Hybrid Intelligent Systems (HIS)

Hybrid Intelligent Systems (HIS) are a combination of two or more IS techniques. Most of the combination of IS will improve on the disadvantages of individual systems, or will optimize the process either in terms of faster data



processing or giving higher efficiency in terms of the performance rate of the technique.

For example, assumes that an FL model has more than 6 inputs, and each input has more than 2 MF, therefore it is necessary to generate more than 100 rules. The large number of rules being generated will make the processing system slower and sometimes can cause the system to run out of memory. Therefore, a combination of other IS techniques is needed to improve the processing time of the method, either by using ANN or GA. For this reason many new HIS have been developed and are being applied to a different applications and areas of study. Table 2.1 shows the capability of each IS based on (Z. Didekova, 2009).

Table 2.1: Capability of FL, NN and GA

Capability	FL	ANN	GA
Knowledge representation	Good	Bad	Rather bad
Uncertainty tolerance	Good	Good	Good
Imprecision tolerance	Good	Good	Good
Adaptability	Rather bad	Good	Good
Learning ability	Bad	Good	Good
Explanation ability	Good	Bad	Rather bad
Knowledge discovery and data mining	Rather bad	Good	Rather good
Maintain ability	Rather good	Good	Rather good

### 2.5.1 Adaptive Neuro-Fuzzy Inference System (ANFIS)

The adaptive Neuro-Fuzzy Inference System (ANFIS) is a combination of the high level reasoning capability of the Fuzzy Inference System (FIS) and the low level computational power of a Neural network (NN) (Negnevitsky, 2005, J.-S.R. Jang, 1997). This combined technique has a number of advantages. For example the NN can recognize patterns and help with the adaptation to the environment, whilst the FIS will incorporate human knowledge and perform decision making as described in table 2.1 and the architecture shown in Figure 2.9.

The ANFIS architecture consists of five layers where the first layer is an ‘If...’ layer, also called the fuzzification layer, and functions as an adaptive node to the structure. This layer converts crisp inputs to fuzzy inputs, and it also generates the MF for each of the inputs (Wang and Elhag, 2008, Fahimifard, 2009). As described in the FL section above, the most common MFs being used are the trapezoidal and triangular functions. In ANFIS however, the most common MF is the generalized bell function, which is shown in equation (2.5), where  $x$  or  $y$  is the input node  $i$ ;  $a_i$ ,  $b_i$  and  $c_i$  are the parameter set and  $O$  is the output for the node (Muhd Khairulzaman Abdul Kadir, 2011). The type of MF can also be varied depending on the data being applied to the system.

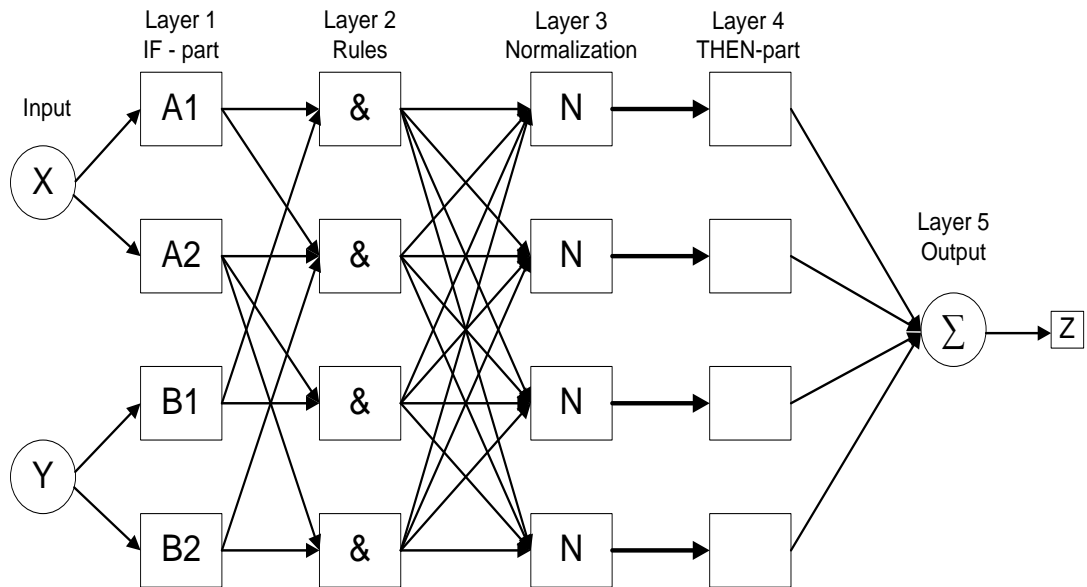


Figure 2.9: Architecture of a 2 input ANFIS. (Muhd Khairulzaman Abdul Kadir, 2011)

$$\mu_A(x) = \frac{1}{1 + \left| \frac{x - c_i}{a_i} \right|^{2b_i}}$$

where

$$O_{1,i} = \mu_{a_i}(x_1), i = 1, 2$$

$$O_{1,i} = \mu_{b_{i-2}}(x_2), i = 3, 4$$

(2.5)

The next layer consists of a simple multiplier to represent the firing strength of each rules that being generated in layer 1. It uses AND as the connective products as shown in equation (2.6) (J.-S.R. Jang, 1997, Negnevitsky, 2005).

$$O_{2,i} = w_i = \mu_{A_i}(x) \cdot \mu_{B_i}(y), i = 1, 2$$

(2.6)

In layer 3, each output of the second layer will be normalized based on equation (2.7) (J.-S.R. Jang, 1997, Negnevitsky, 2005). The normalization is based on the ratio of each of the firing strengths divided by the total rule of the firing strength.

$$O_{3,i} = \overline{w}_i = \frac{w_i}{w_1 + w_2}, i = 1,2 \quad (2.7)$$

In ANFIS, a Takagi-Sugeno type of output is used in layer 4 as shown in equation (2.8). This layer will convert all fuzzy data to crisp data (J.-S.R. Jang, 1997, Negnevitsky, 2005).

$$O_{4,i} = \overline{w}_i f_i = \overline{w}_i (p_i x_1 + q_i x_2 + r_i) \quad (2.8)$$

$[p_i, q_i, r_i] = \text{consequent.parameter}$

In the final layer, the total of all the neurons for each node in layer 4 forms the output of the system.

Although ANFIS has its advantages which come from the combination of FL and ANN, it also has drawbacks. As explained in (Potter and Negnevitsky, 2004), ANFIS will become slow and will use up its available memory if more than 9 inputs are used. This happens because the number of generated rules for the system is more than 2000.

### 2.5.2 Sensitive Genetic Neural Optimization (SGNO)

SGNO is part of the hybrid intelligent optimization technique, the model of which is shown in Figure 2.9. It involves the combination of GA, ANN and sensitivity analysis, which uses sensitivity analysis of high fitness chromosomes and Monte

Carlo simulation. This model is part of supervised learning, and needs a target. It is also relatively good for input or feature selection for certain applications and from there it can also be used as a prediction model.

Referring to Figure 2.10, the process of SGNO is based on few basic steps: -

- 1- The data pre-processing is made based on the normalization of the data with five-fold cross validation.
- 2- The GA will decode each input or feature to its random binary number representation. Each random input will be analyzed by its fitness function using the ANN, and evaluated using the mean square error (MSE) as its fitness function.
- 3- After all of the analysis has been done by the GA, based on the population size and number of generation, all of the data analysis results will be evaluated by the sensitivity module, which will calculate its frequency for a quarter of population from each GA generation and calculate its score for the selected variables.
- 4- Before the ANN is remodelled, the importance of each feature is determined by taking the average of the sensitivity scores in the chromosomes. The higher values will be ranked from most important to least important.
- 5- Finally, the ANN will be remodelled using the multi-layer perceptron (MLP).

The whole process of SGNO was developed by Fu Zhang in 2011 (Zhang, 2011).

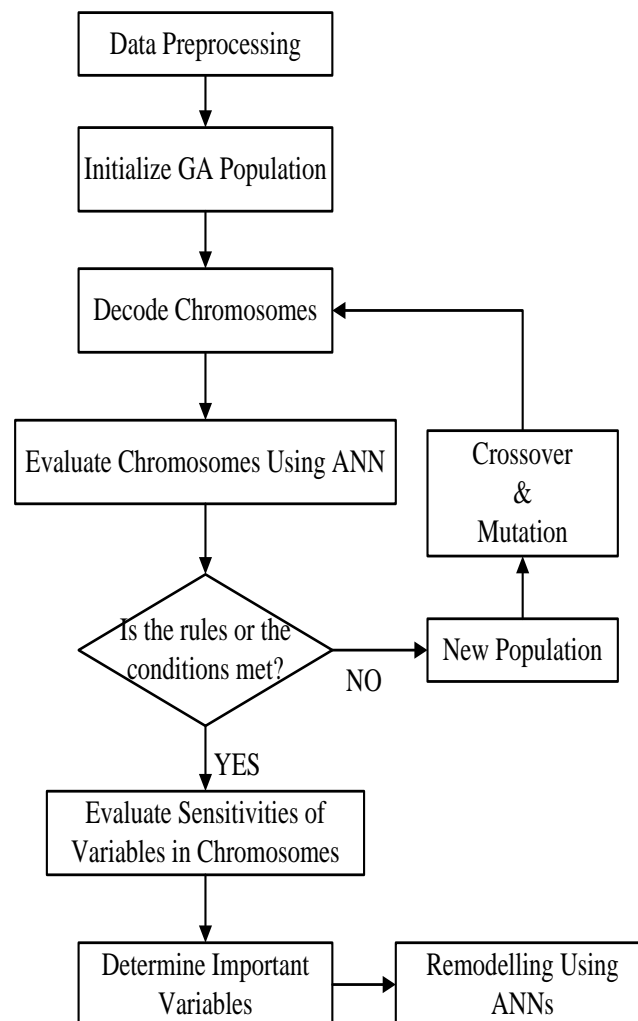


Figure 2.10: Basic SGNO system (Zhang, 2011)

## 2.6 Conclusion

In this chapter, most of the well-known IS techniques that were used in these research are generally explained. These include FL, ANNs, GA, and HIS. In addition a recently developed HIS, the SGNO, was introduced. All of these techniques are powerful tools for different types of application, such as clustering, classification, forecasting or prediction, optimization and feature selection problems. Most of the techniques also have been proven to be successful in different area of study, either in linear or nonlinear applications.

In the following chapter, a new modelling technique will be introduced. The model is for use in general forecasting application or problems. It consists of a two-stage GA-ANN combination, which both methods to generate a model output.

### References

- A.S. SODIYA, S. A. O., AND B. A. OLADUNJOYE 2007. Threat Modeling Using Fuzzy Logic Paradigm. *Informing Science and Information Technology*, 4.
- B. SAFA, A. K., M. TESHNEHLAB, A. LIAGHAT 2004. Artificial neural networks application to predict wheat yield using climatic data. *International Conference*.
- CHEMIN, Y. & HONDA, K. 2006. Spatiotemporal Fusion of Rice Actual Evapotranspiration With Genetic Algorithms and an Agrohydrological Model. *Geoscience and Remote Sensing, IEEE Transactions on*, 44, 3462-3469.
- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. In: DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). Department for Environment, Food and Rural Affairs.
- EDUARD LLOBET, E. L. H., JULIAN W GARDNER AND STEFANO FRANCO 1999. Non-destructive banana ripeness determination using a neural network-based electronic nose *Meas. Sci. Technol.*, 10.
- EL-SEBAKHY, E. A., RAHARJA, I., ADEM, S. & KHAERUZZAMAN, Y. Year. Neuro-Fuzzy Systems Modeling Tools for Bacterial Growth. In: *Computer Systems and Applications*, 2007. AICCSA '07. IEEE/ACS International Conference on, 13-16 May 2007 2007. 374-380.

- FAHIMIFARD, S. M., M. SALARPOUR, M. SABOUHI AND S. SHIRZADY 2009. Application of ANFIS to agricultural economic variables forecasting case study: Poultry retail price. *Journal of Artificial Intelligence*, 2, 65-72.
- FARKAS, I., REMÉNYI, P. & BIRÓ, A. 2000. A neural network topology for modelling grain drying. *Computers and Electronics in Agriculture*, 26, 147-158.
- FOSTER, D., MCCULLAGH, J. & WHITFORT, T. Year. Evolution versus training: an investigation into combining genetic algorithms and neural networks. In: *Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on, 1999 1999. 848-854 vol.3.*
- FU XIAPING, Y. Y., ET AL 2007. Principal Components-Artificial Neural Networks for Predicting SSC and Firmness of Fruits based on Near Infrared Spectroscopy. ASABE Annual meeting Paper.
- GARDNER, J. W., HINES, E. L. & TANG, H. C. 1992. Detection of vapours and odours from a multisensor array using pattern-recognition techniques Part 2. Artificial neural networks. *Sensors and Actuators B: Chemical*, 9, 9-15.
- GEMAN, S. A. B., E. 1992. Neural networks and the bias / variance dilemma. *Neural Computation*, 4, 1-58.
- GOLDBERG, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley Longman, Inc.
- H. DEMUTH, M. B. 2004. *Neural Network Toolbox: For use with Matlab*.
- HOLLAND, J. H. 1992. *Adaptation In Natural And Artificial Systems*, MA, USA, MIT Press Cambridge.
- HUEY-MING, L. 1996. Applying fuzzy set theory to evaluate the rate of aggregative risk in software development. *Fuzzy Sets and Systems*, 79, 323-336.



- J W GARDNER, E. L. H. A. M. W. 1990. Application of artificial neural networks to an electronic olfactory system Meas. Sci. Technol., 1.
- J.-S.R. JANG, C.-T. S., E. MIZUTANI 1997. Neuro-Fuzzy and Soft Computing, Prentice Hall.
- JABBARI, A., JEDERMANN, R. & LANG, W. Year. Neural network based data fusion in food transportation system. In: Information Fusion, 2008 11th International Conference on, June 30 2008-July 3 2008 2008. 1-8.
- JONES, P. Year. Networked RFID for use in the Food Chain. In: Emerging Technologies and Factory Automation, 2006. ETFA '06. IEEE Conference on, 20-22 Sept. 2006 2006. 1119-1124.
- KERMANI, B. G., SCHIFFMAN, S. S. & NAGLE, H. T. 2005. Performance of the Levenberg-Marquardt neural network training method in electronic nose applications. Sensors and Actuators B: Chemical, 110, 13-22.
- LOURAKIS, M. I. A. 2005. A brief description of the Levenberg-Marquardt algorithm. [Online]. Available: <http://www.ics.forth.gr/lourakis/levmar> [Accessed February 11. 2010].
- M. AHMEND, E. D., ET AL. 1999. A general purpose fuzzy engine for crop control. Computational Intelligence, 1625, 473-481.
- MATHWORK. 2004. Genetic Algorithm and Direct Search Toolbox.
- MICHALEWICZ, Z. 1992. Genetic Algorithm + Data Structures = Evolution Programs, Springer-Verlag.
- MUHD KHAIRULZAMAN ABDUL KADIR, E. H., SAHARUL AROF, DACIANA ILIESCU, MARK LEESON, ELIZABETH DOWLER, ROSEMARY COLLIER, RICHARD NAPIER, QADDOUM KEFAYA AND REZA GHAFFARI 2011. Grain Security Risk Level Prediction Using ANFIS. 3rd

---

International Conference on Computational Intelligence, Modelling and Simulation. Langkawi, Malaysia.

MUHD KHAIRULZAMAN ABDUL KADIR, E. L. H., SAHARUL AROF, DACIANA ILIESCU, MARK LEESON, ELIZABETH DOWLER, ROSEMARY COLLIER, RICHARD NAPIER, ARJUNAN SUBRAMANIAN 2012. Neural Network for Farm Household Output Prediction. International Conference on Statistics In Science, Business And Engineering. Langkawi, Malaysia.

NEGNEVITSKY, M. 2005. Artificial intelligence A guide to intelligent systems, Addison-Wesly.

ODETUNJI, O. A. & KEHINDE, O. O. 2005. Computer simulation of fuzzy control system for gari fermentation plant. Journal of Food Engineering, 68, 197-207.

PERROT, N., BONAZZI, C., TRYSTRAM, G. & GUELY, F. Year. Estimation of the food product quality using fuzzy sets. In: Fuzzy Information Processing Society, 1999. NAFIPS. 18th International Conference of the North American, Jul 1999 1999. 487-491.

POTTER, C. & NEGNEVITSKY, M. Year. ANFIS application to competition on artificial time series (CATS). In: Fuzzy Systems, 2004. Proceedings. 2004 IEEE International Conference on, 25-29 July 2004 2004. 469-474 vol.1.

RANDY L. HAUPT, S. E. H. 2004. Practical Genetic Algorithm, John Wiley & Sons, Inc.

ROSS, T. J. (ed.) 2007. Fuzzy Logic With Engineering Applications: John Wiley & Sons, Ltd.

S.A. SHEARER, T. F. B., J.P FULTON, S.F. HIGGINS 2000. Yield Prediction Using A Neural Network Classifier Trained Using Soil Landscape Features and Soil

- Fertility Data. . In: ASAE (ed.) Annual International Meeting. Midwest Express Center, Milwaukee, Wisconsin: ASAE
- TETKO, I. V., LIVINGSTONE, D.J., AND LUIK, A.I. 1995. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. . J. Chem. Info. Comp. Sci., 35, 826-833.
- WANG, Y.-M. & ELHAG, T. M. S. 2008. An adaptive neuro-fuzzy inference system for bridge risk assessment. Expert Systems with Applications, 34, 3099-3106.
- WEIGEND, A. Year. On overfitting and the effective number of hidden units. . In: Proceedings of the 1993 Connectionist Models Summer School, 1994. 335-342.
- XIE, G., XIONG, R. & CHURCH, I. 1998. Comparison of Kinetics, Neural Network and Fuzzy Logic in Modelling Texture Changes of Dry Peas in Long Time Cooking. Lebensmittel-Wissenschaft und-Technologie, 31, 639-647.
- Z. DIDEKOVA, S. K. 2009. Applications of Intelligent Hybrid Systems In Matlab. Mezinárodní konference Technical Computing. Prague.
- ZHANG, F. 2011. Intelligent Feature Selection for Neural Regression Techniques and Application. Doctor of Philosophy, University of Warwick.

# **Chapter 3: Two stage modelling using a Genetic Algorithm-Neural Network (GA-ANN) and an Optimized Artificial Neural Network (ANN)**

## **3.1 Overview**

In the previous chapter, most of the IS techniques currently being used in food security modelling were described and explained. One of the techniques described is the GA method, which is well known as a feature selection and optimization algorithm. In the previous chapter also, the advantages of ANNs in prediction as a fast adaptive modelling tool for complex relationships with non-linear datasets were also explained.

In order to design a good prediction model and at the same time optimize the inputs, GA and ANN are used together as HIS techniques, combining their capability for optimizing, and as a fast prediction tool for complex linear or non-linear datasets. This chapter continues by describing and introducing a two-stage hybrid (TSH) model.

Generally, the TSH model consists of two techniques based on HIS methods. The first stage of GA hybrid model is based on an input selection or feature selection

(FS) technique using a GA and ANN. For the second stage, the GA hybrid model use the Optimized Weight and Threshold Neural Network (OWTNN) described in (Muhd Khairulzaman Abdul Kadir, 2012), where a GA is used to optimize the weights and thresholds of an ANN. Figure 3.1 shows the general structures flow of this model system.

The reason for combining these two HIS into a single two stage process is to provide automatic input selection. The input will be selected based on the importance of the inputs in giving higher impact to the output. After that, with the selected input, the second stage will optimize the ANN architecture, preventing the generalization problem which will give a good performance in prediction. Furthermore, the TSH model can be applied to any application other than the food security, because of the inherent advantages of ANN and GA in handling complex linear and non-linear datasets, as described in the previous chapter.

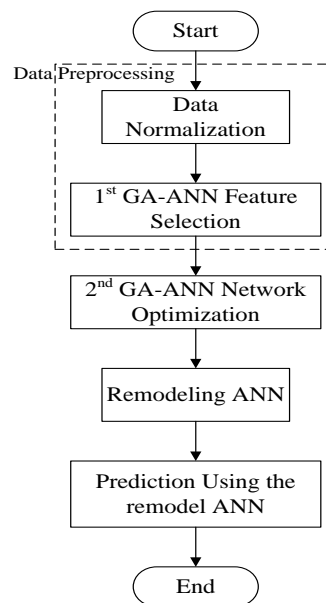


Figure 3.1: General structure of the two stage GA model

### 3.2 Data Pre-processing

As shown in Figure 3.1, the TSH model has two stages of data pre-processing. The first stage is to standardize all of the data into a unified scale to allow faster processing, which in this case transform to have zero mean and unit variance. These datasets will be used in the ANN as the fitness function, or as the objective function in the GA. Usually, in ANNs, datasets are represented as raw data, which will contains some noise in different ranges. Therefore to make the ANN learn faster, as describe earlier, standardization is needed, which will also allows the ANN to give more accurate results (Bishop, 1995).

In applying the standardization, mean standard deviation normalization as shown in equation (3.1), was used, where  $X_{\text{raw}}$  is the raw data that needs to be normalized,  $X_{\text{mean}}$  is the mean of  $X$  and  $X_{\text{std}}$  is the standard deviation of  $X$ . As describe earlier, this will transform the data to have zero mean and unit variance.

$$X_{\text{new}} = \frac{(X_{\text{raw}} - X_{\text{mean}})}{X_{\text{std}}} \quad (3.1)$$

For the second stage of data pre-processing, a feature selection process is performed by using a GA-ANN hybrid algorithm. The GA has good performance in FS or input selection, where it can perform multiple searches simultaneously in any region. In this case, the GA will be used to select which is the most important input for the ANN. This part of data pre-processing will be explained later in this chapter in the first stage of the TSH section of this model.

### **3.3 Determining GA parameters and fitness function**

In the GA, there are a few parameters which are important in making a contribution to find the best solution for each application, such as: crossover, mutation, selection, population and generation. These parameters basically define the stages of the main process of the GA in achieving an effective search through each region consecutively. Each of these parameters has been determined based on its capability and its performance. Not only that, most of these GA parameters will be used by both stages of the GA hybrid model to ensure that the first stage and second stages of the THS model will be optimized improving overall system performance

#### **3.3.1 Selection**

As explained in Chapter 2, after a random generation of initial population data for the chromosome, the GA will select chromosomes to become parents for the following crossover process and mutation process. Generally, one of the earliest types of selection function, and the most well known, is the roulette wheel. In the model developed here however, a stochastic uniform selection function is used instead. This is because the number of selected individuals in the stochastic selection function is proportional to the fitness value, which makes this selection function more accurate in terms of its individual selection performance, compared to the roulette wheel selection function (Elizabeth M. Rudnick, 1997).

The stochastic uniform selection function converts each parent into a section of length of line which is proportional to its scaled value. The parents are then allocated based on a uniform random number less than the step size in the section

(Goldberg, 1997, Randy L. Haupt, 2004, Mathwork, 2004, Chipperfield and Fleming, 1995).

Figure 3.1 demonstrates an example of the stochastic uniform selection process. In this example, it is assumed that a population consists of 5 chromosomes and the GA needs to select 4 chromosomes to become the parents. Here, each of the chromosome fitness scaled values is already being given as shown in the figure, where the highest fitness scaled value will have a higher chance to be selected as the parents. As a result, the 4 individuals being selected as parents are X2, X3 and two times X4 (X4 and X4).

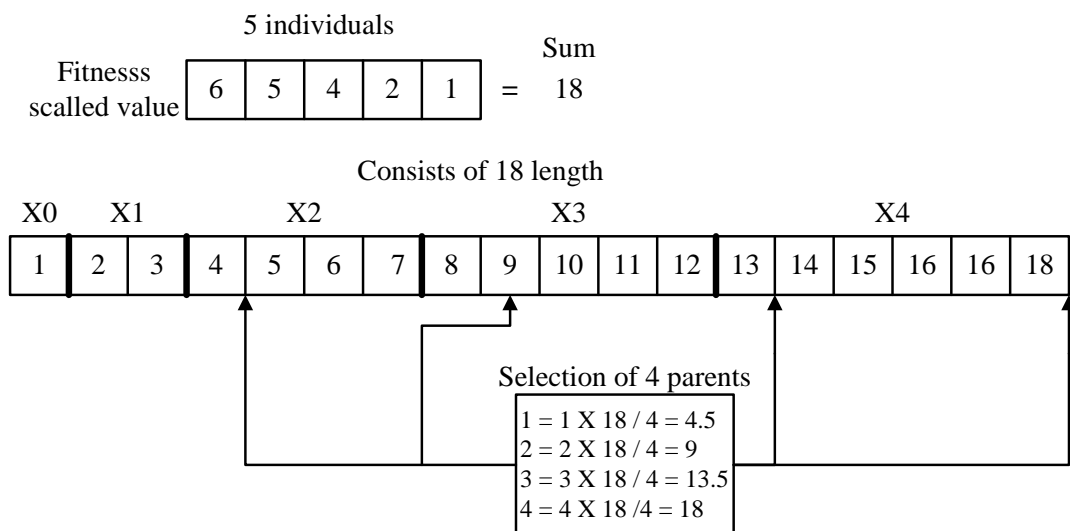


Figure 3.2: Stochastic selection function

### 3.3.2 Crossover

After the operation of the selection function, where the best parent for that cycle of generation and population is selected, a crossover function is used to search the best crossover parents, by referring to the fitness function. The crossover function



selects genes from a parent in the current generation to create a new offspring or a new child, which will be used in the next process.

In this model, to maintain the current best fitness value and to give a better search result, a crossover scattered function is used. This crossover function creates a random crossover point between parents in that generation. The crossover combination is based on a binary vector; if the binary vector is 1, it will take the genes from the first parent and if the vector is 0, it takes the genes from second parent (Michalewicz, 1992, Mathwork, 2004, Popov, 2005). An example of this crossover is shown in Figure 3.3 - the crossover fraction should be between 0 and 1.

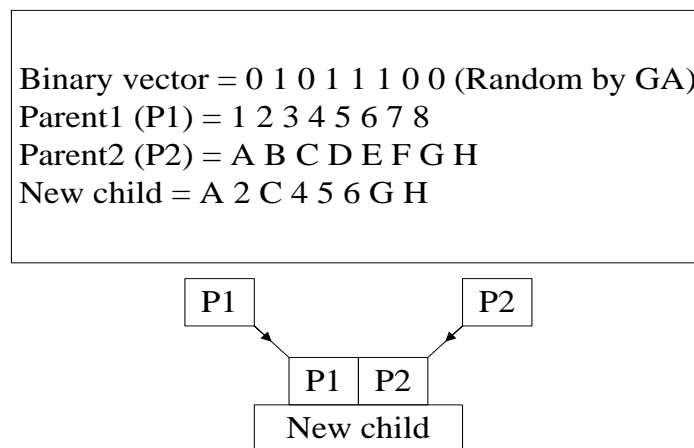


Figure 3.3: Example of scattered crossover (Mathwork, 2004, Michalewicz, 1992, Popov, 2005)

### 3.3.3 Mutation

A mutation process is not a compulsory function in a GA, but is used to implement random changes to a parent in creating a new child for that generation (Mathwork, 2004). The usage of it can be neglected, to give a faster computational result. In this model, the use of mutation is essential to give better search results in

any kind of dataset or application by using it as special solution to any unchanged action.

One of the well known mutations is a uniform mutation which consists of a two-step process, where the first stage selects a fraction of the vector entries of a population (Mathwork, 2004). Each entry has a probability mutation rate which is based on the equation (3.2), where  $L$  represents the number of variables and  $n$  is the population size (Salman and Ong Hang, 2008, Dumitrescu, 2000). In the second stage, the selected entry will be replaced by a random number, selected uniformly in the range of the entry, which in this model is set 0 as the lower range and 1 as the upper range.

$$Pm \approx \frac{1.75}{(\sqrt[2]{L})N} \quad (3.2)$$

### 3.3.4 Population Vs Generation

Many studies have been made to analyze the population size and generation size needed for the GA to work well, and to give the best performance result. Referring to (Gu-Li et al., 2009, Tsoy, 2003), the increase of population size will affect the number of generation, or vice-versa, and it will also improve the performance of the GA, by reducing the genetic drift and allele loss, and by increasing - the parallelism of the algorithm. Therefore, in this model, the determination of population size and the number of generations for the GA is based on these studies. If the population size is big enough for a certain application, the generation number will

be smaller than the population size, as long as it gives good convergence to the model, and it will also save a lot of time in searching for the best solution to the GA.

In the GA process, each time the process starts it randomly initializes the initial population of the chromosomes. The chromosome initialization will be based on the number of variables and the population size, where the population consists of a group of different chromosomes. For example, if there are 5 variables or chromosomes, the population size is 5 and number of generations is also 5, for each generation, the process will randomly assign (5 chromosomes x 5 population size) chromosomes, where each population will consists of 5 variables in each generation, as shown in Figure 3.4, assuming a one generation initialization.

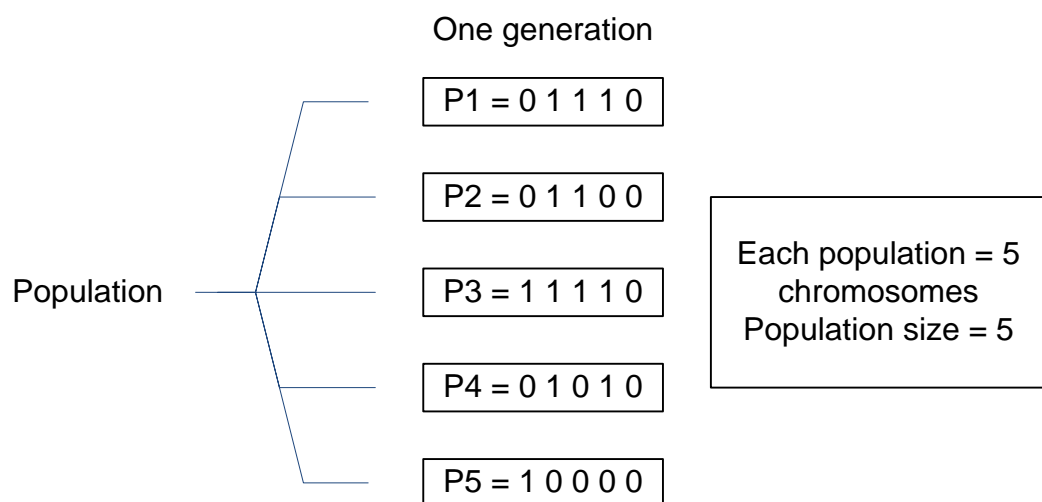


Figure 3.4: Example of population size in one generation

The process determining the population size is usually by trial and error. However, in this model, this seems impractical and will waste a lot of time, because of the complexity of the model. Therefore, for this model, either in the first stage or second stage of the TSH, the population size is calculated based on the number of inputs. In the first stage of GA-ANN, the number of inputs refers to the number of

features available, and for the second stage; the number of inputs refer to the ANN network itself, which is the interconnecting node of the network. Comparing this two-stage GA-ANN, the second stage will consist of larger inputs than the first stage (Gu-Li et al., 2009, Foster et al., 1999). Therefore, the method used in determining the population size is shown in equation (3.3), where  $n$  is the number of inputs of the GA (Gu-Li et al., 2009). Although increasing the population size will make the process slow, it will increase the search performance of the GA, and in addition there is no need to use a larger size of generation number. Therefore in this TSH model, the generation size is fixed to 50 for both stages.

$$n < \text{pop size} < 3n \quad (3.3)$$

### 3.3.5 Fitness function

In achieving the optimum solution for the ANN used in the prediction process, the GA generates random input variables for the GA, either for the feature selection of the first stage TSH model or for the weight and threshold of ANN for the second stage of the TSH model by using the Mean Square Error (MSE) between the actual output and the ANN output as its objective function. This is shown as equation (3.4) where the fitness function of the GA depends on the ANN output ( $NN_{out}$ ), actual output ( $T$ ) and total number of inputs ( $N_{total}$ ) (Muhd Khairulzaman Abdul Kadir, 2012). Equation (3.5) shows the root mean square error (RMSE).

$$MSE = \frac{\sum (NN_{out} - T)^2}{N_{total}} \quad (3.4)$$

$$RMSE = \sqrt{MSE} \quad (3.5)$$

### **3.4 Determining ANN parameter**

In this model, both stages use the MLP-ANN as the fitness function, which increases the probability of prediction as described in Chapter 2. Therefore in addition the need to determine some GA parameters, some of the ANN parameters also need to be adjusted to ensure that the ANN will work well with the GA and give good performance either for optimization or FS. In remodelling the ANN, it will also use the same parameters as in the GA fitness function. The most important parameters that need to be determined in the ANN are the hidden layer, number of hidden neurons, activation function and learning function.

#### **3.4.1 Hidden layer and number of hidden neuron**

Generally, ANN without any hidden layers will only represent a linear function. However, adding a single hidden layer will allow continuous mapping from one region of space to another and the extra hidden layer will also help to shape any discontinuities (Zhang, 2011). In addition, a number of research indicate that just one hidden layer can solve any problem, however, adding an additional hidden layer will increase the processing time (Foster et al., 1999). Therefore in the model described here, the ANN will be based on a single hidden layer, thereby keeping the model simple and giving a fast processing speed, and preventing the development of a complex, time consuming process that may consume a lot of memory and increase computational cost.

In deciding on the number of hidden neurons, there are many method which can be used to determine this, as describe by (Weigend, 1994, Geman, 1992, Tetko,

1995). Based on Weigend, Geman and Tetko, the number of hidden neurons can be determined either by ‘trial and error’ or by ‘rules of thumb’. However, in deciding on the number of hidden neurons for the ANN as the GA fitness function, it is impractical to use trial and error as this would increase the complexity of the model drastically. Therefore, the ‘rule of thumb’ approach is used as guidance in deciding on the number of hidden neurons.

There are a few ‘rules of thumb’ methods which have been developed, such as:-

1 – Hidden neuron size is somewhere between the input layer size and out layer size (Blum, 1992)

2 – Hidden neuron should be twice as large as the number of inputs (Berry, 1997)

3 – The size of hidden neurons should be as large as the dimensions needed to capture 70% - 90% of the variance of the input data sets (Boger and Guterman, 1997)

All of these rules of thumb are general and straight-forward. However, one rule of thumb exists which basically consists of the combination of the above rules. It takes two quarters of the total number of inputs and outputs of the ANN network as shown in equation (3.6) where H is the number of hidden neurons,  $D_{in}$  is the number of inputs and T is the number of outputs (Sarle, 2001).

$$H = \frac{2(D_{in} + T)}{3} \quad (3.6)$$

Although this rule of thumb method simply ignores any changes of the MLP-ANN architecture, such as the number of iterations (training) for all network

parameters, in this model, at the second stage of TSH, the optimization will reflect directly on the network of ANN itself. This optimization demonstrates the ability of the model to avoid any generalization problems, such as over-fitting or under-fitting. Therefore, it will automatically generate the best ANN network, based on the hidden neurons that have been decided by equation (3.6), and is therefore used in this model.

### **3.4.2 Activation function**

As discussed in Chapter 2, there are 4-four commonly used activation functions: step function, sign function, hyperbolic function and linear function (Negnevitsky, 2005). To provide a smooth and nonzero derivative with respect to the input signals, the sigmoidal function, also known as the hyperbolic function, was used in the majority of back-propagation multilayer networks (J.-S.R. Jang, 1997, Gardner et al., 1992, Gardner and et al., 1990). The model developed in this thesis, the back-propagation MLP, a tangent sigmoid activation function is used to improved the performance of the learning process (Negnevitsky, 2005, J.-S.R. Jang, 1997). This was selected for use at the hidden layer and output layers because of its range which is wider than the sigmoidal function, as shown in Figure 3.5.

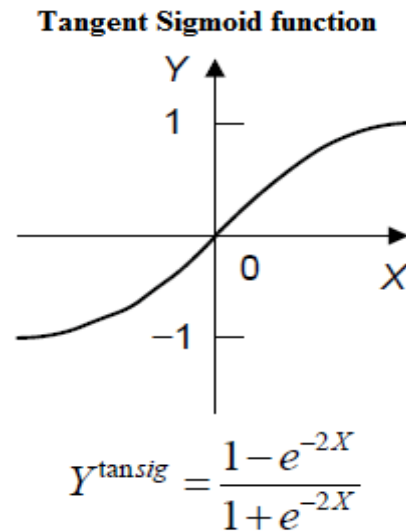


Figure 3.5: Tangent Sigmoid activation function (J.-S.R. Jang, 1997)

### 3.4.3 Learning function

The Levenberg-Marquadt (LM) technique when applied to an ANN, can produce optimum convergence, and can also give better results when compared to other ANN learning techniques such as the Gauss-Newton method and the gradient descent algorithm (Kermani et al., 2005, Lourakis, 2005). This is one of the reasons that the LM technique is selected to be used in this model. If the problem being addressed involves a network which is complex, this learning method can consume a lot of memory (H. Demuth, 2004). However, in terms of the application to this developed model, this will not be an issue, due to the optimization of the network itself by the GA in the TSH model.



### **3.5 First stage process**

The purpose of FS or input selection is to reduce the amount of data being used which will simplify the overall model and process. In an ANN, reducing the number of parameters accordingly will not affect the information or features of the model, which can give better results and good generalization to the ANN architecture. The main steps of input selection are: variable encoding, population evaluation and variable selection.

#### **3.5.1 Variable Encoding**

At the beginning of this stage, variable encoding is initialized by each chromosome, where each input will be represented by several binary numbers known as chromosomes. The bit size of the binary numbers (chromosomes) is dependent on the input size of a particular case or application. In binary number format, the number includes 0's and 1's only, which will be generated randomly by the GA, and represent the chromosomes. If the binary number '0' exists in the input variables as its representation, it means the input will be neglected, but the binary number '1' representation on the input means the variables has been selected. The process of input selection runs continuously until the GA termination criteria are met. This process is also performed using MLP-ANN as its fitness function, in evaluating each of the inputs based on the MSE of the ANN. The lowest value of the MSE in the process will determine the selection of the inputs.

### 3.5.2 Population evaluation

The process for determining the population size is usually based on experiments performed for a certain application, in other words ‘trial and errors’. As described in the previous section, population size for the developed model is instead based on the theorem (3.3). At this stage, in order to ensure that all of the inputs area trained, tested and optimized perfectly by the GA, the population size will be multiplied by 3, suggested by (Gu-Li et al., 2009). For example, if there are 5 chromosomes in binary number format, this will give  $2^5 = 32$  possible states, which means that the population size will be  $5 \times 3 = 15$ , giving  $2^{11} = 2048$  possible states to be trained by the GA and ANN as the fitness function. Different datasets will have different population sizes. The larger the feature size or number of input variables, the larger population size will be. The increased size of the chromosomes will also increase the number of possible states, and the population size that will make the GA-ANN process more complex, and therefore will usually produce a better result, especially as the GA is capable of searching multiple regions simultaneously. However the more complex process will increase computation time.

For the population evaluation, the chromosomes will be evaluated by the objective function of the GA. The performance of the chromosomes is determined by the fitness value of the fitness function, which in this case is the MLP-ANN. The fitness values are evaluated using the error produced by the MLP-ANN.

The process of population evaluation will always be iterative, based on the number of generations of the GA and the number of epochs in the MLP-ANN. This two-fold process produces better results in terms of input or feature selection. The final result for the population evaluation is determined by the lowest MSE value, and

this will be the input or feature that had been selected based on the training in this stage. The second stage of the TSH will process the selected input by building a better MLP-ANN network before further prediction. The overall process flow for this first stage GA-ANN hybrid is shown in Figure 3.6.

### **3.6 Second stage process**

As described in chapter 2, one of the key capabilities of the GA is in optimization. The optimization of the GA can be used to optimize any kind of problems, such as the optimization of the network architecture; or optimization of a dataset for better performance. Furthermore, due to the operation of the GA, which is based on genetics and natural selection, it also offers an effective search technique which gives better performance in optimization (Muhd Khairulzaman Abdul Kadir, 2012). Therefore, because of this capability, an optimization of the ANN architecture has been applied. Here the population evaluation will be the same as in previous stages, except that the population size will not be the same as the input variables for this stage, which will be explained in section 3.6.1.

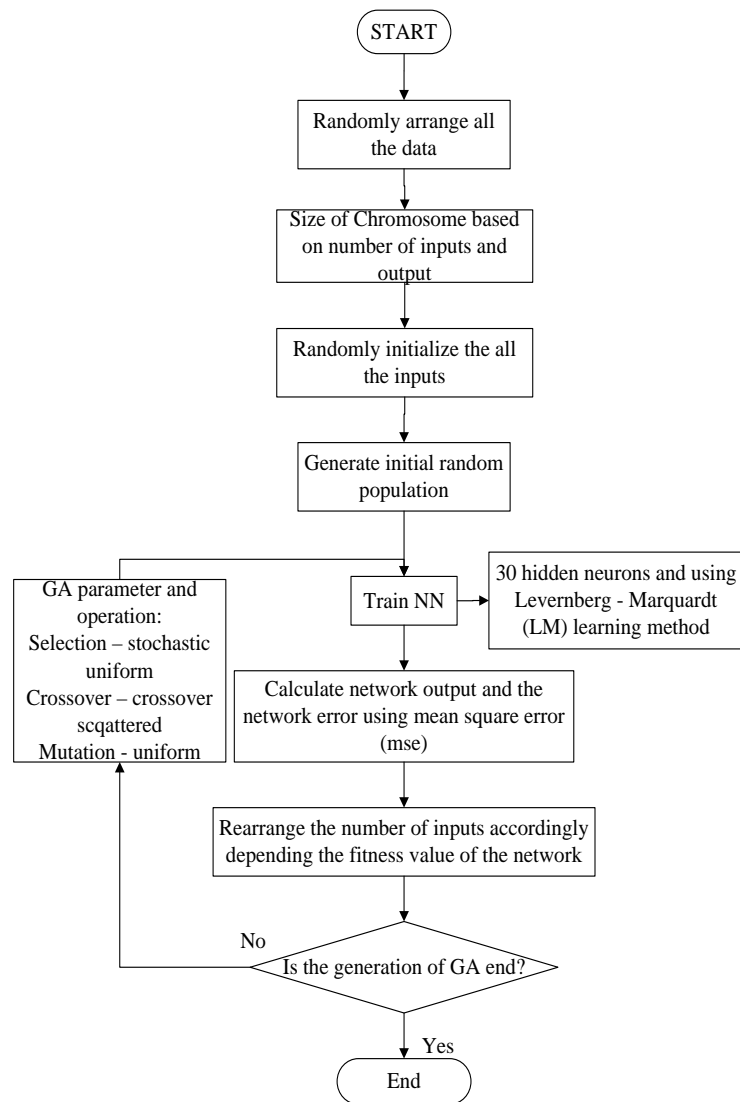


Figure 3.6: First stage process flow

The main objective of this stage is to optimize the weight and threshold of the ANN itself. After selecting the best inputs for the ANN in the first stage of the TSH model, the GA will be used again in deciding the best weight and threshold values for the ANN; this is to ensure that no generalization problem occurs. The overall process flow for this second stage is shown in Figure 3.7.

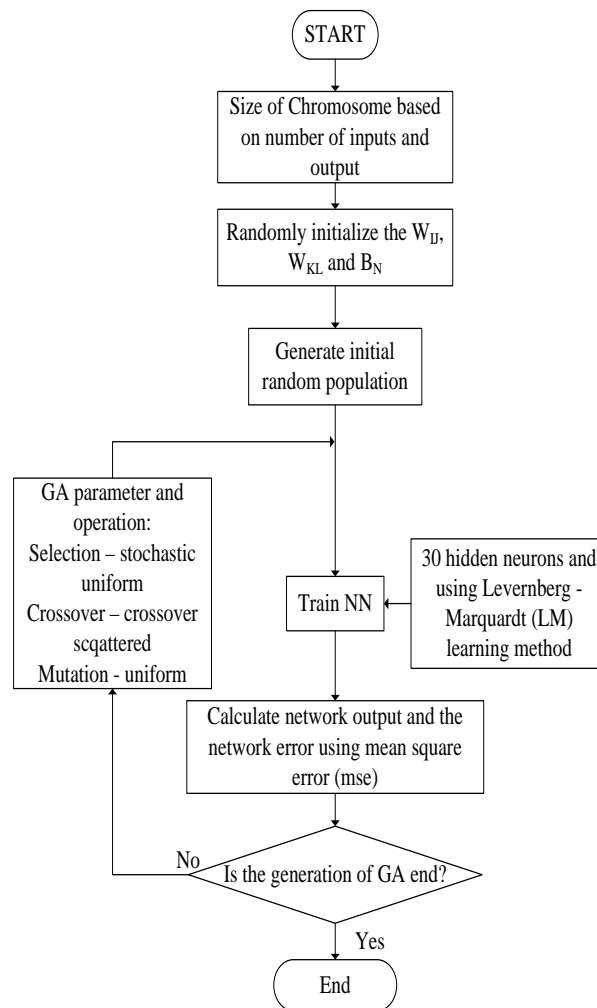


Figure 3.7: Second stage process flow

Generally, an ANN will provide random initialization of the weights and thresholds of the network, and due to this feature; it produces local extremes and sometimes makes the convergence slow. In addition, when the ANN cannot generalize, the dataset involved in the model can be easily subject to over-fitting or under-fitting especially when it is divided into training, validation and testing sections (Muhd Khairulzaman Abdul Kadir, 2012, Hajir Karimi, 2011, Rajendra et al., 2009, Zhang and Wang, 2008, T. Hasangholi, 2010). In (Hajir Karimi, 2011), the optimization of weights and thresholds also indicates that this gives the ANN benefits

by minimizing the mean square error (MSE) between the output of the network and the actual output.

### 3.6.1 Variable presentation

In the second stage of the TSH model, the variables use decimal representation, which is different from that use in the first stage model. Another difference is that the chromosomes are based on the total number of inputs, the total number of layers of the ANN and the total number of hidden neurons being used. The total number of chromosomes is shown in (3.7) (Muhd Khairulzaman Abdul Kadir, 2012).

$$\text{Number of chromosomes} = W_{IJ} + W_{JK} + W_{KL} + B_I + B_K + B_L \quad (3.7)$$

Where,  $W_{IJ}$  = weight between input layer and hidden layer,  $W_{JK}$  = weight between hidden layer and another hidden layer,  $W_{KL}$  = weight between hidden layer and output layer,  $B_I$  = input layer threshold,  $B_K$  = hidden layer threshold,  $B_L$  = output layer threshold. All of these interconnections are shown in Figure 3.8

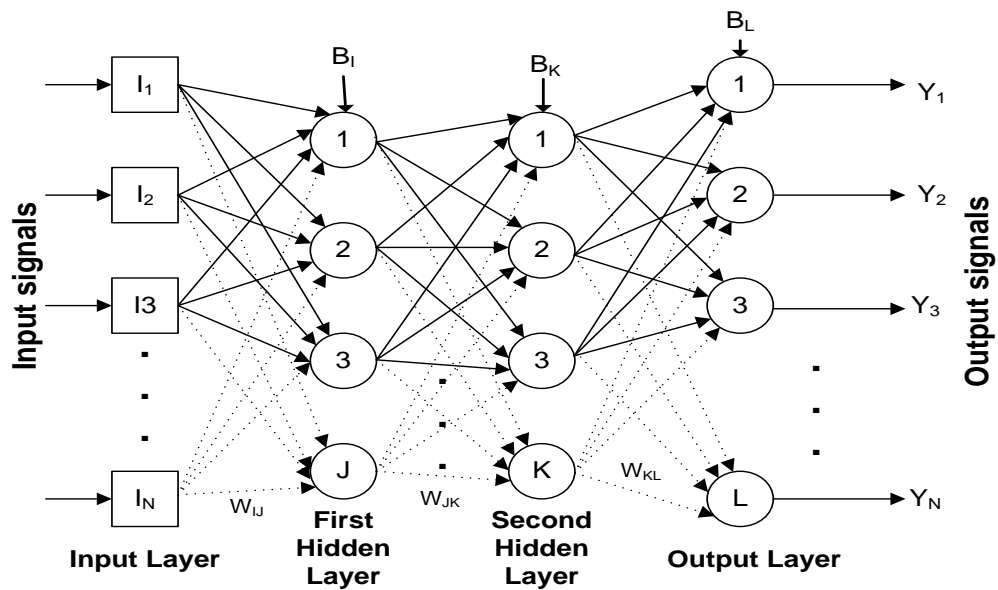


Figure 3.8: Interconnection of weights and thresholds for ANN (Muhd Khairulzaman Abdul Kadir, 2012)

For this model, as for the previous stage, three layers are used: one input layer, one hidden layer and one output layer. This is to ensure that the GA will optimize the same types of network as previous stages, giving better accuracy in terms of the model performance. Therefore, for equation (3.7),  $W_{JK}$  can be removed from the total number of chromosome in the GA. The number of hidden neurons will also be the same as for the previous stage model, as describe in the previous section in equation (3.6).

