

MULTIPHASE ANALYSIS OF COCOA PRODUCTION IN KERALA

By

SHIVAKUMAR M

(2018 19 005)



DEPARTMENT OF AGRICULTURAL STATISTICS

COLLEGE OF HORTICULTURE

VELLANIKKARA, THRISSUR- 680656

KERALA, INDIA

2020

MULTIPHASE ANALYSIS OF COCOA PRODUCTION IN KERALA

By

SHIVAKUMAR M

(2018-19-005)

THESIS

Submitted in partial fulfillment of the requirement for the degree of

Master of Science in Agricultural Statistics

Faculty of Agriculture

Kerala Agricultural University, Thrissur



DEPARTMENT OF AGRICULTURAL STATISTICS

COLLEGE OF HORTICULTURE

VELLANIKKARA, THRISSUR- 680656

KERALA, INDIA

2020

DECLARATION

I hereby declare that the thesis entitled “**Multiphase analysis of cocoa production in Kerala**” is a bonafide record of research work done by me during the course of research and the thesis has not previously formed the basis for the award to me of any degree, diploma, associateship, fellowship or other similar title, of any other university or society.

Place: Vellanikkara

Shivakumar M

Date: 30 10 2020

(2018 19 005)

CERTIFICATE

Certified that the thesis entitled “**Multiphase analysis of cocoa production in Kerala**” is a record of work done independently by **Mr. Shivakumar M** under my guidance and supervision and that it has not previously formed the basis for the award of any degree, diploma, fellowship or associateship to him.

Place: Vellanikkara

Date: 30-10-2020

Dr. Ajitha T K

(Chairman, Advisory Committee)

Associate Professor

Department of Agricultural Statistics

College of Horticulture, Vellanikkara

CERTIFICATE

We, the undersigned members of the advisory committee of **Mr. Shivakumar M (2018-19-005)**, a candidate for the degree of **Master of Science in Agricultural Statistics**, with major field in **Agricultural Statistics**, agree that the thesis entitled **“Multiphase analysis of cocoa production in Kerala”** may be submitted by **Mr. Shivakumar M**, in partial fulfillment of the requirement for the degree.

Dr. Ajitha T K

(Chairman, Advisory Committee)

Associate Professor

Dept. of Agricultural Statistics

College of Horticulture, Vellanikkara

Dr. Laly John C

(Member, Advisory Committee)

Professor and Head

Dept. of Agricultural Statistics

College of Horticulture, Vellanikkara

Dr. N. Mini Raj

(Member, Advisory Committee)

Professor

Dept. of Plantation crops and spices

College of Horticulture, Vellanikkara

Dr. J. S. Minimol

(Member, Advisory Committee)

Associate Professor

Dept. of Plant Breeding and Genetics

Cocoa Research Centre, Vellanikkara

ACKNOWLEDGEMENT

*First and foremost, I would like to thank **God** for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily. Without his blessings, this achievement would not have been possible.*

*With immense pleasure, I wish to express and place on record my sincere and deep sense of gratitude to **Dr. Ajitha T.K.**, Associate Professor, Department of Agricultural Statistics, College of Horticulture, Vellanikkara and Chairman of the advisory committee for the valuable guidance, critical suggestions throughout the investigation and preparation of the thesis. I strongly believe that her guidance will be a light for my future paths.*

*I sincerely thank **Dr. Laly John C.**, Professor and Head, Department of Agricultural Statistics and member of my advisory committee for her unwavering encouragement, timely support and critical examination of the manuscript that has helped me a lot for the improvement and preparation of the thesis.*

*I express my heartiest gratitude to **Dr. S. Krishnan**, Professor (Retired) for his ever-willing help, technical advice, valuable guidance and creative suggestions throughout the period of my study.*

*Words are inadequate to express my sincere gratitude to **Dr. J. S. Minimol**, Assistant Professor, Department of Plant Breeding and Genetics, Cocoa Research Centre, Vellanikara and member of the advisory committee for providing the data, and for the constant inspiration with utmost sense of patience and ever willing help bestowed upon me.*

*I would like to express my extreme indebtedness and obligation to **Dr. N. Mini Raj**, Professor, Dept. of Plantation crops and spices and member of my advisory committee for her meticulous help, expert advice and support throughout my course of study.*