The ANN has a very complex architecture, especially when it involves multiple layers and multiple variables, inputs and outputs. The techniques performed in this stage have the same GA fitness function as in the first stage – the ANN–MLP architecture, and use the same GA parameters for optimizing the input variables. The

reason for maintaining the same parameters is to ensure that the performance of the first stage will be maintained in the next stage process.

### **3.7 Remodelling ANN**

In remodelling the MLP-ANN, all of the parameters of the second stage of the TSH model is maintained ensuring that the performance optimized network made by the GA is also maintained. This remodelling will affect the entire MLP-ANN architecture in terms of its accuracy, and give faster performance. When compared with the original MLP-ANN, it always gives random results and random performance because of the random weights and thresholds initialization by the MLP network itself, which produce under-fitting and over-fitting problems in the training, testing and validation process.

### **3.8 Benchmark Techniques**

The model being developed is mainly intended for use with a food security prediction model; it is specifically intended for predicting the indicators defined in (DEFRA, 2010). However, other applications which involve complex linear or non-linear datasets can also use the TSH model.

In comparing and assessing the performance of this model, a benchmarking technique needs to be selected which includes Principal Component Analysis (PCA) as one of the conventional statistical methods, the traditional MLP-ANN, FS (GA-ANN), Optimized weight and threshold of neural network (OWTNN) and Sensitive Genetic Neural Optimization (SGNO).



### 3.8.1 Principal Component Analysis (PCA)

PCA is a widely used method to identify patterns in a dataset, allowing identification of the differences and similarities in other words the variations within the data (Smith, 2002). This pattern recognition by PCA gives the advantage of allowing it can be used to reduce the dimensionality of the datasets without significant information loss (Smith, 2002, Kadir et al., 2011, Jang, 1996). The reason that this technique is selected as one of the benchmarking techniques is because the capability of the input selection is same as in first stage process of the proposed model.

According to (B. Balasko, 2004, Smith, 2002), PCA consists of multiple mathematical theorems, which show the variations in the data. These representations are based on the correlated variables, known as ‘principal components’. The main objectives of PCA stated by (B. Balasko, 2004) and (Smith, 2002) are to indentify the new meaningful variables and at the same time either discover the dimension of the datasets or reduce the datasets which are considered as non-important for optimum usage, especially in high-dimensional datasets.

In understanding and reducing the feature of a certain dataset, there are five general steps which need to be performed as below:-

1 – All of the data needs to be averaged and subtracted across each of the dimension as in equation (3.8). In conjunction with finding data variations, the standard deviation needs to be calculated. The reason that the denominator is selected as ‘ $n - 1$ ’ rather than ‘ $n$ ’ is because the sample of the data had been converted to a standard deviation which shows the subset of the sample data used and not the entire population (Smith, 2002).

$$\bar{X}_{subtract} = \frac{\sum_{i=1}^n (X_i - \bar{X})}{n-1}$$

where,

(3.8)

$\bar{X}_{subtract}$  = mean subtraction of the input set X

$n$  = number of row or data

$X_i$  = X number of 'i' in the sequence

2 – The variance for each averaged dataset and the covariance matrixes need to be determined as in equation (3.9), where X and Y refer to a dataset of dimensions X and Y (Smith, 2002).

$$A = cov(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
(3.9)

3 – The key process in understanding the data pattern will be based on the eigen-analysis, through the determination of the eigenvalues and eigenvectors as shown in equation (3.10), where A is the covariance matrix,  $\lambda$  is the eigenvalues, I is identity matrix and p is the eigenvectors matrix (Smith, 2002, B. Balasko, 2004).

$$Ap = \lambda Ip$$
(3.10)

4 – The eigenvalues are used to determine the number of components that can be used or reduced. The principle components of the dataset usually come from the highest eigenvalue, which is the value pointing to the centre of the dataset. The next step is to find which component will be left out before finding the final value of the dataset.

5 – Finally, a new data component is determined based on equation (3.11) where it will be used as a new data in the ANN as the prediction model (Smith, 2002). The

Row Feature Vector (RFV) is the matrix of the transpose column of the eigenvectors selected and the Row Data Adjust (RDA) is based on the mean subtraction result as in step 1.

$$\text{Final Data} = \text{RFV} \times \text{RDA} \quad (3.11)$$

### **3.8.2 Artificial Neural Network (MLP-ANN)**

As described in Chapter 2 and in the previous section, a basic ANN consists of 3 layers; the input, hidden and output layers. In the original MLP-ANN method, a LM learning technique, tangent sigmoid activation function and all of the parameter are the same as in the proposed model. The dataset used is from the standardized dataset for each of the application without any data pre-processing.

### **3.8.3 Feature Selection (GA-ANN)**

This is the first stage process of a hybrid GA-ANN in the proposed model. In comparing the capability of the TSH model, performance comparisons need to be made in terms of input or feature selection by using the GA. After that, the feature being selected by the GA is used with the ANN for comparing the prediction performance.

#### **3.8.4 Optimized Weight and Threshold Neural Network (OWTNN)**

Optimized Weight and Threshold Neural Network (OWTNN) is able to ensure that the ANN performs optimally without any generalization problems, especially over-fitting or under-fitting. This technique also ensures that the ANN is not getting any random results, and it also gives faster performance than the original ANN. As in the previous benchmark, the performance is based on the ANN performance as a prediction model.

#### **3.8.5 Sensitive Genetic Neural Optimization (SGNO)**

SGNO is one of the ranking method for feature selection which was developed by Fu Zhang (Zhang, 2011). This technique is developed by using a GA, ANN and sensitivity analysis which measures the frequency of the inputs being selected by GA, and ranks each of the inputs accordingly, by using the sensitivity analysis. The higher frequency of appearances for the inputs is considered preferable. The details of the implementation of SGNO are discussed in Chapter 2.

#### **3.8.6 Benchmark Performance**

Prediction of the output models basically is defined as making assumption on the future action or results based on the actual output of the dataset compare with the output of either the TSH or benchmark models (Pin Chang Chen, 2011, C. Jareanpon, 2004). In validating the prediction performance of the TSH and the other benchmark techniques, two performance plots have been used; the regression plot and the MSE plot.

The regression plot basically shows an indication of the relationship between outputs of the model and the targets of the ANN; this plot shows the fitting of the ANN where if  $R = 1$  indicated as a good fit, whereas if  $R = 0$  this shows a non fit network with  $R$  as the regression value (H. Demuth, 2004). The MSE plot shows the errors for the ANN model over an entire iteration, where the lower MSE values indicate that the target (actual output) and the ANN output are almost the same. These performance outcomes of the model shows that the model tested can be used as a guideline for prediction.

### **3.9 Complexity of TSH**

As described in the previous section of this chapter, the TSH consists of a two-stage GA-ANN process. Each stage's processes are combined with the GA population size, GA generations, ANN training cycles and ANN testing errors, contributing to the TSH complexities required in order to reach a solutions. However, if compared to each benchmark techniques and the TSH itself, the performance evaluation is based on the ANN where the number of ANN training processes can be used to compare the complexities.

For TSH, PCA, FS (GA-ANN), OWTNN and SGNO, the ANN component and architecture used in performance evaluation are similar to each other, with the ANN having the same three layers; one input layer, one hidden layer and one output layer. The number of hidden neurons for each of them is estimated with two third of the total inputs and output for each benchmark and TSH. However, in SGNO, the determination of hidden neurons is performed by halving the numbers of inputs and output.

Referring to each of the structures for each technique, the TSH, FS (GA-ANN), OWTNN and SGNO have the same flow process in terms of GA and ANN. However, in TSH, the complexity is twice that of the processes in the other techniques, due to the two-stage GA-ANN. The GA-ANN complexities depend on the number of generations and the population size of the ANN training iterations. PCA and ANN processes, only involve the ANN iterations; in general, PCA and standalone ANN are less complex compared to the other benchmark techniques. It follows that FS (GA-ANN), OWTNN, SGNO and TSH where the TSH having more complexity and GA-ANN (FS) less complexity.

### **3.10 Potential of this modelling in food security area of study and other study**

As explained in the previous section, this model can be used to predict most of the indicators of food security, both the main indicators and sub-indicators. The proposed model is capable of handling various sizes of dataset dimensions, and is mainly used as a prediction model.

In the next three chapters, the proposed model will be applied in various areas of food security related studies. From these multiple applications, the prediction results of this model can be used as one of the reference points for making decisions related to food security and other research related to prediction models.

### **3.11 Conclusion**

In this chapter, the TSH model is introduced and the general procedures are explained in details. The model consists of two main modules; GA modules and ANN

modules. The modules are combined to become a hybrid model, which is then used twice for different purposes; input selection of the dataset and optimization of thresholds and weights of the ANN.

The GA module works as a main controller for the overall process of the hybrid system, and at the same time employs the ANN module as the fitness function in evaluating the performance of the chromosomes from each generation, which will show the potential solutions of the model. All of the parameters used in the GA module and ANN module will always be the same for each stage, ensuring a good final prediction result. At the end of the TSH process, the MLP-ANN uses all of the inputs selected in the first stage and also uses the optimized thresholds and weights that are optimized in the second stage process, to generate predictions for a particular application.

In order to ensure the performance of the model being developed, benchmarking techniques that include each of the stage of the model are used. The techniques used are: the PCA, original MLP-ANN, stand alone FS (GA-ANN), stand alone OWTNN and SGNO. Each of the result of these techniques then be used by the ANN for comparing the prediction performance.

## **References**

- B. BALASKO, J. A., B. FEIL 2004. Fuzzy clustering and data analysis toolbox: For use with Matlab.
- BERRY, M. J. A., AND LINOFF, G 1997. *Data Mining Techiques*, New York, John Wiley & Sons.

- BISHOP, C. M. 1995. *Neural networks for pattern recognition*, Oxford: Clarendon Press.
- BLUM, A. 1992. *Neural Network in C++: An Object-Oriented Framework for Building Connectionist Systems*, New York, John Wiley & Sons.
- BOGER, Z. & GUTERMAN, H. Year. Knowledge extraction from artificial neural network models. *In: Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation.*, 1997 IEEE International Conference on, 12-15 Oct 1997 1997. 3030-3035 vol.4.
- C. JAREANPON, W. P., R. J. FRANK, N. DAVEY 2004. An Adaptive RBF Network Optimised Using a Genetic Algorithm Applied to Rainfall Forecasting. *International Symposium on Communications and Information Technologies*. Sapporo, Japan.
- CHIPPERFIELD, A. J. & FLEMING, P. J. Year. The MATLAB genetic algorithm toolbox. *In: Applied Control Techniques Using MATLAB*, IEE Colloquium on, 26 Jan 1995 1995. 10/1-10/4.
- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In: DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). DEFRA.*
- DUMITRESCU, D., LAZZERINI, B., JAIN, L.C., DUMITRESCU, A. 2000. *Evolutionary Computation*, The CRC Press International Series on Computational Intelligence.
- ELIZABETH M. RUDNICK, G. S. G., JANAK H. PATEL, THOMAS N. NIERMAN 1997. A genetic Algorithm Framework for Test Generation. *IEEE Transaction on Computer Aided Design of Intergrated Circuits and Systems*, 16, 1034 - 1044.



- FOSTER, D., MCCULLAGH, J. & WHITFORT, T. Year. Evolution versus training: an investigation into combining genetic algorithms and neural networks. *In: Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on, 1999 1999. 848-854 vol.3.*
- GARDNER, J. W. & ET AL. 1990. Application of artificial neural networks to an electronic olfactory system. *Measurement Science and Technology, 1, 446.*
- GARDNER, J. W., HINES, E. L. & TANG, H. C. 1992. Detection of vapours and odours from a multisensor array using pattern-recognition techniques Part 2. Artificial neural networks. *Sensors and Actuators B: Chemical, 9, 9-15.*
- GEMAN, S. A. B., E. 1992. Neural networks and the bias / variance dilemma. *Neural Computation, 4, 1-58.*
- GOLDBERG, D. E. 1997. *Genetic Algorithm in Search, Optimization, and Machine Learning*, Addison-Wesley.
- GU-LI, Z., XIAO-XIA, L. & TONG, Z. Year. The impact of population size on the performance of GA. *In: Machine Learning and Cybernetics, 2009 International Conference on, 12-15 July 2009 2009. 1866-1870.*
- H. DEMUTH, M. B. 2004. *Neural Network Toolbox: For use with Matlab, USA: Mathworks.*
- HAJIR KARIMI, F. Y., MAHMOOD REZA RAHIMI 2011. Correlation of viscosity in Nanofluids using Genetic Algorithm - Neural Network (GA-NN). *World Academy of Science, Engineering and Technology, 73, 531-538.*
- J.-S.R. JANG, C.-T. S., E. MIZUTANI 1997. *Neuro-Fuzzy and Soft Computing*, Prentice Hall.

- JANG, J. S. R. Year. Input selection for ANFIS learning. *In: Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on, 8-11 Sep 1996* 1996. 1493-1499 vol.2.
- KADIR, M. K. A., HINES, E. L., AROF, S., ILLIESCU, D., LEESON, M., DOWLER, E., COLLIER, R., NAPIER, R., KEFAYA, Q. & GHAFARI, R. Year. Grain Security Risk Level Prediction Using ANFIS. *In: Computational Intelligence, Modelling and Simulation (CIMSIM), 2011 Third International Conference on, 20-22 Sept. 2011* 2011. 103-107.
- KERMANI, B. G., SCHIFFMAN, S. S. & NAGLE, H. T. 2005. Performance of the Levenberg-Marquardt neural network training method in electronic nose applications. *Sensors and Actuators B: Chemical*, 110, 13-22.
- LOURAKIS, M. I. A. 2005. *A brief description of the Levenberg-Marquardt algorithm*. [Online]. Available: <http://www.ics.forth.gr/lourakis/levmar> [Accessed February 11. 2010].
- MATHWORK. 2004. Genetic Algorithm and Direct Search Toolbox.
- MICHALEWICZ, Z. 1992. *Genetic Algorithm + Data Structures = Evolution Programs*, Springer-Verlag.
- MUHD KHAIRULZAMAN ABDUL KADIR, E. L. H., SAHARUL AROF, DACIANA ILIESCU, MARK LEESON, ELIZABETH DOWLER, ROSEMARY COLLIER, RICHARD NAPIER, ARJUNAN SUBRAMANIAN 2012. Neural Network for Farm Household Output Prediction. *International Conference on Statistics In Science, Business And Engineering*. Langkawi, Malaysia.
- NEGNEVITSKY, M. 2005. *Artificial intelligence A guide to intelligent systems*, Addison-Wesly.

- PIN CHANG CHEN, C. Y. L., HUNG TENG CHANG, YU LOCHO 2011. A study of applying Artificial Neural Network and Genetic Algorithm in Sales Forecasting Model. *Journal of Convergence Information Technology*, 6, 352-362.
- POPOV, A. 2005. Genetic Algorithms For Optimization.
- RAJENDRA, M., JENA, P. C. & RAHEMAN, H. 2009. Prediction of optimized pretreatment process parameters for biodiesel production using ANN and GA. *Fuel*, 88, 868-875.
- RANDY L. HAUPT, S. E. H. 2004. *Practical Genetic Algorithm*, USA, John Wiley & Sons, Inc.
- SALMAN, Y. & ONG HANG, S. Year. The effect of GA parameters on the performance of GA-based QoS routing algorithm. *In: Information Technology, 2008. ITSIM 2008. International Symposium on*, 26-28 Aug. 2008 2008. 1-7.
- SARLE, W. S. 2001. *Generalization* [Online]. NC: comp.ai.neural. Available: <http://www.faqs.org/faqs/ai-faq/neural-nets/part3/index.html> [Accessed 6/3/12 2013].
- SMITH, L. I. 2002. A tutorial on Principal Component Analysis. Available: [http://www.cs.otago.ac.nz/cosc453/student\\_tutorials/principal\\_components.pdf](http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf) [Accessed 26 February 2002].
- T. HASANGHOLI, F. K. 2010. A Novel Optimized Neural Network Model for Cost Estimation using Genetic Algorithm *Journal of Applied Sciences*, 10, 512-516.

- TETKO, I. V., LIVINGSTONE, D.J., AND LUIK, A.I. 1995. Neural Network Studies. 1. Comparison of Overfitting and Overtraining. . *J. Chem. Info. Comp. Sci.*, 35, 826-833.
- TSOY, Y. R. Year. The influence of population size and search time limit on genetic algorithm. *In: Science and Technology, 2003. Proceedings KORUS 2003. The 7th Korea-Russia International Symposium on, 6-6 July 2003 2003.* 181-187 vol.3.
- WEIGEND, A. Year. On overfitting and the effective number of hidden units. . *In: Proceedings of the 1993 Connectionist Models Summer School, 1994.* 335-342.
- ZHANG, F. 2011. *Intelligent Feature Selection for Neural Regression Techniques and Application.* Doctor of Philosophy, University of Warwick.
- ZHANG, Q. & WANG, C. 2008. Using Genetic Algorithm to Optimize Artificial Neural Network: A Case Study on Earthquake Prediction. *Proceedings of the 2008 Second International Conference on Genetic and Evolutionary Computing.* IEEE Computer Society.

# **Chapter 4: Farm household output prediction using farm household activities and behaviour**

## **4.1 Introduction**

Chapter 3 gives a detailed description of the methods and benchmarking techniques used to evaluate the TSH model, which are; PCA, original MLP-ANN, FS (GA-ANN), OWTNN and SGNO. In Chapter 4, a model of farm household prediction had been developed using the TSH technique, to process most activities and behaviour in the farm, such as: land related criteria, crops related criteria, fertilizer and pesticide usage criteria, and manpower or animal power usage criteria. This is done, without relying on any complex physiological models, in order to achieve better farm household output.

In attempting to optimize farm household output with high quality product, all of the activities and behaviour in the farm will affect its yield. In terms of the land criteria, a larger area of land will give a larger output, if the land is fully used, irrigated and fertile. At the same time, in ensuring the full use of the land, the total number of seeds being used will also depend on this behaviour, where if the seed cost is high, it will also affect the quantity of seed that the farm can buy based on their

total expenses and affordability. To fully utilize the available of the land when sufficient seed is applied to it, an amount of fertilizer is needed to give good fertility to the soil of the farm, and at the same time, a quantity of pesticide is needed to control other external threats. In terms of manpower and animal power criteria, optimization of these factors will ensure that the work on the farm, from the early stage of irrigation to the end product, runs smoothly and without any delays or problems. Therefore, all of these criteria need to be monitored to ensure that the farm output can produce a sufficient supply of food and income for the farmer, which will in turn ensure long-term food security.

Although a large number of predictive models have been developed in the past, few studies of predictive models applied to areas such as farm households and food security exist. Most of the previously developed predictive models tend to give poor performance in handling large and non linear datasets.

## **4.2 Background**

As described in Chapter 1, one of the main food security themes is the maintenance of crops yields, for example in wheat and other cereals. In order to increase the probability of giving better food security management, crop yields should be monitored. Farm household behaviour is one of the areas that need to be assessed, because the management of the farm will always affect its output, which will, in turn, affect the total farm output of the country. A factor influencing this behaviour is the type of technology used, resulting either in a more modern or more traditional method of farm management. Different farms typically have different management styles or behaviours, and this can make it more difficult to monitor the output of each farm (R.P. Singh, 1984).

Referring to (Muhd Khairulzaman Abdul Kadir, 2012) and (DEFRA, 2010), one of the headline indicators of food security is household food security, which states that ideally everyone should be able to access a variety of food, and that each household should also be able afford to buy healthy food. This means that the proportion of people's income available for the purchasing of food should be enough to allow the purchase of healthy food, without considering production of their own food or for profitable sales contributing to their income. (R.P Singh, 1985).

A number of previous studies show that the behaviour or activities in the farm can significantly affect the farm household output (Bhende and Venkataram, 1994, R.P. Singh, 1984, Rao, Sarin, Skoufias, 1993, Skoufias). The types of behaviours or activities that can impact this have been described previously. There is clearly a need to control the amount of food or any kind of food product output, and this needs to be monitored from an early stage of production; this applies to the monitoring of food security worldwide.

As described in an earlier chapter, in terms of prediction or forecasting, few studies have been made to predict the farm household output, especially relating to food security assessment. Farm household management can be very complex because it involves many criteria that need to be managed at the same time.

### **4.3 Dataset**

The dataset used to study the farm household output prediction is based on the village of Aurepalle in India. This dataset was collected from early of 1975 to 1984, and was acquired from Dr. Subramanian, University of Glasgow, where the data was originally given to him by the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) in the form of master data. The data is not publicly available,

and permission is required for its use from either Dr. Subramanian or ICRISAT. Due to the data constraints in studying the food security, this data was used. Although this dataset is old, it is still reliable and relevant for the study of agriculture improvement, and can be helpful to other countries in developing Semi-Arid Tropical (SAT) agriculture as discussed by (Bhende and Venkataram, 1994, Skoufias, 1993, Skoufias).

Generally, the data is based on the village level study (VLS), where the major purpose of these studies is to understand the socioeconomic, agro-biological and, institutional constraints to agricultural development in SAT areas. All of the information that was collected will be used to generate prospective technology that is feasible to be used by the farmers, and each different location can also be useful for testing or modifying the technologies by ICRISAT (R.P Singh, 1985). In this case, the data will be used to predict the farm household output.

Originally, this dataset was collected in five villages in India, where each of the villages was selected based on basic factors such as soil types, pattern of rainfall, and the relative importance of the crops being grown: sorghum, pearl millet, pulses and groundnuts. The selection of the villages studied was also influenced by the presence of a nearby agricultural university or research station for easy access to logistical assistance. In this case study, the Aurepalle village was selected because the dataset is very large and it also has a more complete data filling compared with other villages.

In the village, the households were selected based on their representation of all household categories – labourers, small farmers, medium farmers and large farmers. All of the datasets have been compiled into 9 different files of RAW data which indicate overall household descriptions as below:-



- 1- Five files indicate the general endowment schedule, which is basically the inventory files for animals, farm implements, farm buildings and current physical stock such as food grains, fodder farm inputs etc.
- 2- One file provides data on financial assets and liabilities such as bank accounts, life insurance and loans.
- 3- All of the data on labour, draft animals and machinery utilization is collected in one file.
- 4- Two files contain all of the data on specific plots and cultivation schedules, also the plot summaries.
- 5- All of the household transactions are put into one file, which is used to assess the income position, consumption quantities and expenditure of a household.
- 6- Another two files contain a collection of weather and commodity price data.

In this chapter, to study the farm household output prediction, the combination of two files between plot summaries, based on schedule Y (PS files) and plot and cultivation schedule (Y files) have been made, as in (4) above. These two files consist of an average of 29 different farm households in Aurepalle, as described earlier in this paragraph, and consisting of feature variables as shown in table 4.1. All of the features are categorized as land-related criteria, crop-related criteria, fertilizer-and-pesticide-related criteria and manpower or animal power-related criteria, each of which has its own features, related to each other in terms of farm household behaviour or activities as describe in the beginning of this chapter.

### 4.3.1 ICRISAT

ICRISAT is an institution which operates as a worldwide partnership. It is involved with agriculture research in Asia and sub-Saharan Africa, to aid further development which can be beneficial for the semi-arid or dry land tropics, consisting of regions in 55 countries containing 644 million of the poorest people from over 2 billion people in these countries as a whole.

The headquarters is in Hyderabad, Andhra Pradesh, India, and currently has two regional hubs and four country offices in sub-Saharan Africa. It is a non-profit, non-political organization and a member of Consultative Group on International Agriculture Research (CGIAR) consortium. CGIAR is one of the organizations that is involved in research ensuring future food security (CGIAR).

The vision and mission of ICRISAT is the same as that of this thesis; to try to develop a better food security model for reducing the poverty, hunger, malnutrition and environmental degradation in the dry land tropics (ICRISAT, R.P Singh, 1985).

Table 4.1: Features variables on farm household

<b>Features</b>	<b>Output</b>
<u>Land related</u> Plot Value, Crop Areas, Soil Type, Irrigated area	Crop output
<u>Crops related</u> Total seed value, Total main product, Total by product value, Total Output value, Total Input Value, Net income, Net return	
<u>Fertilizer and pesticide related</u> Total fertilizer value, Total F.Y.M Quantity, Total F.Y.M value, Sheep penning Value, All organic manure value, Nitrogen	

inorganic, Phosphorus, Potash, Total N, Total P <sub>2</sub> O <sub>5</sub> , Total K <sub>2</sub> O, All pesticides value	quantity (kg)
<u>Manpower/animal power related</u> Family male, Family female, Family child, Hired male, Hired female, Hired child, Owned bullock, Hired Bullock, Total family labour value, Total hired labour value, Total owned bullock labour value, Total hired bullock labour value, Total machinery value	

#### 4.4 Data pre-processing

Referring to table 4.1, all feature variables for the dataset consist of several features, each with various units. Different features are recorded on a daily, weekly or annual basis. For example, in the Y files datasets, all of the features are recorded on a daily basis, whereas in the PS files the data is the annual summary of some of the Y files.

In this work, instead of using the daily basis measurements, the yearly averages are calculated, on the basis that the crop output data is used as the growth and the environmental influences from the long seasonal process outputs, which contains data on several types of crops within the year for each of the households.

Table 4.2: Statistics of all feature variables

No.		Unit	Min.	Max.	Mean	Standard Deviation
1	Plot Value	Rs	5	830	134.01	163.46
2	Crop Area	Acres	0.4	51.8	10.17	9.38
3	Soil Type	-	2	75	22.82	15.22
4	Irrigated area	Acres	0	17.5	2.16	3.41
5	Total seed value	Rs	2	2387.5	252.10	342.48
6	Total main product	Rs	0	29265.65	4439.20	6022.51
7	Total by product value	Rs	0	5377	720.97	1100.51
8	Total Output value	Rs	0	33250.65	5169.34	6990.14
9	Total Input Value	Rs	36.6	18745.81	2805.72	3659.92
10	Net income	Rs	-4373.81	18044.21	2363.62	3757.43
11	Net return	Rs	-2074.53	20309.67	3218.66	4357.19
12	Total fertilizer value	Rs	0	3741.68	335.13	621.52
13	Total F.Y.M Quantity	Quintals	0	1026	79.69	125.23
14	Total F.Y.M value	Rs	0	2970	215.00	332.85
15	Sheep penning Value	Rs	0	720	19.56	81.40
16	All organic manure value	Rs	0	3020	237.92	360.85
17	Nitrogen (N) inorganic	Kg	0	409.74	47.88	84.90
18	Phosphorus (P <sub>2</sub> O <sub>5</sub> )	Kg	0	222.18	18.26	39.18
19	Potash	Kg	0	84	1.57	6.67
20	Total N	Kg	0	1123.65	72.00	126.61

21	Total Phosphorus	Kg	0	640.92	31.99	63.93
22	Total Potassium (K <sub>2</sub> O)	Kg	0	1218	30.03	98.10
23	All pesticides value	Rs	0	377.2	20.03	48.12
24	Family male	hours	0	5980.40	585.36	719.19
25	Family female	hours	0	615.5	154.86	132.06
26	Family child	hours	0	364.5	14.47	39.46
27	Hired male	hours	0	5674.65	524.18	1041.45
28	Hired female	hours	0	9306	1113.12	1791.45
29	Hired child	hours	0	304	2.01	19.75
30	Owned bullock	hours	0	4097	440.75	577.04
31	Hired Bullock	hours	0	622.90	64.17	93.62
32	Total family labour value	Rs	0	2400.75	376.67	419.15
33	Total hired labour value	Rs	0	7701.01	729.69	1340.09
34	Total owned bullock labour value	Rs	0	3851.18	478.36	594.44
35	Total hired bullock labour value	Rs	0	798.75	73.71	113.42
36	Total machinery value	Rs	0	2619	300.78	508.23
37	Crop output quantity	Kg	0	47738	7526.61	9490.48

Table 4.2 shows the basic statistical values of the dataset after taking the yearly averages of the total features. These values have various ranges due to the different units for each feature. All these variables will be determined by using the mean standard deviation as shown in equation (3.1), in order to standardize the range to zero

mean and unit variance. This is the first pre-processing operation that needs to be done, and can give optimum performance in terms of FS and optimization in the TSH model. The reason for selecting this standardization is that it will give a wider range and therefore better scope for the data to be processed.

#### **4.5 First stage process**

The reason for the use of this first stage model, as described in an earlier chapter, is that it will determine the number of features which most influence farm household output. This process occurs through the use of the GA module and the ANN module, which combine the optimization by GA with the evaluation of potential solution performance by ANN.

In the ANN module, as described in Chapter 3, the required number of hidden layers is determined as a single layer, and the number of hidden neurons was fixed per equation (3.6). As explained in Chapter 3, a single hidden layer works better as a universal approximation and is more efficient computationally than multiple hidden layers, in fact (Foster et al., 1999) stated that a single hidden layer can solve any cases without any problems. At the same time, one hidden layer will also make the process shorter in terms of its architectural simplicity. The activation function used is also fixed by using tangent sigmoid for each layer, both the hidden layer and output layer. The tangent sigmoid activation function is also a well known function which can reduce the search space, thereby also reducing the training times (Foster et al., 1999, Swingler, 1996).

For the GA module, the chromosomes will be based on the number of features of the dataset. In this application, the number of chromosomes is 36 bits as described in table 4.1 and table 4.2. The initial population size is estimated using equation (3.3),

where the chromosome quantity will be multiplied by 3. So, in this case, the population size is 108 chromosomes with 36 input variables.

In each generation, a fixed crossover probability ratio of 0.7 and a custom mutation as shown in Equation (3.2) was used. Each chromosome in the population will go through the crossover and mutation process, but if the chromosome gives the highest fitness value in the population, it will be considered as part of an elite group and will not go into the crossover and mutation process. In this stage, the GA stopping criteria is set at 50 generations.

Overall, the first stage is simply trying to find the lowest fitness value of the ANN, and from this it will give the optimum feature which will be used in the ANN for the next stage process. As described in the previous chapter, each fitness value is generated by ANN. based on the equation (3.4) of MSE. The lowest value of the fitness value in this stage is 0.8102 and the binary numbers at this time are [1110 1101 1001 0110 0101 0011 0010 1111 110]. The '1' in the chromosomes will be selected by the GA and binary number '0' will be neglected. All of the features selected are as follows:-

Selected input variables = [Plot Value, Crop Area, Soil Type, Total seed value, Total main product, Total Output value, Total Input Value, Total fertilizer value, Total F.Y.M value, Sheep penning Value, Phosphorus, Total N, All pesticides value, Family male, Hired female, Owned bullock, Hired Bullock, Total family labour value, Total hired labour value, Total owned bullock labour value, Total hired bullock labour value]

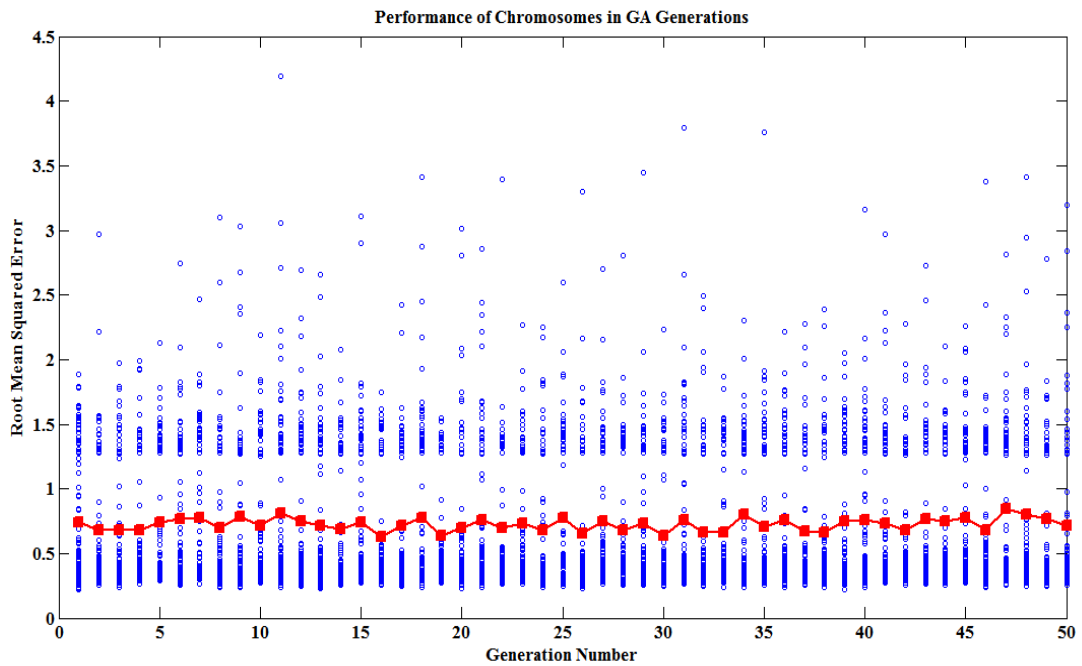


Figure 4.1: First stage performance via GA generation

Figure 4.1 shows the performance of the first stage GA-ANN via the RMSE by each generation number. It also shows the performance of each individual chromosome, which represents the fitness value. The lines in this figure represent the mean population of each individual chromosome's fitness value, represented by the blue dots by each generation. It also illustrates the diversity and the variance of the first stage process population.

#### 4.6 Second stage process

All of the 21 feature variables that have been selected in the first stage process are used again in this second stage process. It has the same modules as the first stage modules – the GA module and the ANN module. Although this stage is using the same modules as the first stage, it works differently, with the GA performing optimization on the weights and thresholds of the ANN. The process ensures a good



generalization of the ANN and this also gives better predictive performance as described in Chapter 3.

The chromosomes for the first stage process were based on the binary number of each feature, whereas in this stage, the GA module chromosomes consist of the thresholds and weights of the ANN, being represented as decimal numbers based on equation (3.7). The range of the numbers set in the GA for random initialization is between -1 and 1. This range is specified because the weight value and the threshold value can be either negative or positive.

As explained earlier, in this application, to optimize the ANN, the GA randomly selects the weights and thresholds for each neuron as shown in the flow diagram of Figure 3.7. It then tests each of the chromosomes for that population, for each generation, through multiple training of the ANN network. The number of chromosomes for this stage is 576, with twice this value, 1152, as the population size. The stopping criterion in this module is the same as first stage process - 50 generations. Other parameters such as the mutation probability ratio and crossover probability ratio are also the same as the first stage process.

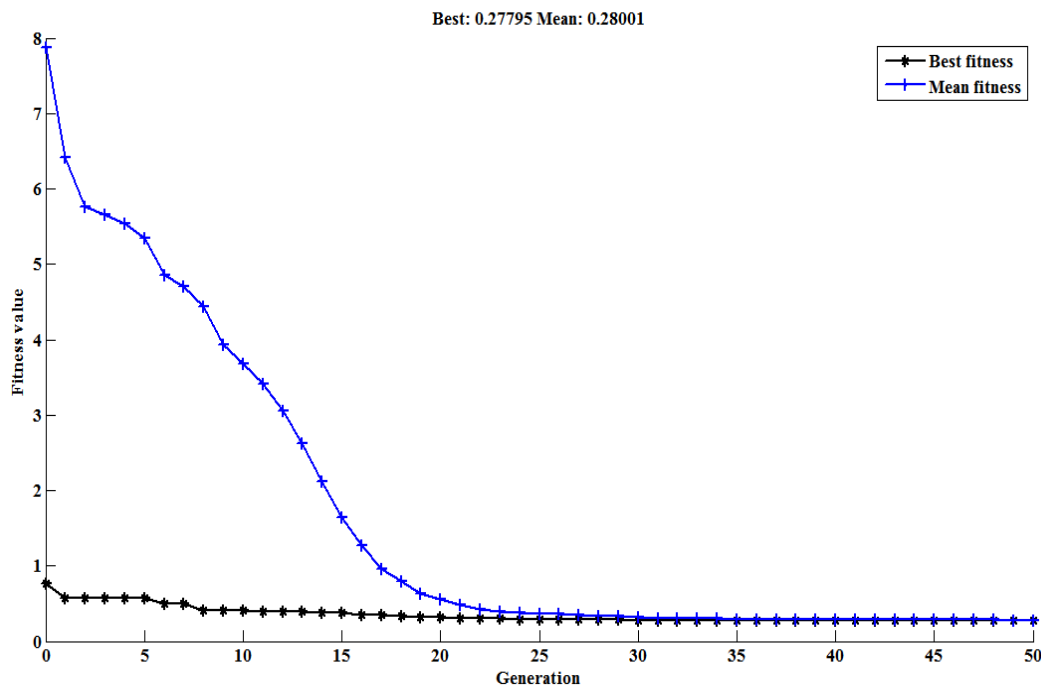


Figure 4.2: Performance of second stage via number of generation

Figure 4.2 shows the fitness value decreasing as the results of the GA finding the best weight and threshold values for the ANN. As described in the first stage, the fitness value of this stage also using the MSE function between the target of the ANN and the actual output. The lower MSE value shows that better optimization has been achieved for the weights and thresholds of the ANN. The performance slopes of the GA process for the first stage and the second stage show a totally different curve. This is because, in this stage, the optimization is applied directly to the ANN architecture, which changes concurrently with the change of the thresholds and weights. In the first stage, the input variables are tested using the same network with the random weight and threshold values, with the ANN learning parameter trying to achieve fewer errors for each epoch.

The ANN module also uses a single hidden layer, the same number of hidden neurons as given in equation (3.6), and a tangent sigmoid as its activation function for

each layer. This same ANN architecture is used to ensure that the same performance is achieved for the features being selected and also to ensure it is optimized for its network.

As described in Chapter 3 and earlier in section 4.6, at this stage, optimization ensures better generalization, which will help to overcome the under-train and over-train problems in the ANN-MLP module. This can be achieved because of the use of the fixed weights and fixed thresholds which are being optimized by the GA module, and is used in remodelling the ANN.

#### **4.7 Remodelling ANN**

After the TSH model has successfully generated the optimum feature variables and the optimum ANN weight value and threshold value for each neuron, these parameters are fed into the MLP-ANN. This remodels the ANN, using the optimum parameter values to get the optimum prediction performance.

Figure 4.3 and Figure 4.4 show the performance of the TSH model. These figures show the performance plots based on the regression values and the MSE value. For figure 4.3, the regression values are compared between the actual outputs and the TSH model outputs in each epoch or ANN iteration. This shows the overall performance to be around 96%. To show the different performance results, a MSE performance graph is shown in Figure 4.4. In the first epoch, it shows a low MSE value and the overall ANN process is taking up to 11 epochs only. Further discussion on the MSE performance will be discussed later in the summary section.

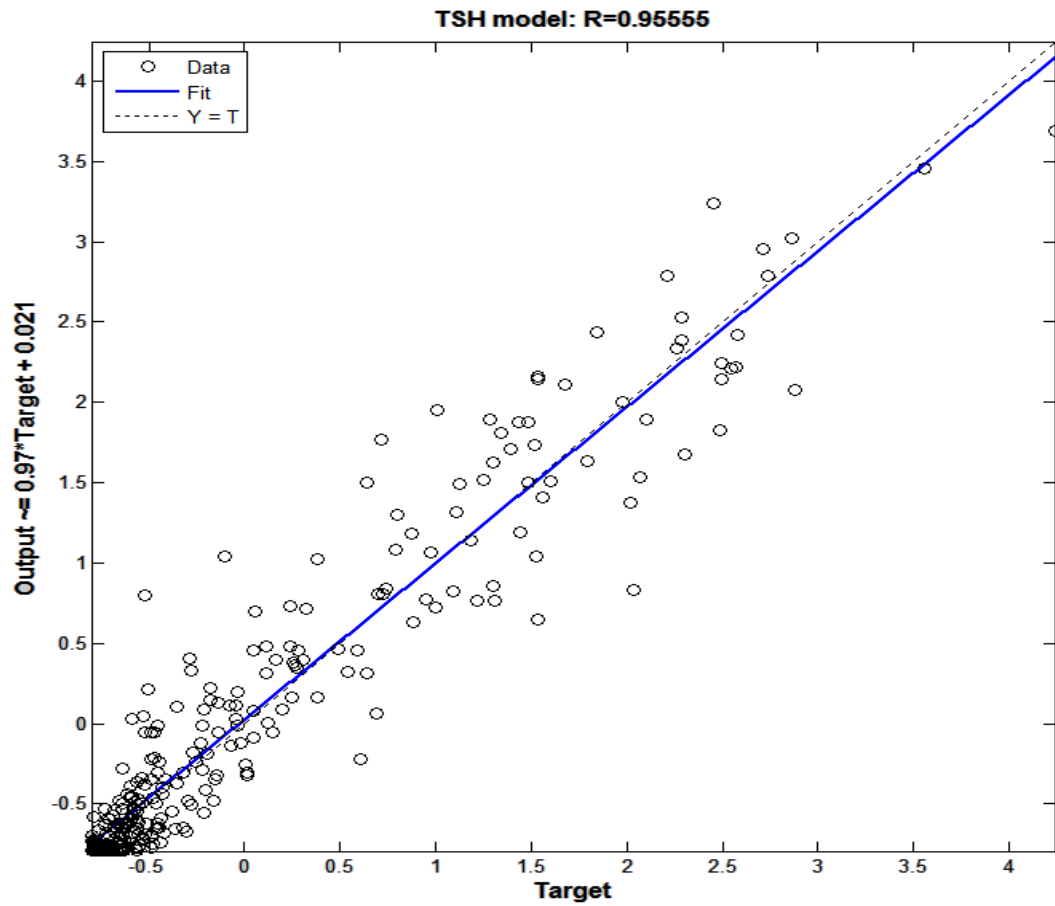


Figure 4.3: Regression of TSH model

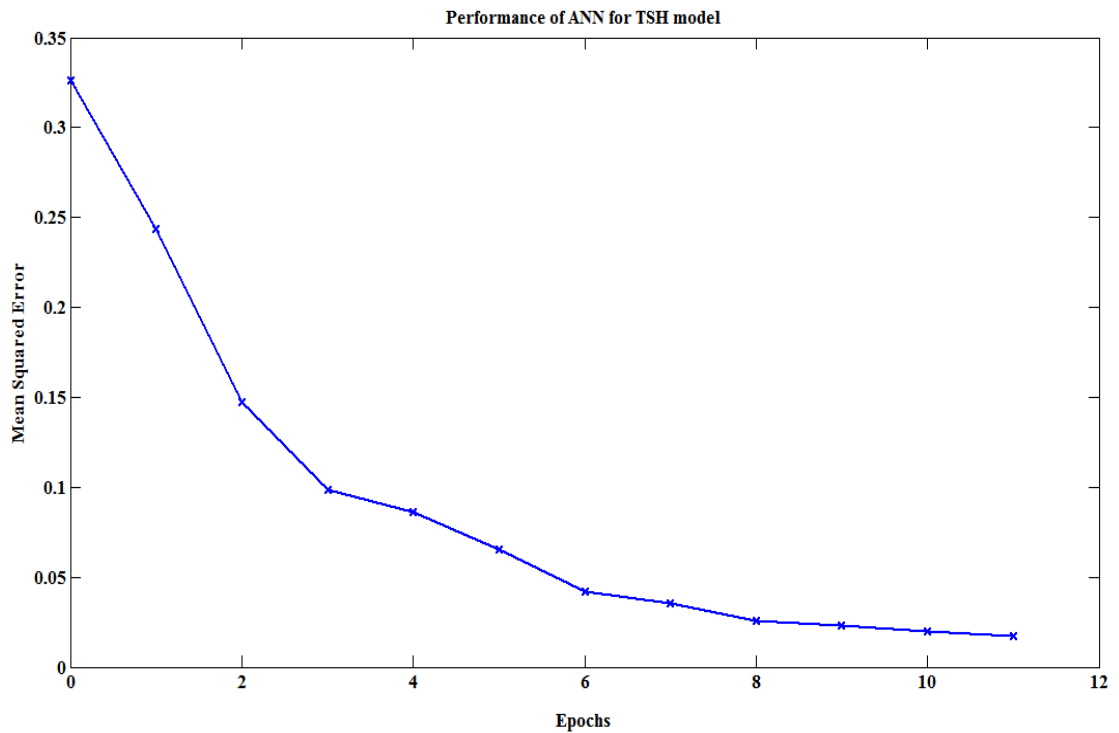


Figure 4.4: Performance of ANN based on MSE vs Epochs

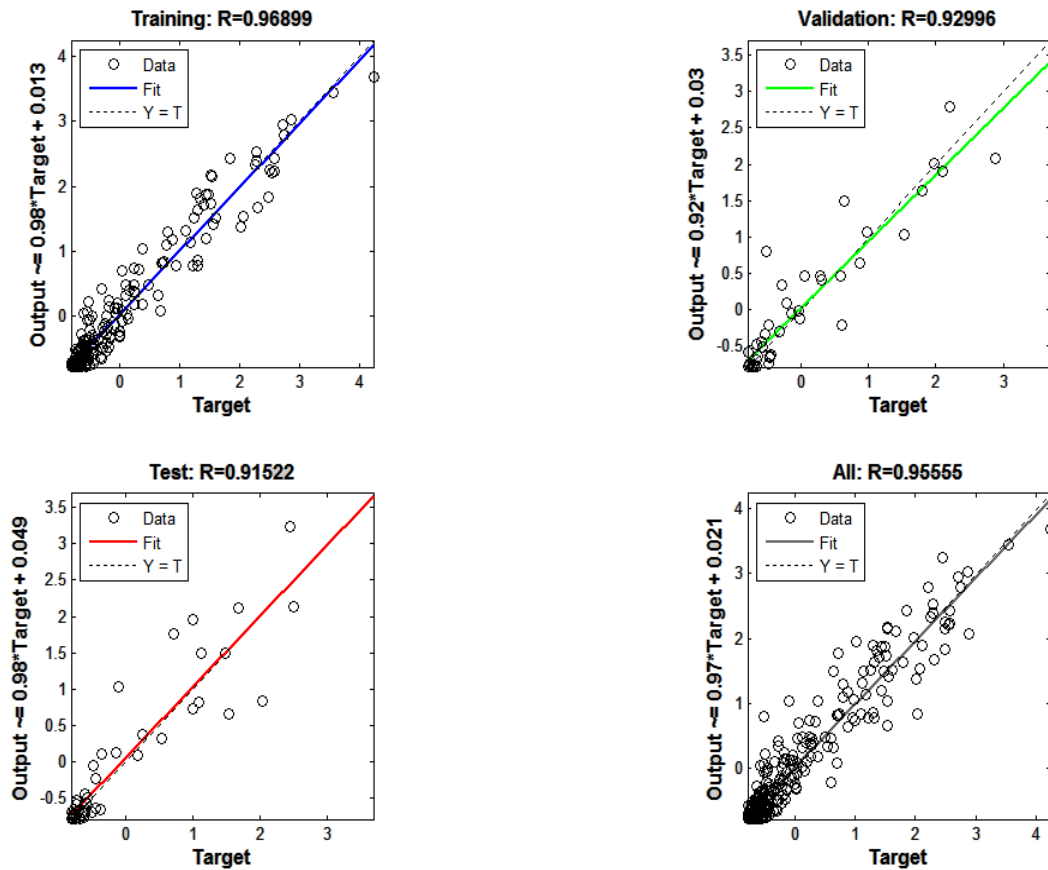


Figure 4.5: Overall regression for each part – training, validation and testing

Figure 4.5 also shows the regression performance, but the difference between this figure and Figure 4.3 is that it shows the variation of regression performance for each part – training, validation and testing. Each part shows good performance results, which are above 91%, and gives the overall regression performance at about 96%. All of these performance values will be compared with each stage of the model via prediction performance. At the same time, the MSE and regression performances will also be compared with the PCA, original MLP-ANN and SGNO to determine the prediction accuracy.

## **4.8 Benchmarking and discussion**

As briefly described in Chapter 3 and in the previous section, the proposed model, the TSH model, will be compared with the PCA, original MLP-ANN, SGNO and with each individual stage of the proposed model (FS and OWTNN). The evaluation will be based on the regression for each part of the dataset division in the ANN and the MSE of the overall performance of the MSE of number of iterations in the ANN.

### **4.8.1 Principal Component Analysis (PCA)**

PCA is capable of analyzing the data in terms of its principal components, where the dataset is transformed to a new dataset based on its 'eigen-analysis'. The data can be reduced based on the less significant number of variations in terms of the smallest eigenvalues, as described by (Jolliffe, 2002, Smith, 2002).

In this case, PCA is used to reduce the number of features variables of the farm household activities from 36 input variables to 21 input variables. In reducing the variables, the number of principal components is analyzed and selected, which is then turned into a new final data set as shown in Equation (3.11).

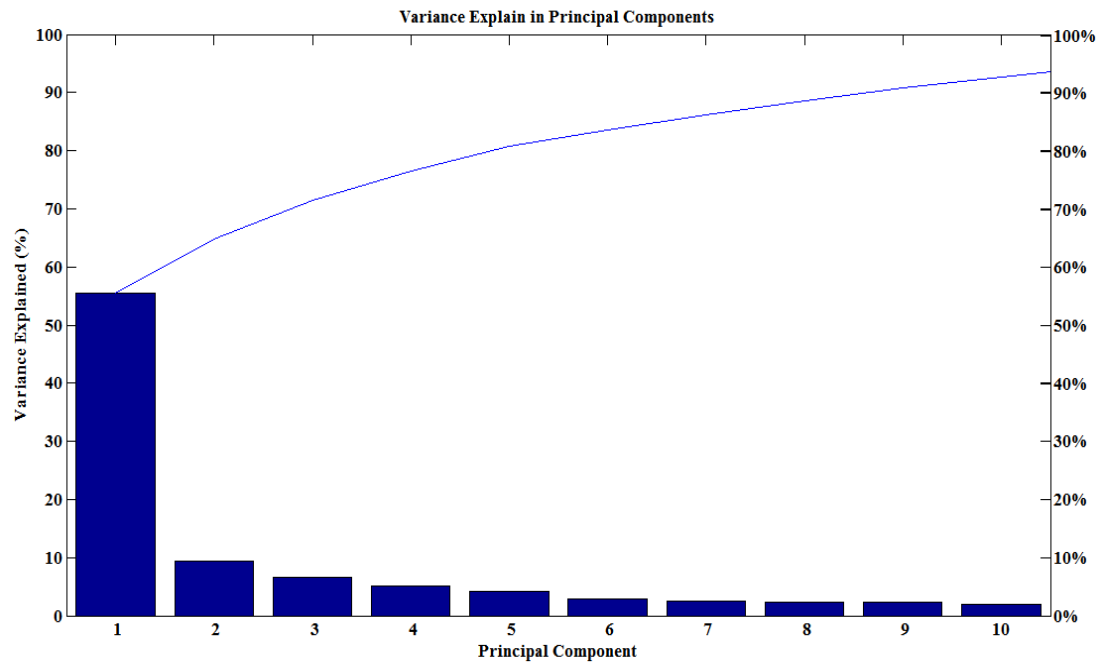


Figure 4.6: Variance in Principal component

Figure 4.6 shows the variations of the principal components where only 10 principle components are present. As shown in the figure, the cumulative value of the first five components, represented by the thin blue line, shows an 80% variance, and the first three components show a 70% variance. As in the first stage process, the lowest MSE is taken for the feature being selected. In this PCA technique, the first 21 components will be selected. These 21 selected components will be converted to a final data set following equation (3.11), and are then used as the inputs to the ANN model to predict the farm household output, and compared with the prediction performance between the proposed models.

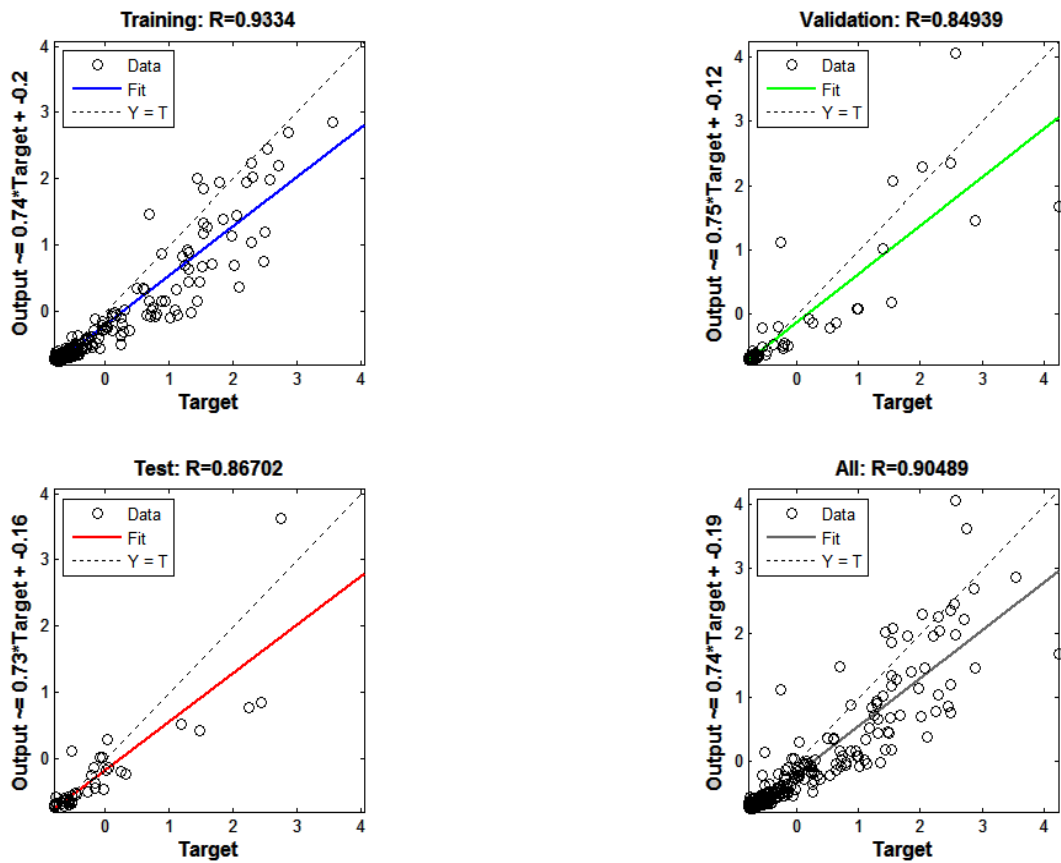


Figure 4.7: Regression for PCA using ANN

Figures 4.7 and 4.8 show the performance of the PCA, as a regression value and MSE value. In the regression plot, the proposed model is outperforming the PCA-ANN prediction model, with a small difference of 0.05. However, referring to the figure of the data fitting for the training, validation and testing parts shows a poor fit of data to the line of  $Y = X$ . This illustrates a generalization problem with the network.



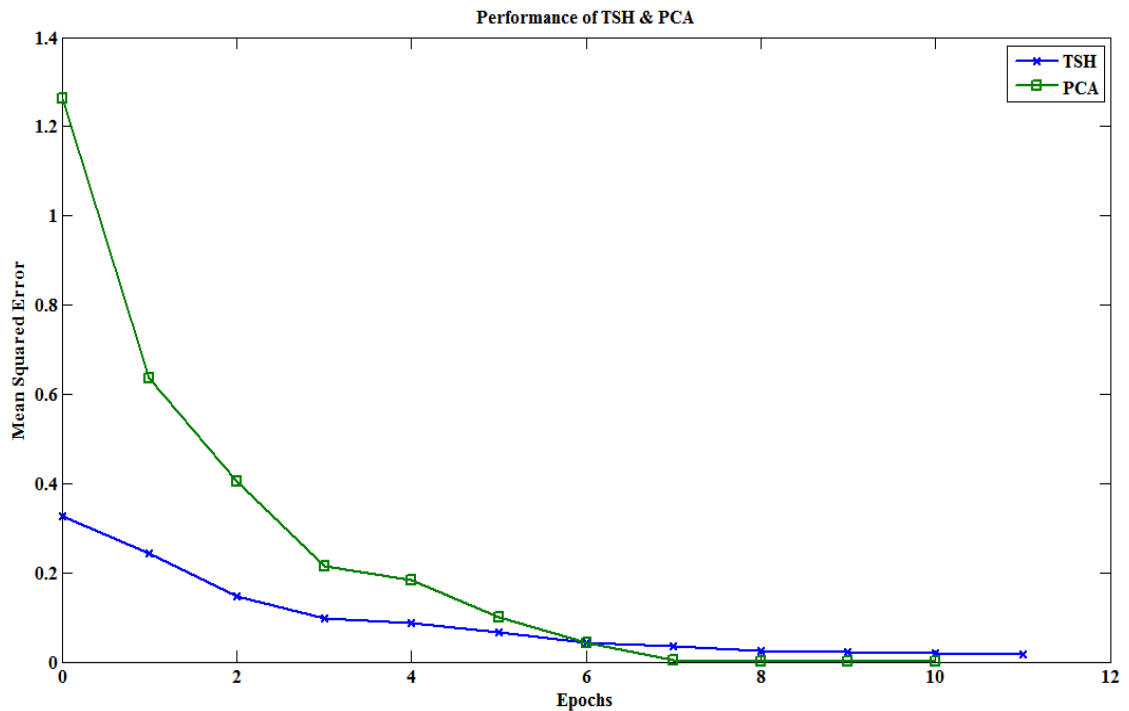


Figure 4.8: MSE between PCA and TSH model

Referring to the MSE plot in Figure 4.8, although the number of iterations of the PCA is less than the proposed model, it still shows that the PCA give better errors than the proposed model. Although the PCA error is better than the proposed model, the start error of the proposed model is less than the PCA, and both error performances gradually decrease. However, the final error difference between the PCA and the proposed model is only 0.0168, which is a small difference. The advantage of the proposed model in this case gives a small start errors but at the sixth epoch, the PCA errors become smaller than the proposed model.

#### 4.8.2 Multi Layer Perceptron – Artificial Neural Network (MLP-ANN)

MLP-ANN is one of the best existing prediction methods, and is widely used by other researchers. However, sometimes weaknesses in the ANN exist, especially when it involves a complex and high dimensional dataset. A detailed explanation of

this MLP-ANN technique is discussed in Chapters 2 and 3. In this case, MLP-ANN is used to predict the farm household outputs, in comparison with the proposed model. All of the parameters used in this MLP-ANN will be the same as the TSH model. This is to ensure that the MLP-ANN will have the same architecture as the remodelling of the ANN, allowing direct comparison of both models. The only difference for this MLP-ANN is the dataset used, which will be based on all features of the Aurepalle dataset, which has 36 input variables and does not involve any data pre-processing as in the proposed model.

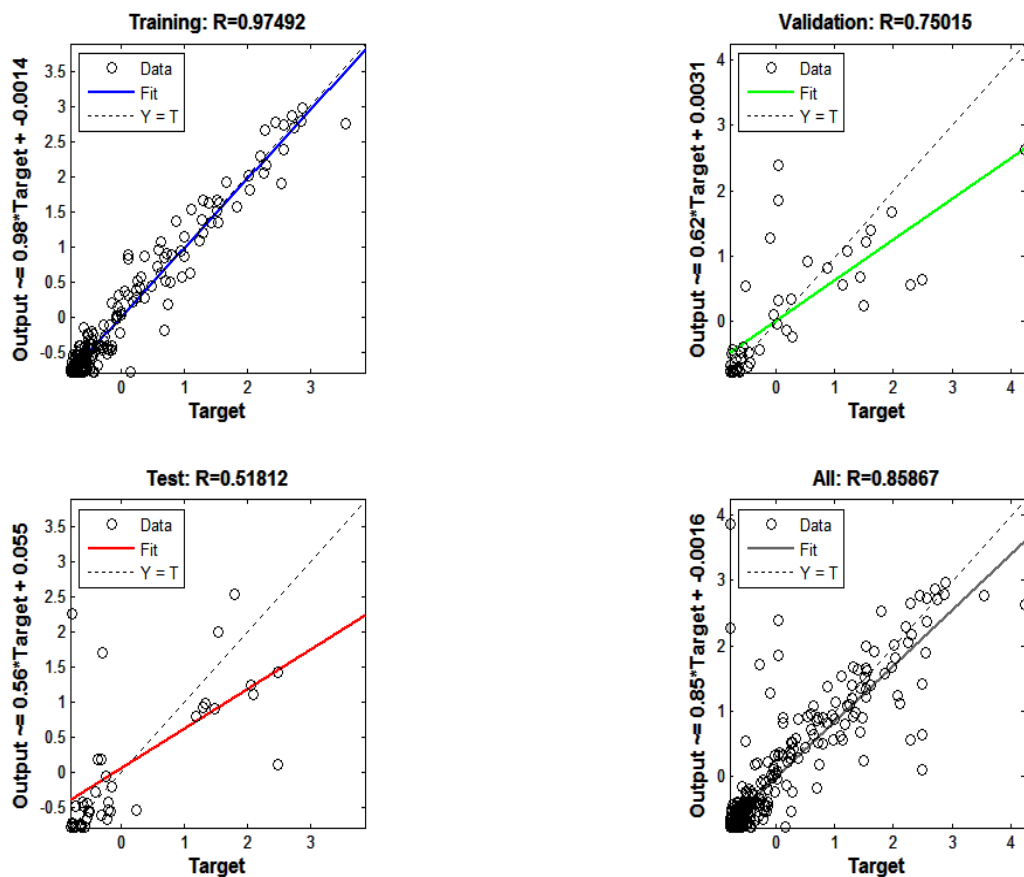


Figure 4.9: Regression for MLP-ANN

Figure 4.9 shows the regression performance of the prediction on the farm household output between the actual output and the MLP-ANN model output. The overall regression value is 0.86, less than 10% different to the performance of the TSH model which show TSH model has higher prediction than MLP-ANN. By referring to the figure, the problems in generalizing the data for the testing part and validation part can be seen; the data is scattered and not a good fit to the regression line. However the training part shows very good performance. This may be due to a random weight and threshold being given to the network and at the same time may be due to the amount of data being divided for the training part being larger than the validation and testing parts, each of which is 20%, compared with 60% for the data training part.

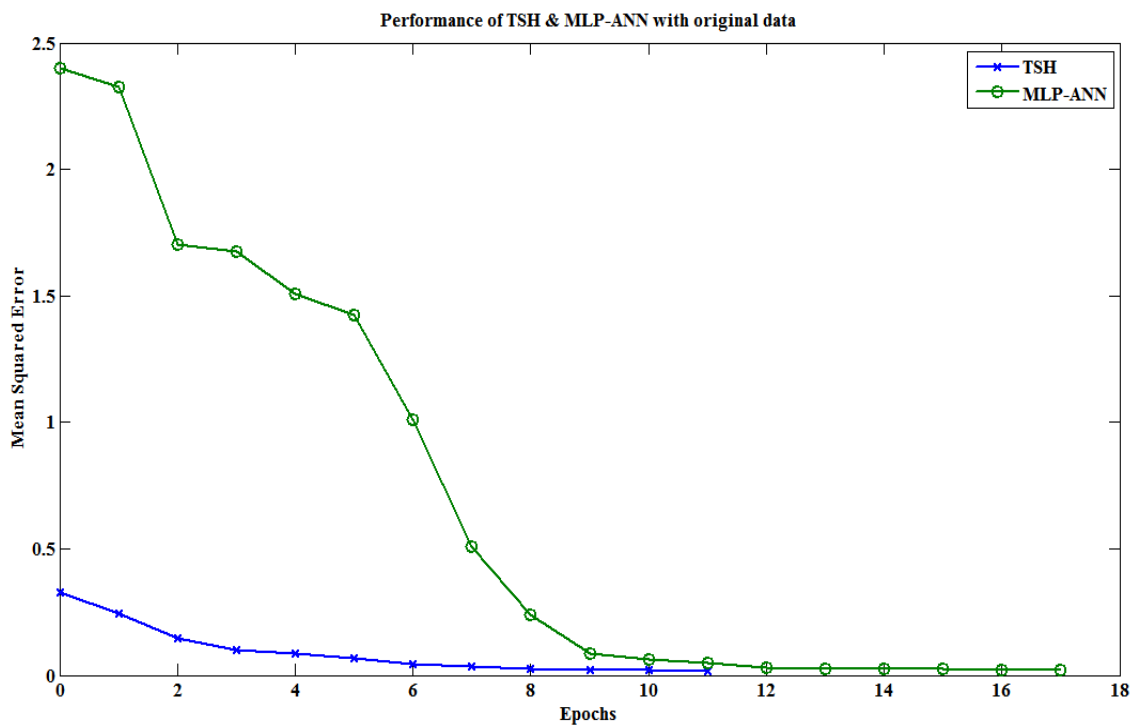


Figure 4.10: MSE performance of ANN models using TSH against MLP-ANN with original data

The error performance in each epoch for the MLP-ANN model and proposed model is shown in Figure 4.10. Here, the plot shows a very good decreasing MSE value curve, and both models converge at the higher number of epochs. If compared to the proposed model, the MSE for the first epochs starts at between 2 and 2.5, which are higher than the TSH model, and it also finishes at the much higher 17<sup>th</sup> epochs. The final error of the MLP-ANN is 0.0223, which is significantly higher than the value of 0.01728 for the TSH model. This shows that the proposed model has the advantage of a small error at the starting epochs, finishes earlier at the 11<sup>th</sup> epoch, and gives a lower error compared with the MLP-ANN prediction model.

### **4.8.3 Feature Selection (GA-ANN)**

Sometimes, to ensure good prediction, it is necessary to analyze the features used in the network. In comparing the TSH model, the first stage result, which is the feature being selected, will be used to predict the target by using the ANN. All of the parameters for the GA module and the ANN module will be the same as for the first stage process. As described in section 4.5, 21 features were selected, which were taken from the lowest fitness value of the GA. By using the same selected feature and parameters, the generalization problem still occurs, which is almost the same as for the MLP-ANN model, as shown in Figure 4.9. However, the regression value for the FS model is much better than that for the MLP-ANN model, where the difference between them is 0.057. The validation part and the testing part still show the same problem of generalization as the MLP-ANN model.

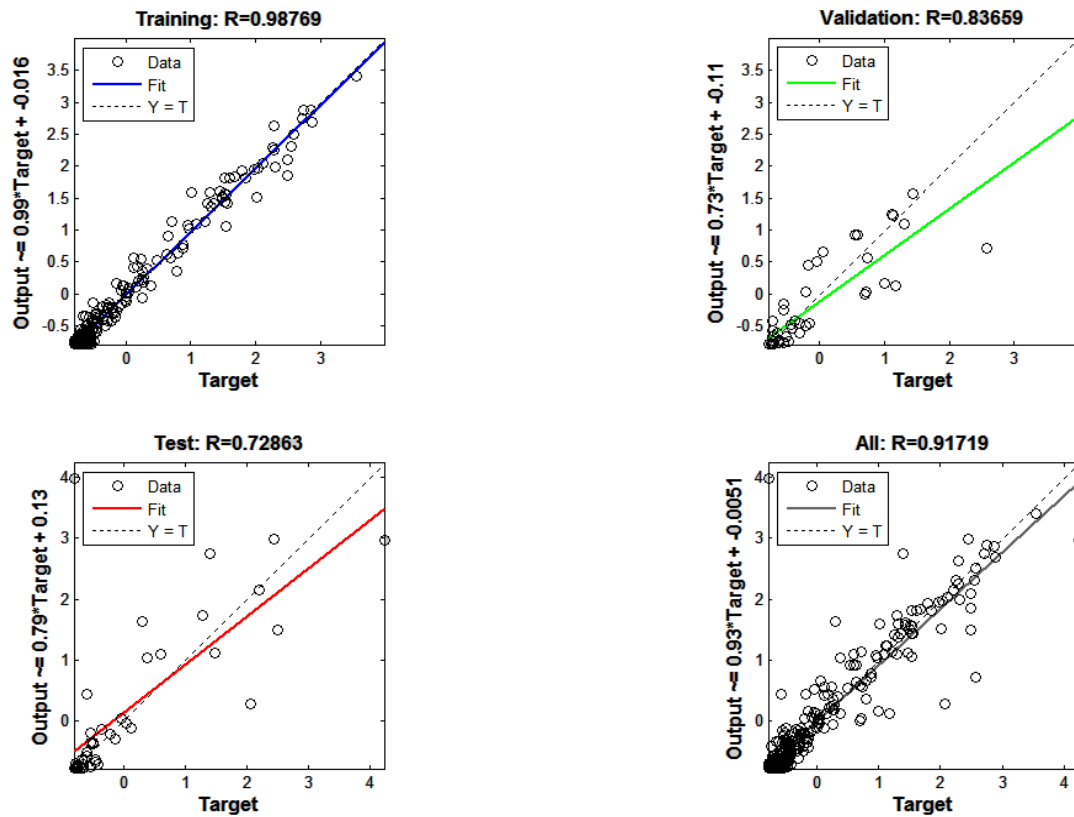


Figure 4.11: Regression for feature selection using ANN

For the MSE performance, FS (GA-ANN) gives a much better MSE value at the first epoch, but it takes a longer time to give a solution at 25 epochs, as shown in Figure 4.12. This may be because of the random weight and threshold generated, and it gives a worse prediction performance than the TSH model. The final error at the end of the epochs still shows that the proposed model is better than the first stage process only.

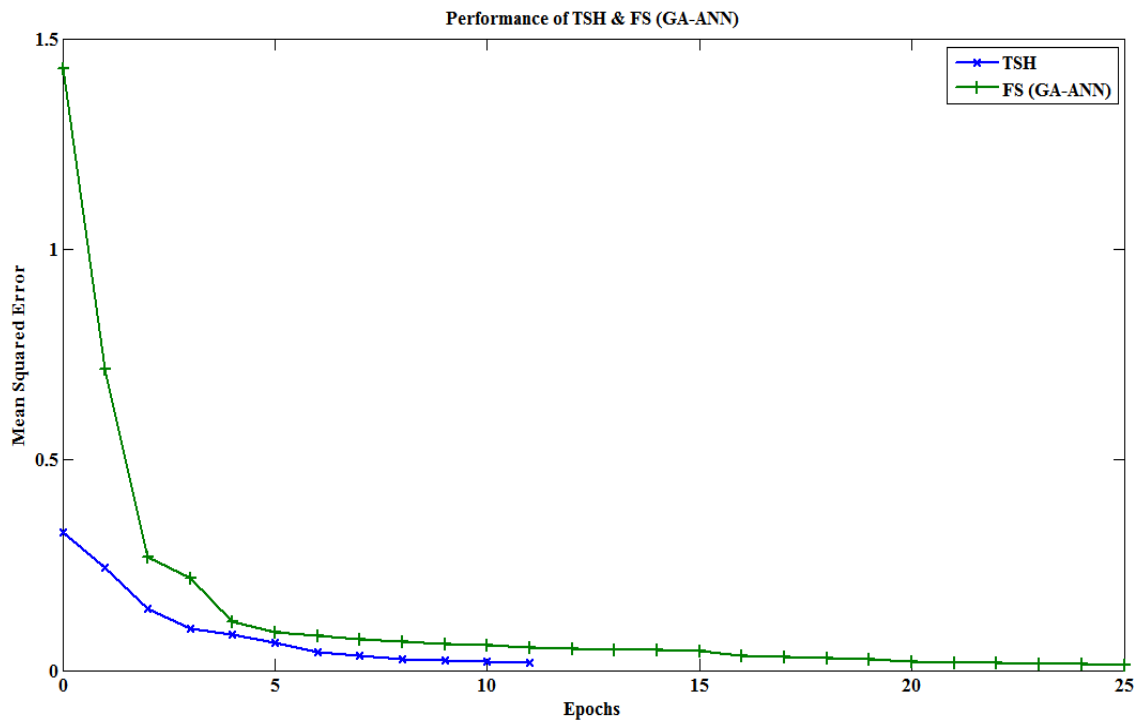


Figure 4.12: MSE performance of ANN models using TSH against feature selection (GA-ANN)

#### 4.8.4 Optimized Weight and Threshold Neural Network (OWTNN)

In the TSH model, as described earlier, OWTNN is used as one of the processes. Therefore, in ensuring the reliability of the model, a standalone OWTNN model needs to be compared with it. By using the same parameters and the same ANN layer architecture, the GA will optimize all of the weights thresholds for the ANN, which is called OWTNN. This optimization will be done using the entire set of features available in Aurepalle dataset - 36 input variables, and will then be re-applied to the ANN with the optimized weight and optimized threshold as an added parameter, rather than using a random weight and threshold in the network. The performance of the ANN is plotted based on the regression plot and MSE plot at each epoch.

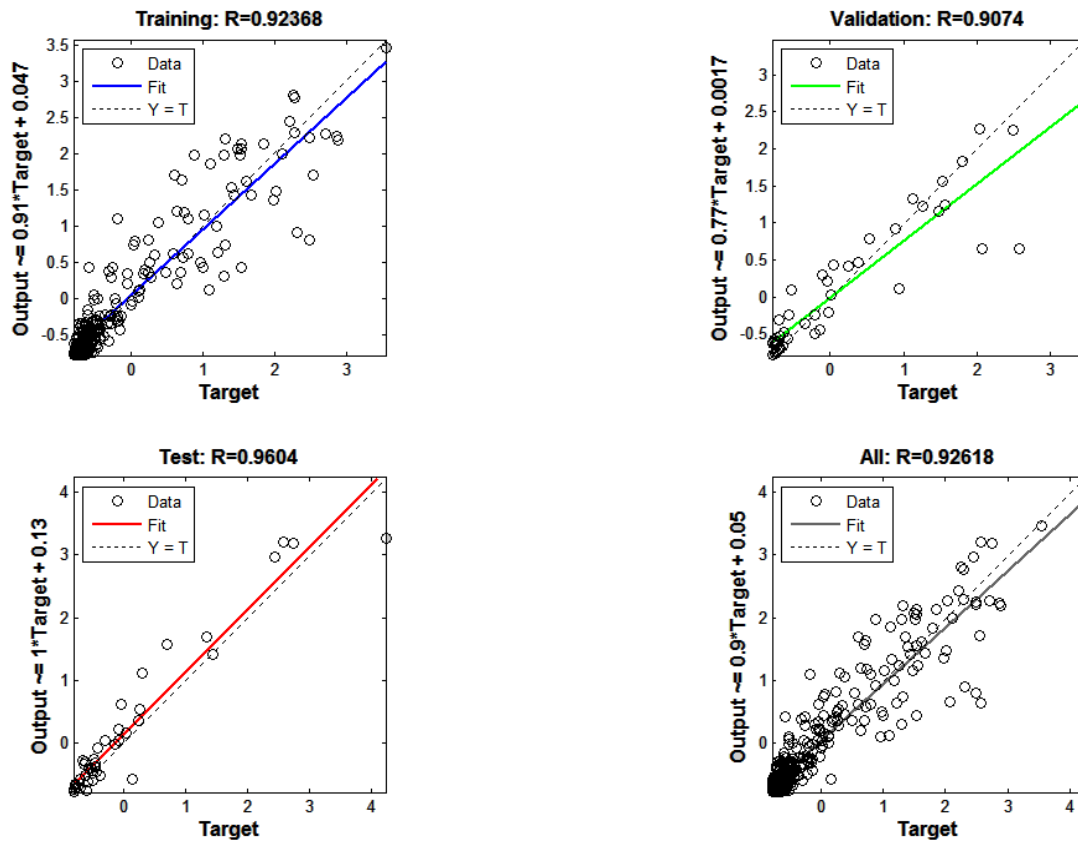


Figure 4.13: Regression for OWTNN using ANN

The improvement of generalization for the ANN of the applied optimized weight and threshold, where each part – training, validation and testing – gives a regression value of more than 0.92, which is almost the same as that for the TSH model. However, the overall regression value still has less than a 4% difference, and this shows the TSH model had better prediction than this model.

For MSE performance at each epoch, the OWTNN performs better than the TSH model in terms of the lowest MSE in the first epochs. It also gives faster solution than the TSH model, which achieves the best solution at epoch 7, compared to epoch 8, to achieve almost the same MSE value. However the final error of the proposed model is better than that for the OWTNN. The TSH model shows an advantage in the regression for better prediction, but the OWTNN can operate much faster than the

TSH model. The variations of the MSE performance curve plot between the TSH model and OWTNN are shown in Figure 4.14.

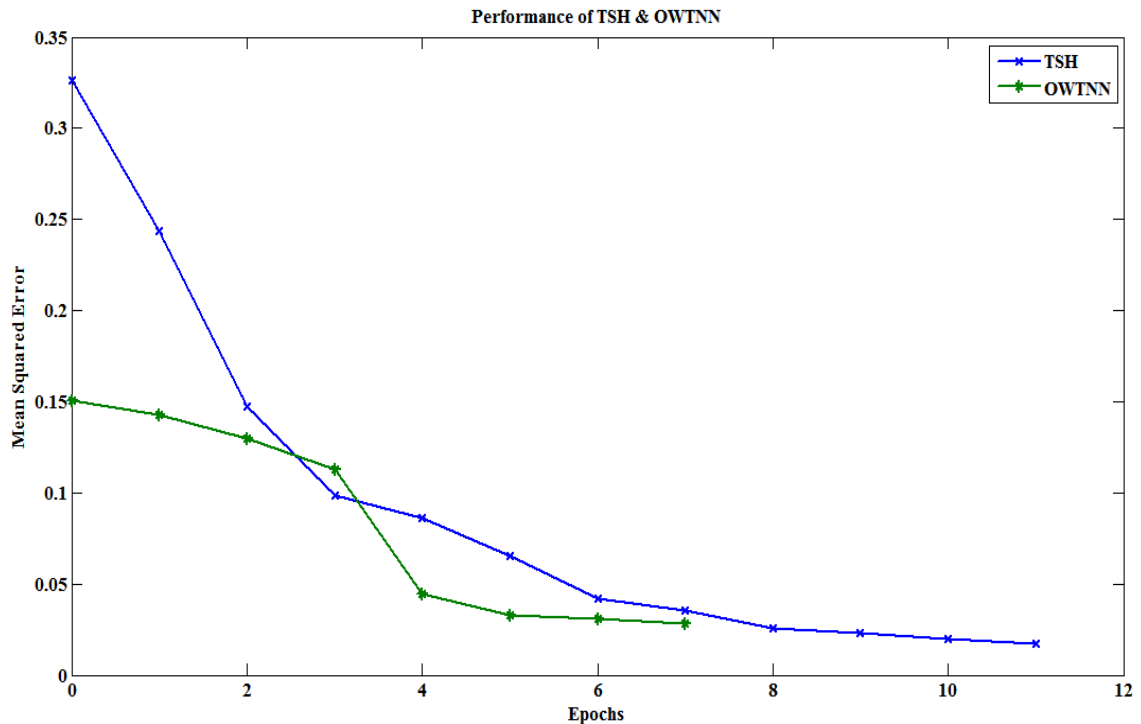


Figure 4.14: MSE performance of ANN models using TSH against OWTNN

#### 4.8.5 Sensitive Genetic Neural Optimization (SGNO)

SGNO is a ranking type of feature analysis, which analyzes each feature based on the frequency of that being used in the GA, and the sensitivity of the feature behaviour towards the actual output. As described in Chapter 3, it consists of 3 modules – a GA module, an ANN module and a sensitivity analysis module, where two modules are the same as those in the TSH model.

Although it uses two of the same modules, SGNO has a different data pre-processing module and uses different parameters. For the GA module, it involved a five-fold cross validation pre-processing, which divided all of the datasets into five groups. The fitness function is the same as the TSH model.



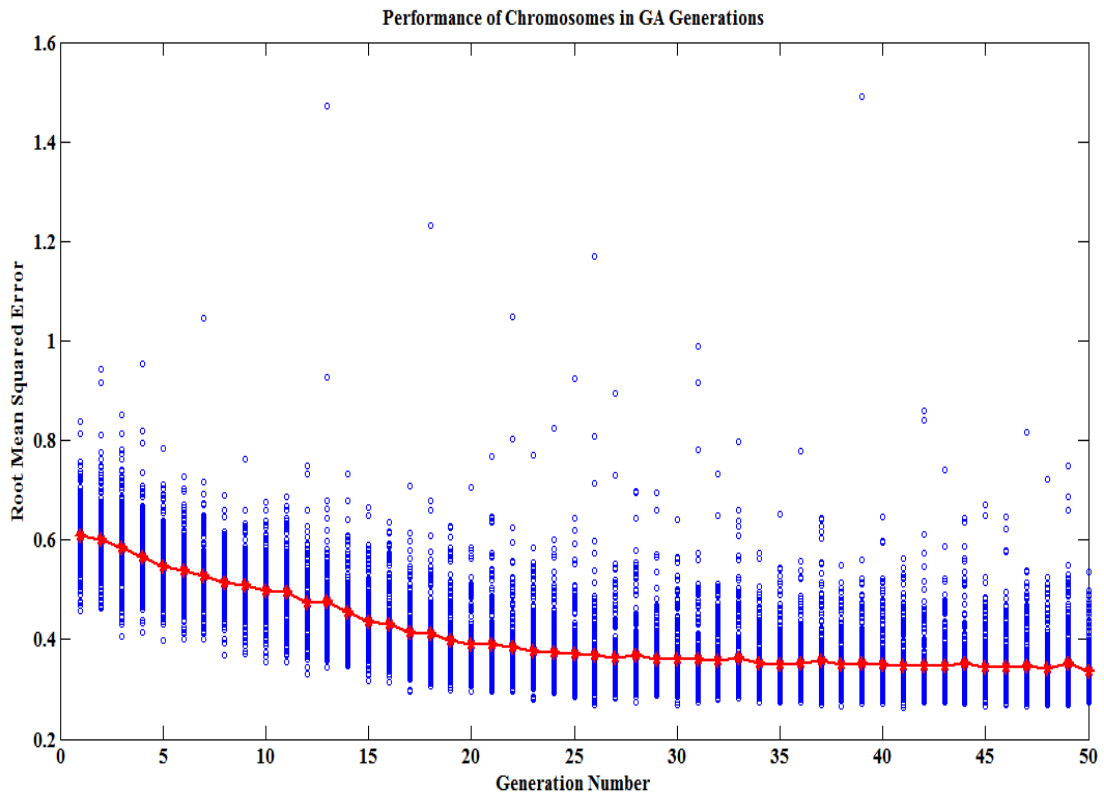


Figure 4.15: Performance of SGNO chromosomes via number of generation

For the ANN module, SGNO used the same layers in the ANN, but the number of hidden neurons is different, and is estimated by using the half-sum of the input and output variables. The weight and threshold for this model is randomly generated by the ANN. The activation function for the hidden layer is a tangent sigmoid, which is the same as for the TSH model, but uses a different output activation function, the pure linear function.

The performance of the SGNO, referring to each of the generation numbers of the GA as in Figure 4.15, shows a reduction of RMSE value which is different from the TSH model results. After the GA-ANN process, each feature is ranked based on its importance, and then the sensitivity analysis module will identify again the global influences of each feature, by analyzing the importance of the input parameters in each generation. Each input parameter will then be given a score, based on the global

sensitivity, by taking the mean of the sensitivity score, derived from all of the selected chromosomes. The scores for each parameter are shown in Figure 4.16.

Each score, as shown in Figure 4.16, was then rearranged based on its importance, where the highest mean value was considered as having more influence or importance than the lower score. The importance rank table is shown in Table 4.3, while the feature variable number can be seen in Figure 4.16.

Table 4.3: Ranking selection for each of the features

<b>Rank</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>Feature Variable</b>	36	1	35	28	32	31	2	27	33
<b>Rank</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	<b>16</b>	<b>17</b>	<b>18</b>
<b>Feature Variable</b>	26	24	29	25	34	16	18	7	13
<b>Rank</b>	<b>19</b>	<b>20</b>	<b>21</b>	<b>22</b>	<b>23</b>	<b>24</b>	<b>25</b>	<b>26</b>	<b>27</b>
<b>Feature Variable</b>	4	17	21	23	22	9	10	14	6
<b>Rank</b>	<b>28</b>	<b>29</b>	<b>30</b>	<b>31</b>	<b>32</b>	<b>33</b>	<b>34</b>	<b>35</b>	<b>36</b>
<b>Feature Variable</b>	30	3	15	12	20	11	5	8	19

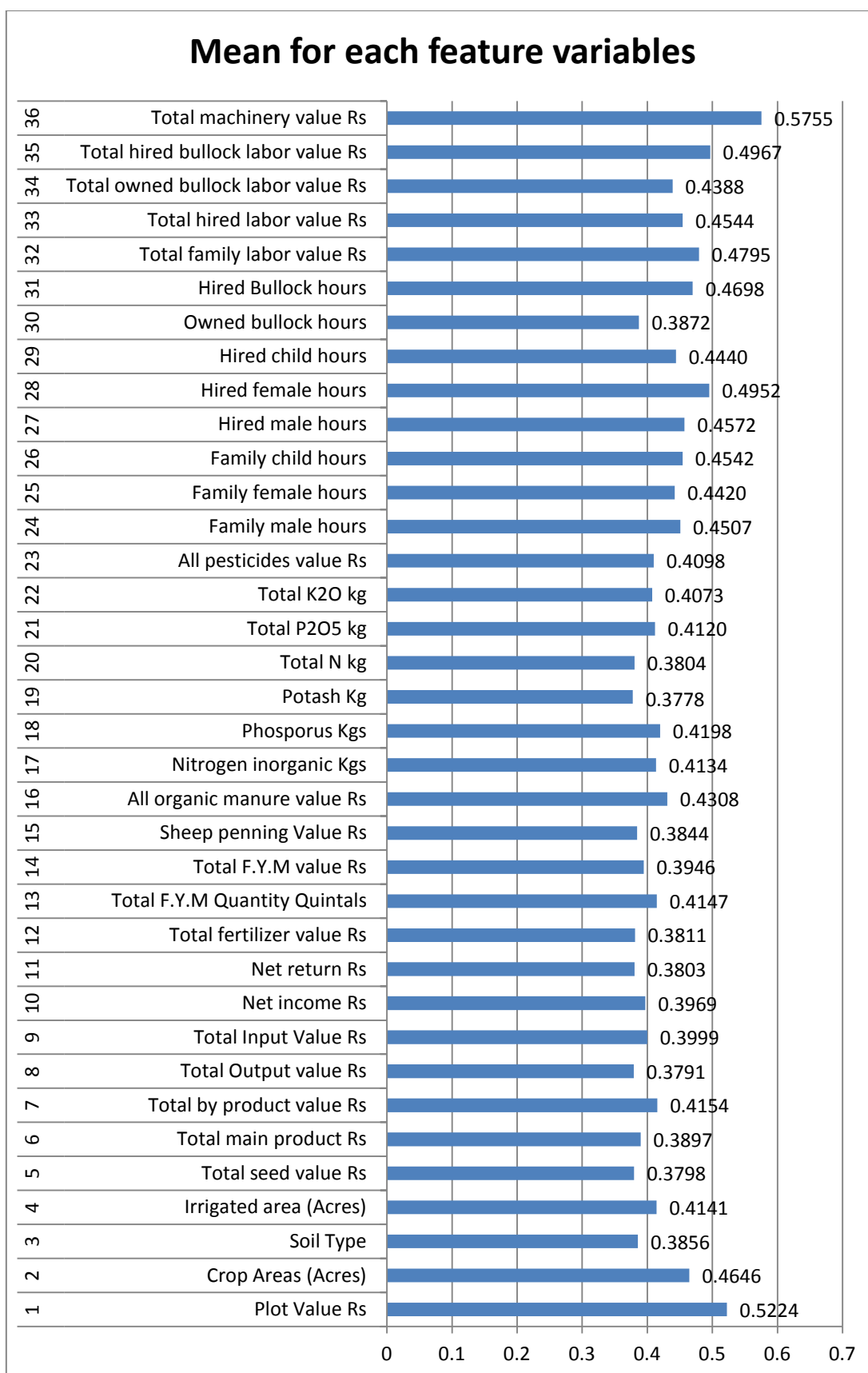


Figure 4.16: Mean for each of the feature variables

The rank table can also be express as the element below:-

[Total machinery value, plot value, Total hired bullock labour value, Hired female, Total family labour value, Hired Bullock, Crop Areas, Hired male, Total hired labour value, Family child, Family male, Hired child, Family female, Total owned bullock labour value, All organic manure value, Phosphorus, Total by product value, Total F.Y.M Quantity, Irrigated area, Nitrogen inorganic, Total P2O5, All pesticides value, Total K2O, Total Input Value, Net income, Total F.Y.M value, Total main product, Owned bullock, Soil Type, Sheep penning Value, Total fertilizer value, Total N, Net return, Total seed value, Total Output value, Potash]

In this table of importance, 21 features will be selected, the same size of input variables as in the TSH model. Then, by using the same ANN parameter and architecture as the TSH model, the performance benchmark or regression plot and MSE plot is performed.

As in the previous benchmarking, the ANN performance graph is simulated, as shown in Figure 4.17, where it shows almost a perfect regression plot for the training part. However, it lacks the data concentration in the validation part and testing part. As with the MLP-ANN, this is because the SGNO only performs data analysis for use in the testing division rather than in the overall ANN network analysis. Therefore by referring to this generalization problem in the validation part and the testing part, the overall regression is lower than the TSH model, which gives better prediction than SGNO.

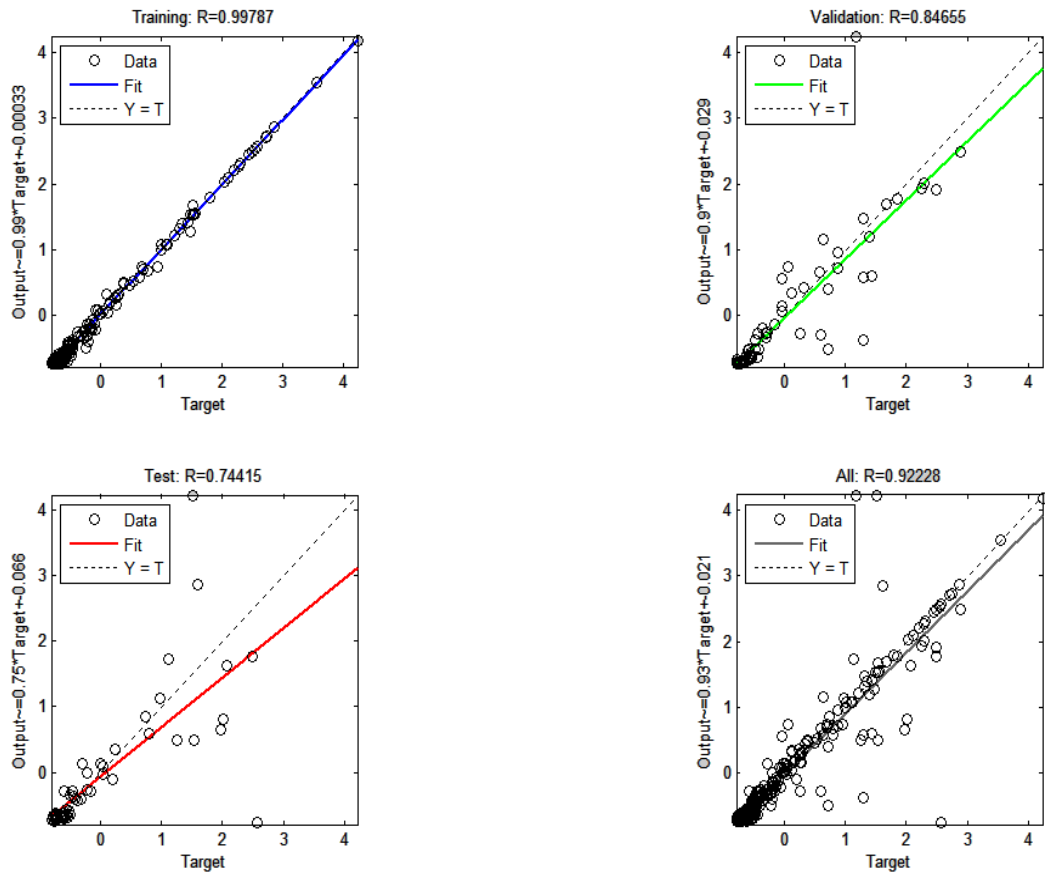


Figure 4.17: ANN performance based on SGNO

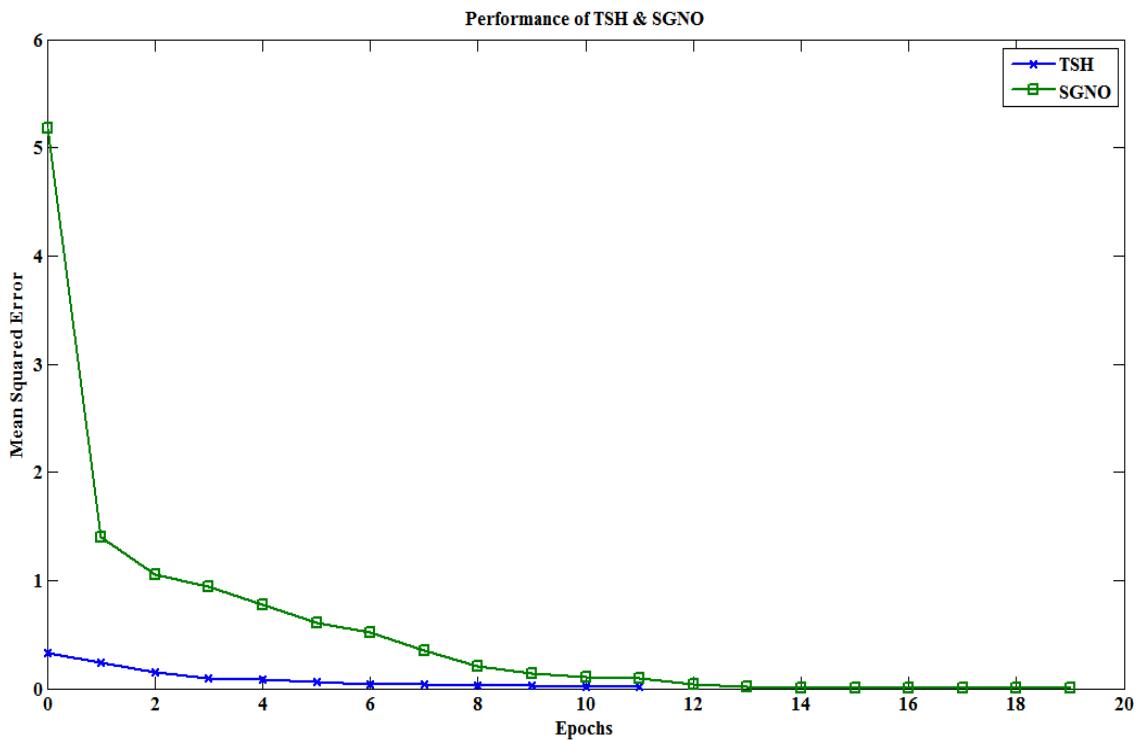


Figure 4.18: MSE performance of ANN models using 2-stage against SGNO

As in previous benchmark techniques, the proposed model gives lower errors at the start of epochs, which is the same with SGNO. This is shown in Figure 4.18. It appears that SGNO takes more time to finalize the solution, ending at the 19th epoch, compared with only 11 iterations for TSH model.

#### **4.8.6 Summary**

In this chapter, one dataset from the years 1975 – 1984, for Aurepalle village in India, is used to measure the prediction performance of the proposed model. The proposed model is then compared with PCA, the original MLP-ANN, each stage of the proposed model (FS and OWTNN) and SGNO. The comparisons are based on regression and MSE performance plots, with the overall performances plots being shown in Figure 4.19 and 4.20. By referring to these performance plots, it can be seen that, in terms of the overall regression value, the TSH model outperforms other model in terms of prediction. The regression for each part – training, validation and testing, shows that the TSH model also gives better generalization than the other models, although in the training part the SGNO perform better than the others models.

In terms of MSE performance, although the overall errors show that the TSH model is better than other benchmark models, the OWTNN gives a lower MSE value in the first epoch, and it also finds faster solutions than the other benchmarked models. However, the final MSE value for the TSH model is 0.0173, which is lower than that for OWTNN (0.028). Most of the MSE plot show gradually decreasing curves as the number of epochs increases, and at the end, although the proposed model did outperform other techniques, there is only a small difference between them.