*I wish to express my sincere gratitude to **Mr. Ayoob K.C**, Assistant Professor, Department of Agricultural Statistics, for the valuable suggestions, critical evaluation and fruitful advice, for which I am greatly indebted.*

*With immense pleasure and deep respect, I express heartfelt gratitude to **Mr. Vishnu B.R**, for his judicious and timely suggestions.*

*I wish to express my sincere thanks to my seniors **Chittira, Shilpa** and **Athira** for their support and help in providing the necessary data during my course of research work.*

*I am thankful to **ICAR** for providing me Non JRF fellowship to complete my master degree. I am also grateful to **ICAR** which extended financial support by recognizing my potential and zeal towards the research, without their assistance it would have been very difficult to complete my research.*

*I intend to place on record my sincere heartfelt thanks to **College of Horticulture, Kerala Agricultural University, Thrissur** for the generous assistance, help, support and peace for the completion of my study. I thankfully remember the services rendered by all the **staff members of Student's computer club, College Library, Office of COH and Central library, KAU**.*

*I am very thankful to my senior **Jasma V.A, Sharanabasappa M.G, Lokesh S** and lovely classmates **Pooja A, Yogeesh G, Apeksha, Pooja B.N, Anjana, Sisira, and Deenamol** for making two years an unforgettable journey.*

*I express my gratitude for all the wisdom, love and support given to me by my parents **Shri. Markandesh marali** and **Smt. Hampamma marali** who have been the backbone of my success and the sole inspiration of my life. I deep heartedly acknowledge their love and affection and I owe it for my lifetime. I am very much thankful to God for blessing me with such a wonderful brother **Mr. Vinaykumar**, sister **Ms. Ashwini** who are there always for me in all situations with loads of love and affection.*

SHIVAKUMAR M.

CONTENTS

Chapter	Title	Page No.
1	INTRODUCTION	1-5
2	REVIEW OF LITERATURE	7-26
3	MATERIALS AND METHODS	28-69
4	RESULTS AND DISCUSSION	71-147
5	SUMMARY AND CONCLUSION	149-155
6	REFERENCES	157-164
7	ABSTRACT	

LIST OF TABLES

Table No.	Title	Page No.
3.1	Description of variables in the SEM model of cocoa production	64
4.1	Descriptive statistics for the time series data of cocoa cultivation in Kerala from 1980 to 2017	71
4.2	Quadratic regression model to test the presence of trend in area under Cocoa of Kerala from 1980-2017	72
4.3	Regression model to test the presence of linear trend in cocoa production of Kerala from 1980-2017	73
4.4	Linear regression model to test the presence of trend in cocoa productivity of Kerala from 1980-2017	74
4.5	Unit root test for area, production and productivity of cocoa in Kerala from 1980 to 2017	75
4.6	Parameters of the exponential smoothing coefficients of the Holt's model for area under Cocoa in Kerala for the period from 1980 to 2011	77
4.7	Accuracy measures of Holt's exponential smoothing model	77
4.8	Validation of predicted area under cocoa in Kerala using Holt's model for 2012-2017	77
4.9	Comparison of accuracy measures of ARIMA (0,2,2) and Holt's exponential model	78
4.10	Parameters of the Holt's exponential smoothing model to forecast area under Cocoa in Kerala for the period from 1980 to 2017	79
4.11	Forecasted values of area under Cocoa (ha) in Kerala for the period 2018-2022	81
4.12	Parameters of Simple exponential model of cocoa production for the years 1980 to 2011	81
4.13	Accuracy measures of Simple exponential smoothing model	82
4.14	Validation of cocoa production for the period 2012-2017 using Simple exponential model	82
4.15	Comparison of accuracy measures of ARIMA (0,1,1) and Simple exponential model	83
4.16	Parameters of the model ARIMA (0,1,1)	83
4.17	Forecasted values of Cocoa production (tonnes) in Kerala for the period 2018-2022	85
4.18	Parameters of the Simple exponential smoothing model for Cocoa productivity in Kerala	85
4.19	Statistical measures of Simple exponential smoothing model for prediction of Cocoa productivity	86

4.20	Validation of cocoa productivity in Kerala for the period 2012-2017 using Simple exponential smoothing model	86
4.21	Comparison of accuracy measures of ARIMA (0,1,1) and Simple exponential model for Cocoa productivity in Kerala	87
4.22	Parameters of Simple exponential smoothing for Cocoa productivity in Kerala	87
4.23	Forecasted values of Cocoa productivity (tonnes/ha) in Kerala using simple exponential smoothing model	89
4.24	Parameters of ARIMAX (0,1,0) model	90
4.25	Accuracy measures of ARIMAX model	91
4.26	Comparison of actual and predicted values of cocoa production for the years 2012 to 2017 in Kerala	91
4.27	The parameter estimates of ARIMAX (0,1,0) model for Cocoa production (tonnes) in Kerala	92
4.28	Statistical measures of ARIMAX model for Cocoa production (tonnes) in Kerala	93
4.29	Forecasted values of Cocoa production (tonnes) in Kerala for the period 2018-2022	94
4.30	Descriptive statistics for the average monthly cocoa yield from 100 plants for the period from 2003 to 2017	95
4.31	Parameter estimates of SARIMA (1,0,0)(1,1,0) ₁₂ model	97
4.32	Statistical measures for SARIMA (1,0,0)(1,1,0) ₁₂ model	98
4.33	Validation of the SARIMA model for average monthly cocoa yield for the next 12 months for the year 2017	99
4.34	Parameter estimates of SARIMA (1,0,0)(1,1,0) ₁₂ model	99
4.35	Statistical measures for SARIMA (1,0,0)(1,1,0) ₁₂ model	100
4.36	Forecasted values of average monthly cocoa yield for the next 12 months for the year 2018	101
4.37	The dependent variables in GLM	105
4.38	Between subject factors	105
4.39	The accession ids of cocoa hybrids included in the GLM repeated measures analysis	106
4.40	Mean and S.D of yearly production of 5 groups (Factors) of Cocoa hybrids used for GLM	107
4.41	Multivariate Tests to compute F values in GLM	109
4.42	Mauchly's Test of Sphericity in GLM	110
4.43	Tests of Within-Subject Effects	111
4.44	Tests of Within-Subjects Contrasts	112
4.45	Tests of Between-Subjects Effects	112
4.46	Multiple comparison of means of different factors	113

4.47	Goodness of fit test	117
4.48	Summary of Kolmogorov-Smirnov test	117
4.49	Summary of Anderson Darling test	118
4.50	Parameters of Geometric distribution	119
4.51	Correlation of previous five months cumulative weather variables with average monthly Cocoa yield of 100 plants	123
4.52	The variables entered or removed in doing the Step wise regression	125
4.53	Coefficients of the Step wise regression for Model1 and Model 2	125
4.54	Model summary of the Step wise regression	126
4.55	Correlation of current month's weather variables with average monthly Cocoa yield of 100 plants	127
4.56	Demographic characteristics of the Respondents	130
4.57	Model fit summary of base model of SEM on cocoa production	131
4.58	Coefficients of variables in Structural Equation Model (SEM)	133
4.59	Covariances between the independent variables and between the error terms in Structural Equation Model (SEM)	136
4.60	Model fit summary of SEM	137
4.61	Estimation of Yield gap of cocoa in Kerala	141
4.62	Probit model on decision making to make use of Plant protection measures in cocoa cultivation	141
4.63	Diagnostic measures of the Probit regression model	142
4.64	Ranks of different factors perceived by farmers	145
4.65	Test statistic of Kendall's Coefficient of concordance	146