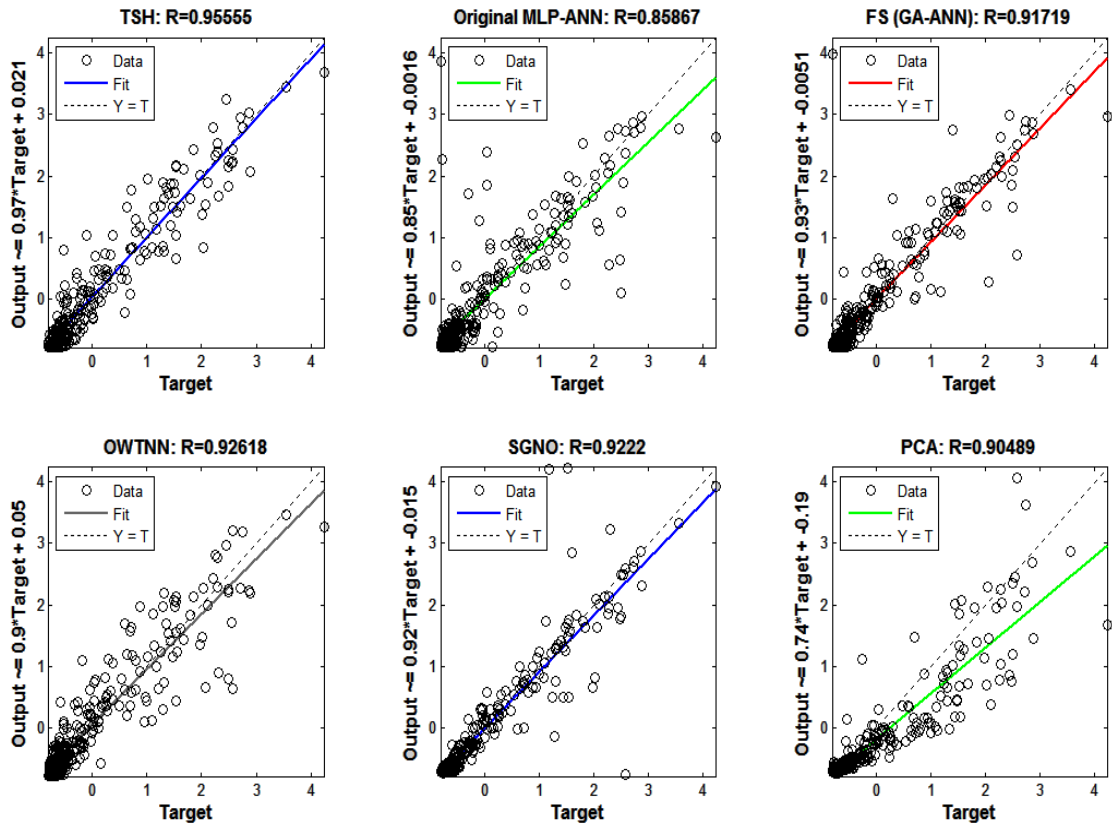


Figure 4.19: Benchmarking on overall ANN regression performance

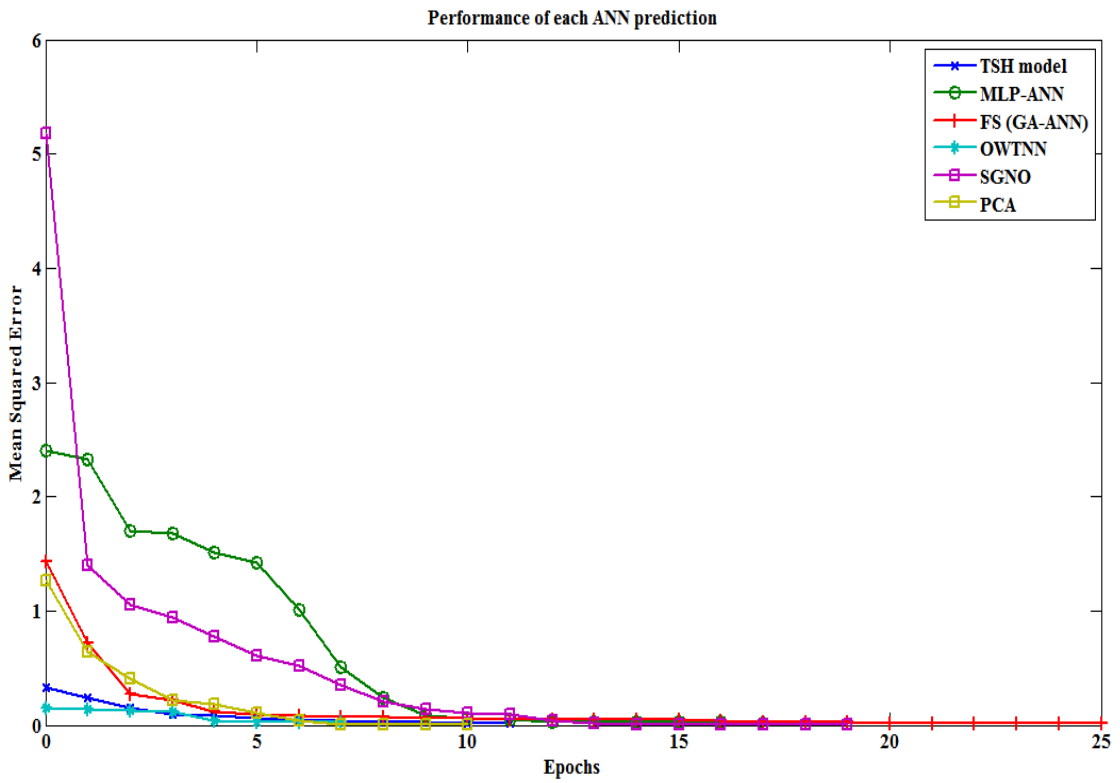


Figure 4.20: Benchmarking on MSE performance of ANN

## 4.9 Conclusion

In this chapter, the farm household output prediction model has been developed. This model consists of 36 features with 292 samples, where the data was from the ICRISAT Aurepalle village dataset. The proposed model of TSH, when used for modelling the prediction, can generate very good results based on the farm household behaviours and activities.

The TSH model combines the FS model and OWTNN model, which focus on the entire dataset analysis – training, validation and testing. This model is benchmarked against the PCA, original MLP-ANN, FS, OWTNN and a recently developed optimization technique, SGNO. Each of the input selection techniques, such as the PCA, FS (GA-ANN) and SGNO, are set to 21 features, the same as those for the proposed model, and each of these selected features is then used in the ANN in plotting the performance in terms of regression and MSE.

By referring to each of the benchmark analysis results, it can be seen that the TSH model outperforms the entire benchmarked model in terms of prediction, by regression value only. However the MSE value shows quite different results, where PCA and SGNO gives lower final errors than the proposed model, but only show a slight difference of between 0% and 1%, compared to the TSH model.

## References

- BHENDE, M. J. & VENKATARAM, J. V. 1994. Impact of diversification on household income and risk: A whole-farm modelling approach. *Agricultural Systems*, 44, 301-312.
- CGIAR. *A Global Agriculture Research Partnership* [Online]. Available: <http://www.cgiar.org/who-we-are/> [Accessed 24/09/12 2012].



- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In:* DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). DEFRA.
- FOSTER, D., MCCULLAGH, J. & WHITFORT, T. Year. Evolution versus training: an investigation into combining genetic algorithms and neural networks. *In:* Neural Information Processing, 1999. Proceedings. ICONIP '99. 6th International Conference on, 1999 1999. 848-854 vol.3.
- ICRISAT. *International Crops Research Institute for Semi-Arid Tropics* [Online]. Available: <http://www.icrisat.org/Icrisat-aboutus.htm> [Accessed 24/09/12 2012].
- JOLLIFFE, I. T. 2002. *Principal Component Analysis*, New York, Springer.
- MUHD KHAIRULZAMAN ABDUL KADIR, E. L. H., SAHARUL AROF, DACIANA ILIESCU, MARK LEESON, ELIZABETH DOWLER, ROSEMARY COLLIER, RICHARD NAPIER, ARJUNAN SUBRAMANIAN 2012. Neural Network for Farm Household Output Prediction. *International Conference on Statistics In Science, Business And Engineering*. Langkawi, Malaysia.
- R.P SINGH, H. P. B. A. N. S. J. 1985. Manual of Instructions for Economic Investigators In Icrisat's Village Level Studies. Andhra Pradesh: ICRISAT, International Crops Research Institute for the Semi-Arid Tropics, Patancheru P.O., Andhra Pradesh 502 324, India.
- R.P. SINGH, B. L. J. A. V. B. R. 1984. Feature of traditional Farming Systems in Selected Villages of Madhya Pradesh. *Economics Program Progress report 61*.

- RAO, T. S. W. A. K. V. S. Yield and Net Return Distributions in Common Village Cropping Systems in the Semi-Arid Tropics of India. *Economics Program Progress Report*. Andhra Pradesh: ICRISAT.
- SARIN, D. J. A. R. An analysis of levels, patterns and determinants of fertilizer use on farms in selected regions of Semi-Arid Tropical India *Economics Program Progress Report*. Andhra Pradesh: ICRISAT.
- SKOUFIAS, E. Risk and seasonality in Empirical Model of the Farm Household. *Journal of Economic Development*, 19, 93 - 116.
- SKOUFIAS, E. 1993. Seasonal Labor Utilization in Agriculture: Theory and Evidence from Agrarian Households in India. *American Journal of Agricultural Economics*, 75.
- SMITH, L. I. 2002. A tutorial on Principal Component Analysis. Available: [http://www.cs.otago.ac.nz/cosc453/student\\_tutorials/principal\\_components.pdf](http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf) [Accessed 26 February 2002].
- SWINGLER, K. 1996. *Applying neural networks : a practical guide*, London : Academic Press.

# **Chapter 5: Trends in global output per capita prediction and modelling using FAOstat, USDA and World Bank database**

## **5.1 Introduction**

Each key component of food security consists of many indicators, including sub-indicators and main indicators. In the previous chapter, one of these sub-indicators, farm household output has been predicted using the TSH model. In this chapter, it will be demonstrated that the TSH model can be employed in optimizing the ANN for predicting the global output per capita of food availability, one of the food security themes for 13 European countries.

In the DEFRA food security assessment, it is indicated that food quantity per capita is one of the main components which can affect food security patterns in terms of food global availability (DEFRA, 2010). Within this food security theme, there are 7 supporting indicators which yield growth by region: real commodity prices, stock to consumption ratios, share of production traded, concentration in world markets, R&D expenditure and impact of animal disease. If these indicators fail to be monitored, this can result in threats to the population and to economic growth, harvest shortages, breakdown in trade, lack of investment, global warming and a more volatile climate (DEFRA, 2010, Tacio). These factors will continue to affect every one of 9 billion

people globally through to 2050, as stated in the People and Planet report by (Tacio). Therefore, in attempting to prevent this behaviour happening through to 2050, a representation of each sub-indicator as the feature of predicting the food growth per capita will be explained in a later section. This prediction can be used as a stepping stone in monitoring the security of our foods availability for the future.

## **5.2 Background**

Food quantity per capita studies have been made empirically and theoretically by researchers and economists. However there have been very few studies on the relationship between the food availability and the trends in global output per capita in food security, especially in terms of its prediction for further analysis of future trends.

In 2011, a report on the state of food insecurity in the world was prepared by multiple organizations. It considers food security in its entirety, and this third edition report is more concerned with the price of food, which can affect people's lifestyle and ability to acquire quality and nutritious food. One of the key messages of the report was that high food prices can worsen food insecurity in the short-term (FAO, 2011). This behaviour involves the production prices or values, which is indicated in (DEFRA, 2010) as one of the sub indicators for food growth per capita, and is included as one of the features used in predicting it.

In the report by (DEFRA, 2010, DEFRA, 2009), it is also stated that global output per capita can be affected by multiple external and multiple internal behaviours, such as: calories needed by each person globally, food production per capita change, food production and agriculture production stock difference,

technology change for agriculture, cereal usage by various people and animals, net harvest lost caused by environmental issues or disease, and increases in animal changes. Although the need for food to be available to all without any constraint is important, the amount of food being wasted by each household can also affect the overall output per capita in terms of food security, due to an overstated actual consumption analysis, as described in (Nelleman, 2009, Godfray et al., 2010).

In 2050, the population around the world is expected to increase to 9 billion people. However, by 2010 there was already an increase of 35% in the projected population levels (DEFRA, 2010). The FAO also state in their reports that, to ensure enough food for everyone, including a more urban and richer population, all food production needs to be increased by 70% (FAO, 2009). It seems that the increase of population in each country or region will have a major impact on food usage. Although dietary changes have had effect on consumption, from eating a smaller variety of crops to consuming more meat and dairy products (DEFRA, 2010), overall the need for any kind of food is still import, in ensuring sufficient food and maintaining food security.

### **5.3 Dataset**

The dataset had been compiled based on (DEFRA, 2010) food security themes and sub-indicators, with the main indicator being the food availability. The compilation of this data is from the FAOStat online database (FAO, 2012), World Bank online database (Bank, 2012) and United States Department of Agriculture (USDA, 2012) where all of these datasets are available online. The idea of this compilation is to show the prediction of the growth per capita of food which can be

used as a stepping-stone in monitoring food security. At the same time, it can be used to plan the prevention of any problems relating to food, especially there not being enough food for everybody. Among the European countries, 13 countries have been selected from well studied countries, to be represented in the dataset, from 1969 to 2007. The crops being considered in this dataset are cereals, because they are widely used in Europe as the main food, either for production or as end products.

Cereal can also be called grain, and the term is interchangeable in many publications. It can have various definitions, for example in (Chapman, 1976), it is defined as a grass grown for its small and edible seed, but Lantican (2001) states that cereal or grain crops belong to the grass family, and are used mostly as staple foods (Bareja, 2010-2012). In the FAO report, cereal is defined as a combination of any plant resembling grass that produces grains, which are used either for seed, food, production and feed, such as maize, corn, rice, wheat, barley, oats and rye.

These cereal datasets have 18 features and 507 samples, including: yield, real commodity prices, stock to consumption ratio, share of production trade, concentration of cereal in world market and the country's population. The details of the datasets are shown in Table 5.1. As described earlier, the input variables were based on sub-indicators of the food availability theme in food security (DEFRA, 2010). In each sub-indicator was assumed to be in relation to the main indicator of food supply per capita as the output variables.

Table 5.1: Features categories for trends in global output per capita

Input Variables	Output Variable
<p style="text-align: center;"><u>Yield growth</u></p> <p>Area Harvested (Ha), Yield (Hg/Ha), Production (tonnes), Seed (tonnes)</p>	<p>Food supply growth per capita (kg/capita/yr)</p>
<p style="text-align: center;"><u>Producer and production prices / value</u></p> <p>Gross Production Value (1000 Int. \$), Net Production Value (1000 Int. \$), Gross Production Value (SLC), Gross Production Value (USD), Producer Price (Local Currency/tonne) (LCU)</p>	
<p style="text-align: center;">Food supply quantity (tonnes)</p>	
<p style="text-align: center;"><u>Share of food trade</u></p> <p>Import Quantity (tonnes), Import Value (1000 \$), Export Quantity (tonnes), Export Value (1000 \$)</p>	
<p style="text-align: center;"><u>Concentration of world market</u></p> <p>Raw materials exports (% of merchandise exports), Agricultural raw materials imports (% of merchandise imports), Food exports (% of merchandise exports)</p>	
<p style="text-align: center;">Population</p>	

In Table 5.1, the yield growth categories represent all of the relevant components, such as harvested area, production of cereal crops, seed quantity being used and crop yield itself. Each of these features were selected and assumed as the key factors towards giving a good productivity and played an important role in increasing the food supply entirely (DEFRA, 2010).

For production and producer prices or values, this mainly reflects the cereal commodities, showing either shortages or supply quantities itself. In Global Economic Prospects 2009, it is expected that commodity prices will be decline until 2030, showing either the improved supply or oversupply (WorldBank, 2009). This means a weaker demand for growth will give more time to develop any unused land for future agricultural use if no problems in temperature factors or water supply exist in the future (DEFRA, 2010, DEFRA, 2009). The unit of this producer and production value is based on the US dollar and Standard Local Currency (SLC), where each value was referred to the output prices at the farm gate (FAO, 1986-2007).

To represent the stock to consumption ratio of cereal, the outcome is provided based on the food supply by the production of a cereal end product, in this case, Maize oil, which excludes any type of beer or wine. However, certain European countries did not provide data on Maize oil production because they do not produce it via cereal, so the food supply dataset shown in Table 5.1 entirely is assumed to be representative of this indicator. It is hard to identify the optimal stock ratio due to drastic changes in the production quantity over the years, and policy changes which have contributed to reducing the cereal stocks for each of the 13 European countries as discussed in (DEFRA, 2010).



In terms of the share of food trade represented by cereal, the units presented here are in US Dollars, represented by 1 unit being 1000 US Dollars, and rounded to the nearest 1000 Dollars. The sub indicator will show the imports and exports made by each of the 13 European countries to any of the other 13 countries, or from other countries around the world. The trade not only involves trading between these countries, it also involves investment in agriculture in other countries. However the investment factor is being neglected in this case due to the data constraints.

In the case of concentration on commodity markets for the cereal, it is assumed that the markets consist of the raw material of import, raw material of exports and food exports. Each of these features is based on the percentage of the merchandise compared to the raw materials.

Finally, the population of the 13 countries is also included in the datasets, because each person needs to have access to food, as stated by FAO (FAO, 2006), so, each of the available indicators need also to be related to the number of people that need the food, to ensure sufficient food growth per capita. In (Godfray et al., 2010) it is also stated that the population will always affect the trend for using the food in terms of its consumption; either the food is totally used or it just wasted as described previously.

In considering the food availability, there are still two additional sub-indicators that should be considered; the agriculture research spending and the impact of animal disease. However, due to constrain in obtaining the data, it is being assumed that these factors will not affect the overall growth of food per capita.

In this chapter also, the TSH model will be used to study and explore the relationship between the trends of global output per capita and food global output per

capita for cereals, by each demand growth trends category, which act as supporting indicators in (DEFRA, 2010, DEFRA, 2009) for 13 European countries: United Kingdom (UK), Sweden, Spain, Portugal, The Netherlands, Italy, Ireland, Greece, Germany, France, Finland, Denmark and Austria. As mention earlier, the dataset will consist of 18 X 507 (39 X 13) samples over 39 years and 13 countries, where each of the 18 features shown in Table 5.1 contains the demand growth trends of each country for each year.

#### **5.4 Data pre-processing**

Table 5.2 lists the basic statistics of the data with the actual units given in Table 5.1 of the 18 input variables. It can be seen that the table consists of a range of data with different units. Generally, in any kind of IS technique or model, it is easy for the model to process these datasets. However, in terms of finding the local minimum solution, it will take some time and the performance of the model will be different each time the training process is started because of the random weight and threshold initialized by ANN. For example, in the ANN, if the number of input variables in the network is increased, it will also increase the network size, which leads to a higher computational cost, especially in estimating the weights connection efficiently (D'Heygere et al., 2003).

As described in the previous chapter, each data set will be standardized to have a zero mean and unit variance by following the equation in (3.1). This will make the process of FS and optimization faster for each generation of GA, and for each ANN epoch in finding the solution.

As indicated earlier, each of the 18 input variables will be represented as the GA chromosomes in the TSH process, and it also will be used as the input variables for the ANN, either in the fitness functions of the GA or in remodelling the ANN.

Table 5.2: List of basic statistic for input variables

<b>Variables</b>	<b>Mean</b>	<b>Minimum</b>	<b>Maximum</b>	<b>Standard Deviation</b>
<b>Area Harvested</b>	3064338.69	170284.00	9893542.00	2949557.95
<b>Yield</b>	44155.36	8908.00	84109.00	16986.48
<b>Production</b>	13521292.69	789561.00	70516553.00	15514201.46
<b>Seed</b>	495894.18	32000.00	1618846.00	495894.18
<b>Gross Production Value1</b>	1916801.19	128041.00	10383179.00	2238372.75
<b>Net Production Value</b>	1847475.74	120034.00	10211190.00	2185307.47
<b>Gross Production Value2</b>	2257.08	117.00	7898.00	2238.44
<b>Gross Production Value3</b>	1921.86	146.00	9161.00	2080.82
<b>Producer Price</b>	191625.74	107.00	4000000.00	619553.03
<b>Food supply quantity</b>	3089621.58	369565.28	10599600.55	3085234.62
<b>Import Quantity</b>	2915013.10	17173.00	14388106.00	3181949.53

<b>Import Value</b>	543717.07	4888.00	3106985.00	573978.45
<b>Export Quantity</b>	3310886.38	55.00	34859028.00	6605249.27
<b>Export Value</b>	576581.07	20.00	6695291.00	1153440.58
<b>Agricultural raw materials exports</b>	3.99	0.42	25.95	3.95
<b>Agricultural raw materials imports</b>	3.57	0.84	11.92	1.97
<b>Food exports</b>	13.49	1.64	50.61	10.76
<b>Population</b>	27316055.48	2925600.00	82504552.00	25843485.59
<b>Food supply growth per capita</b>	111.79	65.60	189.40	26.11

For the gross production, values 1, 2 and 3 represent the gross product values where 1 equals 1000 units of international currency, value 2 same as value 1, and value 3 is in United States Dollars (USD). International currency is representing as the standard currency used in each of the 13 European countries.

## 5.5 First stage process

The purpose of this stage is to decide which features give the greatest impact on the outputs prediction of the food growth per capita. As described earlier, the features consisting of the trends of food availability have been assumed and determined based on the sub-indicators of the food growth per capita of cereals as described in (DEFRA, 2010). In achieving this objective, the combination of the GA module and the ANN module is developed. These combinations protect against the

weaknesses of the ANN, such as the fact that the ANN can only perform a local search (Kitano, 1990, Whitley et al., 1993), and the problem that the ANN often experiences in generalization problems (over fitting). Therefore the GA tries to find the best solution randomly, by searching all global regions, based on using the training errors of the ANN module as its fitness function.

The ANN module uses a three-layered network, where the number of layers, the activation function for each layer and the learning function will be the same as those used in the previous chapter. This is also applied to the GA module, where the fitness function, selection function, crossover function and mutation function used are the same as those in the previous chapter, and it is terminated when it reaches 50 generations. The only difference in this stage, in predicting the food per capita, is the mutation probability ratio, which depends on the size of chromosomes and population, see (Salman and Ong Hang, 2008), as described in Chapter 3 Equation (3.2).

In this first stage of the TSH, the number of the selected variables depends on the lowest RMSE values that result from the global search, by the GA trying to find the optimum features. Each of the GA chromosomes is a representation of each feature variable of the model. The final bit string of the chromosomes is the lowest RMSE value, where a string of value '1' will be selected and a string of value '0' will be abandoned. In this case, the number of input variables is 18, so the chromosome will be 18 bits, which is '0101 0100 1011 1000 11'. By referring to the chromosomes, the features selected as having an effect on the output of the model are as below:-

Features selected = [Yield, Seed, Net Production Value, Producer Price, Import Quantity, Import Value, Export Quantity, Food exports, Population ]

Figure 5.1 shows the performance of the first stage model in input selection by each generation. The chromosomes fitness value by each generation is shown as a blue dot and the red line represents the mean of each chromosome in that generation.

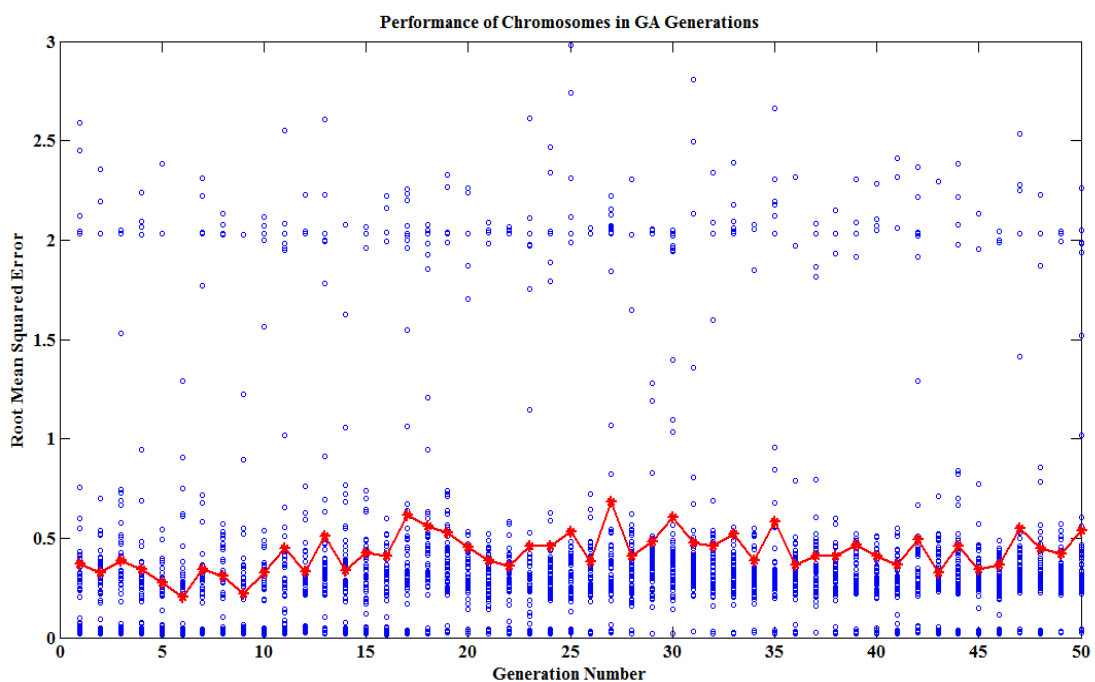


Figure 5.1: First stage performance via GA generation

## 5.6 Second stage process

After getting rid of the unwanted features in the first stage of the model, the selected input variables will be used in this second stage. In this stage, the GA module is used to optimize the ANN weight and threshold. Therefore the chromosome determinations are dependent on the number of inputs, number of layers and number of neurons in each layer, as shown in equation (3.7).

The parameter for the GA module and ANN module will be the same as the first stage of the TSH model. If in the first stage it concentrates on finding the optimum feature variables, in this stage it is instead focussing on finding better solutions in the local search region, and trying to prevent the generalization of the ANN by using optimum weight values and optimum threshold values. Usually, the original ANN will randomly generate its own weight and threshold and this often tends to give poor generalization, especially when it involves a very large and complex dataset. In this case however, each weight value and each threshold value is fixed and an optimum value for the ANN.

Figure 5.2 shows the GA-ANN process for finding the optimum value for threshold and weight. The line curve shows the decreasing fitness value, which it maintains to the 50<sup>th</sup> generation - this is different to the first stage. This is because the training being done is directly affecting the overall network, which also means that the GA successfully applied the selected input, with optimum weight and threshold values, into the ANN network. The fitness value in this stage is represented by the MSE value of ANN by Equation (3.4).

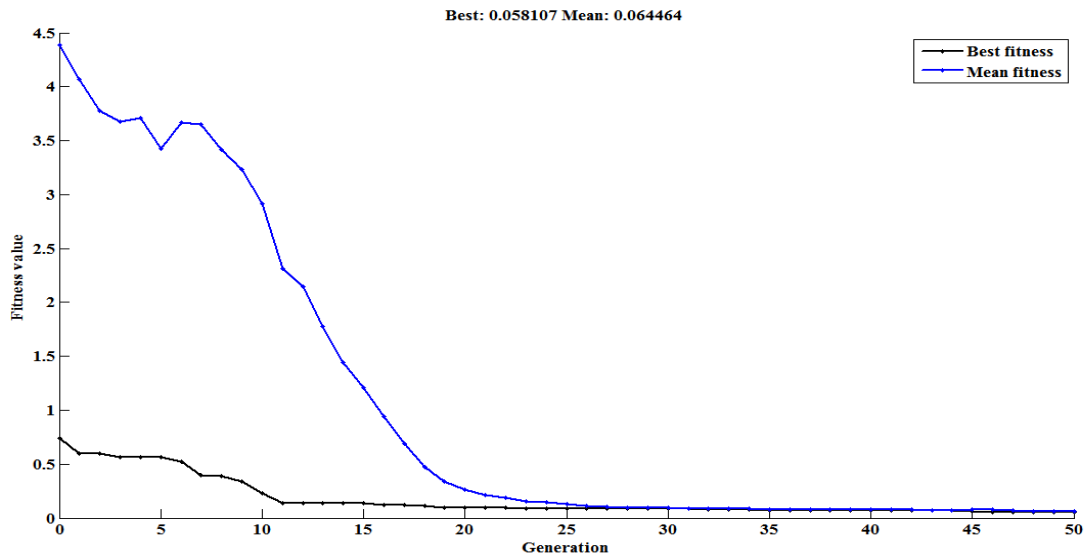


Figure 5.2: Performance of second stage via number of generation

As previously stated, all of the parameters for this stage use the same parameter as the first stage of the TSH. This also applies to the size of the hidden neurons in the ANN module. In Chapter 3, Equation (3.7), the sizes of hidden neurons were described as dependent on the total number of input and output variables, which are also applied to the same size of neurons for each stage of this model. Generally, in this stage, the number of hidden neurons should also be reduced due to the reduction of the number features being selected in the first stage process. To maintain the performance of the feature being selected in the first stage, the second stage hidden neurons will also be using the same size, ensuring that the optimum values of weight and threshold are achieved.

## 5.7 Remodelling ANN

Next, all of the selected input variables, the optimized threshold and the optimized weight will be used in remodelling the ANN. The same ANN parameter is



used as in the second stage process, using a three layer network – one input layer, one hidden layer and one output layer. In addition, the ANN is remodelled using the same hidden neurons, using the same division of datasets – 60% for training, 20% for validation and 20% for testing, using the LM learning method algorithm. The same activation function – the tangent sigmoid function – is used at each layer, which is capable of finding the solution in wide range of  $-1$  to  $1$ . The number of iterations of the ANN was set to 1000 epochs, but it will terminate when the result of the validation error check reaches 0.01 and this occurs 6 times. The purpose of the validation error check is to prevent the over-fitting of the network by stopping the training at the minimum of the validation error (H. Demuth, 2004).

To measure the performance of this model, Figure 5.3 and Figure 5.4 show the regression performance plot and the MSE performance plots respectively. In terms of the regression, the TSH model gives a value of around 0.99967, which equates to an error almost as low as 0.033%. This also applied to Figure 5.4, where it takes until 52 epochs for the ANN to stop searching for the solution, and shows an error of 0.0004. This shows that the model gives great accuracy in predicting the cereal food growth per capita, based on the trends of its activities or behaviour.

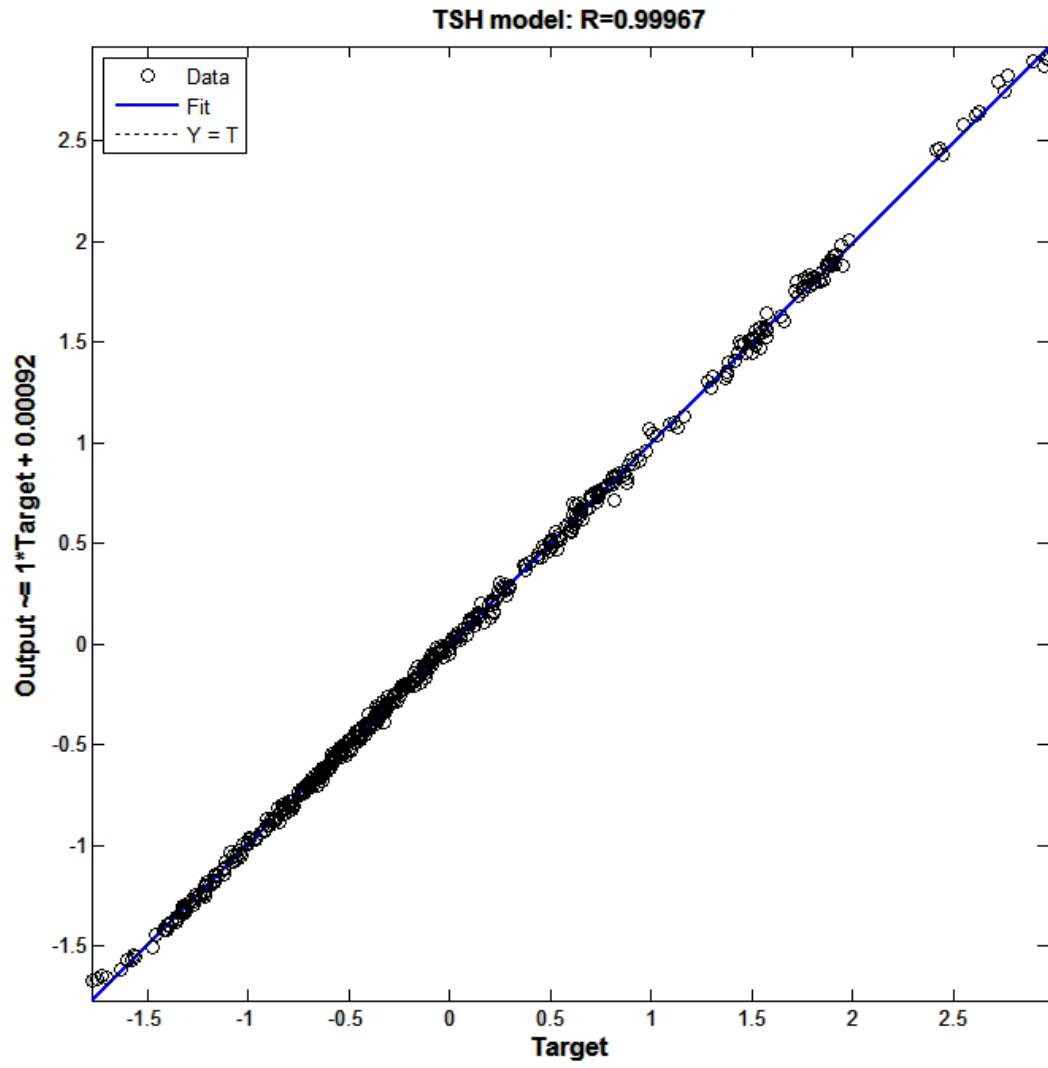


Figure 5.3: Regression of TSH model

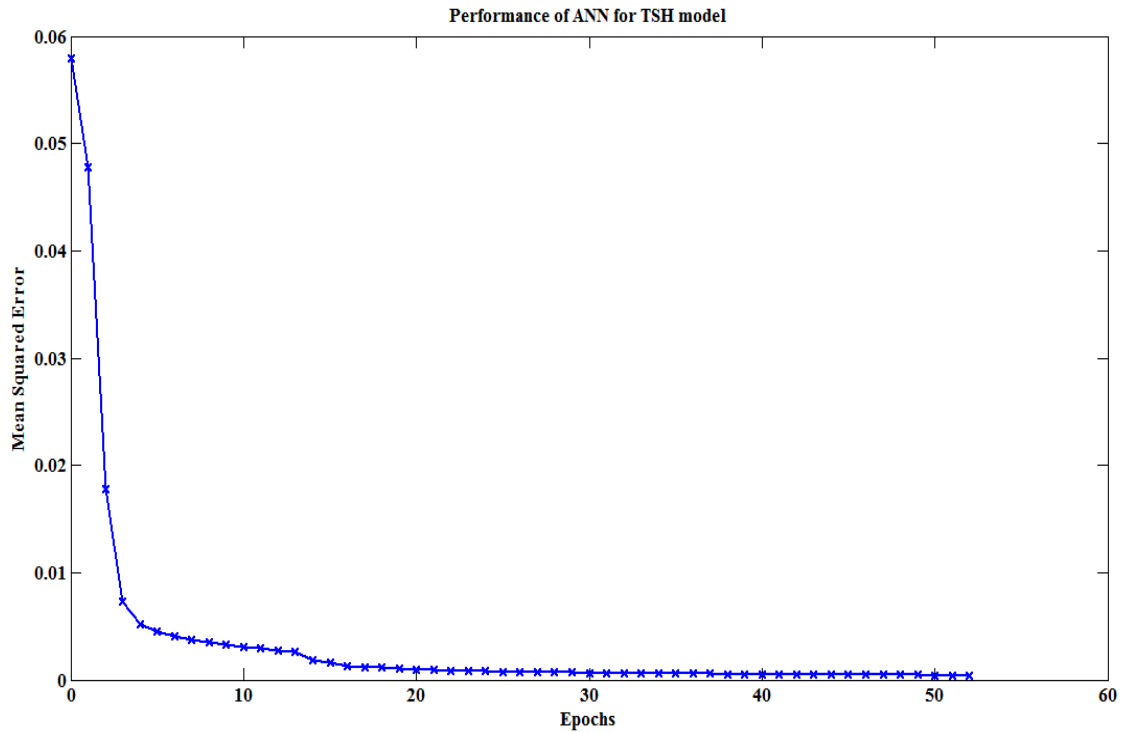


Figure 5.4: Performance of ANN based on MSE vs Epochs

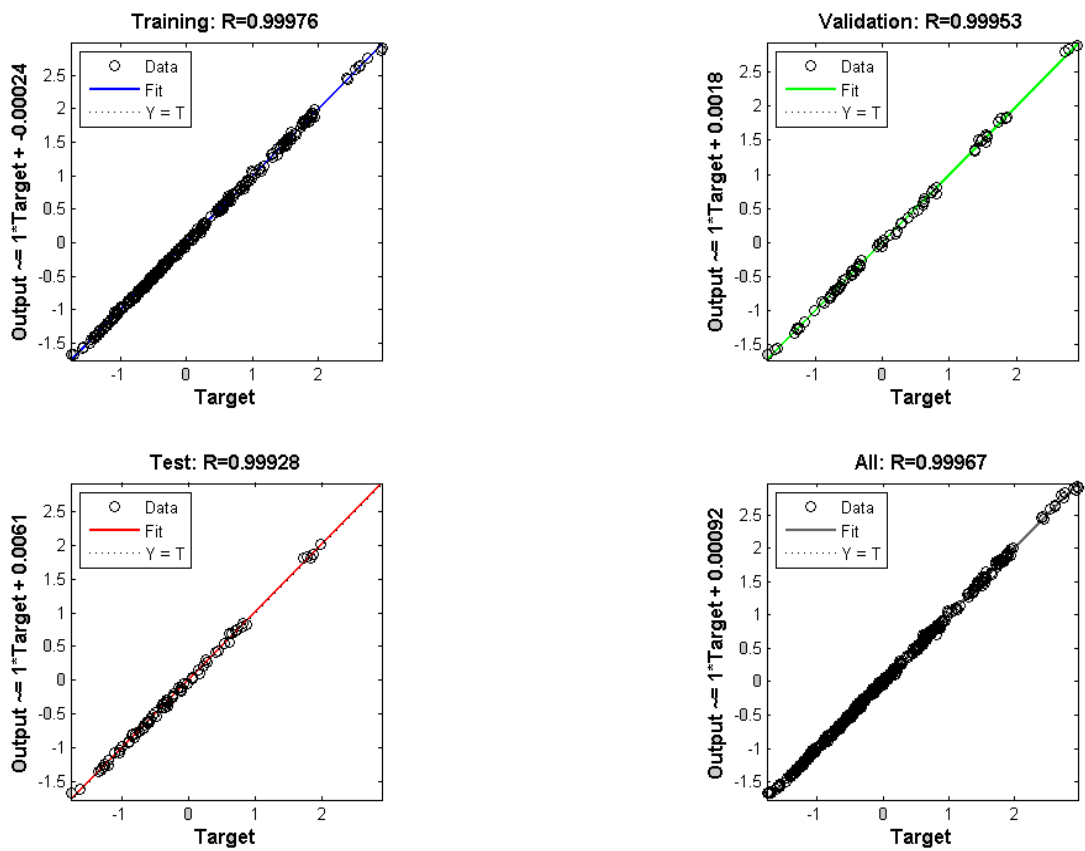


Figure 5.5: Overall regression for each part – training, validation and testing

Figure 5.5 shows the performance of each part in terms of the division of the datasets into 60% training, 20% validation and 20% testing. It also shows that the ANN achieves very good generalization, where most of the overall regression percentage is around 99.9%, and all data fits through the line between the actual output and the TSH model output.

## **5.8 Benchmarking and discussion**

In this section, the TSH model is compared with PCA, the original MLP-ANN, FS (GA-ANN), OWTNN and SGNO. This is done by evaluating the performance of the prediction output of the ANN model against each selected feature or network optimization technique.

### **5.8.1 Principal Component Analysis (PCA)**

PCA is a well known statistical method for data reduction, especially for highly dimensional data. The method is described in detail in Chapter 3. In this chapter an attempt is made to develop a prediction model for food growth per capita, based on the indicators in (DEFRA, 2010), consisting of 18 original variables. Each of these variables will be created in a new data space, with new data input variables, by using PCA. The newly created data will then be used to predict the food growth per capita using an ANN as the prediction model. Figure 5.6 shows the percentage variance of the principal components, where the thin blue line increases to more than 80% at the first four components, and the first principal components already show over 50% variance in the original dataset compared with other principal components. As seen in the previous chapter, the data transformed using PCA will be used as input

variables for further prediction analysis, with the first nine components selected to be the same as the proposed prediction model.

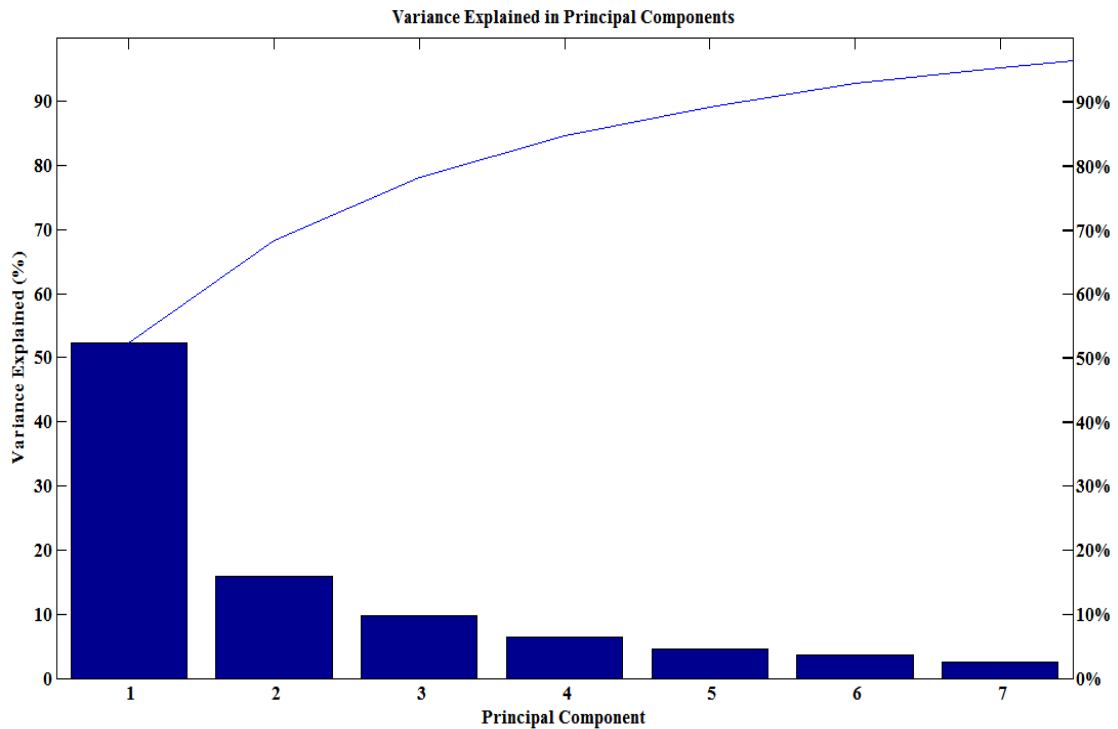


Figure 5.6: Variance in Principal component

Figure 5.7 shows that the new PCA data gives quite a good regression value of 0.98, and at the same time, it tends to give good generalization in all data divisions – training, validation and testing. However, the TSH model still gives better generalization and the prediction outcome is better than the PCA. In the figure shown, it seems that in the data testing division, the data almost does not fit to the line, and the red line tends to move outside of the  $Y = T$  region. For the MSE performance, as in Figure 5.8, it can be seen that the proposed model gives lower errors in the starting epochs, but both errors achieve almost the same MSE after the 10<sup>th</sup> epoch. In the final result, the proposed model takes a longer time to reach the final solution, at the 52<sup>nd</sup>

epoch. However, the error given by both models shows that the proposed model outperforms the PCA by using the new data components as the inputs to the ANN.

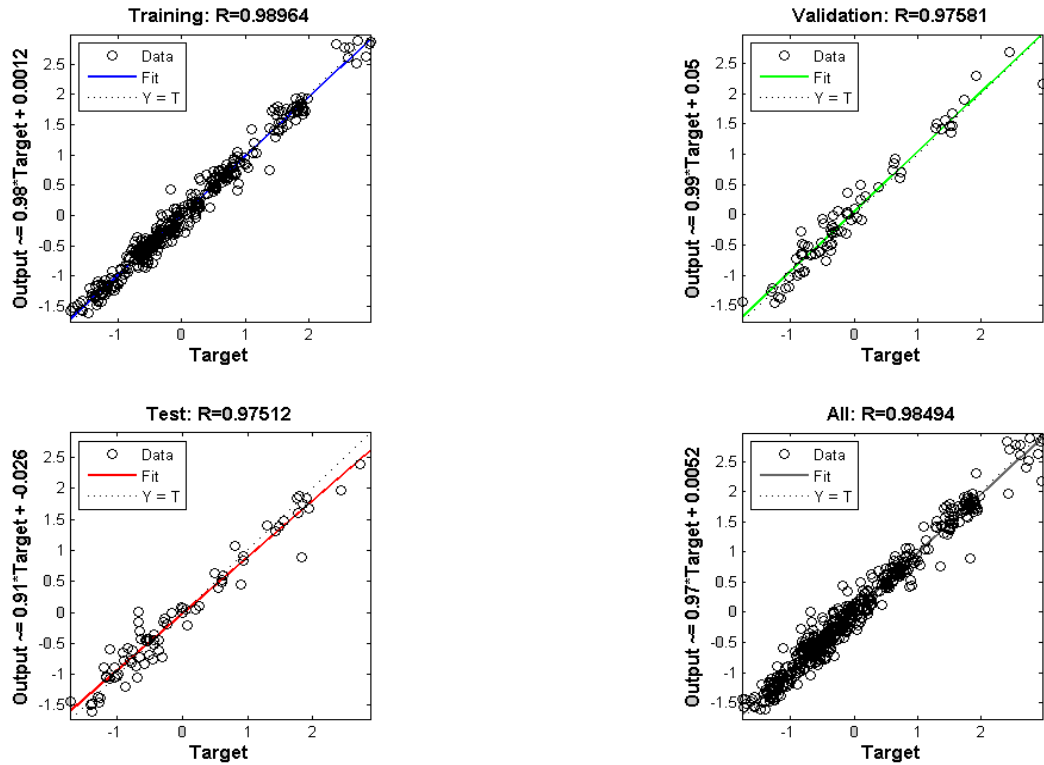


Figure 5.7: Regression for PCA using ANN

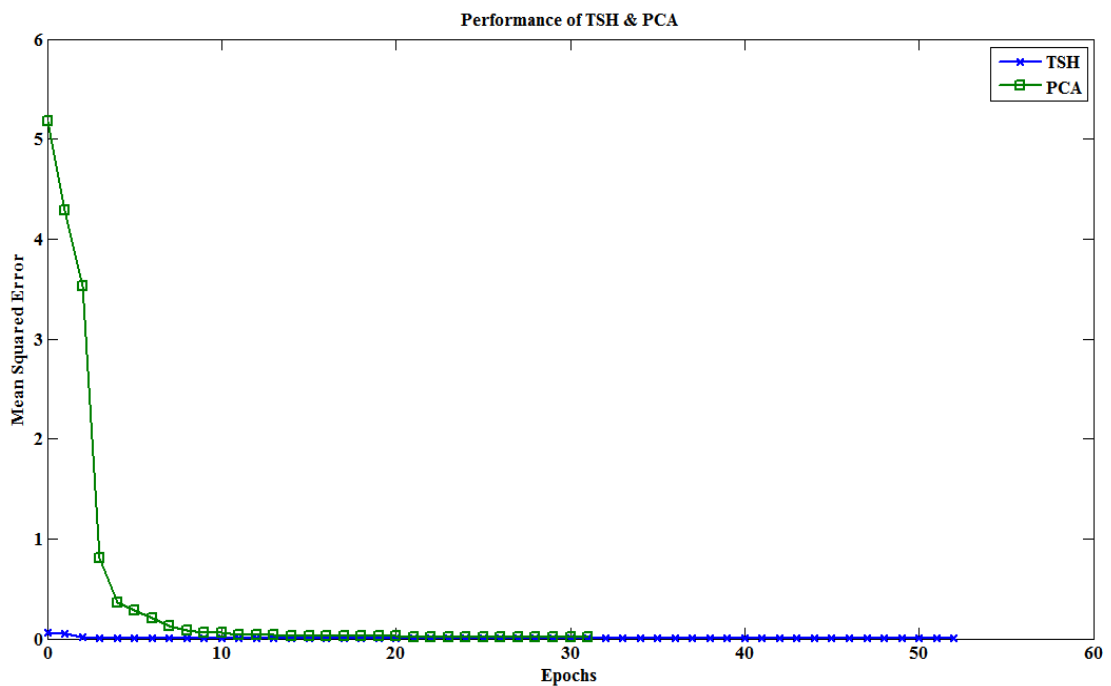


Figure 5.8: MSE between PCA and TSH model

### 5.8.2 Artificial Neural Network (MLP-ANN)

As one of the popular prediction models indicated in (Eduard and et al., 1999, Gardner et al., 1992, Negnevitsky, 2005, Gardner and et al., 1990), MLP-ANN is used as one of the benchmarks for this model. The parameters of the ANN are the same as for the ANN module in the TSH model, which consists of a 3-layer network (input layer, hidden layer and output layer), the hidden neuron as in Equation (3.6) - in this case it is 13, and using tangent sigmoid for all layers and LM as its learning method. The data used in the MLP-ANN comes purely from the datasets of 18 features with 507 samples, without any data pre-processing, in predicting the food growth output per capita.

The performance of the original MLP-ANN prediction model gives a 0.99837 regression value, which is quite good. The overall performance of each part of training, validation, testing and the overall regression is shown in Figure 5.9. In terms of generalization, it also shows a good generalization for each division of the dataset. However, in training and validation, it shows some data at the end of the plot for both divisions which are almost outside of the fit line. Overall, as described in an earlier paragraph, the regression result between the actual output and the MLP-ANN model output almost shows a perfect regression value of 1.

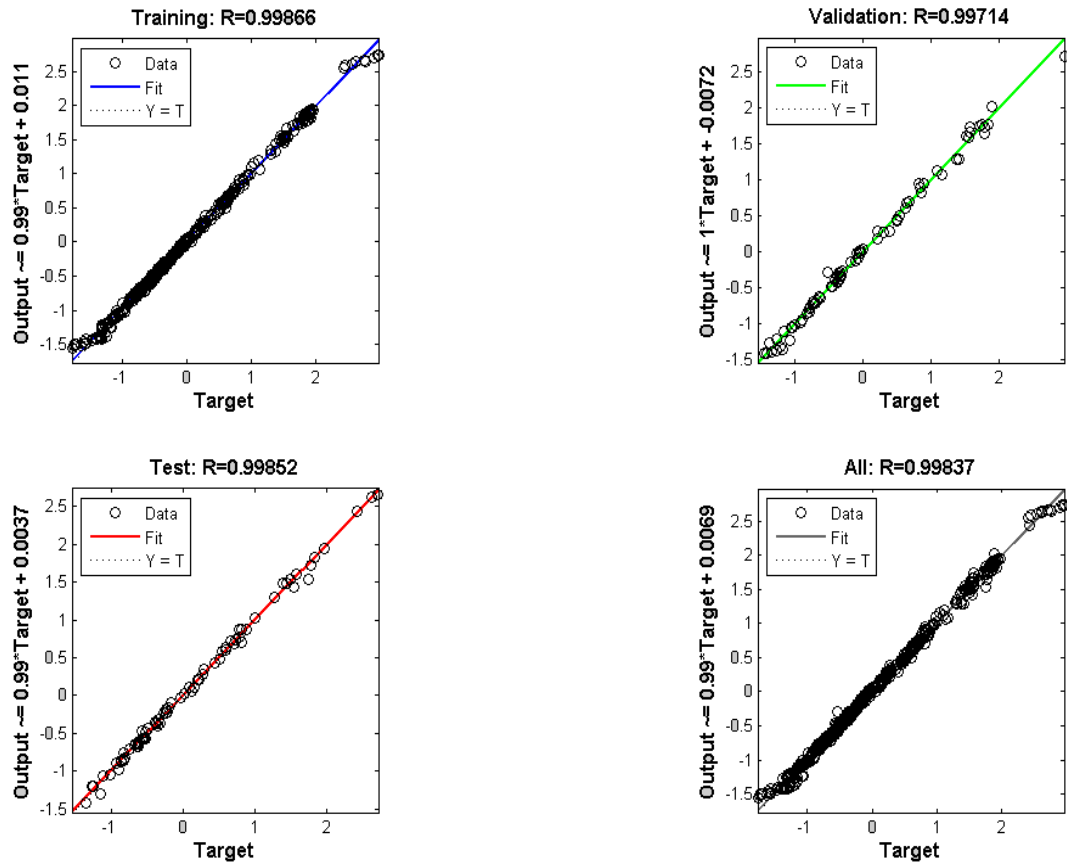


Figure 5.9: Regression for original ANN

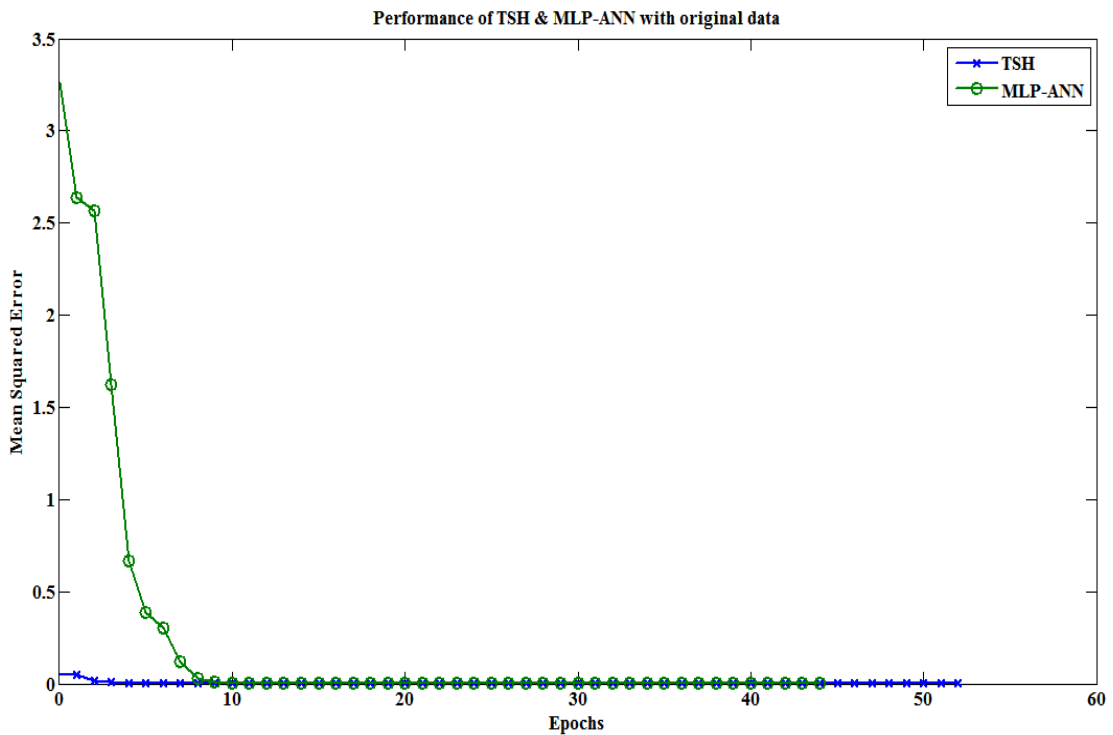


Figure 5.10: MSE for original ANN and TSH



As shown in figure 5.10, the MSE starts with quite a high value, compared with the TSH model, but it seems to gradually decrease as the number of epochs increases. The difference in the MSE value is quite large - 3.256 in the first epoch. Both models have almost the same MSE at the 9<sup>th</sup> epoch. Finally, the TSH model took some time to complete its final local search, at epoch 52, with an MSE value of 0.0004, compared with a 0.002 MSE at epoch 44 for the MLP-ANN prediction model. The proposed model has a lower error than the MLP-ANN, which differs by around 0.067%.

### **5.8.3 Feature selection (GA-ANN)**

In this benchmark, a GA algorithm is used to globally search for the best inputs for the ANN, having the same process as the first stage of the proposed model. After this, the ANN will use the selected inputs to predict the cereal growth per capita. In making judgements of its performance, the same parameters as used in the TSH model for the GA and ANN models will be used. Generally, the first stage of TSH input selection will be used directly in making predictions using the ANN for the cereal growth per capita. As described previously in section 5.7, there are 9 selected features - this will be the same in this section.

As shown in Figure 5.11, the results are shown as a regression plot for all data divisions – training, validation, testing and overall dataset. The result shows a good regression performance with 99% accuracy for all parts, which is the same as for the MLP-ANN benchmark. It also shows a very good generalization when compared with to MLP-ANN, where all of the data fit to the  $Y=T$  line.

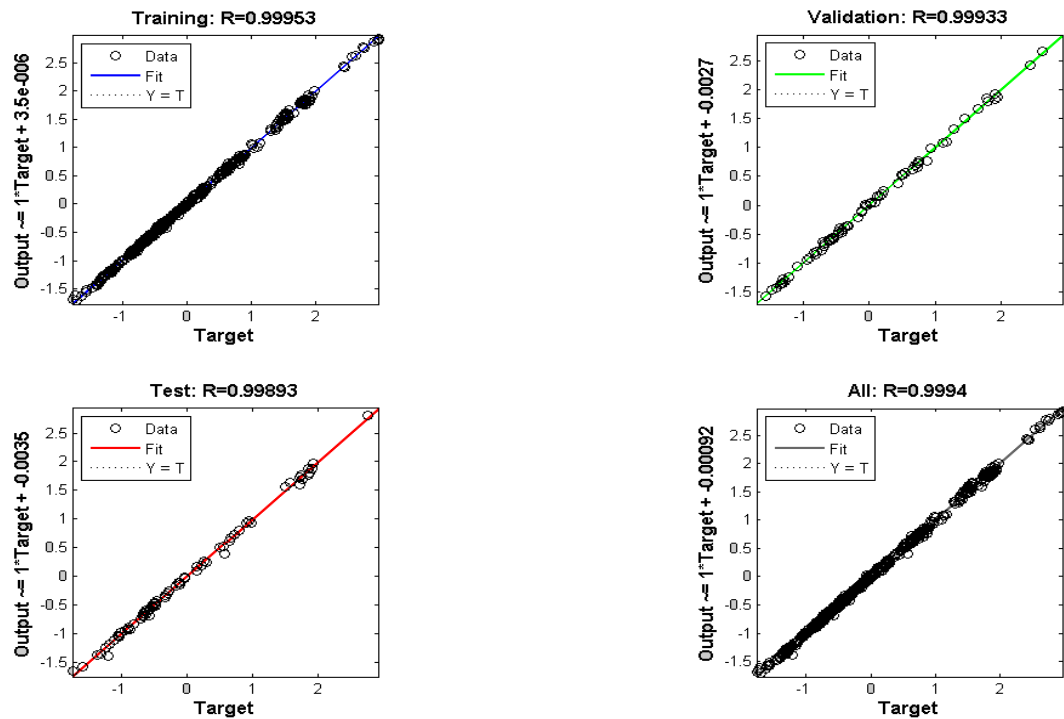


Figure 5.11: Regression for feature selection using ANN

In terms of errors, the MSE plot is used to show the error performance for each epoch, for the TSH model and the FS model. The MSE curve pattern shows the same pattern as the MLP-ANN, which starts at 4.712 in the first epochs, 0.058 higher than the TSH model. The MSE values drastically decrease as the epochs increase. The MSE maintains the error value for quite some time, until it reaches its final solution at the 58<sup>th</sup> epoch. This differs from the TSH prediction model, which ends at the 55<sup>th</sup> epoch, and it still shows a lower MSE value than the prediction model, using GA-ANN selected features only.

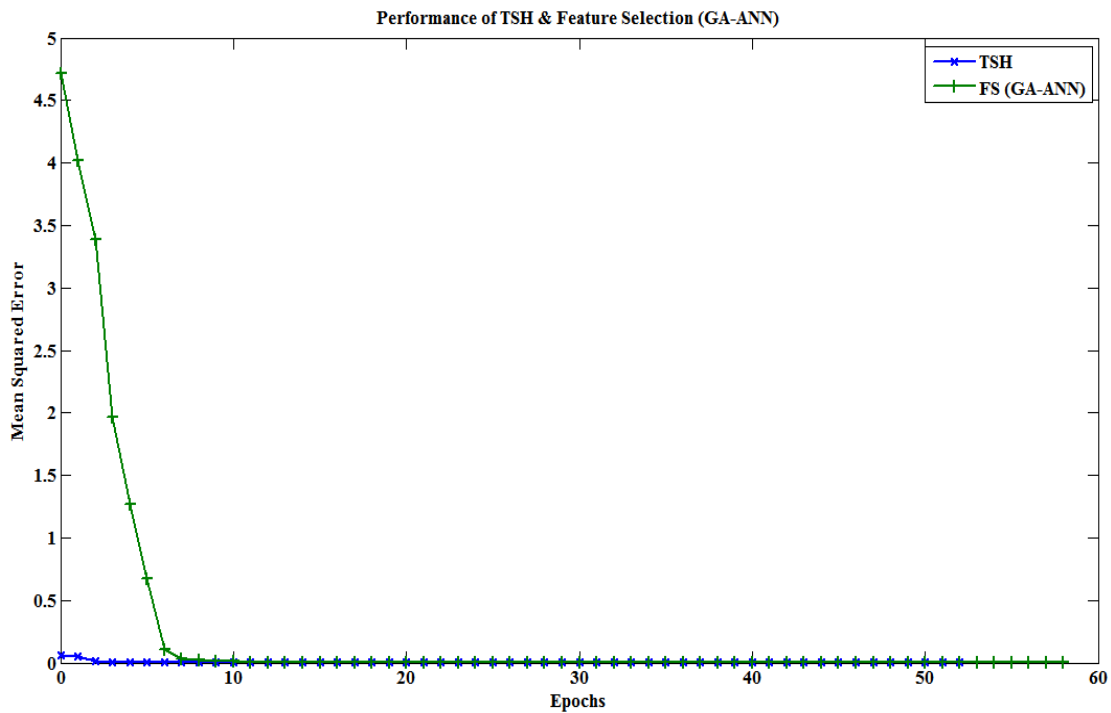


Figure 5.12: MSE for feature selection and TSH using ANN

#### 5.8.4 Optimized Weight and Threshold Neural Network (OWTNN)

When considering the TSH model, each stage needs to be compared and benchmarked. In this section, the performance of the OWTNN compared with the performance of the proposed model, with both having the same parameters in the GA module and the ANN module. The only difference is that the OWTNN will be based on the 18 inputs of the datasets and then remodel the ANN by using the optimized weights and thresholds, to compare its performance in the prediction of cereal growth output per capita.

In the same way as in the other benchmark sections, the performance of the OWTNN is shown in Figure 5.13 for the regression plot performance and Figure 5.14 for the MSE plot performance. For the OWTNN, the regression shows almost the same performance as the MLP-ANN, FS (GA-ANN) and TSH models, but also much

closer to that of the TSH model in terms of the regression value and errors. The OWTNN model terminated at epoch 51 and the TSH model ended at epoch 52.

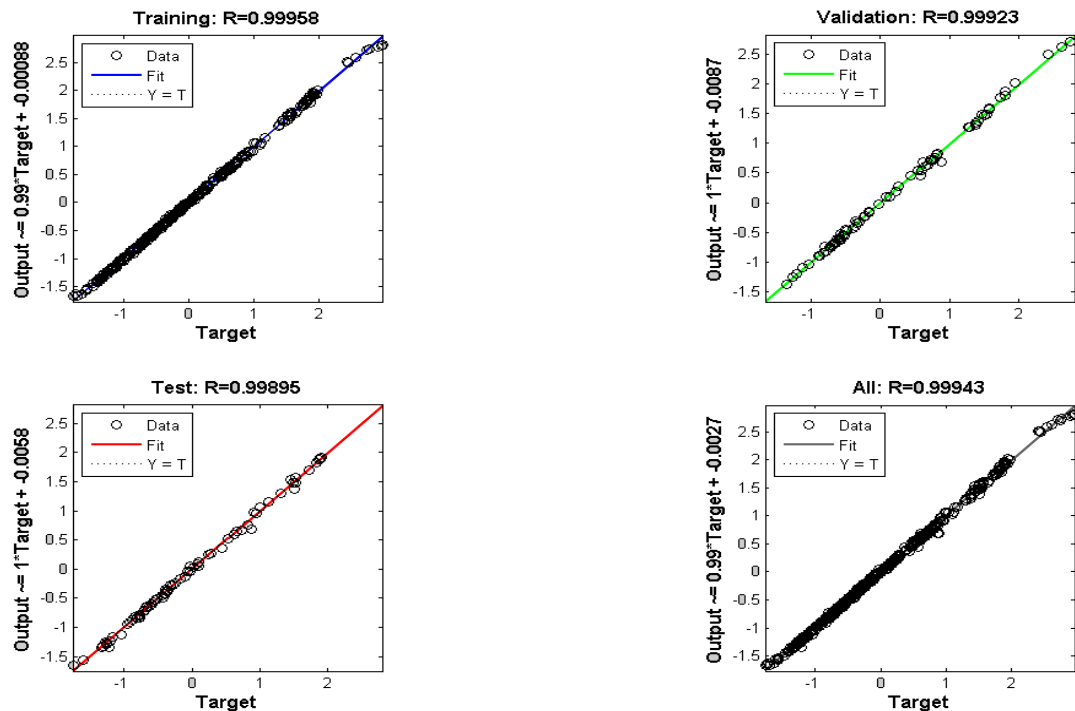


Figure 5.13: Regression for OWTNN using ANN

In comparing the MSE, initially the OWTNN model shows a higher MSE value than the proposed model. However, both plots show a smooth decrease in errors and intersect each other at the 4<sup>th</sup> epoch, and the final error value shows that the proposed model gives a lower error than the OWTNN. The overall performance, in terms of the prediction regression plot and errors produced by both models shows a very small difference, which indicates that using OWTNN in the proposed model did provide some benefits in terms of the combination of its local search and global search capabilities.

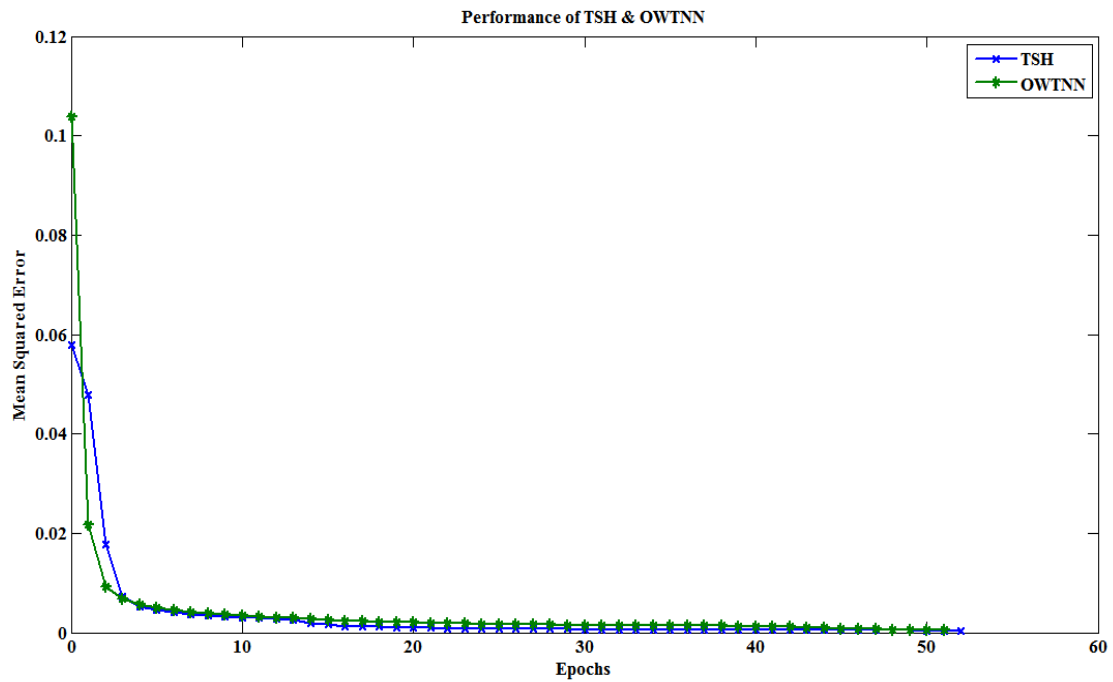


Figure 5.14: MSE for OWTNN and TSH using ANN

### 5.8.5 Sensitive Genetic Neural Optimization (SGNO)

As described in Chapter 3, SGNO is a combination of three modules – a GA module, an ANN module and a sensitivity analysis module (Zhang, 2011). The GA module uses an ANN module as the training function, where the RMSE values used as the fitness values are shown in Figure 5.15. In the figure, the line represents the mean value of the fitness function, and the blue dots represent the chromosome population. Both model, the SGNO and the proposed model, use a GA as the global search algorithm, with both models trying to find the best features for the ANN.

In earlier description, the combinations of the GA module and the ANN module have almost the same functions as the first stage of the TSH model, which is used for feature selection. The difference between these two models is in the pre-processing part, which includes the random five-fold cross validation for the dataset, and also the addition of the sensitivity analysis module for SGNO.

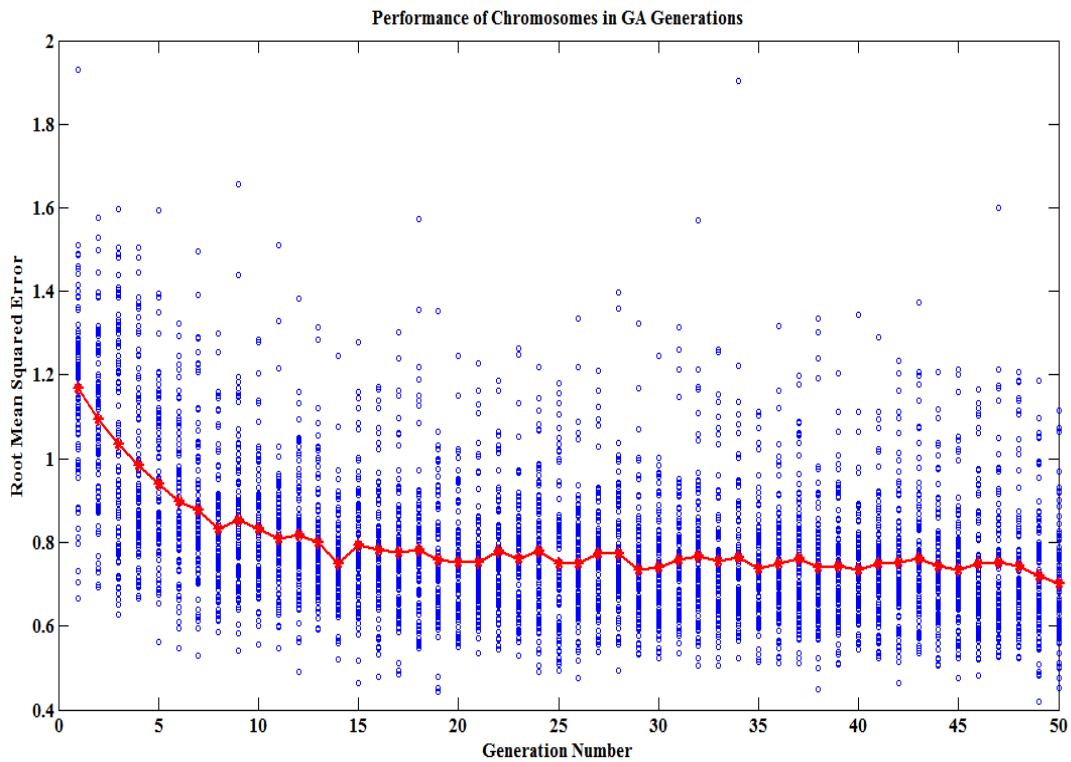


Figure 5.15: Performance of SGNO chromosomes via number of generation

In comparing the TSH model for the first stage process with SGNO, it can be seen that there is a totally different feature selection produced by each model. This possibly happens because of the thorough five-fold cross validation pre-processing operations, and also because of the sensitivity analysis performed in the SGNO model which gives a precise input variable selection.

Figure 5.16 is based on the mean of each feature frequency which appeared in the sensitivity analysis module. Generally, this module will re-evaluate each of the features commonly used by the GA-ANN, and rank them by their mean values. The highest mean value is the most important feature, and the lowest mean value represents the lowest ranking (Zhang, 2011). The rank of each of the features, shown in Figure 5.14, has been rearranged using the feature variables numbers in Table 5.3.

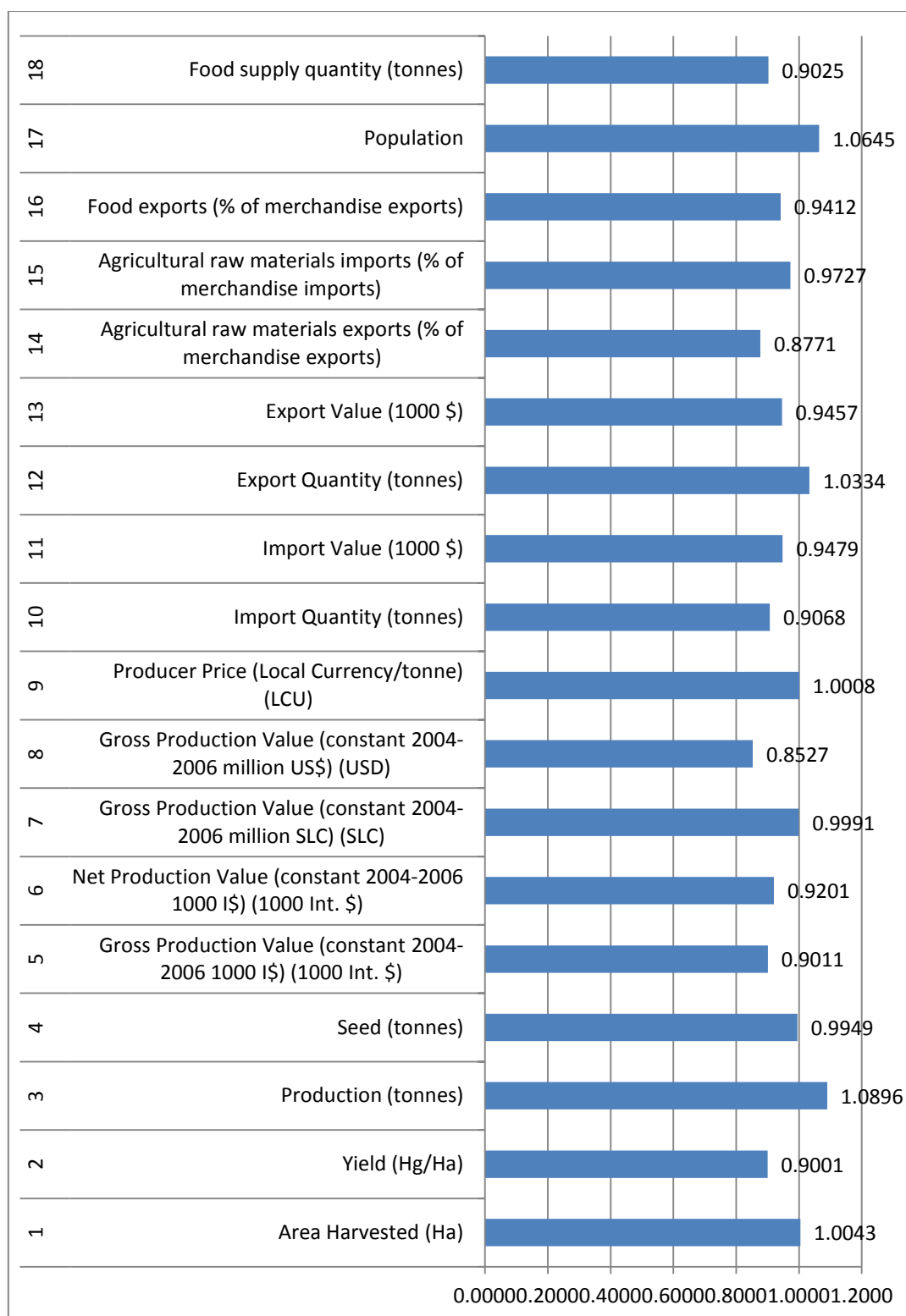


Figure 5.16: Mean for each of the feature variables

Table 5.3: Ranking selection for each of the features

<b>Rank</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>
<b>Feature Variable</b>	3	17	12	1	9	7	4	15	11
<b>Rank</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	<b>16</b>	<b>17</b>	<b>18</b>
<b>Feature Variable</b>	13	16	6	10	18	5	2	14	8

This arrangement of importance can also be show as below form:-

Feature variable ranking = [Production, Population, Export Quantity, Area Harvested, Producer Price, Gross Production Value2, Seed, Agricultural raw materials imports, Import Value, Export Value, Food exports, Net Production Value, Import Quantity, Food supply quantity, Gross Production Value1, Yield, Agricultural raw materials exports, Gross Production Value3 ]

In giving the same feature as in the TSH model, nine features were selected based on the highest ranking of SGNO – from rank 1 to rank 9 per Table 5.3, which is shown below:-

Nine Selected features = [Production, Population, Export Quantity, Area Harvested, Producer Price, Gross Production Value2, Seed, Agricultural raw materials imports, Import Value ]



The selected features were then used in the ANN as the prediction model, to compare its performance. Figure 5.17 shows the regression performance and Figure 5.18 shows the MSE performance of the SGNO.

In Figure 5.18, the dataset for each part – training, validation and testing, shows some scattered values when compared to other models in this section. It also seems to have a generalization problem, but eventually achieved a 95% accuracy of regression, which is quite good. For the MSE performance, initially the MSE value is quite high, compared with the TSH model, but it stops searching for the local optimum at iteration 24, which is much faster than the other model.

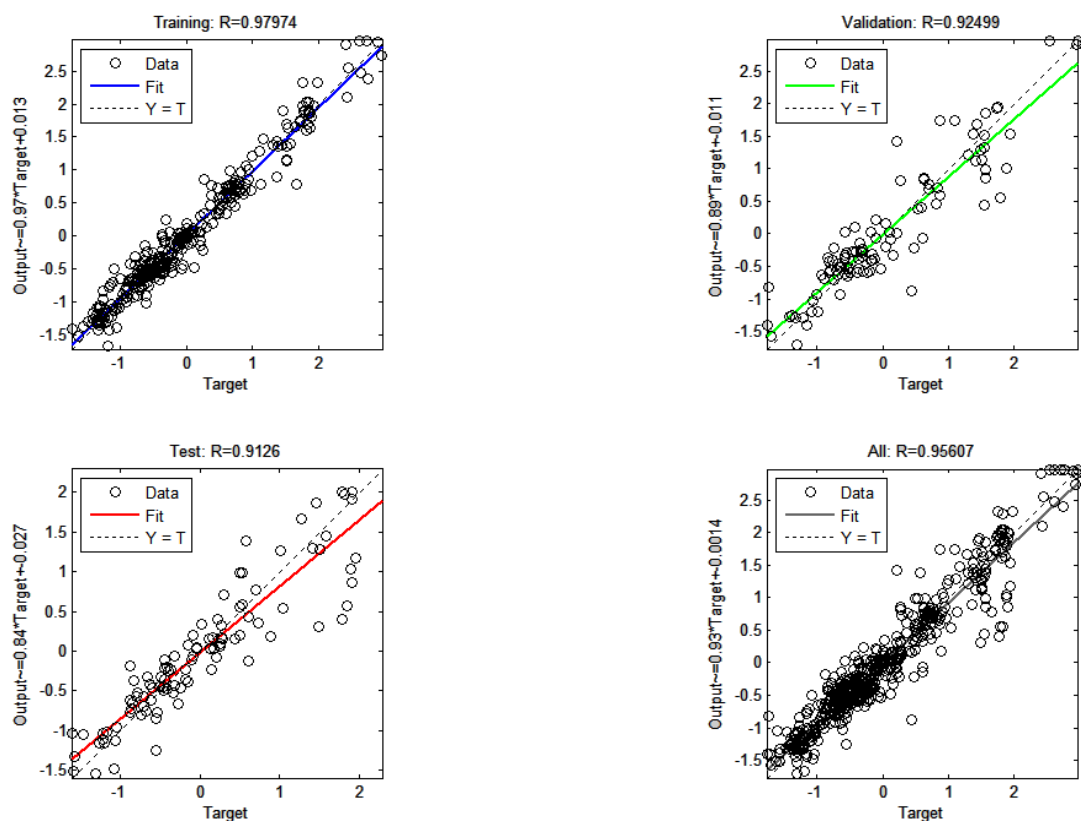


Figure 5.17: ANN performance based on SGNO for 9 input selections

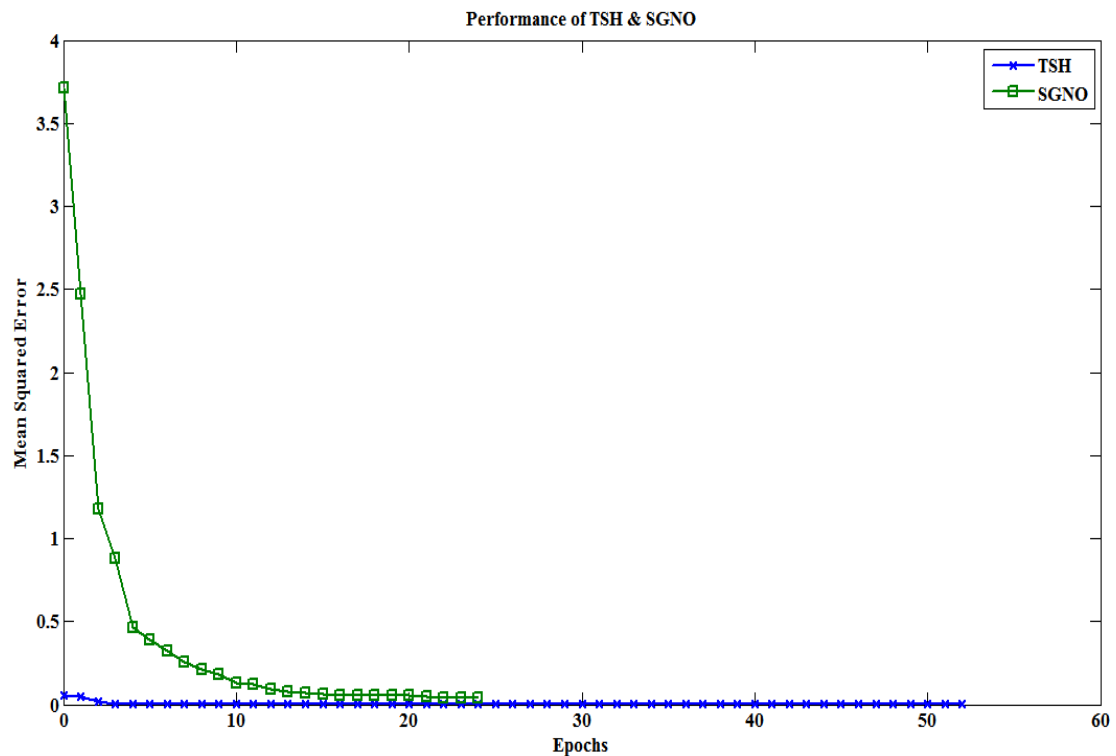


Figure 5.18: MSE performance of ANN models using TSH, compared with SGNO

Although having a high MSE at the first epochs, it still achieves quite good errors with just 24 epochs. In the final resolution, the proposed model shows a better prediction model than SGNO, referring to the regression plot and the MSE plot.

### 5.8.6 Summary

The overall model performance is shown in Figure 5.19 for the regression plot and Figure 5.20 for the MSE plot. It can be seen that the TSH model outperforms the other benchmarked techniques. However, the difference in performance between the models is very small. In terms of generalization, the result shows that almost all of the models give good generalization, except the SGNO, for which the dataset is a bit scattered, but is still acceptable.

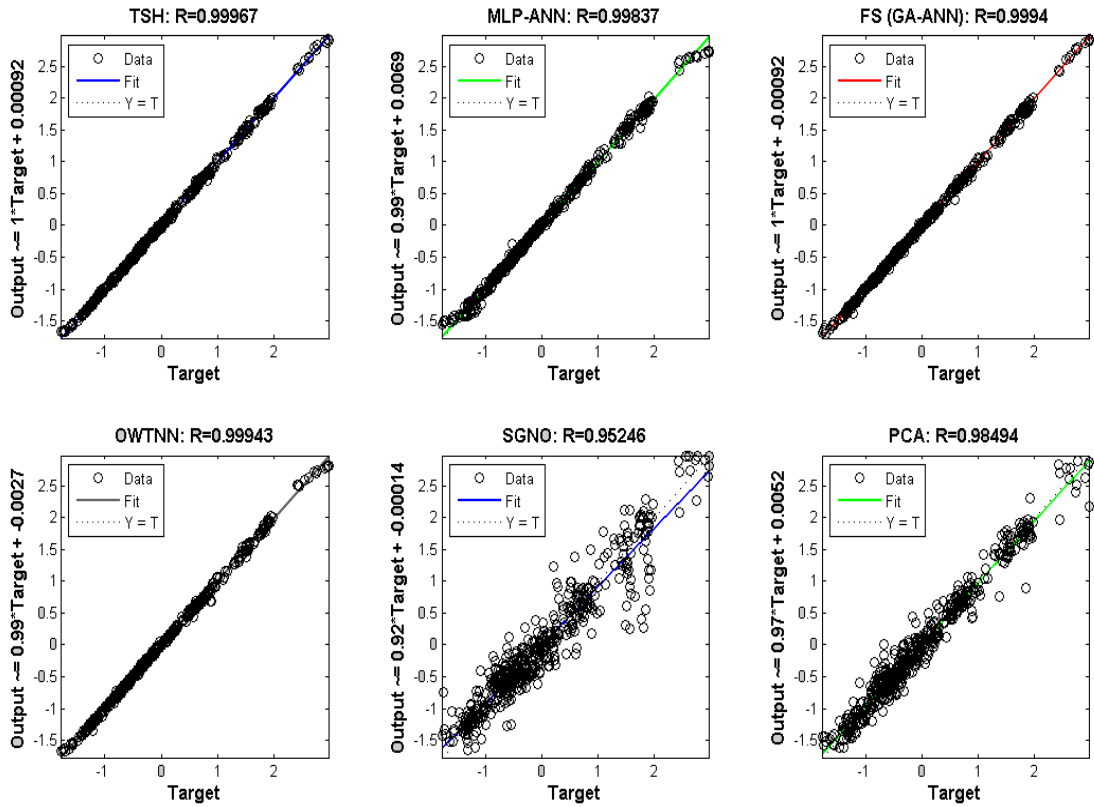


Figure 5.19: Benchmarking on overall ANN regression performance

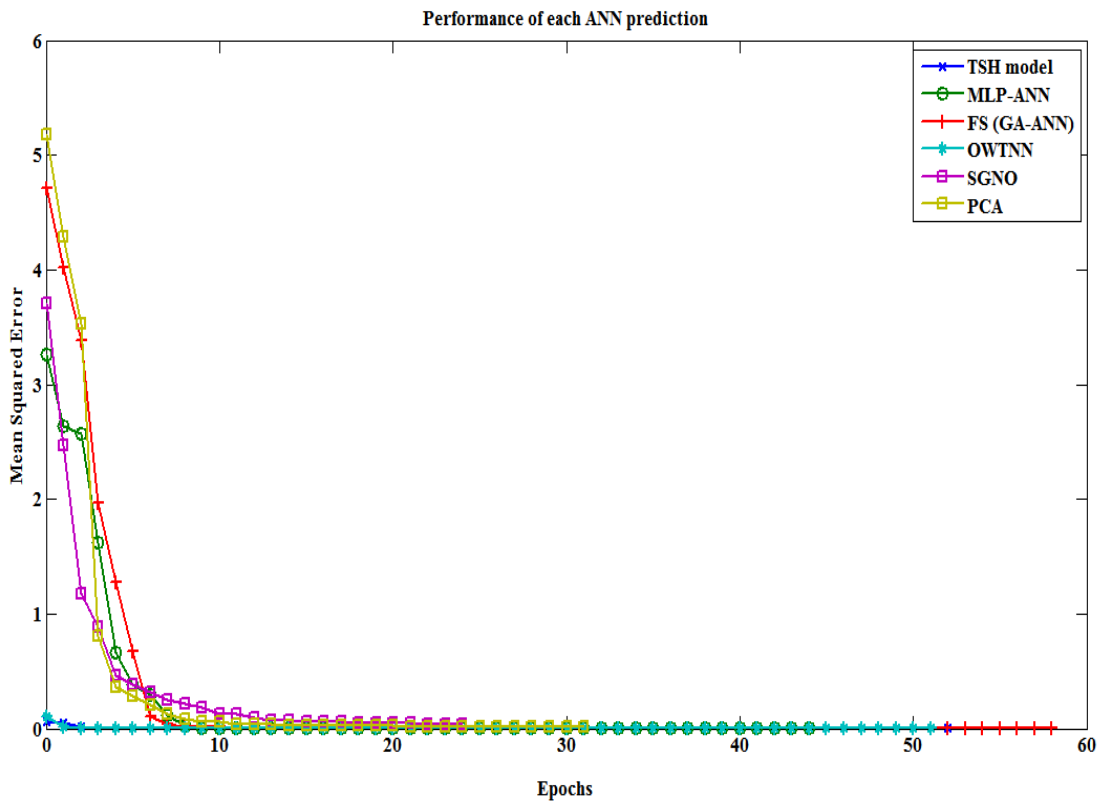


Figure 5.20: Benchmarking on MSE performance

If compared with all benchmarked models, the GA-ANN for FS takes quite some time in reaching the generalization values, terminating at epoch 58 and the SGNO terminates slightly early at epoch 24, which probably accounts for it having the lowest performance compared to the other benchmarked models. Among all of these techniques, the TSH prediction model produces the best performance.

## **5.9 Conclusion**

In this chapter, the TSH model is tested with 18 features and 502 samples of a dataset to predict cereal growth per capita, which is created based on a compilation of FAOstat, World Bank and USDA databases. The 18 features of the dataset consist of a presentation of: yield growth, production and producer prices or value, food supply quantity, share of food trade, concentration of world markets and the population of the 13 European countries that the data relates to.

In this model, prediction is performed using feature selection by a GA-ANN, with the selection of 9 out of 18 input variables as the first stage process. Using the selected input variables, it is then optimized again, but this time the optimization is done for the weights and thresholds of the ANN itself. After this, the ANN is remodelled with the optimized input, optimized weights and optimized thresholds, by using the same parameters of the ANN used in the first and second stage processes.

In considering the overall result, benchmarking is performed with five techniques, and most of the prediction models give a very good performance, better than 95%, of prediction accuracy especially for the TSH model and the OWTNN model, which give better prediction performance compared to the other models. These

two models have a regression value difference of only 0.00024, which shows that the TSH model outperforms the OWTNN in overall regression. The proposed model also shows the lowest MSE values compared to the other benchmark techniques.

## References

- BANK, T. W. 2012. *World Bank Data* [Online]. The World Bank. Available: <http://databank.worldbank.org/ddp/home.do#ranking> [Accessed 2011].
- BAREJA, B. 2010-2012. *Cereal Crops, Pseudocereals* [Online]. Available: <http://www.cropsreview.com/cereal-crops.html> [Accessed 24/10/12].
- CHAPMAN, S. R. C., LARK P. 1976. *Crop Production: Principles and Practices*, San Francisco, USA, W. H. Freeman.
- D'HEYGERE, T., GOETHALS, P. L. M. & DE PAUW, N. 2003. Use of genetic algorithms to select input variables in decision tree models for the prediction of benthic macroinvertebrates. *Ecological Modelling*, 160, 291-300.
- DEFRA 2009. UK Food Security Assessment: Our approach. Department for Environment Food and Rural Affairs.
- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In*: DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). DEFRA.
- EDUARD, L. & ET AL. 1999. Non-destructive banana ripeness determination using a neural network-based electronic nose. *Measurement Science and Technology*, 10, 538.
- FAO 1986-2007. Value of Agriculture Production. *In*: FAOSTAT (ed.). FAO Statistics Division 2012.

FAO 2006. Policy Brief : Food Security. *In: ECONOMICS*, A. A. D. (ed.). FAO's Agriculture and Development Economics Division (ESA) with support from the FAO Netherlands Partnership

Programme (FNPP) and the EC-FAO Food Security Programme.

FAO. 2009. 2050: A third more mouths to feed. Available: <http://www.fao.org/news/story/en/item/35571/icode/> [Accessed 24/10/12].

FAO 2011. The State of Food Insecurity in the world. 3rd ed. Rome: Food and Agriculture Organization of The United Nation.

FAO. 2012. *FAOSTAT* [Online]. Available: <http://faostat3.fao.org/home/index.html#DOWNLOAD> [Accessed 2012].

GARDNER, J. W. & ET AL. 1990. Application of artificial neural networks to an electronic olfactory system. *Measurement Science and Technology*, 1, 446.

GARDNER, J. W., HINES, E. L. & TANG, H. C. 1992. Detection of vapours and odours from a multisensor array using pattern-recognition techniques Part 2. Artificial neural networks. *Sensors and Actuators B: Chemical*, 9, 9-15.

GODFRAY, H. C. J., BEDDINGTON, J. R., CRUTE, I. R., HADDAD, L., LAWRENCE, D., MUIR, J. F., PRETTY, J., ROBINSON, S., THOMAS, S. M. & TOULMIN, C. 2010. Food Security: The Challenge of Feeding 9 Billion People. *Science*, 327, 812-818.

H. DEMUTH, M. B. 2004. *Neural Network Toolbox: For use with Matlab*, USA: Mathworks.

KITANO, H. 1990. Empirical studies on the speed of convergence of neural network training using genetic algorithms. *Proceedings of the eighth National conference on Artificial intelligence - Volume 2*. Boston, Massachusetts: AAAI Press.

- NEGNEVITSKY, M. 2005. *Artificial intelligence A guide to intelligent systems*, Addison-Wesly.
- NELLEMAN, C., MACDEVETTE, M., ET AL. 2009. *The Environmental Food Crisis: The Environment's Role In Averting Future Food Crises* Birkeland Trykkeri AS, Norway.
- SALMAN, Y. & ONG HANG, S. Year. The effect of GA parameters on the performance of GA-based QoS routing algorithm. *In: Information Technology, 2008. ITSIM 2008. International Symposium on, 26-28 Aug. 2008* 2008. 1-7.
- TACIO, H. *Feeding a world of 9 billion* [Online]. People & the Planet. Available: <http://www.peopleandplanet.net/?lid=26107&section=34&topic=44> [Accessed 24/10/10 2012].
- USDA. 2012. *International Macroeconomic Data Set* [Online]. United States Department of Agriculture. Available: <http://www.ers.usda.gov/data-products/international-macroeconomic-data-set.aspx> [Accessed 2011].
- WHITLEY, D., DOMINIC, S., DAS, R. & ANDERSON, C. W. 1993. Genetic Reinforcement Learning for Neurocontrol Problems. *Mach. Learn.*, 13, 259-284.
- WORLDBANK 2009. *Global Economic Prospects*.
- ZHANG, F. 2011. *Intelligent Feature Selection for Neural Regression*. Doctor of Philosophy, University of Warwick.

# **Chapter 6: Food security risk level assessment prediction using Grain China dataset**

## **6.1 Introduction**

This chapter will examine a prediction model for food security risk level, by using the China grain dataset. In ensuring the stability of food security, many forums, conferences, and discussions between world leaders and global food organizations have taken place, such as stated in (FAO, 2006, FAO, 2009, FAO, 2011, Godfray et al., 2010, WorldBank, 2009). In each of these discussions, the factors of food security were discussed, such as: food prices, growth per capita, and development of technology related to increasing food production or farming output.

In reports and paper by (Kadir et al., 2011, WorldBank, 2009, DEFRA, 2010), food security has been defined as the access to sufficient nutritious food without any hardship, to maintain a healthy and active lifestyle. Generally, referring to (DEFRA, 2010, FAO, 2006, DEFRA, 2009), the descriptions of food security are related to the following areas: food availability, resource sustainability, access to food, food chain resilience, household food access and the confidence among the public in their domestic food systems. These key components should be monitored individually and



as a whole to allow for complete assessment of the risk levels of food security. All of the above descriptions were also discussed in Chapter 1 and Chapter 3.

## **6.2 Background**

As explained in the previous section, each of the key components of food security has their own indicators and sub-indicators, which are used in the monitoring of food security. It is important to ensure that the level of risk is at an acceptable level and will not be affected by any other factors.

In this study, a food security risk assessment prediction model is described. The model uses three main criteria as the factors significantly affecting food security in China. The criteria are productive indexes, consumptive indexes and disaster indexes. Each of these criteria is assumed to have an impact on food security and is used as a warning-based level indication.

A few studies had been performed on the modelling of food security risk assessments, example by (Men et al., 2009, Jianling and Yong, 2010a, Yong and Jianling, 2010), but in terms of predictive models, there is little research being done. In (Kadir et al., 2011), a prediction model was developed using PCA and ANFIS, and compared with an ANN; the result shows a good prediction performance.

The problem with using the ANFIS model is that it generates a complex network structure, and it can easily result in the system running out of memory. At the same time, in PCA data reduction, five of the features were reduced in order to prevent the memory outage. These features were selected based on the lowest Eigen values from the 'Eigen analysis' of the principal component where the percentage of accumulated variance is more than 95%. Therefore in this chapter, a model is

described which is used to select the inputs automatically, based on the lowest MSE value between the actual output and the model output.

### 6.3 Dataset

The dataset used in this chapter was based on China's grain security studies by (Men et al., 2009, Jianling and Yong, 2010a, Jianling and Yong, 2010b, Yong and Jianling, 2010, Kadir et al., 2011) and it is available for use by others. All of these studies used multiple techniques such as: Analytical Hierarchical Process with Gray multi level (AHP-GRA) technique, generalized fuzzy numbers, linguistic ranking models and fuzzy sets in analyzing the grain security in China. However, as described in Section 6.2, only one model developed by (Kadir et al., 2011) uses this dataset with a prediction model.

Table 6.1 shows the index category for each feature variable assumed to affect grain security in China; a total 11 of features and therefore 11 variables. As the features are in a productivity index, each feature is based on production related categories; therefore it can be assumed that each feature can have a major impact on production output. The consumptive indexes contain features of grain usage by production industries, and the grain price itself. In terms of food security, or specifically in this study, grain security, a disaster which occurs either in China, or in other countries, will always have a big impact either on farming output or food production. Therefore the disaster index is included in the study, providing two additional features for the datasets.

This dataset is taken from the grain security raw data from years 1997 to 2007, which is available in the papers mentioned above. The output of the dataset is based on the correlation values of equations (6.1) to (6.3) in the report by (Men et al., 2009),

where it was used to rank China's grain security warnings as: quite safe, safe and unsafe. By referring to the datasets, the report gives 11 samples of data with 11 features variables as shown in Table 6.1.

$$R = P \cdot E^T \quad (6.1)$$

$$r_i = (p_{i1}, p_{i2}, \dots, p_{im}) \cdot \begin{pmatrix} \xi_i(1) \\ \dots \\ \xi_i(m) \end{pmatrix} \quad (6.2)$$

$$r_i = \sum_{k=1}^m p_k (k) \quad (6.3)$$

Table 6.1: Features categories for trends in global output per capita (Kadir et al., 2011)

<b>Productive index</b>	<b>Consumptive index</b>	<b>Disaster index</b>
Grain Production (tons)	Per capital occupation of grain (kilogram)	Disaster affected area (hectares)
Grain seeding area (hectares)	Food imports for the proportion of total agricultural	Proportion of disaster-affected area (%)
Per capital seeding area (square meter per person)	Food exports for the proportion of total agricultural	
Effective irrigation area (hectares)	Growth of grain price index	
Proportion agriculture value (GDP - %)		

## 6.4 Data pre-processing

As in the previous chapter, each dataset will be pre-processed by making it standardized to a zero mean and a unit variance, based on Equation (3.1). The standardization was applied to the input variables and output variable. This equation is intended to make each variable have the same range, and this will shorten the process time. The data consists of multiple-range variables as shown in Table 6.2; the table shows the basic statistics for each input variable in terms of the minimum, maximum, mean and standard deviation.

Table 6.2: List of basic statistic for input variables

Features	Min	Max	Mean	Standard Deviation
Production	43069	51229	47913.64	2642.94
seeding area	99410	113787	107155.64	4936.27
Per capital seeding area	770	910	833.55	52.65
Effective irrigation area	51238.50	56518.30	54115.15	1529.50
Proportion agriculture value	9.90	18.30	15.04	2.68
Per capital occupation of grain	334	411	374.73	23.53
Food imports for the proportion of total agricultural	8.55	39.65	24.61	10.70
Food exports for the proportion of total	6.57	25.11	16.07	6.92

agricultural				
Growth of grain price index	-0.10	6.60	1.06	2.33
Disaster affected area	37106	54688	48008.18	6287.33
Proportion of disaster-affected area	43.90	62.90	55.31	5.75

### 6.5 First stage process

In deciding on the features for this model, a GA-ANN model is used, which combines the capabilities of the GA and ANN algorithms in finding the local and global regions for the solution. The population size in the GA module is dependent on the number of input variables of the dataset. In this case, the dataset input variables is 11, which is then multiplied by 3 to create a population size of 33, based on the theorem in (3.3). In this case, GA chromosomes are also based on binary numbers as described in Chapter 3, which represent each of the input variables for training in the ANN module. The GA module then assigns a random population of chromosomes to the ANN module for training. In each ANN training process, an error of MSE is produced for the fitness value, and the GA then rearranges the population of chromosomes, using a crossover ratio of 0.7. The mutation ratio depends on the number of chromosomes and the population size; the bigger the input variables and the population size, the smaller the mutation ratio, as in Equation (3.2).

As previously described, the ANN module is used as the objective function, where the error between the actual output and the ANN model output is considered to be the fitness function, used in deciding the best chromosomes for the ANN to give the optimum prediction of grain security in China. The ANN module uses 3-layered

network consisting of an input layer, a hidden layer and an output layer. Each of these layers has a different number of neurons; there are 11 neurons for the input layer, 8 for the hidden layer, and one for the output layer. The numbers of hidden neurons are determined by Equation (3.6), which is dependent on the number of inputs and outputs.

In this chapter, the input variables are equivalent to the number of chromosomes, and there is only one output, which is the grain security in China. The learning method used in the ANN is LM, which provides fast learning but can use a significant amount of memory. In this case however, the LM memory issue does not occur, because the dataset only consists of 11 X 11 matrices. Each dataset will be divided into: 60% for training, 20% for validation and another 20% for testing.

The overall performance of the chromosomes being trained and evaluated in this first stage process is shown in Figure 6.1. The line in the graph represents the overall mean RMSE for all chromosomes in each generation, where the lowest RMSE value is considered to represent the best performance for this stage. The lowest RMSE value is for the binary number 1100 1101 001, which represents the 11 features variables of this model.

The following list is the features being selected in this stage:-

Selected features = [Production, Seeding area, Agriculture proportion value, Per capital occupation of grain, Proportion of total agriculture exports, Disaster proportion in effected area]

The final feature being selected is six input variables based on the '1's of the binary numbers and '0's binary numbers shows the neglected features.

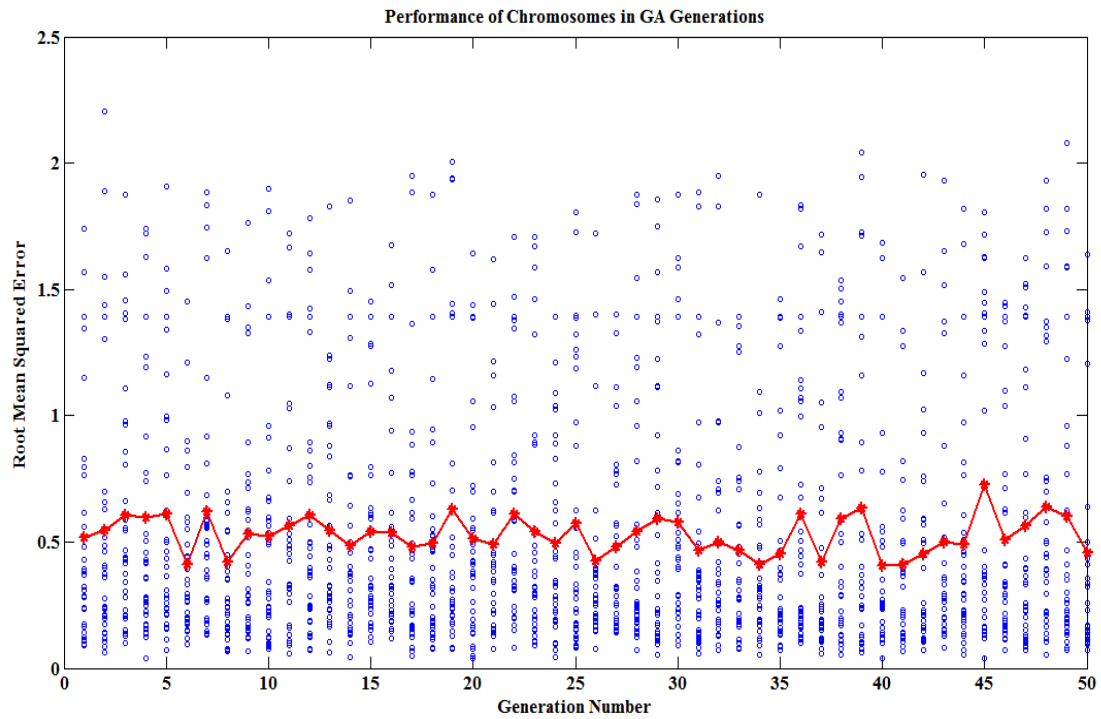


Figure 6.1: First stage performance via GA generation

## 6.6 Second stage process

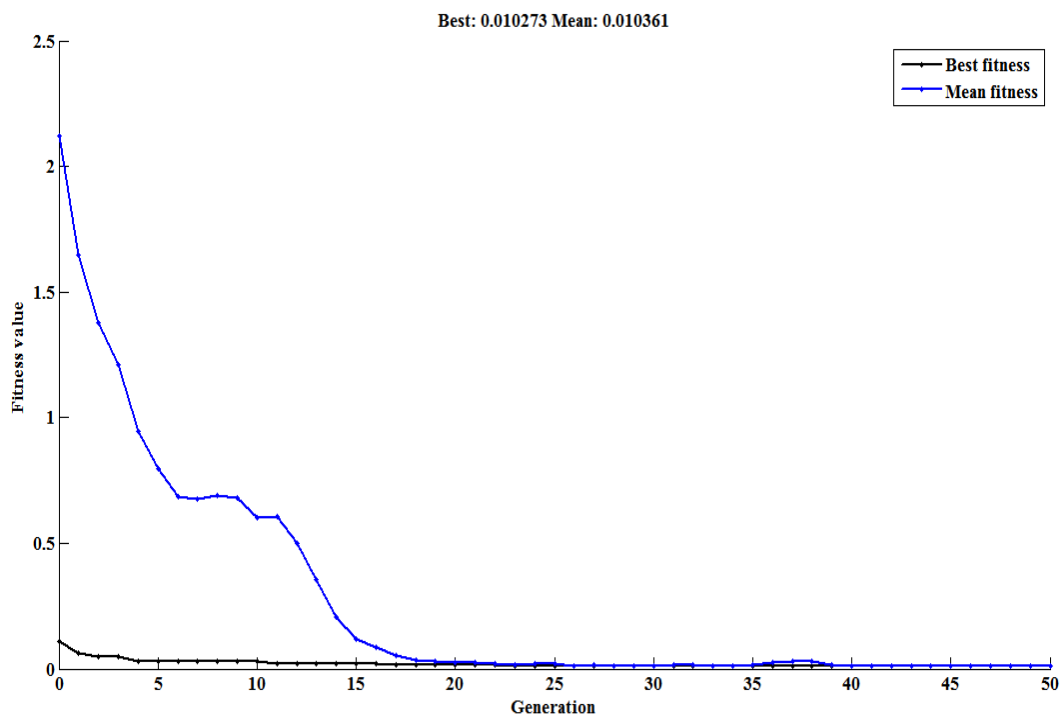


Figure 6.2: Performance of OWTNN second stage via number of generation

Figure 6.2 shows the performance of the OWTNN where it used the selected features in the first stage to define the required number of chromosomes. The mean fitness value rapidly decreases over the first five generations, and it shows a second decrease from the 10<sup>th</sup> generation. The convergence trend of the MSE shows convergence at the 6<sup>th</sup> and 18<sup>th</sup> generations, which shrinks with each generation as the GA module evolves.

The second stage also used two modules in its main design - a GA module and ANN module. In maintaining the performance of the input selection process of the first stage, all of the parameters used in the GA module and ANN module previously were also maintained, such as the number of layers used by the ANN, the learning method for the ANN, the dataset division for the ANN, the size of hidden neurons for the ANN and the crossover ratio of the GA.

The chromosomes in the second stage are represented as a decimal number, which is different in the first stage, and it also has different input sizes, because the input selection is done in the previous stage. The mutation probability ratio will also be different; it is determined based on Equation (3.3), where the ratio value changes to 0.0017 from 0.0160. This also affects the population size, which is multiplied by 2 in this stage, to become 22. The decreasing population size can reduce the computation time, and in this second stage, the quantity of chromosomes is higher than the first stage. Figure 6.2 shows that the performance of this stage seems to be reaching an optimum convergence at the 20<sup>th</sup> generation; however the termination of this stage is also based on the number of generations, which is set at 50.



## 6.7 Remodelling the ANN

The TSH model is used to reduce the feature size of feature and to optimize the weight and threshold of the ANN. The result of the selected feature and optimized threshold and optimized weight are used in remodelling the ANN, to design the optimum prediction model to predict the grain security assessment of China.

As described in the previous section, to ensure the optimum performance of the prediction model, the ANN uses the same parameters as the ANN module in the second stage process. To illustrate the prediction performance, a regression graph is plotted between the actual output and the TSH output, as shown in Figure 6.3. The graph shows a regression value of 0.9949, which is represents an error difference of 0.55% compared to the actual output. This is a good performance result, and it seems that TSH model can be used to predict the grain security assessment, although it only consists of 11 samples of data with 11 inputs.

Figure 6.4 shows the detailed plot of the model performance, categorised by the number of iterations. The starting MSE at the first epoch is 0.0095, which is quite a low value. At the third epoch, the MSE converges and is reduced to 0.00000496 at the sixth epoch. The pattern shows a rapid reduction of MSE between the first and fourth epoch.

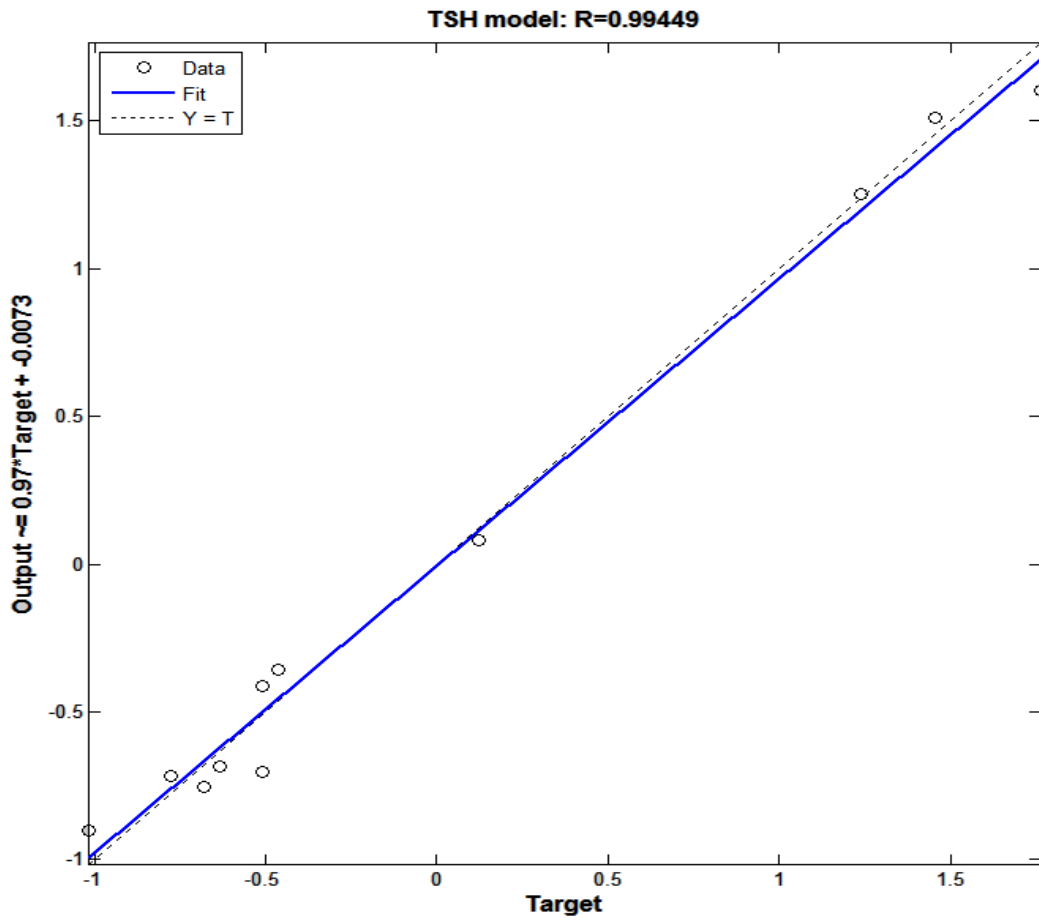


Figure 6.3: Regression of TSH model

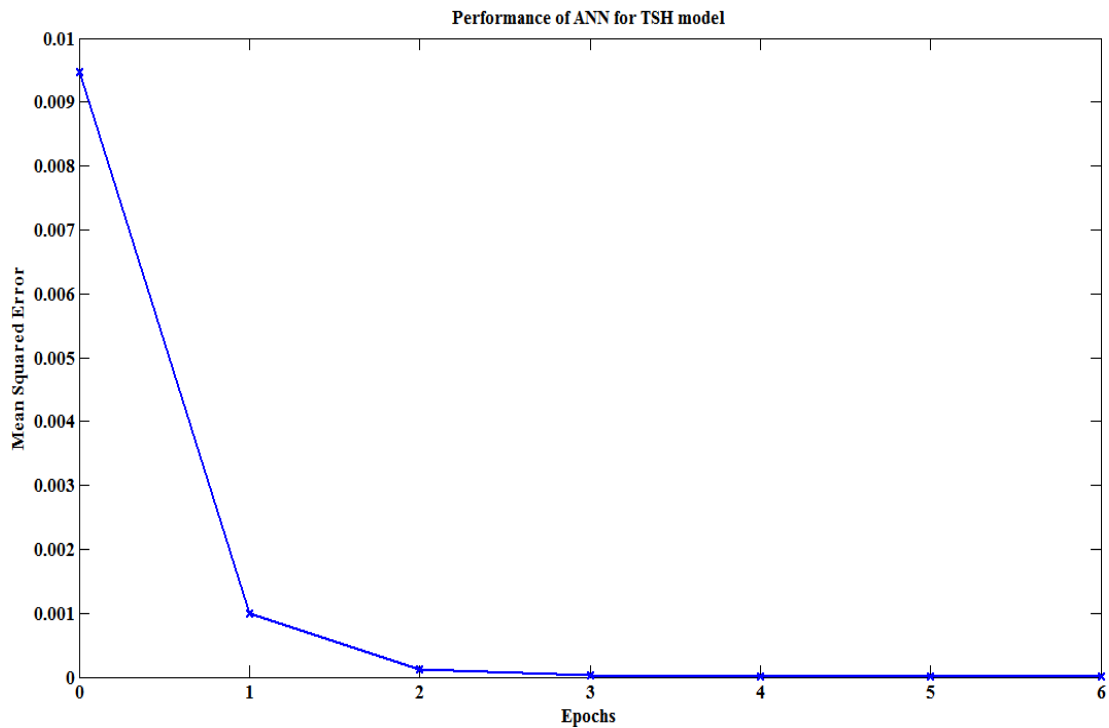


Figure 6.4: Performance of ANN based on MSE vs Epochs

As mentioned in Section 6.5, the data is being divided into three divisions; 60% for training, 20% for validation and the last 20% for testing. This data division is applied in the same way for all stages, and also in the remodelling of the ANN. Figure 6.5 shows the performance of each of these divisions. The regression values are above 0.99 for all divisions, and two divisions, validation and testing, achieve a regression value of 1. The generalization for each division shows a poor fit to the  $Y=T$  line, but, overall it is still a good generalization.

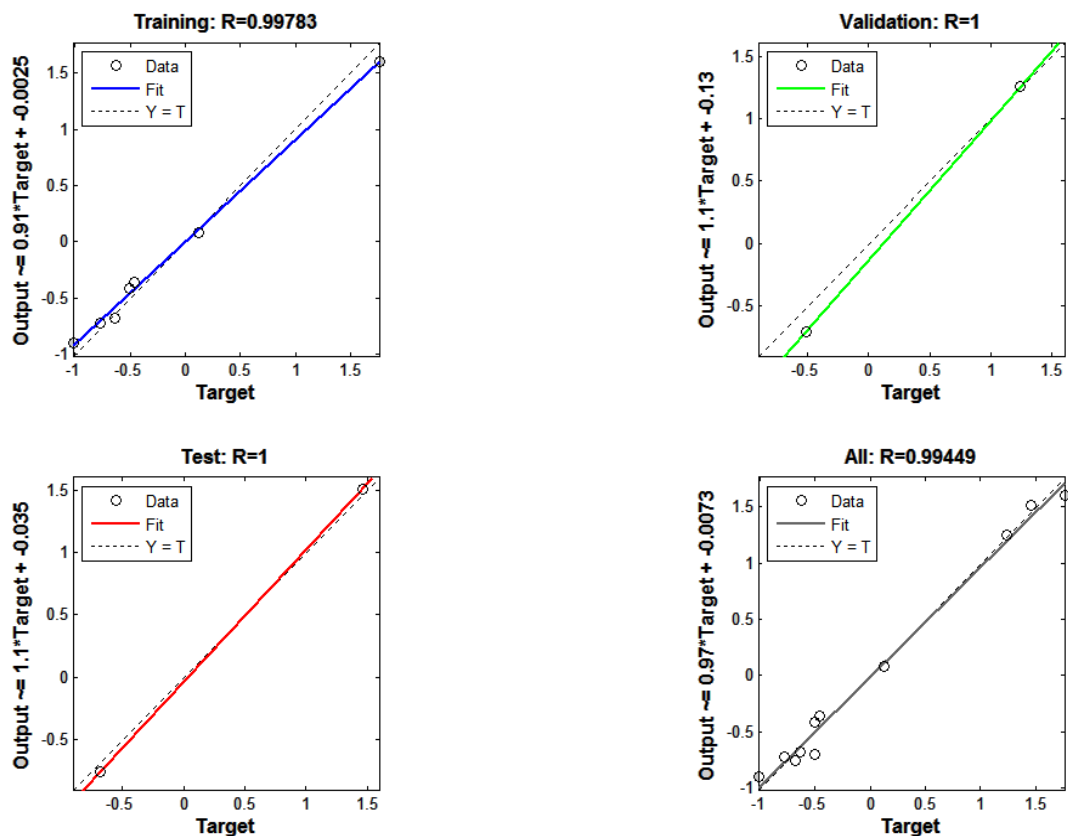


Figure 6.5: Overall regression for each part – training, validation and testing

## 6.8 Benchmarking and discussion

In order to ensure the prediction model being developed gives a good performance result, benchmarking is performed by comparing the regression and MSE of the ANN for each of the following methods: PCA, MLP-ANN, FS (GA-ANN), OWTNN on original dataset ( $11 \times 11$  matrices), and SGNO.

### 6.8.1 Principal Component Analysis (PCA)

In (Kadir et al., 2011, Jang, 1996, Dong Hyun Jeong), PCA is considered to be a well known method for dimension reduction; it converts the original dataset to another new component set of data. A detailed description of PCA is given in Chapter 3.

This case study uses 11 input variables which are transformed into new principal components. The new component data is then used in the ANN to measure the prediction performance, which is then compared with the proposed prediction model. Figure 6.6 illustrates the pattern behaviour of each principal component after the transformation. In Figure 6.6, only four components exist, each of which has a different percentage variance. The first component shows more than 50% variance, and if this is combined with the second and third components, the total variance is 80%, which is very good. The thin blue line in the plot shows the accumulated variance from the first component to the fourth component, resulting in a final value of over 95%.

As described previously, these new components are then used in the ANN as the new input variables, with a regression plot and MSE plot showing the resultant performance. Furthermore, the ANN architecture comprises a 3-layered network, consisting of an input layer, hidden layer and output layer, where the number of

hidden neurons is 8, which is the same as in the ANN module of the proposed model.

Figure 6.7 and Figure 6.8 show the performance of the ANN in predicting the results of the China grain security assessment, based on the three categories in Table 6.1.

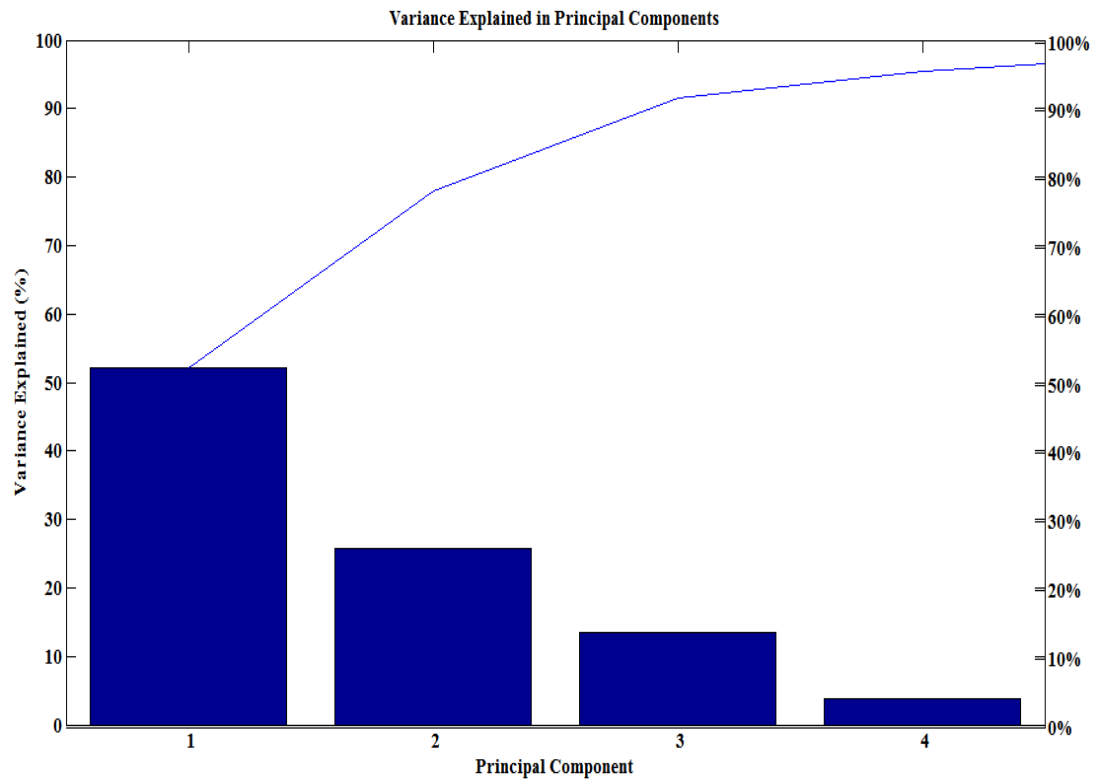


Figure 6.6: Variance in Principal component

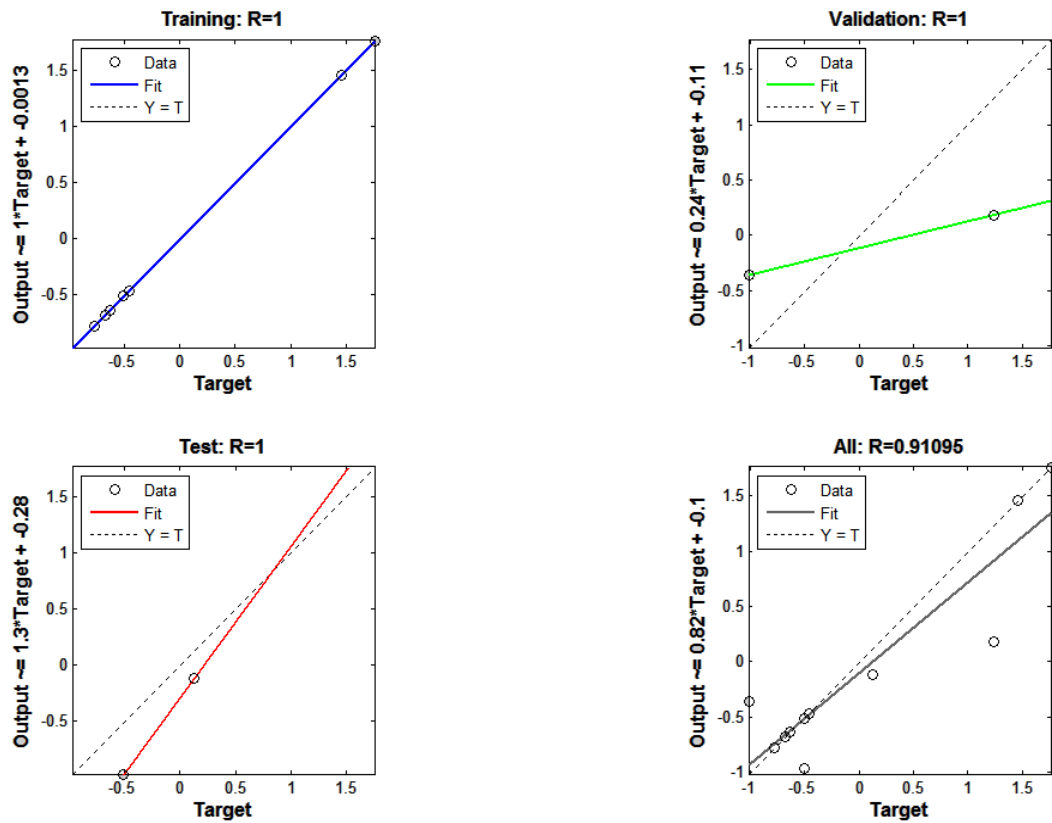


Figure 6.7: Regression for PCA using ANN

Figure 6.7 shows the regression performance of the prediction, between the actual output and the PCA-ANN output. It shows a 0.91 regression value for the overall regression, which is good. However, the result shows a regression value of 1 for all dataset divisions – training, validation and testing, but the fitting line is not equal to  $Y=T$  for the validation part and testing part, which means that there is a generalization problem for both of these parts.

In terms of the error plot as shown in Figure 6.8, PCA reach its final solution at the 24<sup>th</sup> epoch, compared with the proposed model at the 6<sup>th</sup> epoch where the difference is by 18 epochs. The high starting error also drastically reduced with the increasing of number of epochs. The final MSE values for both models have a difference of only  $4 \times 10^{-6}$ , which shows that the PCA-ANN model has smaller errors

than the proposed model, however, the difference is small, and the proposed model ends more quickly than the PCA-ANN. In addition, the errors patterns for each epoch give the TSH model an advantage.

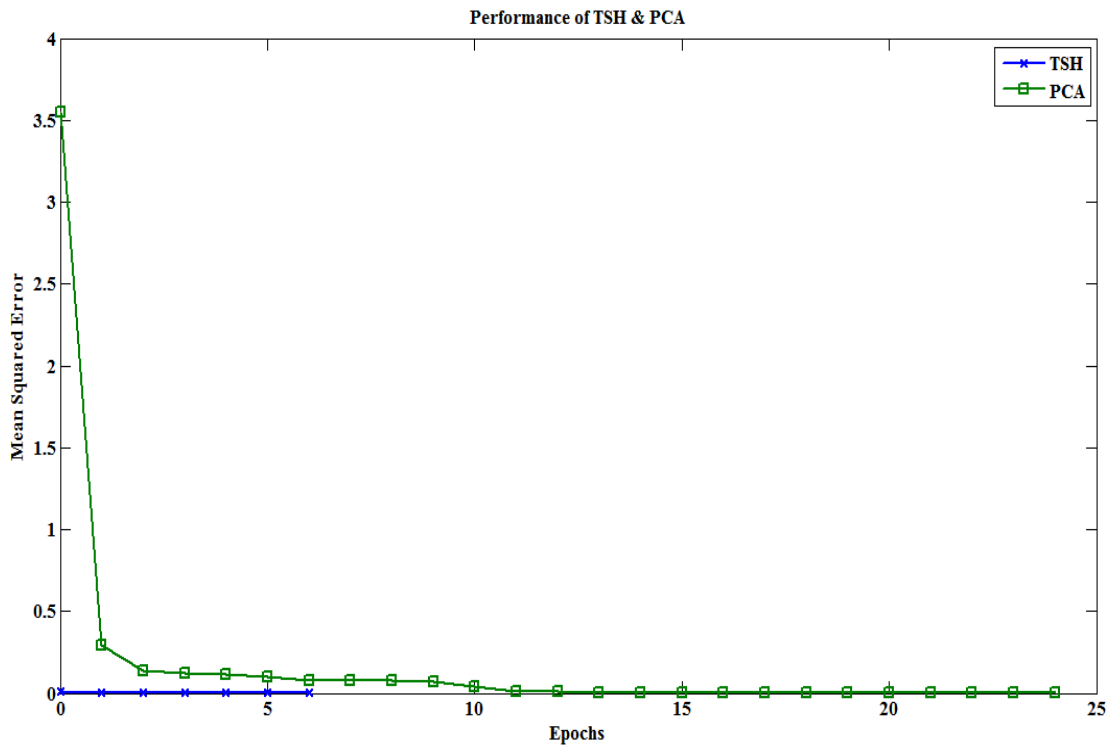


Figure 6.8: MSE between PCA and TSH model

### 6.8.1 Multi Layer Perceptron - Artificial Neural Network (MLP-ANN)

MLP-ANN is one of the best prediction models for dealing with any kind of data – linear or non-linear, as described by (Eduard and et al., 1999, Gardner et al., 1992, Gardner and et al., 1990). To benchmark the MLP-ANN model in similar conditions to the proposed model, all of the parameters in this model use the same parameters as in the remodelled ANN in Section 6.7. The dataset used was a dataset of 11 input variables with 11 samples, using a three-layer network consisting of one input layer, one hidden layer and one output layer. The details of the MLP-ANN process are described in Chapter 2 and Chapter 3.

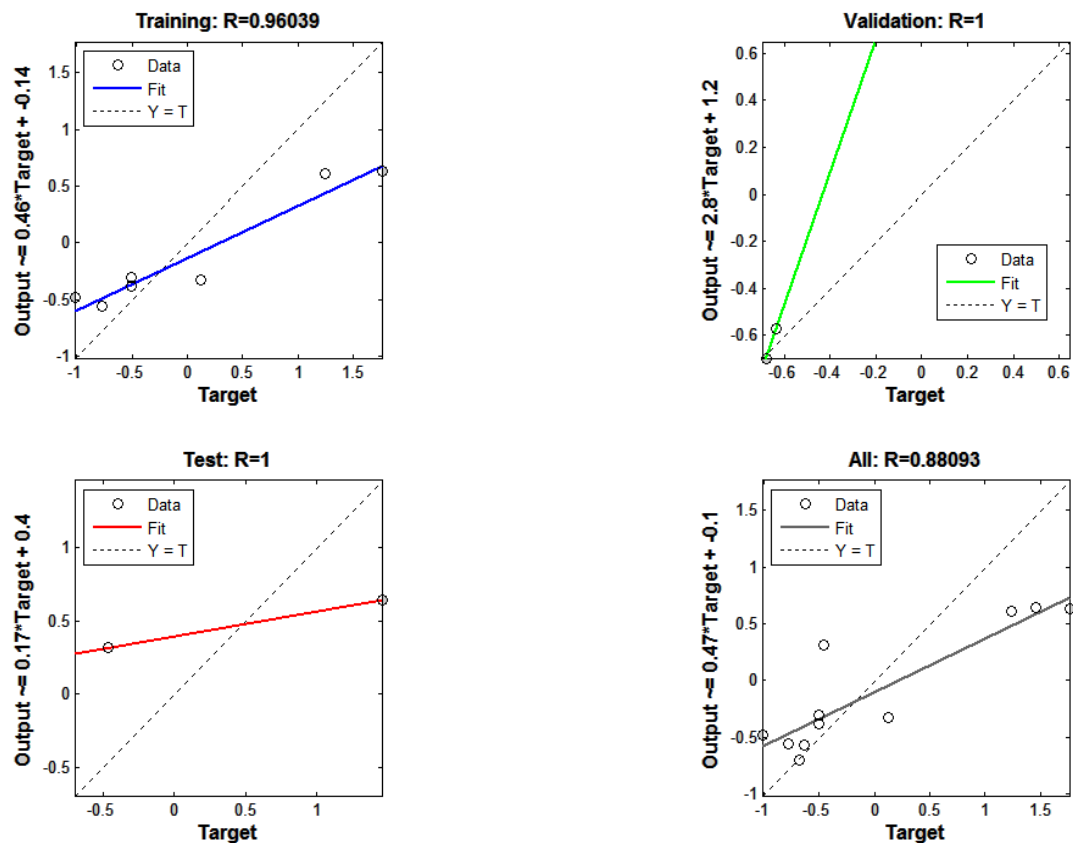


Figure 6.9: Regression for original ANN

The first performance plot, as shown in Figure 6.9, is the regression plot of all the dataset divisions, with an overall regression value of 0.88, which represents a 12% difference when compared with the TSH model. Based from the regression result, it can also be concluded that the TSH model outperformed the original MLP-ANN model in terms of prediction. In terms of the generalization for the testing, validation and test divisions, the plot illustrates that all of the fitting lines are not equal to  $Y = T$ , which shows a poor generalization. This occurs because of the random initialization of the threshold and weight values.

The MLP-ANN model gives a high MSE value at the beginning of the epoch. This is totally different when compared to the proposed model, where the errors at the



beginning of the epochs have a low value. However, in terms of reaching a solution, the plot shows a rapid reduction of the MSE with each epoch, until it converges between the third and fourth epochs, and reaches its final value at the seventh epoch. The difference in error values between the MLP-ANN model and the TSH model is 8.83, where the TSH model performs better than the MLP-ANN model. The MSE plot is shown in Figure 6.10.

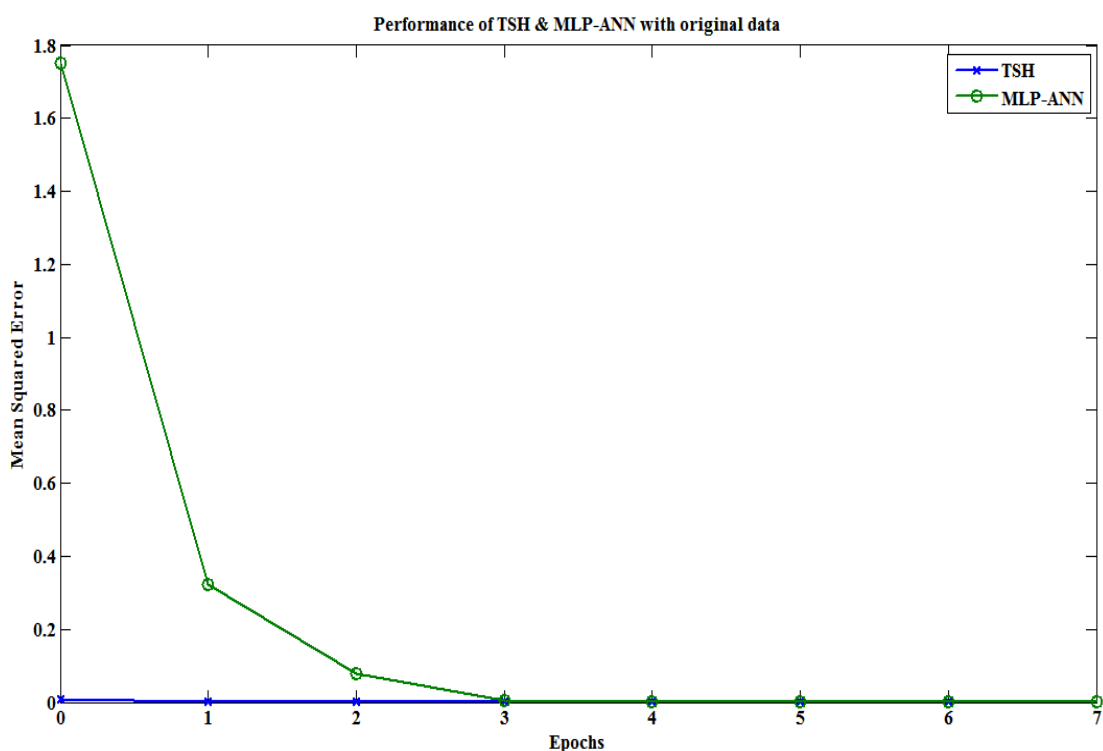


Figure 6.10: MSE for original ANN and TSH

### 6.8.2 Feature selection (GA-ANN)

The hybrid GA-ANN is capable of feature selection, where it is being used as the first stage process in the TSH model. To study the capability of this first stage, it has been specified as one of the benchmark techniques, where its prediction performance is compared with the TSH model. In the same way as for the MLP-ANN model, the performance for this model is compared based on the ANN prediction

model regression plot and MSE plot as shown in Figure 6.11 and Figure 6.12, for six input selections, which are selected by the GA-ANN optimization model.

Due to the small amount of sample data, the model can still easily be trained, validated and tested for each of the data sets via the actual output of the dataset, as shown in Figure 6.11. The regression plot also shows an overall regression value of 98% accuracy, but the GA-ANN feature selection model shows a generalization problem, where each of the dataset divisions do not fit to  $Y = T$ . In the final result, the prediction regression plot indicates a scattered set of data, caused by the generalization problem in each dataset division. If compared to the proposed model, this model still gives quite good prediction performance, but the generalization problem results in a regression value of less than 1 for the overall prediction.

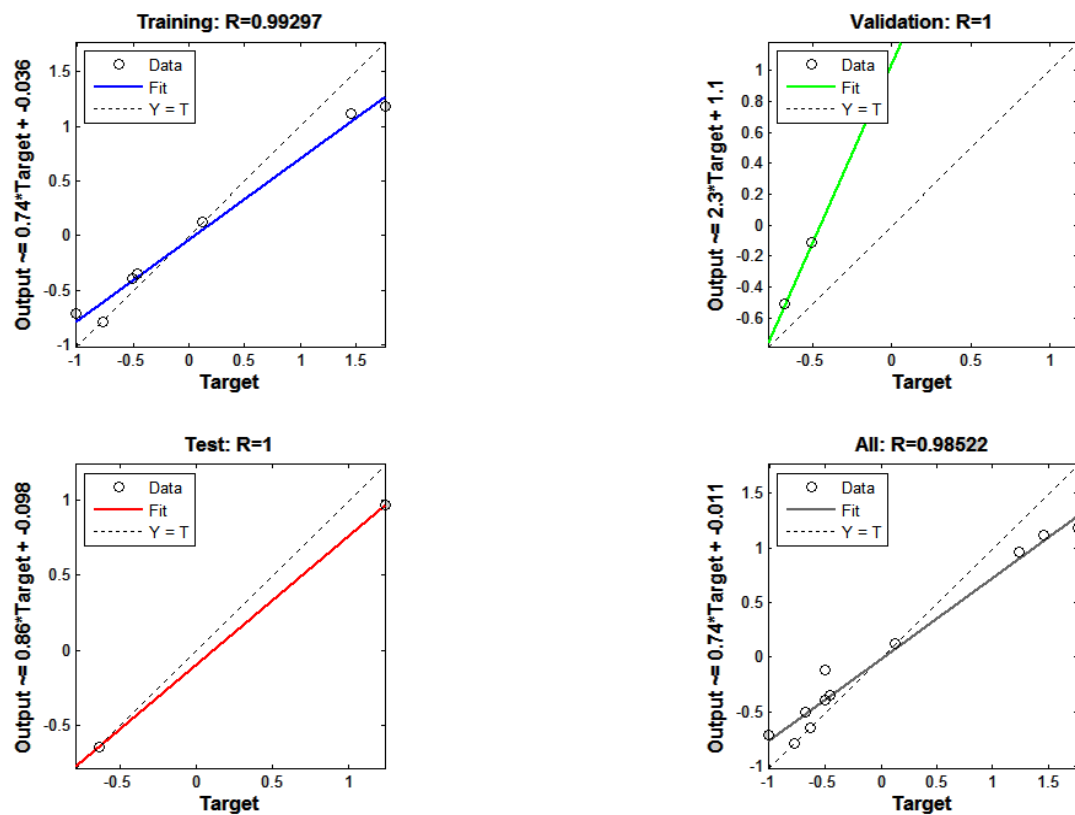


Figure 6.11: Regression for feature selection using ANN

The MSE plot shown in Figure 6.12 shows a rapidly decreasing error values until the third epoch. It then converges from the fourth epoch, with a steady pattern until the eighth epoch. The MSE starts at 2 for the first epoch, is significantly different to the 0.0095 error difference for the TSH model. The final error shows a difference of 0.0075, where the proposed model terminates early when compared to the FS (GA-ANN) model. This shows that the TSH model gives better prediction than the FS (GA-ANN) model, due to its advantages in terms of input selection and the generalization of the ANN.

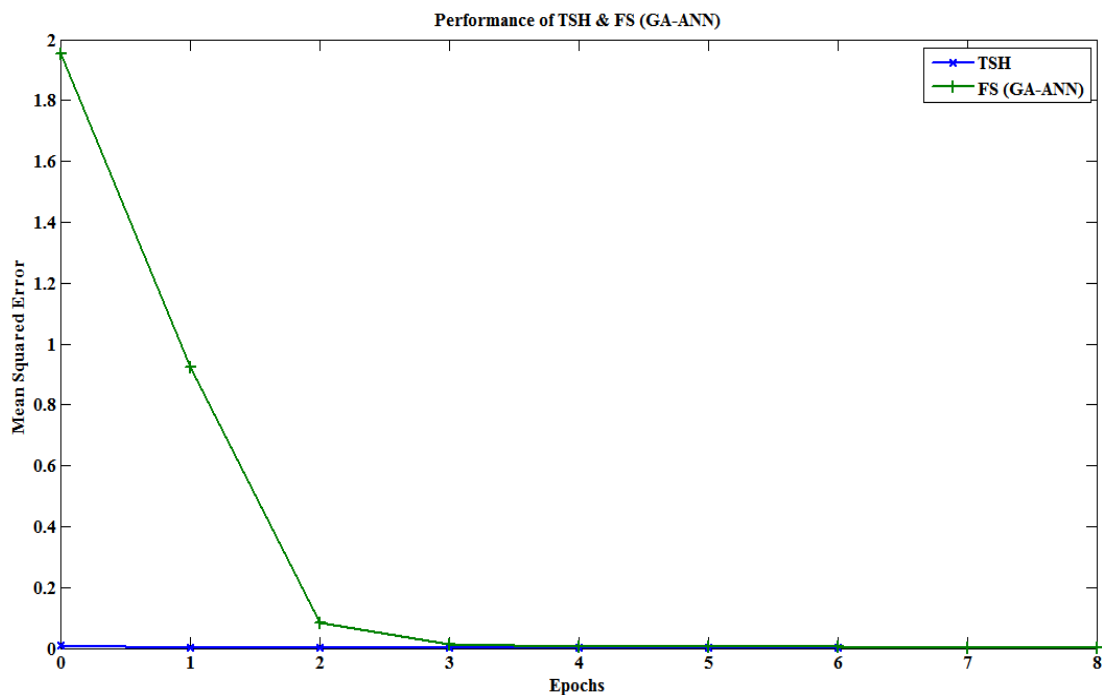


Figure 6.12: MSE for feature selection and TSH using ANN

### 6.8.3 Optimized Weight and Threshold Neural Network (OWTNN)

In conjunction with the TSH model, which uses a GA-ANN in optimizing the threshold and the weight of the ANN, benchmarking using a stand-alone GA-ANN also needs to be performed. This is done using the original dataset with input

variables for all 11 inputs. As with the previous techniques, the ANN module uses the same parameters both in the hybrid GA-ANN model and in remodelling the ANN for comparing the prediction performance. This technique is used specifically to give the optimum value of the threshold and weight for the ANN network, giving better generalization, thereby contributing to better prediction performance. All ANN parameters used are the same as those for the proposed model ANN module.

The regression plot in Figure 6.13 shows the performance for each dataset division, and the overall regression value for the prediction model, based on the optimized weight and optimized threshold in the ANN.

Overall performance is better than that of the TSH model by 1.99%. It seems that the generalization performance of the OWTNN is better than that of the proposed model, where the line fitting is almost equivalent to  $Y = T$  for all data set parts.

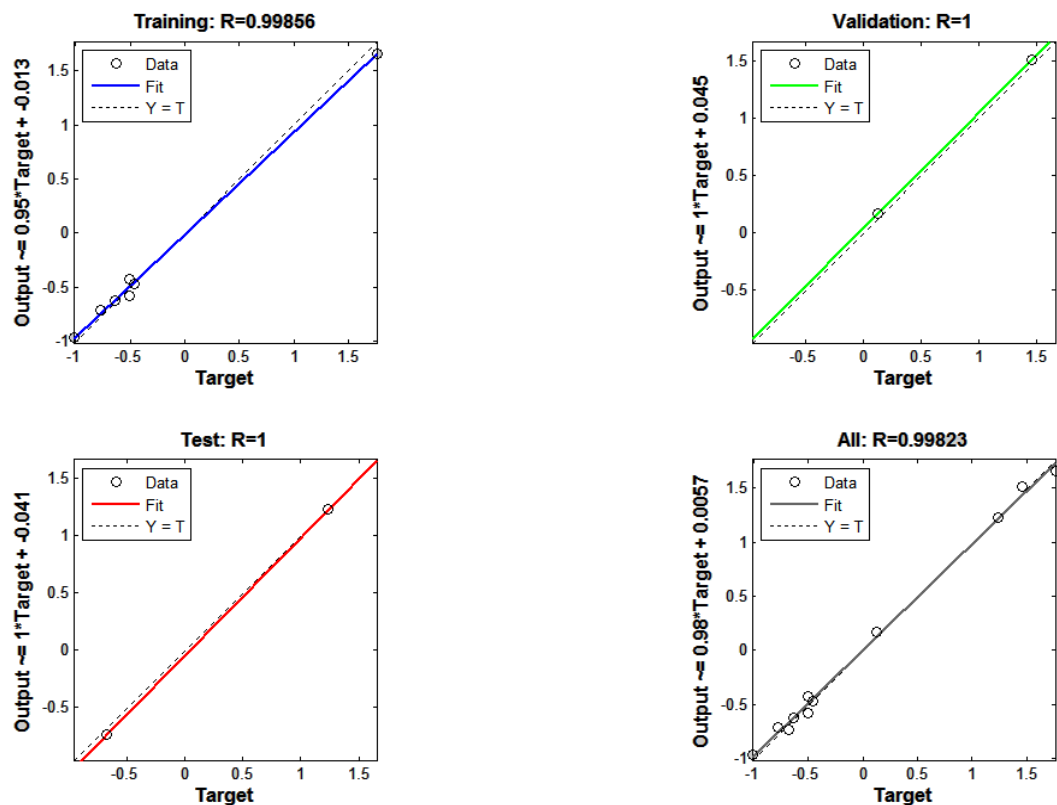


Figure 6.13: Regression for OWTNN using ANN

The error value of the OWTNN also shows that it outperforms the TSH model, where the MSE of the OWTNN model at the first epoch is 0.004072, compared with 0.009462 for the developed prediction model. Both models show a rapidly decreasing error value, which converge at the third epoch, and both also finish at the sixth epoch, but with an error difference of 0.00000473, where the OWTNN error is lower than the TSH model. This error analysis is based on the MSE plot, as shown in Figure 6.14.

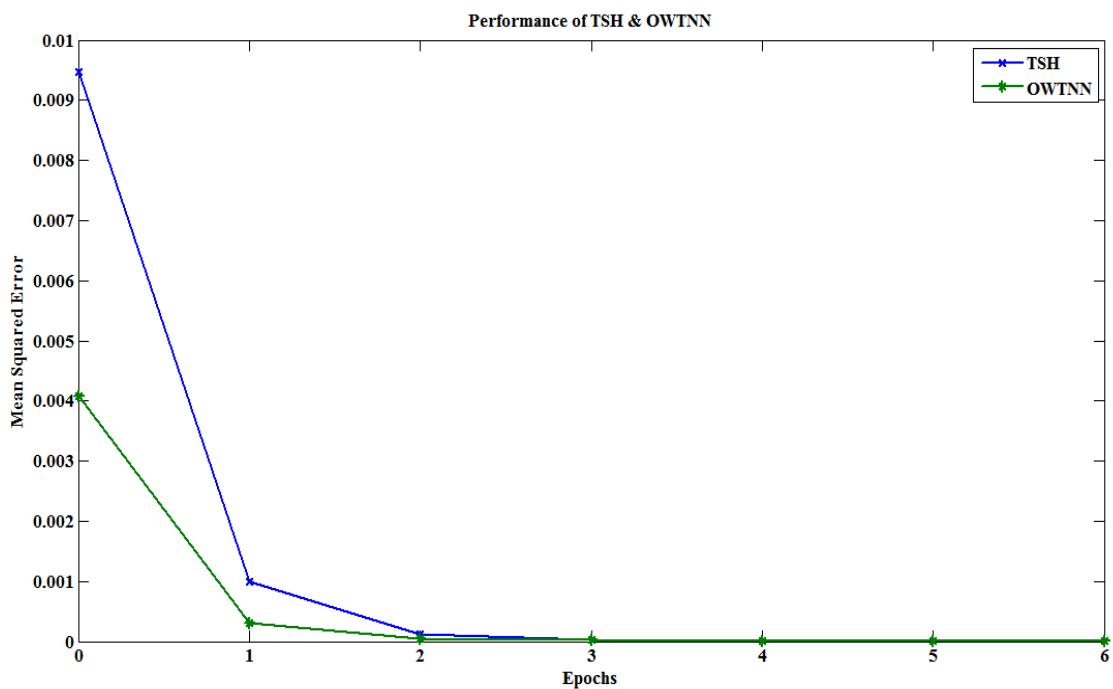


Figure 6.14: MSE for OWTNN and TSH using ANN

#### 6.8.4 Sensitive Genetic Neural Optimization (SGNO)

SGNO consists of three modules as describe in Chapter 3 and in (Zhang, 2011). Two of the modules are the same as those in the TSH model; the GA module and the ANN module. However, the chromosome tabulation for each generation shows better accumulation than the developed model, as shown in Figure 6.15. This is due to the 5-fold random cross-validation of the dataset, which is used because the chromosomes will widen the global search of the GA with multiple training by the

ANN, and because using the 5-fold chromosomes results in a larger population size than the proposed model. In this technique, the most important features are selected to be used in the ANN as the input variables, in comparing the prediction performance with the proposed model. A detailed description of the SGNO model is given in Chapter 2 and Chapter 3.

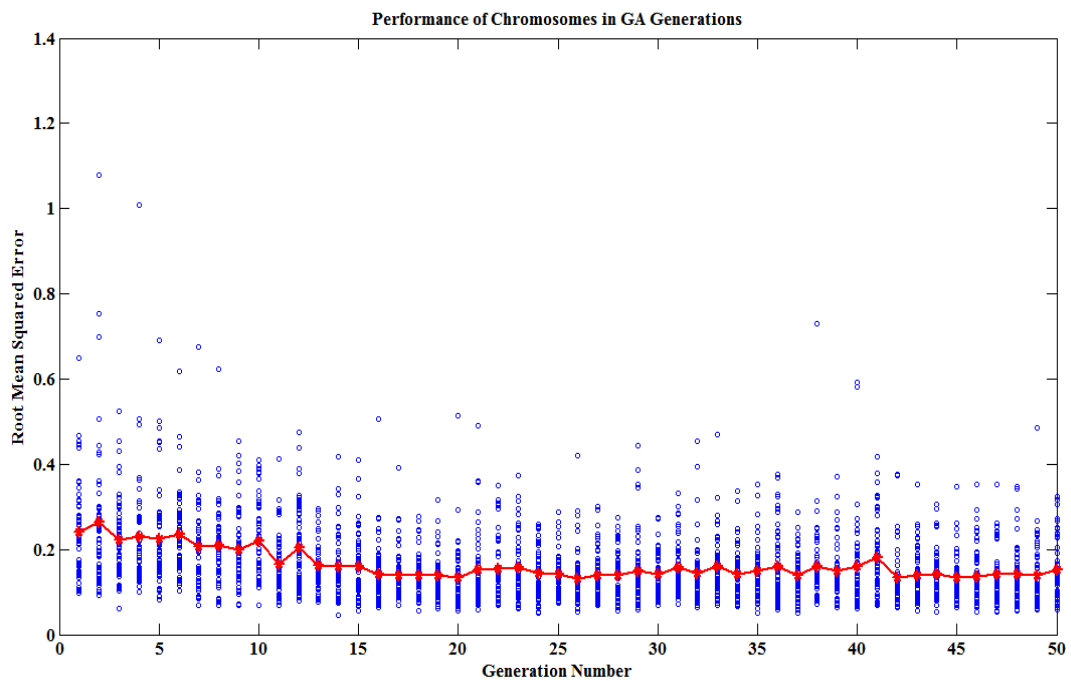


Figure 6.15: Performance of SGNO chromosomes vs number of generations

As each feature is generated randomly by the GA module to train the ANN module, the sensitivity analysis module is designed to analyze the global importance of each feature variable according to the variables most frequently appearing. The GA-ANN module then combines the chromosomes by groups, based on the lowest MSE for each generation. Next, each of the chromosomes in each group is selected based on the level of importance as quantified by the sensitivity module. Finally, the sensitivity analysis module ranks each of the features by taking the mean value of the

sensitivity scores in all selected chromosomes, as shown in Figure 6.16, which illustrates the global sensitivity scores of all the features.

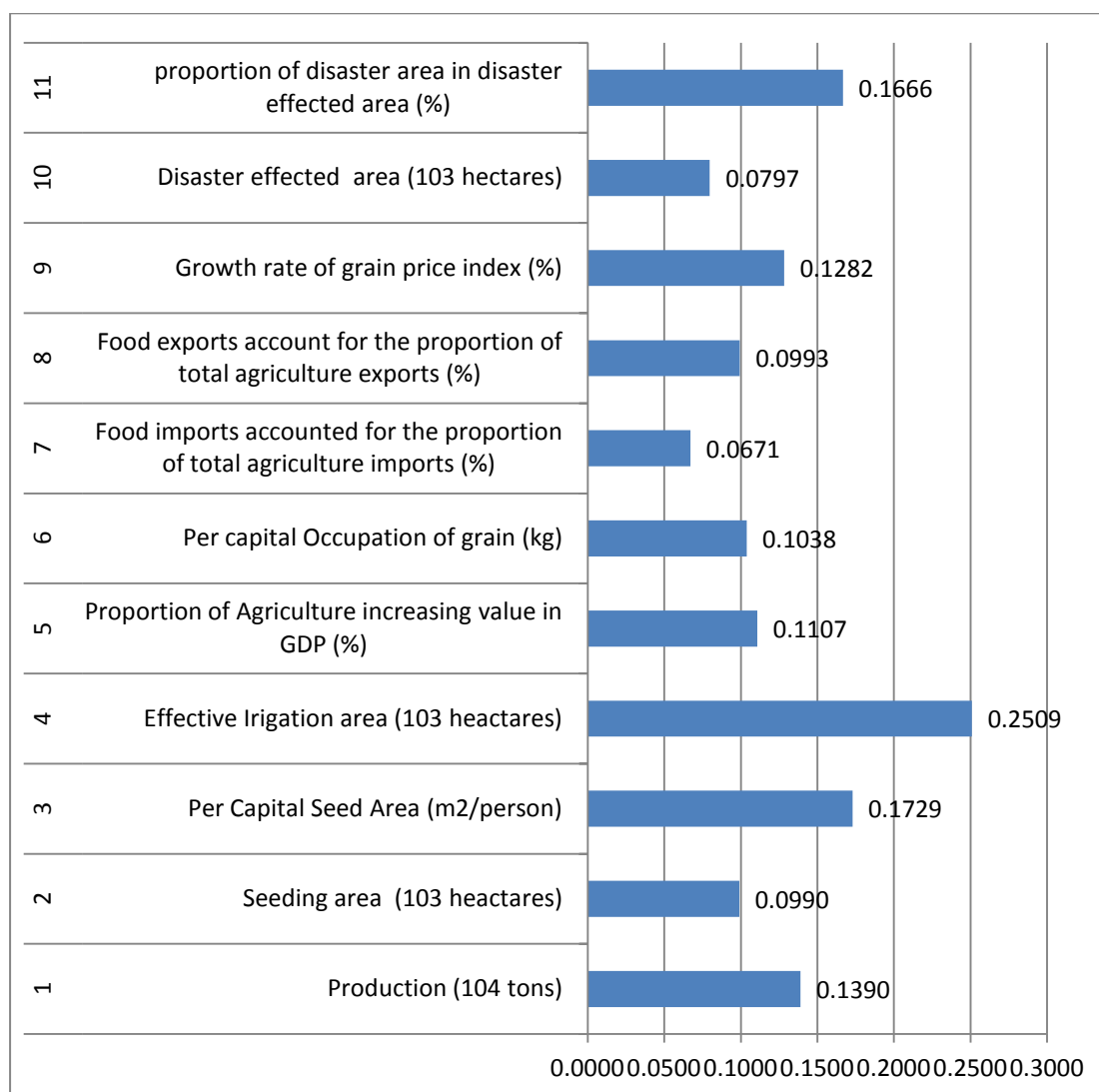


Figure 6.16: Mean for each of the feature variables

Each of the global sensitivity scores, also known as the mean feature scores, as shown in Figure 6.16, is then rearranged in descending order, where the higher scores represent the most important features. The rank list is shown in Table 6.3, referring to Figure 6.16 for the feature unit number and its description. From Table 6.3, six

features are selected to make comparisons with the performance of the prediction model, between the SGNO and the TSH model.

The 6 features are shown below:-

Six selected features = [Effective Irrigation area, Per Capital Seed Area, Proportion of disaster area in disaster effected area, Production, Growth rate of grain price index, Proportion of Agriculture increasing value in GDP]

Comparing the arrangement of the selected features with the TSH model, there is a different selection order, but two selected features are the same - [Per Capital Seed Area, Production]. The six selected features of the PCA will be used in the ANN to compare the prediction performance.

Table 6.3: Ranking selection for each of the features

Rank	Feature Unit	Feature description
1	4	Effective Irrigation area
2	3	Per Capital Seed Area
3	11	Proportion of disaster area in disaster effected area
4	1	Production
5	9	Growth rate of grain price index
6	5	Proportion of Agriculture increasing value in GDP
7	6	Per capital Occupation
8	8	Food exports account for the proportion of total agriculture exports
9	2	Seeding area
10	10	Disaster effected area
11	7	Food imports accounted for the proportion of total agriculture imports



The performance benchmark is compared based on the regression plot and MSE plot, as shown in Figure 6.17 and Figure 6.18. The overall regression shows a difference of 0.0395 when compared with the TSH model, where the proposed model gives a better regression value than the SGNO model. In terms of each dataset division performance, as shown in the previous chapter, the SGNO model has a good training regression value, but in the validation and testing parts, it seems that both plots do not fit to  $Y = T$ , although it gives a regression value of 1, and this is the reason for the scattered data in the overall regression plot.

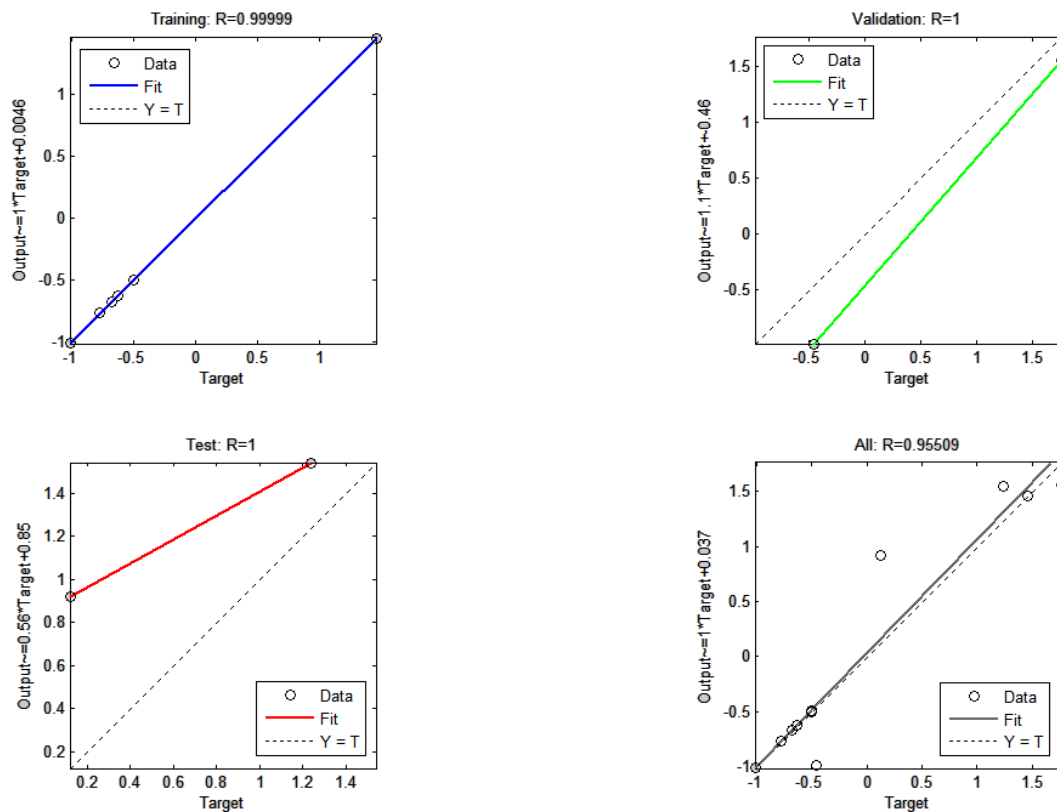


Figure 6.17: ANN performance based on SGNO for 6 input selections

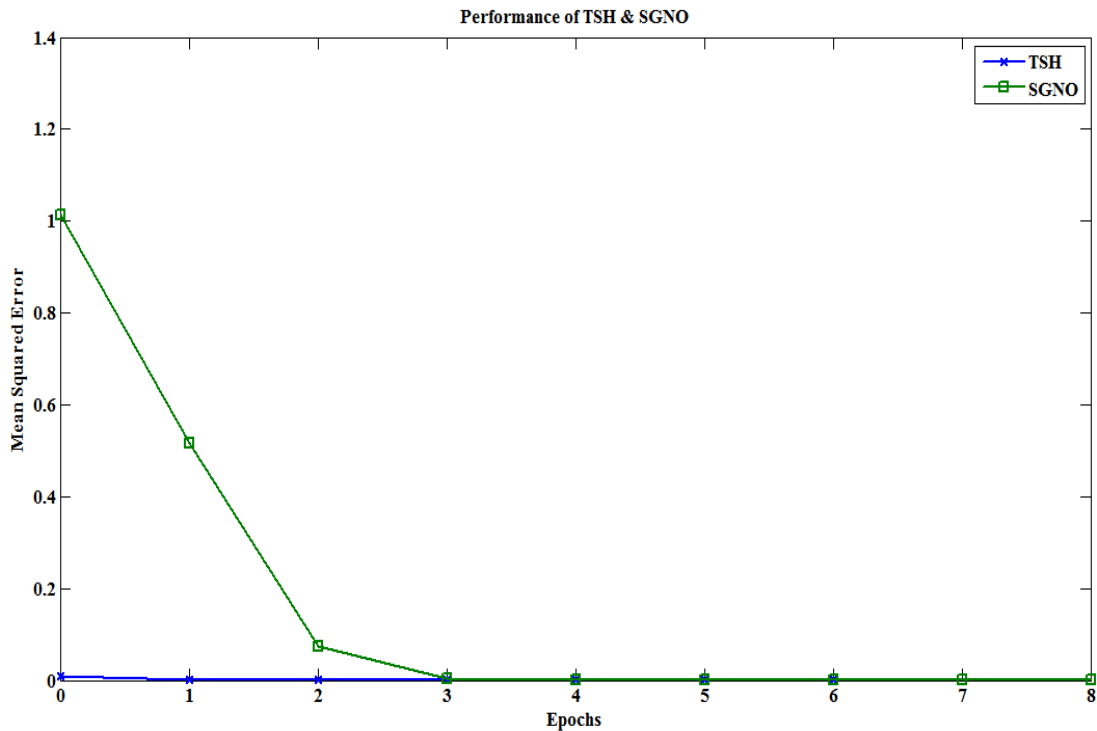


Figure 6.18: MSE performance of ANN models using TSH against SGNO

Figure 6.18 shows the MSE plot for each epoch. It illustrates that the SGNO takes up to eight epochs to find the best solution, compared with the TSH model which terminated at the sixth epoch. The error values of the SGNO model start at quite a high level, but rapidly decrease and converge at the fourth epoch. This shows that the TSH model also outperforms the SGNO model in terms of its errors and optimum speed of finding a final solution for the prediction.

### 6.8.5 Summary

In summarizing the benchmark of PCA, MLP-ANN, FS (GA-ANN), OWTNN and SGNO, an overall performance plot is shown in Figure 6.16 and Figure 6.17 for the regression plot and MSE plot respectively. The regression value for all techniques, except MLP-ANN, shows a prediction performance above 95%, which is very good. However, the OWTNN model produces a better regression value of 0.99823, whereas

the difference between the TSH model and OWTNN is only 0.00374. In term of the generalization of the ANN, the OWTNN model and the proposed model show very good data fitting at  $Y=T$ .

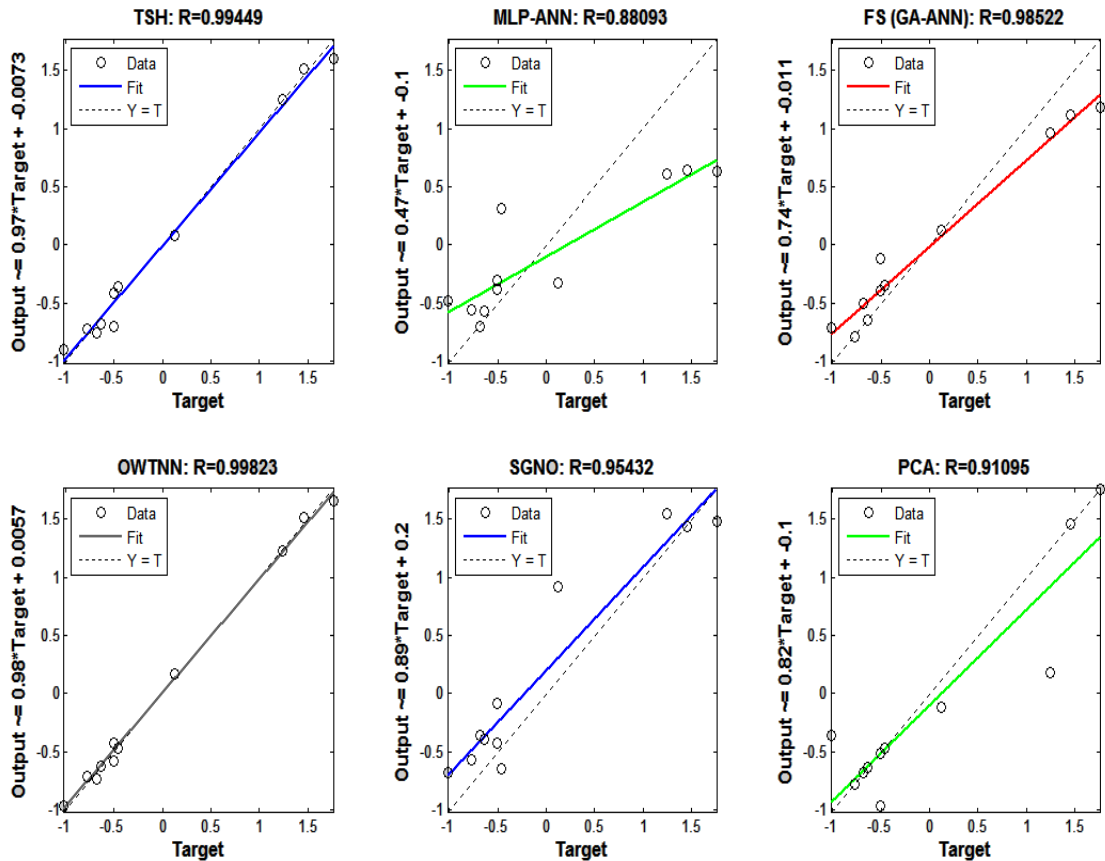


Figure 6.19: Benchmarking on overall ANN regression performance

In terms of the MSE, the OWTNN model outperforms the other techniques, either by showing the lowest final MSE value, or the lowest starting MSE value. Although the pattern for all techniques shows decreasing errors at each epoch, the OWTNN model and the proposed model both find the final prediction solution as early as the sixth epoch. The TSH model also shows a good MSE value, comparable with the OWTNN MSE value.

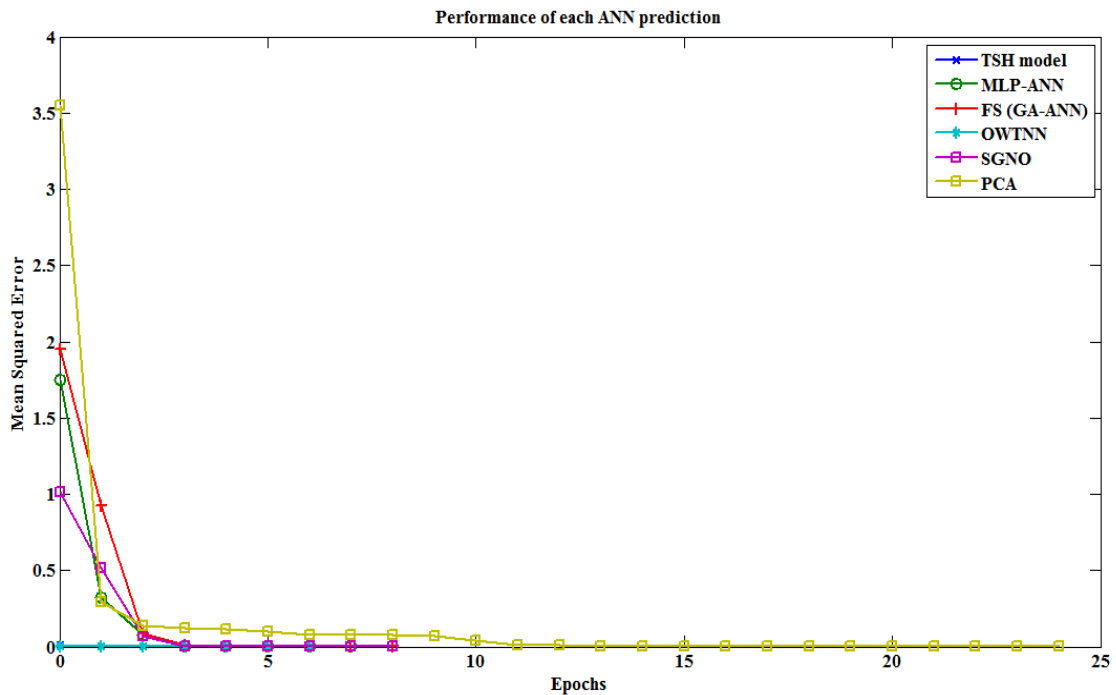


Figure 6.20: Benchmarking on MSE performance

## 6.9 Conclusion

In this chapter, a grain security risk assessment prediction model is developed based on the papers by (Jianling and Yong, 2010a, Jianling and Yong, 2010b, Kadir et al., 2011, Men et al., 2009, Yong and Jianling, 2010), which indicate a dataset with 11 features and 11 samples. This dataset contains three categories: productive indexes, consumption indexes and disaster indexes, where each index has features with multiples ranges of values.

Generally, the TSH model is used to establish a model for demonstrating the prediction of the grain security. It is generally used select the best inputs for use with the ANN in the first-stage process. The TSH model can also be used to find the best weight and threshold values for the selected inputs to be used in the ANN in the second-stage process. Finally, with the best features selected and the optimum weight

and threshold values defined, the ANN is remodelled using these values and input variables, for predicting the grain security.

The analysis of this prediction model shows the contribution of the optimum input variables, weight and threshold values to the prediction of the grain security. The prediction model uses the prediction accuracy performance assessment, in terms of the regression plot and MSE plot between the actual output and the developed model output, for its analysis and comparisons. The proposed model includes the analysis of each dataset division; training, validation and testing, which give the generalization results.

In making a final judgment on the model performance, the TSH model is benchmarked with other techniques including PCA, MLP-ANN, FS (GA-ANN), OWTNN and SGNO. The results for each of these techniques will be used in the ANN model, to compare the prediction performance via the regression plot and the MSE plot. Among these benchmark techniques, the OWTNN model has the lowest MSE value and lowest regression value. The OWTNN technique also gives very good generalization. Most of the techniques show a rapidly decreasing MSE value, and all of them converge between the third and fourth epoch. Although the OWTNN model shows the best prediction performance, the difference between the TSH model and OWTNN model is very small either in both the MSE performance plot and regression performance plot.

## **References**

DEFRA 2009. UK Food Security Assessment: Our approach. Department for Environment Food and Rural Affairs.

- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In:* DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). Department for Environment, Food and Rural Affairs.
- DONG HYUN JEONG, C. Z., WILLIAM RIBARSKY AND REMCO CHANG  
Understanding Principal Component Analysis Using a Visual Analytics Tool.
- EDUARD, L. & ET AL. 1999. Non-destructive banana ripeness determination using a neural network-based electronic nose. *Measurement Science and Technology*, 10, 538.
- FAO 2006. Policy Brief : Food Security. *In:* ECONOMICS, A. A. D. (ed.). FAO's Agriculture and Development Economics Division (ESA) with support from the FAO Netherlands Partnership Programme (FNPP) and the EC-FAO Food Security Programme.
- FAO. 2009. 2050: A third more mouths to feed. Available: <http://www.fao.org/news/story/en/item/35571/icode/> [Accessed 24/10/12].
- FAO 2011. The State of Food Insecurity in the world. 3rd ed. Rome: Food and Agriculture Organization of The United Nation.
- GARDNER, J. W. & ET AL. 1990. Application of artificial neural networks to an electronic olfactory system. *Measurement Science and Technology*, 1, 446.
- GARDNER, J. W., HINES, E. L. & TANG, H. C. 1992. Detection of vapours and odours from a multisensor array using pattern-recognition techniques Part 2. Artificial neural networks. *Sensors and Actuators B: Chemical*, 9, 9-15.
- GODFRAY, H. C. J., BEDDINGTON, J. R., CRUTE, I. R., HADDAD, L., LAWRENCE, D., MUIR, J. F., PRETTY, J., ROBINSON, S., THOMAS, S. M. & TOULMIN, C. 2010. Food Security: The Challenge of Feeding 9 Billion People. *Science*, 327, 812-818.

- JANG, J. S. R. Year. Input selection for ANFIS learning. *In: Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on, 8-11 Sep 1996* 1996. 1493-1499 vol.2.
- JIANLING, X. & YONG, D. Year. Food safety risk analysis based on generalized fuzzy numbers. *In: Advanced Management Science (ICAMS), 2010 IEEE International Conference on, 9-11 July 2010 2010a.* 699-702.
- JIANLING, X. & YONG, D. Year. Linguistic ranking model and its application in food management. *In: Computer Design and Applications (ICCD), 2010 International Conference on, 25-27 June 2010 2010b.* V5-208-V5-212.
- KADIR, M. K. A., HINES, E. L., AROF, S., ILLIESCU, D., LEESON, M., DOWLER, E., COLLIER, R., NAPIER, R., KEFAYA, Q. & GHAFARI, R. Year. Grain Security Risk Level Prediction Using ANFIS. *In: Computational Intelligence, Modelling and Simulation (CIMSIM), 2011 Third International Conference on, 20-22 Sept. 2011 2011.* 103-107.
- MEN, K., WEI, B., TANG, S. & JIANG, L. Year. China's Grain Security warning based on the integration of AHP-GRA. *In: Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on, 10-12 Nov. 2009 2009.* 655-659.
- WORLD BANK 2009. Global Economic Prospects.
- YONG, D. & JIANLING, X. Year. Fuzzy evidential warning of grain security. *In: Advanced Management Science (ICAMS), 2010 IEEE International Conference on, 9-11 July 2010 2010.* 703-706.
- ZHANG, F. 2011. *Intelligent Feature Selection for Neural Regression*. Doctor of Philosophy, University of Warwick.

# **Chapter 7: Food Security Risk Level Assessment by Using Fuzzy Logic**

## **7.1 Introduction**

In the previous chapters, the TSH model is applied to multiple food security prediction modelling problems, specifically to monitor and predict the outcome of the indicators of food security described in (DEFRA, 2009, DEFRA, 2010, FAO, 2006). Usually, to preventing food insecurity, the monitoring process should also start from starting point from which a raw food either as crops or livestock the end-product results. However, in considering all of the indicators which can impact on food security, an assessment model needs to be developed for further monitoring the whole system. In this chapter, a food security risk level assessment model is developed using Fuzzy Logic (FL), with the aim of ensuring that all of the indicators are within safe levels for food security.

In this study, three indicators are assumed to have a major impact on food security, indicating a safe (good) food security condition, an acceptable condition or a severe condition for a certain country. The three indicators use are crop yield, crop production and economic growth, which can be used as stepping stone to effective risk assessment of food security.



## 7.2 Background

As described in Chapter 1, 'food security' can be defined as 'the availability of food, and one's access to sufficient and affordable food'. The definition of food security has itself been greatly debated; for example the Food and Agriculture organization (FAO) states that "Food security exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life" (Organization, 2006). This definition is clear for the purposes of this research. Sometimes however, food security is also linked with a wider range of sustainability issues drawn from economic, social and environmental agendas.

The main factors affecting food security are: food availability, referring to the needs of an individual (encompassing food prices, distance to shops, income available for food purchasing), food affordability, nutritional content, food safety, food system resilience and consumer confidence (DEFRA, 2010, Peihong and Jiaqiong, 2009). Each of these factors can be represented by various indicators, such as: trends in the global output of food (from farm to end products), land-use changes, diversity of supply, energy dependency of food chain, income factors. Trends in food-borne pathogen cases are also considered, although the monitoring of these cases can be difficult due to the long-term effects of pathogens such as Salmonella, Listeria, E. Coli O157 and Campylobacter (DEFRA, 2010). Most of these factors are related to consumer demand and supply chain channels, which are not fully controllable; therefore it is difficult to use a conventional data-based approach, which would require precise information to describe every single interaction. As indicated in (Ding et al., 2007), each of the indicators used in this chapter, which relate to the food chain, are based on imprecise inputs relating to the food chain. Therefore a method which

can translate this information, either in quantitative or qualitative terms, needs to be used in modelling the risk level assessment.

Before considering the development of the model, the risk level assessment itself should be explored. The study of risk level assessment is a common topic amongst researchers. Such studies usually involve consideration of the safety precautions in a working environment, or of the modelling of decision-making process. In (Meltzer et al., 2003, Xiaojun et al., 2008), risk is defined as “the probability of a negative effect which attempts to describe the possible adverse pressure on the system caused by a hazard”.

As previously described, many advanced risk assessment methods have been used in aiding the development of decision making modelling, and risk assessments are considered to be a valuable tool in relation to food security projections and their associated decision support systems. For example, in the construction industry, where the practice is comparatively mature, Fault tree analysis, event tree analysis, Monte Carlo analysis, scenario planning and sensitivity analysis are prevalent as the conventional risk assessment techniques (Peihong and Jiaqiong, 2009).

Although risk assessment methods are popular among scientists and engineers, there are commonly having problems with the accuracy of the results. For example, the Analysis Hierarchy Process (AHP) decision making technique is used in the construction industry. AHP is an improved and advanced risk assessment model, but it still has a drawback, in that it can only operate with definite scales and measured commodities.

Fuzzy Logic (FL) can work effectively with many parameters and non-uniform variables, suggesting that it can address most of the drawbacks in the

previously used and more conventional techniques. The FL technique is an alternative method that is becoming more frequently used to improve the performance of risk assessment systems (Zeng et al., 2007).

There are some previous studies where FL has been successfully applied to specific elements of the food chain and to food security. These include: China's grain security warning study (Jianling and Yong, 2010a, Jianling and Yong, 2010b, Yong and Jianling, 2010, Muhd Khairulzaman Abdul Kadir, 2013), and crop control (M. Ahmend, 1999) and Gari fermentation plant (Odetunji and Kehinde, 2005). It can therefore be stated that FL systems offer a number of advantages when compared to conventional data-based approaches. One of the main advantages of FL systems is that they can be easily implemented and tuned; FL systems use 'IF-THEN' rules that generate outputs based on imprecise inputs. Therefore as described earlier in this section, the previous research on FL described in section 2.2 shows that the choice of the FL selected was also based on its capability to convert any fuzziness into expert language, and at the same time the FL can interpret each term used based on the knowledge of human understanding.

The work in this chapter will concentrate on creating a food security risk level modelling system by using an FL technique as described in (Muhd Khairulzaman Abdul Kadir, 2013), for five countries, including the use of a producer price performance index in testing each risk level result from the proposed model. This model should be able to determine the overall level of prevailing food security risk by monitoring various risk elements within a food supply system.

### **7.3 Dataset**

The data has been selected based on the previously defined inputs to the model; crop yield, crop production and economic growth. These input variables are assumed to have a major impact on the food security risk level, for which the analysis is assumed to start from the roots - farming and manufacturing, and to conclude with impact on the overall economic environment. Each input parameter is taken from each of the following five countries United Kingdom (UK), Germany, Australia, China and India. These countries use a lot of cereal crops both directly as food and for other purposes, therefore cereal is used as the crop input to this model. The period over which the data used was recorded is between 1988 and 2008, and it has been taken from two online databases – World Bank (Bank, 2010) and Food and Agriculture Organization (FAO) (FAO, 2010) which are available for public usage. The cereal yield unit is in hectograms/hectare (Hg/Ha) and the cereal production unit is in tonnes. The economic growth input variable, which is based on growth percentage of GDP for the same period, is taken from the same sources. These three input variables are used in testing the FL model of food security risk level assessment for each of the five countries, and the outcome is then be compared with the performances index that being decided.

### **7.4 Data Pre-processing**

Each of the data compilations has various ranges of values because of the different units and different types of data presentation used. In Table 7.1, the basic statistics of each data input are shown for each of the five countries. To make the fuzzy model evaluation more simple, especially in the determination of fuzzy sets in

assessing multiple countries; the data needs to be standardized to a range from 0 to 1, using Equation (7.1), where  $X$  is the input variables matrix,  $X_i$  represent each of the input,  $X_{min}$  is the minimum input and  $X_{max}$  is the maximum input. Equation (7.1) is applied to all input variables for the dataset. Equation (7.1) is selected as the standardization of the dataset rather than Equation (3.1), because the fuzzy sets range for the fuzzification process has been determined as 0 to 1 as shown in table 7.2 and table 7.3. The datasets for each country will be processed separately to the FL model, and Equation (7.1) will ensure that each countries' datasets will have the same range of fuzzy sets as the FL model of food security risk level assessment. Rather than Equation (3.1), the normalization range will be in negative (-) and positive (+) values.

$$X = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (7.1)$$

## 7.5 Fuzzy Logic Modelling

FL is widely used as a reasoning process because of its ability to express outputs in quantitative terms which can be easily understood by many people (Perrot et al., 2006). The main contributor to the modern era of FL, and most of the application studies, is Lotfi Zadeh. FL was introduced to cope with vagueness in linguistics and the challenges in expressing human 'knowledge' in a natural, but generally imprecise way (Haslum et al., 2007). This is also the reason for using FL as a technique in modelling the food security risk level assessment.

One study by (Xiaojun Wang, et al) used fuzzy set theory and an analytical hierarchy process, which can be applied specifically to the food industry (Xiaojun et

al., 2008). The principles in this paper, and those in another paper by (Muhd Khairulzaman Abdul Kadir, 2013), are combined and applied to the proposed model, to determine the level of risk in food security. The block diagram of the proposed model is shown in Figure 7.1.

Table 7.1: Basic statistic of the dataset for each country

<b>Country</b>	<b>Input Variables</b>	<b>Min</b>	<b>Max</b>	<b>Mean</b>	<b>Standard Deviation</b>
<b>UK</b>	Yield	53993	74192	67105.24	5163.61
	Production	18959000	24576000	21907220.67	1616310.14
	GDP	-1.39	5.03	2.49	1.46
<b>Australia</b>	Yield	10544	22192	17545.86	3534.09
	Production	15410973	41630823	28561831.33	7881868.91
	GDP	-0.64	5.10	3.41	1.44
<b>Germany</b>	Yield	51737	73572	61867.90	6216.44
	Production	34758462	51110273	42037923.33	4908567.70
	GDP	-0.80	5.26	2.08	1.59
<b>China</b>	Yield	39365	55243	47689.05	4299.19
	Production	351824290	480053700	417995246.52	33673713.70
	GDP	3.80	14.20	9.95	2.86
<b>India</b>	Yield	17757	26515	22175.19	2323.34
	Production	183867008	267022200	222317949.33	22416615.32
	GDP	1.06	9.82	6.42	2.29

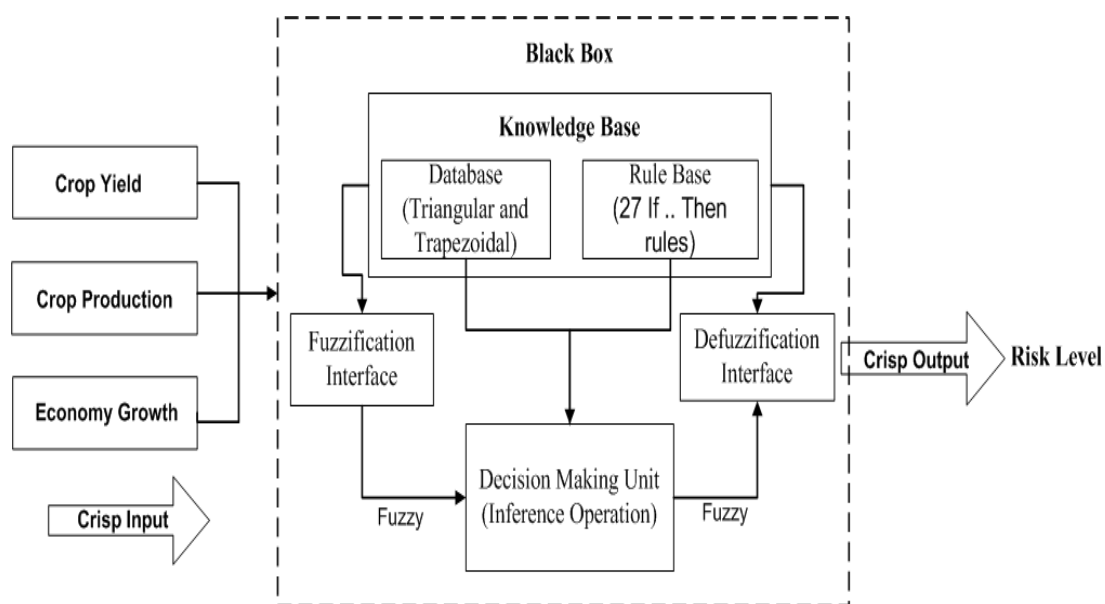


Figure 7.1: Food security risk assessment model

In determining the risk level of food security, the developed model uses three inputs from the dataset described in the previous section: crop yield (defined as the monthly farm gate crop output), crop production (defined as crops which are processed into food products), and economic growth (defined as the growth as a percentage of the gross domestic product, GDP). These three indicators are assumed to represent all of the key components in the DEFRA food security discussion report; this work by Monty P. Jones (Jones, 2008), describing a study in sub-Saharan Africa (DEFRA, 2009, DEFRA, 2010, Odetunji and Kehinde, 2005), shows that the first two inputs are good indicators of overall food availability. The paper by (Jones, 2008) also indicates that economic growth is strongly associated with the food security, which is the reason for the use of GDP as one of the inputs to the model.

Figure 7.1 also shows the following five functional blocks: 1) rule base, 2) database, 3) decision making unit, 4) fuzzification interface which converts the crisp

input values into linguistic values, also called fuzzy values, and 5) defuzzification (J.-S.R. Jang, 1997). The objective in using the FL technique is to turn these semi-precise or qualitative measures into quantitative assessment outcomes.

In this model, each of the inputs has been chosen to have three fuzzy sets that will determine the degree of each of the inputs, as shown in Table 7.2. The fuzzy set ranges are referred to the standardization of the corresponding crisp input value, based on its universe of discourse, by using Equation 7.1. For this study, as previously described in the dataset section, cereal data is used as an example for the type of crop being modelled. Each of the FL blocks will be explained in detail in next section.

### **7.5.1 Fuzzification Interface**

In FL, there are two fuzzification methods; the Mamdani method and the Sugeno method. In this model, the fuzzification process involves rule evaluation and aggregation using the Mamdani method. The Mamdani method is used because of its capability in capturing expert knowledge, whereas the Sugeno method mostly uses a singleton rule output, which only work well with linear techniques (J.-S.R. Jang, 1997, Jones, 2008, Negnevitsky, 2005). A very important part of this process is the determination of the fuzzy sets. Each rules needs to be defined based on grades of importance for each of the inputs and the outputs of the system that is being modelled (Huey-Ming, 1996). The overall fuzzy sets ranges are shown in Table 7.2 and Table 7.3.



### 7.5.2 Knowledge Base and Decision Making Unit

In the Database block, the input grading is divided into three grades, from grade 1 to grade 3, which are determined based on their highest to lowest value, and then divided into 3 sets. The input grading is applied to all inputs and output; cereal yield, cereal production and economic growth, as shown in Table 7.2.

Table 7.3 shows the grade of risk, which indicates the rate of food security risk level; this grade of risk is also scaled from 1 to 3. The range of each input and output is determined by the maximum and minimum value for each input parameter for the specific year, as specified in Equation 7.1. The purpose of the standardization of the data is to ensure that the values from all countries fall within the range of the fuzzy number for the fuzzy set of the FL model.

Table 7.2: The grade of fuzzy inputs

<b>Input</b>	<b>Fuzzy set</b>	<b>Fuzzy number</b>
Cereal Yield	1: High 2: Medium 3: Low	0 – 1
Cereal Production	1: High 2: Medium 3: Low	0 – 1
Economic Growth	1: High 2: Medium 3: Low	0 – 1

Table 7.3: The grade of fuzzy output

<b>Output</b>	<b>Fuzzy set</b>	<b>Fuzzy number</b>
Risk Level	1: Good 2: Acceptable 3: Severe	0 – 1

The Database block not only determines the fuzzy sets, but it also defines the membership function used with each fuzzy set. In designing and implementing the FL control system (Negnevitsky, 2005), the option exists to choose which of the three most popular membership functions should be used; triangular, Gaussian or trapezoidal functions. For this model, a triangular function has been selected for the inputs and a trapezoidal function for the outputs of the model. The reason for using the triangular and trapezoidal functions is that the leniency of these theorems (Negnevitsky, 2005) means that they can perform translation of the rules very quickly, although the level of accuracy will be lower than that possible with either of the other membership functions. This is consistent with the normal speed versus complexity scenario (Xie et al., 1998). As a starting point in developing the risk level assessment model, the triangular and trapezoidal functions were selected by considering the fuzzy set dataset range between 0 and 1, which simplifies the complexity of the FL.

In the Rule Base block, a Bayesian rule (Jang, 1993) is used in determining the rule relationship for each of the inputs and the output. This rule relationship is also known as ‘If...then’ rules. In this model, each of the 3 inputs has been assigned to three membership functions, which will generate 27 rules ( $3^3$ ) as shown in Figure 7.2. This number of rules is then put into the Rule Base block as shown in Figure 7.1.

1. If (Cereal Yield is high) and (Cereal Production is high) and (Economic Growth is High) then (Risk Level is Good) (1)
2. If (Cereal Yield is high) and (Cereal Production is high) and (Economic Growth is Medium) then (Risk Level is Good) (1)
3. If (Cereal Yield is high) and (Cereal Production is high) and (Economic Growth is Low) then (Risk Level is Acceptable) (1)
4. If (Cereal Yield is high) and (Cereal Production is medium) and (Economic Growth is High) then (Risk Level is Acceptable) (1)
5. If (Cereal Yield is high) and (Cereal Production is medium) and (Economic Growth is Medium) then (Risk Level is Good) (1)
6. If (Cereal Yield is high) and (Cereal Production is medium) and (Economic Growth is Low) then (Risk Level is Acceptable) (1)

7. If (Cereal Yield is high) and (Cereal Production is low) and (Economic Growth is High) then (Risk Level is Severe) (1)
8. If (Cereal Yield is high) and (Cereal Production is low) and (Economic Growth is Medium) then (Risk Level is Severe) (1)
9. If (Cereal Yield is high) and (Cereal Production is low) and (Economic Growth is Low) then (Risk Level is Severe) (1)
10. If (Cereal Yield is medium) and (Cereal Production is high) and (Economic Growth is High) then (Risk Level is Good) (1)
11. If (Cereal Yield is medium) and (Cereal Production is high) and (Economic Growth is Medium) then (Risk Level is Good) (1)
12. If (Cereal Yield is medium) and (Cereal Production is high) and (Economic Growth is Low) then (Risk Level is Acceptable) (1)
13. If (Cereal Yield is medium) and (Cereal Production is medium) and (Economic Growth is High) then (Risk Level is Acceptable) (1)
14. If (Cereal Yield is medium) and (Cereal Production is medium) and (Economic Growth is Medium) then (Risk Level is Good) (1)
15. If (Cereal Yield is medium) and (Cereal Production is medium) and (Economic Growth is Low) then (Risk Level is Good) (1)
16. If (Cereal Yield is medium) and (Cereal Production is low) and (Economic Growth is High) then (Risk Level is Severe) (1)
17. If (Cereal Yield is medium) and (Cereal Production is low) and (Economic Growth is Medium) then (Risk Level is Acceptable) (1)
18. If (Cereal Yield is medium) and (Cereal Production is low) and (Economic Growth is Low) then (Risk Level is Acceptable) (1)
19. If (Cereal Yield is low) and (Cereal Production is high) and (Economic Growth is High) then (Risk Level is Severe) (1)
20. If (Cereal Yield is low) and (Cereal Production is high) and (Economic Growth is Medium) then (Risk Level is Acceptable) (1)
21. If (Cereal Yield is low) and (Cereal Production is high) and (Economic Growth is Low) then (Risk Level is Severe) (1)
22. If (Cereal Yield is low) and (Cereal Production is medium) and (Economic Growth is High) then (Risk Level is Severe) (1)
23. If (Cereal Yield is low) and (Cereal Production is medium) and (Economic Growth is Medium) then (Risk Level is Acceptable) (1)
24. If (Cereal Yield is low) and (Cereal Production is medium) and (Economic Growth is Low) then (Risk Level is Acceptable) (1)
25. If (Cereal Yield is low) and (Cereal Production is low) and (Economic Growth is High) then (Risk Level is Severe) (1)
26. If (Cereal Yield is low) and (Cereal Production is low) and (Economic Growth is Medium) then (Risk Level is Severe) (1)
27. If (Cereal Yield is low) and (Cereal Production is low) and (Economic Growth is Low) then (Risk Level is Acceptable) (1)

Figure 7.2: Rule list showing the connection between the inputs and the output

In the Decision Making unit, an AND function is used as the fuzzy operator to compare each of the inputs. This function is also known as the algebraic product function (Negnevitsky, 2005), and the output function will always be '0' if one of the fuzzy inputs is '0' or both of the fuzzy inputs is '0'. This block will also perform all inference operations between the Fuzzification Interface, Database and Rule Base blocks.

### **7.5.3 Defuzzification**

Defuzzification is the final process in a FL 'black box', where it converts all of the fuzzy values back to the original values for the model. For example, to get the crop value (crisp value), the fuzzy output value needs to be defuzzified, or to be aggregated at the rule output (A.S. Sodiya, 2007). In order to perform the defuzzification, a number of different approaches may be used, for example as described by (J.-S.R. Jang, 1997, Negnevitsky, 2005), a detailed discussion of which is given in Chapter 2. Here, the centre of gravity or centroid, as describe in Equation 2.1, is used as the defuzzification technique, which will converts the fuzzy value to a crisp value in the interface block.

## **7.6 Risk Level Assessment Model Analysis**

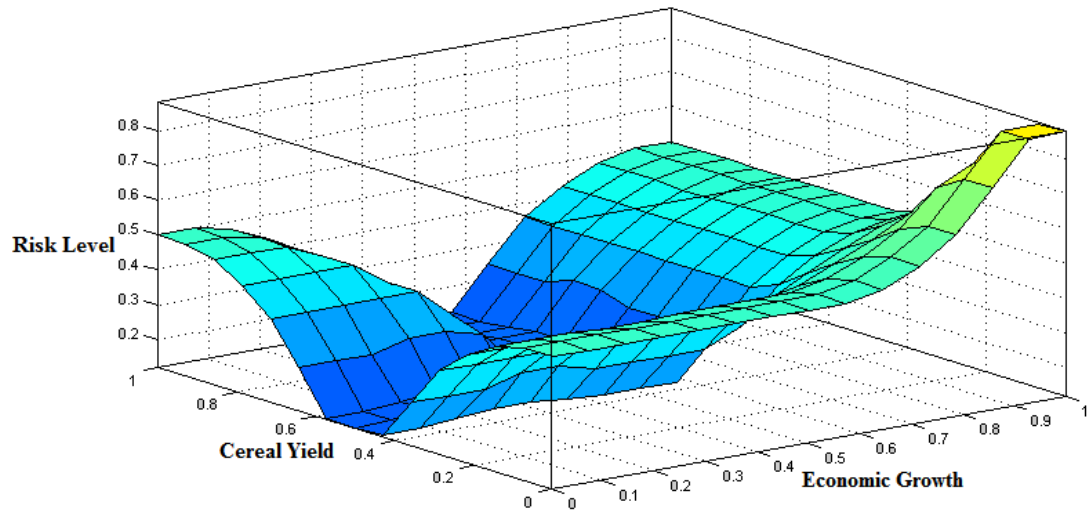
In analyzing the risk model, the FL model is developed as shown in Figure 7.1, with three inputs and one output. A FL 'black box' is created based on each process, as described in earlier in this chapter.

Figure 7.3(a), (b) and (c) shows the control surface for the qualitative risk assessment being discussed. It describes the possible relationships between the input variables in accordance with the overall risk level. The three-dimensional curve represents the mapping from the input to the overall risk level. Although only 2 inputs (x and y-axis) and 1 output (z-axis) can be shown in the three-dimensional plots, it is possible to plot multiple diagrams to illustrate the relationship between different input factors, and their impact on the overall risk level. The output quantifies the level of risk factors affecting overall food security level, which can be useful to monitor the outcome of the model before it is used to process real data.

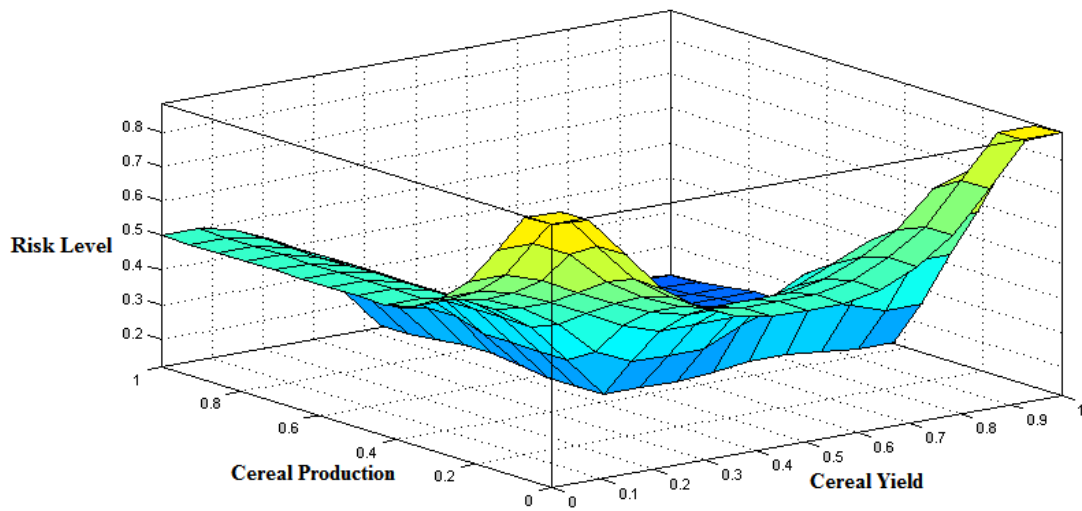
To aid understanding of the relationship between the fuzzy sets and the fuzzy value range, Figure 7.4 and Figure 7.5 have been plotted as diagrammatical representation of these values. Figure 7.4 shows the representation of each input membership function and Figure 7.5 is the membership function for the output. This diagram can be used in translating the fuzzy sets as discussed in an earlier section and as shown in Table 7.2 and Table 7.3.

In order to verify the results, the datasets in Section 7.3 are used as the inputs for five countries. The results show that every year, the assessment of risk level of food security will change, depending on cereal yield, cereal production and economic growth. For example, in 1988, as shown in Table 7.4, the results show that the UK, Australia, Germany, China and India gives food security risk level values of 0.87, 0.6, 0.74, 0.86 and 0.88 respectively. Therefore, from the output of this model, the values of food security risk level for the UK, China and India become almost 0.9, which is defined as a risky (severe) condition, whereas, for Australia and Germany the value is between 0.2 and 0.8, which is an acceptable condition. The levels for the good,

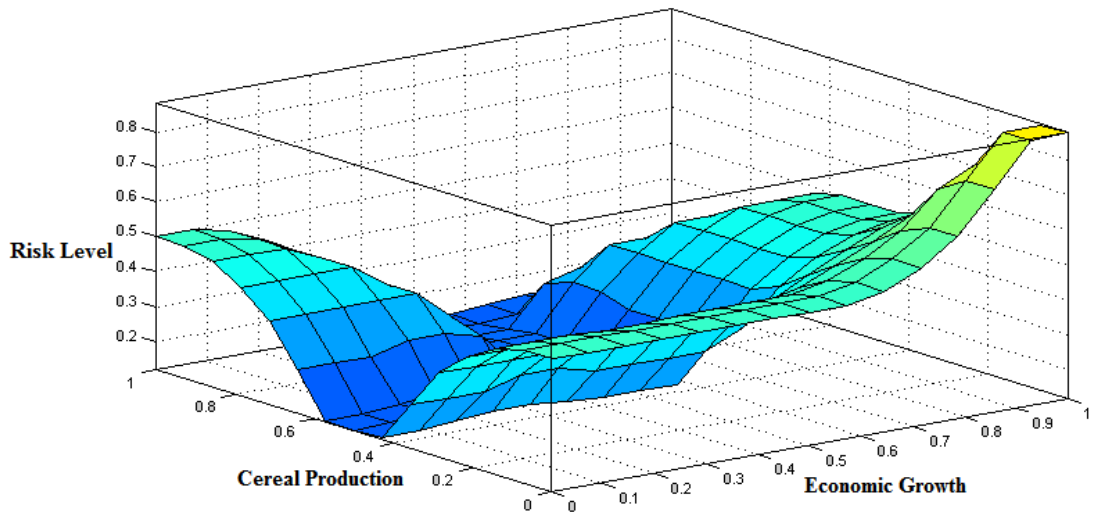
acceptable and severe conditions are based on the assumptions of food security risk level determined in the previous section.



(a)



(b)



(c)

Figure 7.3: Control surface diagram

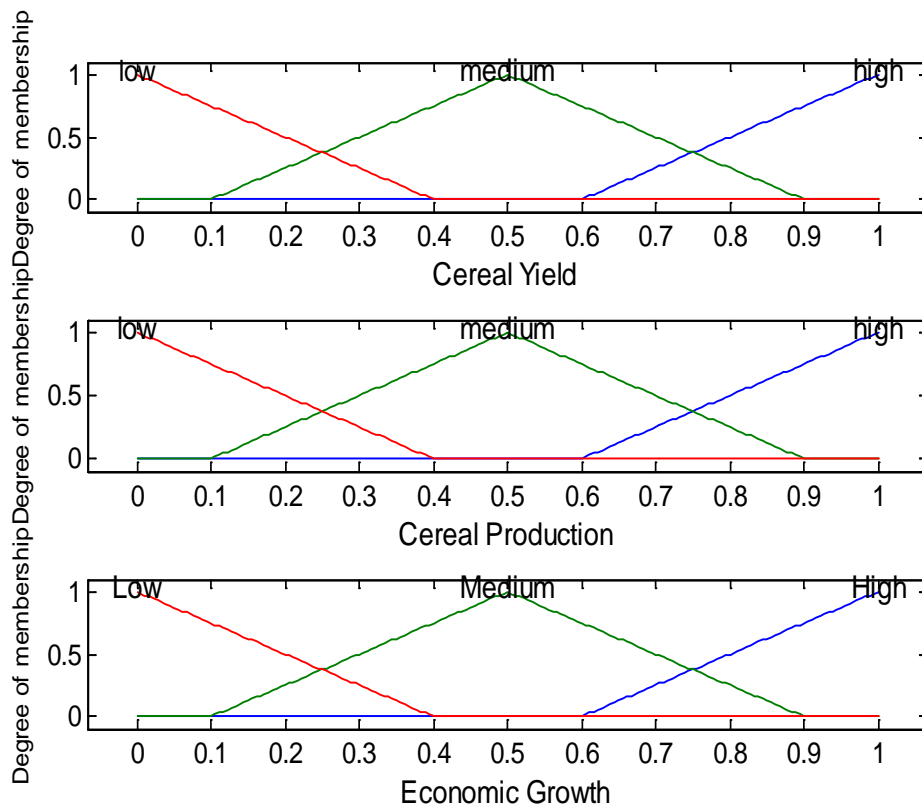


Figure 7.4: Plots for each input membership function and its range

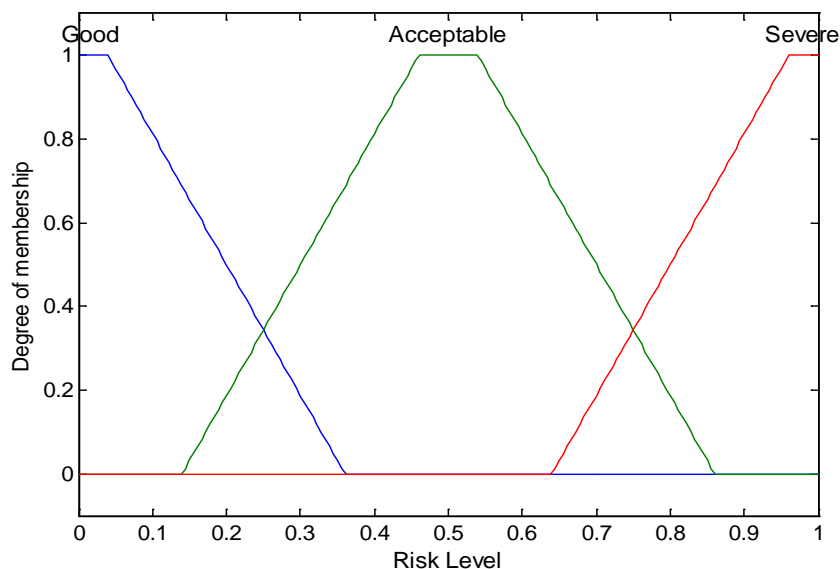


Figure 7.5: Plots for the output membership function and its range

Table 7.4: Summary of input and output for the year 1988

	Cereal Yield (Hg/Ha)	Cereal Production (tons)	Economic Growth (%)	Food Security Risk Level
UK	53993	21063000	5.03	0.87
Australia	16294	21891006	5.16	0.60
Germany	51737	36931897	3.71	0.74
China	39365	351824290	11.30	0.86
India	17757	183867008	9.64	0.88

Based on the results of table 7.4, the model shows that a high yield and a high level of production should lead to high food security. However, when economic growth is low, people will tend to pay the lowest price for their food, which means that the least possible resources are expended by the consumer in order to get the best food. Although some people will often be prepared to pay for a given commodity



even if they cannot afford it, this will not affect the entire system, because this was assumed to be a minor issue. Therefore, an observation based on this result is that the food is likely to be wasted, especially the highest quality food which is generally the most expensive. However, if economic growth is high and the cereal yield is low, the food security is low and there may not be enough food for everyone. The overall risk assessment pattern for each of the five countries is shown in Figure 7.7, which is based on the real data. For most of the years the result for each countries shows an acceptable (0.2 – 0.8) risk level.

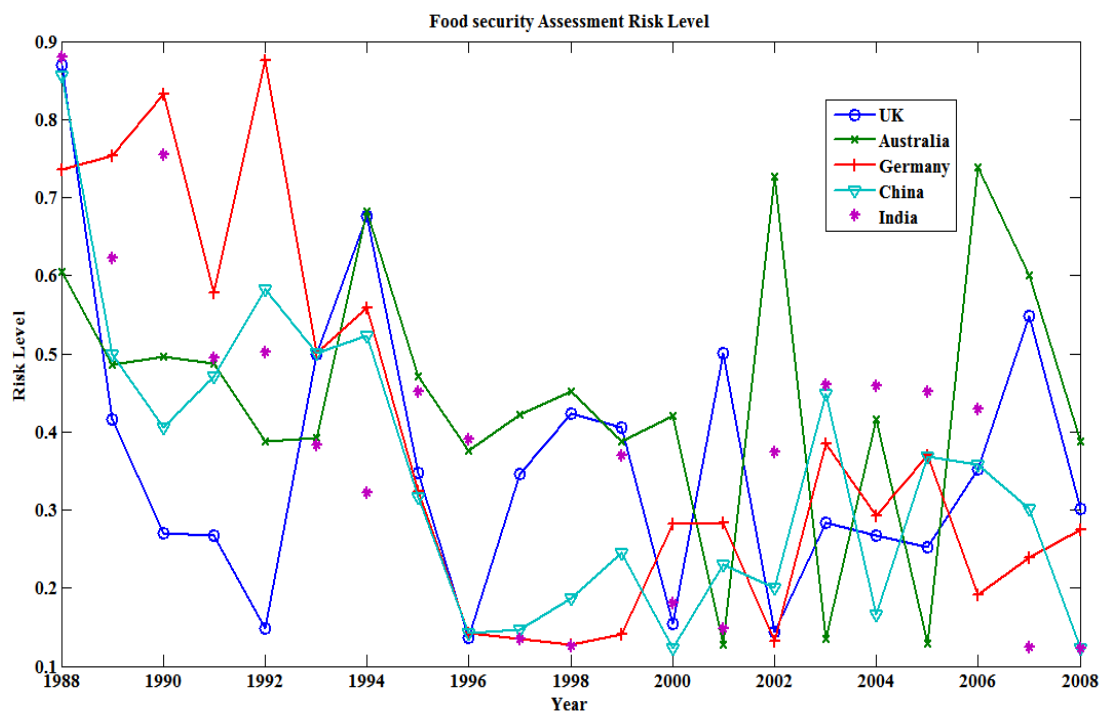


Figure 7.7: Food security risk level for year 1988 – 2008

## 7.7 Performance Index

In supporting the risk level assessment of FL, as shown in Figure 7.7, the performance index plots are made for each country, based from the producer prices for cereal for the years from 1988 to 2008. The comparisons for all five countries are shown in Figure 7.9 to Figure 7.13. Based on all of these figures, it can be seen that if the cereal price increases, the food security risk level will also increase, especially if the production and yield are also low.

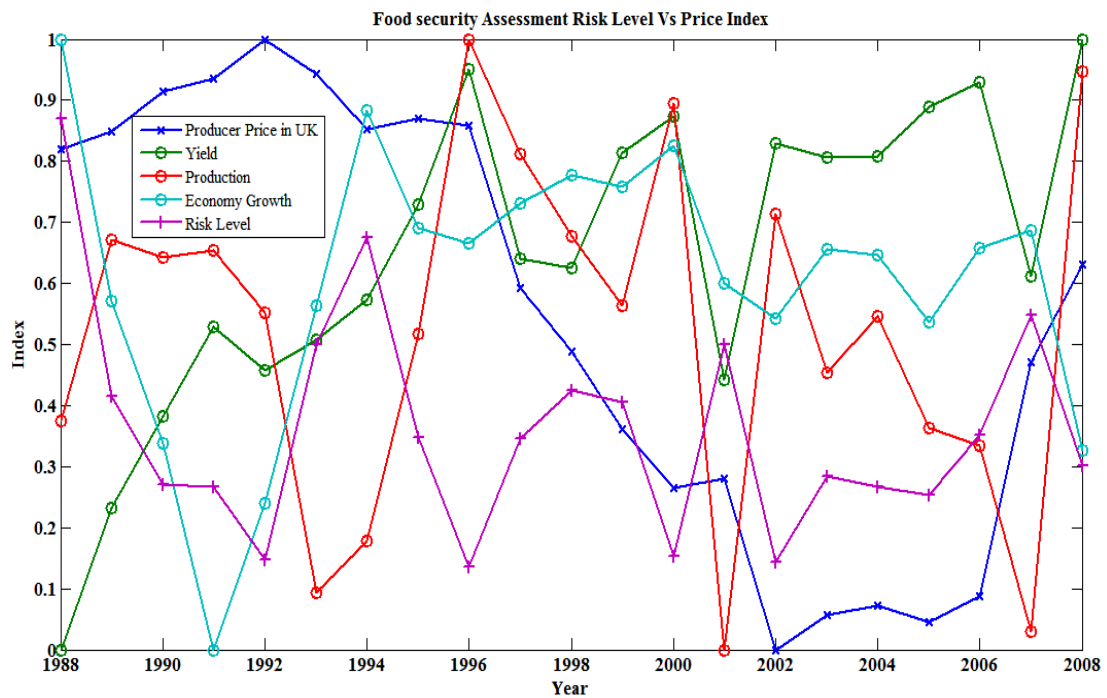


Figure 7.8: UK performance index for risk level via producer price

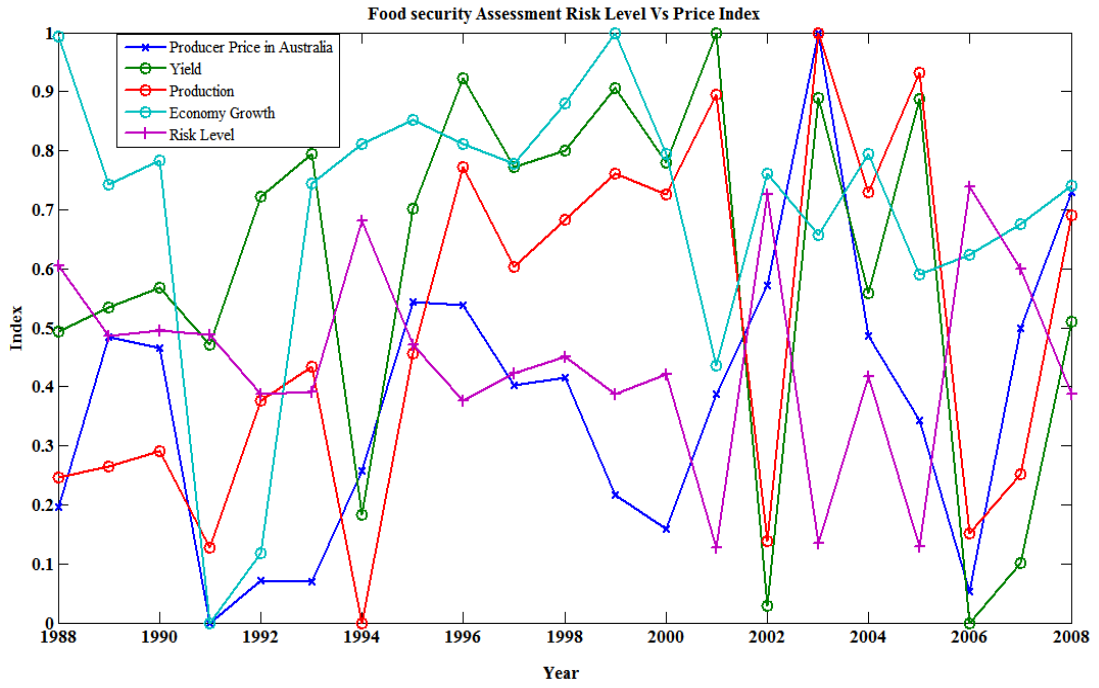


Figure 7.9: Australia performance index for risk level via producer price

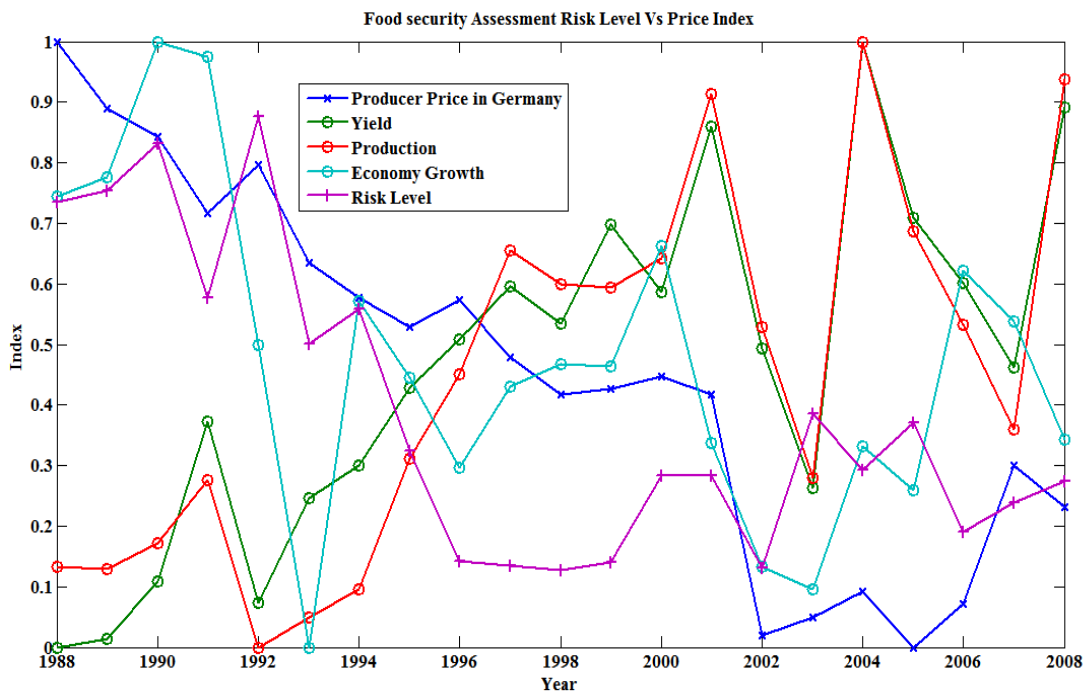


Figure 7.10: Germany performance index for risk level via producer price

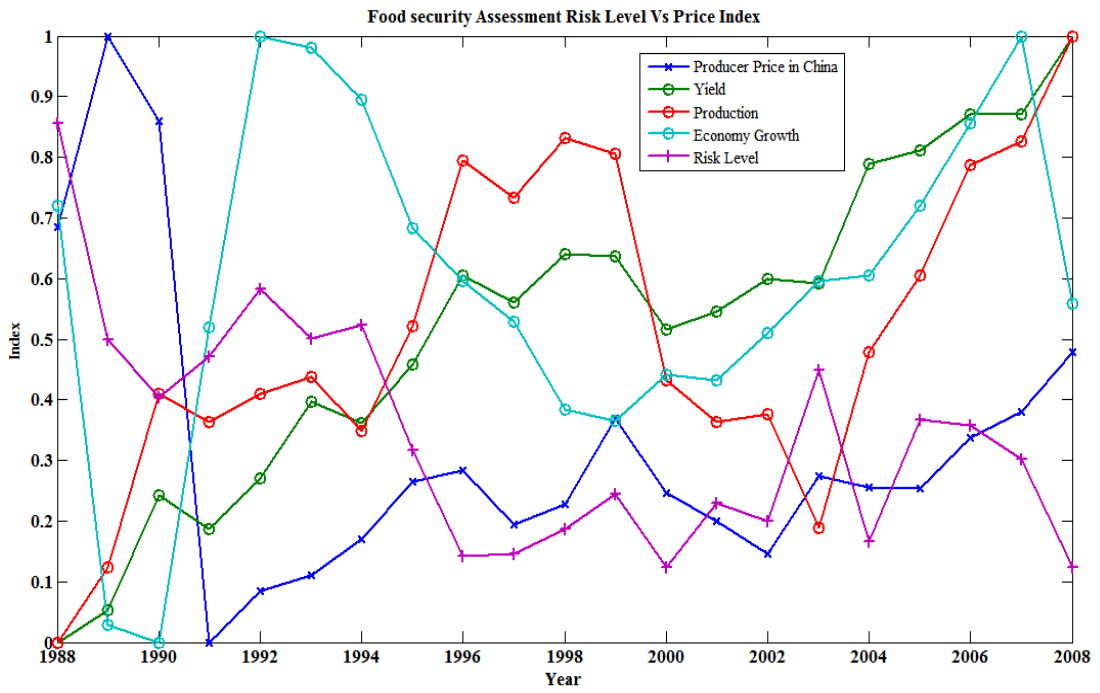


Figure 7.11: China performance index for risk level via producer price

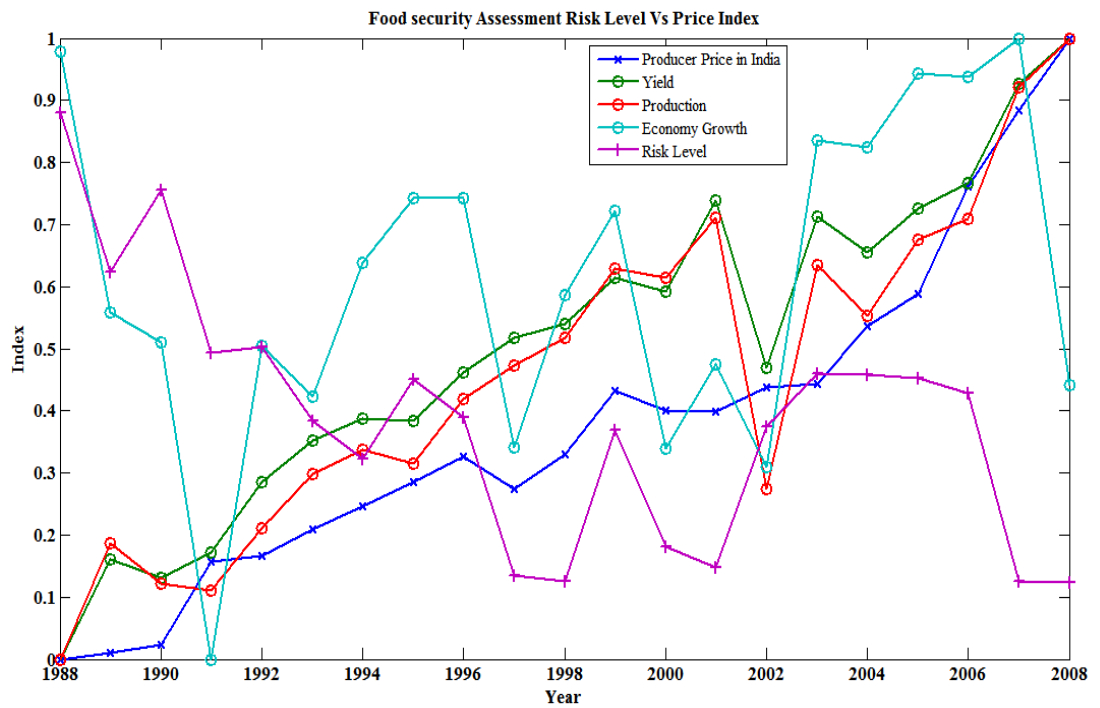


Figure 7.12: India performance index for risk level via producer price

In some of the performance plots in Figure 7.9 to Figure 7.13, the price plot is not the same as the risk level, especially for India and China in Figure 7.11 and Figure 7.12. For example, in India, for the year 1988 where the risk level is high, although the cereal yield and cereal production is low, at the same time the economy is at a medium level. In this case, the risk should theoretically be high, but, at that time the food price is low, which shows that most people should be able afford to buy any available food. It can be assumed that this pattern of behaviour occurs due to the effect of the low cereal yield and cereal production, and it also because of the effect of epidemics in that year.

It can be concluded that, in certain years, the yield, production and economic growth are not the only factors which can affect the food security risk level. The consideration of natural disasters and product prices need to be assessed in terms of their impact on food security. Table 7.5 shows all of the natural disasters and epidemics that occurred between 1988 and 2008 in the five countries. This data should be considered as another performance index comparison for this model. However, due to the data constraints, all of the natural disaster that happened at the time is assumed to impact on food security as a whole.

Table 7.5: List of natural disasters from year 1988 – 2008 (PreventionWeb, 1988 - 2008)

<b>Year</b>	<b>UK</b>	<b>Australia</b>	<b>Germany</b>	<b>China</b>	<b>India</b>
<b>1988</b>	-	-	-	-	Epidemic
<b>1989</b>	-	-	-	-	-
<b>1990</b>	Storm	-	Storm	-	-
<b>1991</b>	Storm	-	-	Flood	Storm
<b>1992</b>	-	Drought	-	-	-
<b>1993</b>	-	-	-	-	Storm
<b>1994</b>	-	-	-	Drought	-
<b>1995</b>	-	-	-	Flood	-
<b>1996</b>	-	-	-	Flood	Storm
<b>1997</b>	-	-	-	Flood	-
<b>1998</b>	Storm	-	-	Flood	-
<b>1999</b>	-	Storm	Storm	-	Storm
<b>2000</b>	Storm & flood	-	-	-	-
<b>2001</b>	-	-	-	-	Earthquake
<b>2002</b>	-	Drought	Flood & storm	-	-
<b>2003</b>	-	-	Heat wave	Flood	
<b>2004</b>	-	-	-	-	Flood
<b>2005</b>	-	-	-	-	Flood
<b>2006</b>	-	Storm	-	-	Flood

<b>2007</b>	Storm & flood	Flood	Storm	-	Flood
<b>2008</b>	-	Flood	Storm	Earthquake & heat wave	-

## 7.8 Conclusion

The purpose of this study is to demonstrate a model of a quantified framework for evaluating food security risk level. A fuzzy logic approach is explored as the modelling technique. The inference approach used is a Mamdani-type fuzzy inference method, used to interlink different aspects concerned with food security. These methods provide some indication of a means by which to elicit expert knowledge, and to enable a practical implementation of fuzzy evaluation of the food supply chain.

In summary, fuzzy logic control can be used to assess the risks to food security. The proposed fuzzy models provide useful tools, while avoiding the unreasonable and unverifiable assumptions embedded in conventional numerical scoring. The risk index can be aggregated over the various factors, which is then set as the input to the future study to estimate an overall level of food security. This can help to track the risks at each stage, and provides a quantitative evaluation of all parts of the supply chain.

---

**References**

- A.S. SODIYA, S. A. O., AND B. A. OLADUNJOYE 2007. Threat Modeling Using Fuzzy Logic Paradigm. *Informing Science and Information Technology*, 4.
- BANK, W. 2010. World Bank data.
- DEFRA 2009. UK Food Security Assessment: Our approach. Department for Environment Food and Rural Affairs.
- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In: DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). Department for Environment, Food and Rural Affairs.*
- DING, Z.-H., LI, J.-T. & FENG, B. Year. Radio Frequency Identification in Food Supervision. *In: Advanced Communication Technology, The 9th International Conference on, 12-14 Feb. 2007 2007. 542-545.*
- FAO 2006. Policy Brief : Food Security. *In: ECONOMICS, A. A. D. (ed.). FAO's Agriculture and Development Economics Division (ESA) with support from the FAO Netherlands Partnership Programme (FNPP) and the EC-FAO Food Security Programme.*
- FAO 2010. FAOStat. Food and Agriculture Organization United Nation website.
- HASLUM, K., ABRAHAM, A. & KNAPSKOG, S. Year. DIPS: A Framework for Distributed Intrusion Prediction and Prevention Using Hidden Markov Models and Online Fuzzy Risk Assessment. *In: Information Assurance and Security, 2007. IAS 2007. Third International Symposium 29-31 Aug. 2007 2007. 183-190.*
- HUEY-MING, L. 1996. Applying fuzzy set theory to evaluate the rate of aggregative risk in software development. *Fuzzy Sets and Systems*, 79, 323-336.



- J.-S.R. JANG, C.-T. S., E. MIZUTANI 1997. *Neuro-Fuzzy and Soft Computing*, Prentice Hall.
- JANG, J. S. R. 1993. ANFIS: adaptive-network-based fuzzy inference system. *Systems, Man and Cybernetics, IEEE Transactions on*, 23, 665-685.
- JIANLING, X. & YONG, D. Year. Food safety risk analysis based on generalized fuzzy numbers. *In: Advanced Management Science (ICAMS), 2010 IEEE International Conference on*, 9-11 July 2010 2010a. 699-702.
- JIANLING, X. & YONG, D. Year. Linguistic ranking model and its application in food management. *In: Computer Design and Applications (ICCD), 2010 International Conference on*, 25-27 June 2010 2010b. V5-208-V5-212.
- JONES, M. P. 2008. *Achieving Food Security And Economic Growth in Sub-Saharan Africa: Key Institutional Levers*. [Online]. Africa FARA. Available: [http://www.fara-africa.org/library/browse/ACHIEVING\\_FOOD\\_SECURITY\\_AND\\_ECONOMIC\\_GROWTH\\_IN\\_SUB\\_SAHARAN\\_AFRICA.pdf](http://www.fara-africa.org/library/browse/ACHIEVING_FOOD_SECURITY_AND_ECONOMIC_GROWTH_IN_SUB_SAHARAN_AFRICA.pdf) [Accessed August 2010].
- M. AHMEND, E. D., ET AL. 1999. A general purpose fuzzy engine for crop control. *Computational Intelligence*, 1625, 473-481.
- MELTZER, H. M., ARO, A., ANDERSEN, N. L., KOCH, B. & ALEXANDER, J. 2003. Risk analysis applied to food fortification. *Public Health Nutrition*, 6, 281-290.
- MUHD KHAIRULZAMAN ABDUL KADIR, E. H., KEFAYA QADDOUM, ROSEMARY COLLIER, ELIZABETH DOWLER, WYN GRANT, MARK LEESON, DACIANA ILIESCU, ARJUNAN SUBRAMANIAN, KEITH RICHARDS, YASMIN MERALI & RICHARD NAPIER 2013. Food

Security Risk Level Assessment: A Fuzzy Logic Based Approach *Applied Artificial Intelligence*.

- NEGNEVITSKY, M. 2005. *Artificial intelligence A guide to intelligent systems*, Addison-Wesly.
- ODETUNJI, O. A. & KEHINDE, O. O. 2005. Computer simulation of fuzzy control system for gari fermentation plant. *Journal of Food Engineering*, 68, 197-207.
- ORGANIZATION, F. A. A. 2006. Food Security : Policy brief *In: ORGANIZATION, F. A. A. (ed.) Issue 2 ed.: FAO*.
- PEIHONG, C. & JIAQIONG, W. Year. Application of a Fuzzy AHP Method to Risk Assessment of International Construction Projects. *In: JIAQIONG, W., ed., 2009. 459-462*.
- PERROT, N., IOANNOU, I., ALLAIS, I., CURT, C., HOSSENLOPP, J. & TRYSTRAM, G. 2006. Fuzzy concepts applied to food product quality control: A review. *Fuzzy Sets and Systems*, 157, 1145-1154.
- PREVENTIONWEB. 1988 - 2008. *Disaster statistics* [Online]. Available: [http://www.preventionweb.net/english/countries/statistics/index\\_region.php?rid=3](http://www.preventionweb.net/english/countries/statistics/index_region.php?rid=3) [Accessed 27/11/12].
- XIAOJUN, W., DONG, L. & XIANLIANG, S. Year. A fuzzy enabled model for aggregative food safety risk assessment in food supply chains. *In: Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on, 12-15 Oct. 2008 2008. 2898-2903*.
- XIE, G., XIONG, R. & CHURCH, I. 1998. Comparison of Kinetics, Neural Network and Fuzzy Logic in Modelling Texture Changes of Dry Peas in Long Time Cooking. *Lebensmittel-Wissenschaft und-Technologie*, 31, 639-647.

YONG, D. & JIANLING, X. Year. Fuzzy evidential warning of grain security. *In:* Advanced Management Science (ICAMS), 2010 IEEE International Conference on, 9-11 July 2010 2010. 703-706.

ZENG, J., AN, M. & SMITH, N. J. 2007. Application of a fuzzy based decision making methodology to construction project risk assessment. *International Journal of Project Management*, 25, 589-600.

## **Chapter 8: Conclusions and Future Work**

### **8.1 Overview**

The aim of the research is to investigate the relationship between the food security and the effect of food security using the IS technique have been successfully met. In addition, both of the TSH and the FL methods have successfully developed the new prediction model and the risk level assessment model that can be used for the general purpose modelling. Each of these models has also been used and tested in the food security research area such as prediction in farm household output, food growth per capita and food security risk level assessment. The other advantage of the models is the capability to measure and analyse its own food security risk level to determine the acceptance level of food insecurity such as good, acceptable or severe.

In term of the TSH model development, it has significantly shown a good prediction capability, where the number of the inputs can be reduced automatically. For example, from the best inputs that were selected by the TSH, it has optimised the threshold and the weight of ANN in order to prevent the generalisation issues and has subsequently produced a better prediction performance. In the food security application, the TSH model has considerably better than other approaches with more

reliable prediction results. For example, three prediction models have been performed using the TSH model that consisted of three different dataset. Each of the models has been tested and benchmarked against PCA, ANN, GA-ANN (FS), OWTNN and SGNO. Even though the different dimension sizes of the dataset were applied into the TSH model, the performance of each application has shown that the model were significantly better than other techniques. In summary, the TSH model can be used for different type of application to analyse variable datasets either with a small dimension size or a high dimension size.

With regard of the food security risk level assessment model, the FL method has been explored to demonstrate the quantified framework for evaluating the risk level. The FL method also performs really well in eliciting expert knowledge and enabled the practical implementation of fuzzy evaluation of the food supply chain or the food security itself. The proposed fuzzy model can be a starting point as the developing tools to assess the current food security risk level of the impending pressure on the food supply.

In conclusion, the TSH model and the food security risk level assessment model using the FL method, have shown a great potential to derive the relationship of the issues concerning the food security that has the impact on the overall food production, food quality and food access for all people. It can also be used as an evidence for a government to make a decision to change the food policy and the food standard in order to improve the safety of food security.

## **8.2 Research Summary**

In this thesis, a Two-stage Hybrid (TSH) model has been developed to provide accurate predictive performance in the area of food security. This includes the implementation of the prediction model for three different datasets, based on a food security study by DEFRA, where each dataset represents different sample dimension sizes. In addition, a food security risk assessment model was also developed using Fuzzy Logic (FL) to show the food security risk level for five different countries.

### **8.2.1 Two-Stage Hybrid (TSH) model**

The main components of this model are described in Chapter 3. It is based on two modules; a Genetic Algorithm (GA) module and an Artificial Neural Network (ANN) module. The main purpose of developing this model is to give a better prediction performance. The TSH model is evaluated using three practical implementations in three novel research areas in food security modelling, including; prediction of farm household crop output, prediction of one of the main components in food security - food growth per capita, and prediction of grain security warnings.

In evaluating the performance of the developed prediction model, five techniques are used to benchmark the model: PCA, MLP-ANN, feature selection (GA-ANN), OWTNN and a recently developed hybrid IS called SGNO.

As previously explained, in developing the TSH model, there are two main modules used, the GA module and the ANN module. All of the processes relating to each of these modules are as follows:-

- The GA module is used twice. The first use is in optimizing the input variables or reducing the number of feature variables. The development of this module is based on the standard GA, where the evolution of each chromosome is performed using standard GA operations including selection, crossover and mutation. The GA will measure the performance of fitness for each population, represented by binary chromosomes in the first stage process. The second GA module is used in the second stage process to find the optimum weights and optimum thresholds for the ANN. The second GA module is almost identical to that used in the first stage process. The only difference is in the chromosome declaration, which in the second module is based on the actual values of the chromosome populations.
- The ANN module is used three times in the proposed model, in the first stage process, second stage process and as the prediction model, where the ANN module is remodelled based on the optimum input variables selection from the first stage process and the optimum weights and thresholds from the second stage process. In the TSH model, the ANN is built mainly for determining the performance of the selection of the best inputs variables, the best weight values and the best threshold values for the ANN. Each of the performance evaluations is based on a regression model, which is trained using the available data, separated into three divisions; training, validation and testing. The fitness of the GA is determined from the difference in errors (MSE) between the ANN output and the actual output data.

- All of the performance metrics for the proposed model are based on the remodelled ANN, which is used as a prediction model. Two performance plots are produced - the regression plot and the error plot. The regression plot is used to show the prediction performance and the generalization of the ANN itself. The MSE plot shows the error curve for each epoch and the optimum speed achieved by the model in finding the final solution.

In developing the food security risk level assessment model, FL is used as the method for evaluating the risk level of five countries. The reason that FL was selected as the base technique is because of its ability to translate any imprecise inputs to quantitative conditions. It can also be easily tuned for any changes in the model itself.

### **8.2.2 Application chapter summary and results**

In this thesis, a total of four datasets are used, where each dataset is described in the chapter in which it is used, from Chapter 4 to Chapter 7. Chapters 4 to 6 demonstrate the implementation and the performance evaluation of the TSH model in three novel application scenarios. In Chapter 7, another novel application is described - a risk level assessment model of food security.

Chapter 4 describes the construction of prediction model for farm household output in India, where the data used was recorded on a daily-basis and then converted to an annual-basis. The dataset dimensions in this application are  $36 \times 292$  matrixes, where the number of features is 36, with 292 samples. The results show that the prediction of the farm household output using the proposed model outperforms other benchmark techniques in terms of its regression value, MSE value from the first



epoch to final epoch, and a very good ANN generalization for all data divisions; training, validation and testing.

Chapter 5 describes the construction of another prediction model, for predicting cereal growth per capita, using  $18 \times 507$  matrices that includes 18 features with 507 samples. The features were based on the sub-indicators of food growth per capita, which is one of the key components of the DEFRA food security assessment report. The overall prediction result for the proposed model also outperforms other benchmark techniques, but almost all of the techniques being compared achieve at least 90% regression. In terms of MSE, the developed model also shows a low error of prediction, and the number of required iterations is also lower for the proposed model compared to other techniques.

Chapter 6 describes another dataset of different dimension size, which consists of 11 features and 11 samples ( $11 \times 11$  matrices). This dataset is applied to the TSH model, in this case being used to predict the China grain security warning assessment. This dataset has the lowest dimension size of the datasets considered. Although, the size of the features is small, the developed model gives very good results. In comparison with other techniques, most of the benchmark techniques, including the proposed technique, show regression values of 0.95 and above. In this case study, the proposed model is outperformed by OWTNN, but with a regression difference of only 0.00374. The errors given by the developed model also have a very small difference compared with the OWTNN technique, being outperformed by it. The generalization of the ANN for the proposed model shows a very good generalization, almost the same as for the OWTNN model.

Chapter 7 demonstrates another novel application; the modelling of risk level assessment in food security. This is different from the applications described in the three earlier chapters. Here, an unsupervised learning method needed to be used to measure the risk level of food security, based on three factors; yield, production and economic growth. This model uses FL as its core to describe the qualitative and quantitative values involved. All of the reasoning included in the model is based on the expert judgment, and is converted to risk levels for five different countries, for data from a period of 20 years. In comparing the reasoning of the rules being used, a performance index of food prices is plotted. The final result is quite useful as a stepping-stone towards developing a model which can be used in the near future for monitoring the impact on food security of any changes made.

### **8.3 Advantages and disadvantages of the TSH model and food security risk level assessment**

From the above analysis and summary for each case study application, it can be concluded that the TSH model has advantages or disadvantages which depend on the problems being predicted or estimated, and for the food security risk level assessment it can also be seen that the developed model also has its own advantages and disadvantages which depend on the problems being assessed or monitored.

The TSH model is generally capable of giving very good prediction results in conjunction with the ANN, because of the selection of optimized input variables, and optimized weights and thresholds, which are selected based on the ANN itself. This is also due to the same parameters being tested by the earlier stage process for the remodelling of the ANN as the prediction model.

This combination model benefits from the global search capability of the GA and the local search capability of the ANN, in searching for the best variables for the ANN. At the same time the combination model reduces the computation time required by the ANN to generate the final prediction result. However, in getting the best revised model of the ANN, the GA optimization is based on population iterations with the ANN training for a number of epochs, which is very computational expensive.

In terms of the food security risk level assessment model using FL, its capability to convert any uncertainties in discovering the risk levels, by using the required linguistic information from experts is expected to offer advantages. However, the process of determining the rules for each of the inputs and the output highlights the disadvantages of the model, if the model has a higher number of input variables, which will increase the number of rules and the complexity of the model.

### **8.3 Future work**

This thesis concentrates on food security modelling either as a prediction model or risk assessment model. In terms of prediction modelling, a TSH model has been developed, giving better performance compared with other prediction models. The proposed model combines both a feature selection module and a module for the optimization of weight and threshold values, where the optimization concentrates on the ANN architecture parameters only. In the future, the combination of these optimizations with the optimization of the structure of the connectivity of the networks could be performed. This could create ANN topologies which are likely to have higher performance compared to random or fully connected ANN topologies (A. Fiszlelew, 2007).

Another possible improvement to the GA module itself may lie in concentrating the analysis of studies towards GA operations such as initialization, crossover and mutation, for faster computation and to increase the overall efficiency of the GA (Hermadi et al., 2010).

In terms of food security modelling, a larger dataset needs to be acquired to permit the study of the overall food security themes defined in the DEFRA report (DEFRA, 2009, DEFRA, 2010). As discussed in Chapter 7, a detailed dataset, which includes other features which can impact on the risk level of food security, needs to be included in order to achieve the best precision in decision making based on the food security risk level values. This will also produce a better analysis, contributing to achieving the best performance result, which can then be used to monitor the risk level of food security, or to predict the impact of other environmental factors in maintaining food security from the farm to the end product.

## References

- A. FISZELEW, P. B., A. OCHOA, H. MERLINO, E. FERNÁNDEZ, R. GARCÍA-MARTÍNEZ 2007. Finding optimal neural network architecture using genetic algorithms. *Advances in Computer Science and Engineering Research in Computer Science*, 15-24.
- DEFRA 2009. UK Food Security Assessment: Our approach. Department for Environment Food and Rural Affairs.
- DEFRA 2010. UK Food Security Assessment: Detailed Analysis. *In*: DEPARTMENT FOR ENVIRONMENT, F. A. R. A. (ed.). Department for Environment, Food and Rural Affairs.

HERMADI, I., LOKAN, C. & SARKER, R. Year. Genetic Algorithm Based Path Testing: Challenges and Key Parameters. *In: Software Engineering (WCSE), 2010 Second World Congress on, 19-20 Dec. 2010 2010. 241-244.*