LIST OF FIGURES

Figure No.	Title	Page No.
3.1	Box plot	29
3.2	Base model of cocoa production	61
4.1	Area under cocoa in Kerala during 1980-2017	72
4.2	Annual cocoa production of Kerala from 1980-2017	73
4.3	Cocoa productivity of Kerala during 1980-2017	74
4.4	Validation of forecasted values of area under cocoa in Kerala for the years 2012 to 2017	78
4.5	Residuals of ACF and PACF plots of Holt's exponential smoothing model	80
4.6	Actual and forecasted area under Cocoa in Kerala	80
4.7	Validation of cocoa production for the period 2012 to 2017 using Simple exponential model	82
4.8	Residual plots of ACF and PACF of ARIMA (0,1,1) model	84
4.9	Actual and forecasted production of Cocoa in Kerala	84
4.10	Validation of cocoa productivity in Kerala for 2012 to 2017 using simple exponential smoothing model	86
4.11	Residual plots of ACF and PACF of Simple exponential smoothing model for cocoa productivity in Kerala	88
4.12	Forecasting of Cocoa productivity (tonnes/ha) in Kerala using simple exponential model	88
4.13	Validation of cocoa production in Kerala for 2012 to 2017 using ARIMAX (0,1,0) model	92
4.14	Residual plots ACF and PACF in the ARIMAX model	93
4.15	Forecasting of cocoa production of Kerala through ARIMAX (0,1,0) model	94
4.16	Monthly average no. of cocoa pods for 100 trees from 2003 to 2017	96
4.17	Residual plots of ACF and PACF of SARIMA (1,0,0)(1,1,0) ₁₂ model	98
4.18	Residual plots of ACF and PACF of SARIMA (1,0,0)(1,1,0) ₁₂ model	100
4.19	A twelve-month forecast of average Cocoa yield using SARIMA (1,0,0)(1,1,0) ₁₂ for the year 2018	101
4.20	Box plot for total number of cocoa pods of 100 cocoa hybrids for 2003 - 2017	103
4.21	The average yearly pod yield of 5 groups of Cocoa hybrids for 2003 - 2015	108
4.22	Box plot for percentage of infected Cocoa pods	115

4.23	Geometric probability distribution function of monthly infected cocoa pods	120
4.24	Cumulative distribution function of Geometric distribution	121
4.25	Final SEM model based on standardized coefficients for Cocoa production	132

LIST OF ABBREVIATIONS

ACF	:	Autocorrelation Function
ADF	:	Augmented Dickey-Fuller
AGFI	:	Adjusted Goodness of Fit Index
AIC	:	Akaike Information Criterion
AKIS	:	Agriculture Knowledge Information System
AMOS	:	Analysis of Moment Structures
ANN	:	Artificial Neural Network
ANOVA	:	Analysis of Variance
AR	:	Auto Regressive
ARDL	:	Auto Regressive Distributed Lag
ARMA	:	Auto Regressive Moving Average
ARIMA	:	Auto Regressive Integrated Moving Average
BIC	:	Bayesian Information Criterion
CFA	:	Confirmatory Factor Analysis
CFI	:	Comparative Fit Index
CV	:	Coefficient of Variation
DCCD	:	Directorate of Cashewnut and Cocoa Development
DW	:	Durbin Watson
GDM	:	Gamma Distribution Model
GLM	:	General Linear Model
GOF	:	Goodness of Fit

ICCO	:	International Cocoa Organisation
IQR	:	Inter Quartile Range
MA	:	Moving Average
MAD	:	Mean Absolute Deviation
MAPE	:	Mean Absolute Percentage Error
MLR	:	Multiple Linear Regression
MM	:	Method of Moments
MPZC	:	Method of Proportion of Zeroth Cell
MSD	:	Mean Squared Deviation
NFI	:	Normed Fit Index
NLSVR	:	Non-Linear Support Vector Regression
OLS	:	Ordinary Least Square
PACF	:	Partial Autocorrelation Function
R^2	:	R square
RMSE	:	Root Mean Squared Error
RMSEA	:	Root Mean Squared Error of Approximation
S.D	:	Standard Deviation
SEM	:	Structural Equation Model
SLR	:	Simple Linear Regression
SPSS	:	Statistical Package for the Social Sciences
TLI	:	Tucker Lewis Index
VAR	:	Vector Auto Regressive

A graphic of a scroll with a black outline and a light gray shadow. The scroll is unrolled, showing the word "INTRODUCTION" in a bold, black, serif font. The scroll has a vertical strip on the left side and a small circular detail at the top right corner.

INTRODUCTION

Chapter 1

Introduction

Chocolate knows no boundaries; speaks all languages; comes in all sizes; is woven through many cultures and disciplines ... it impacts mood, health, and economics, and it is a part of our lives from early childhood through the elderly years.

Herman A. Berliner

It is believed that over 3 million tons of cocoa beans are consumed in each year as per the World Cocoa Foundation. Chocolate has become one of the most widely used popular sweets in the world which is used as an ingredient in bakery, beverage and confectionary item. The primary source of this chocolate is the fruit of cocoa trees. Cocoa is botanically called, *Theobroma cacao L.* It is considered as a food-industrial crop which is gaining importance in the regions of humid tropics. Cocoa cultivation provides income to millions of small and marginal farmers across the world. It is believed that the origin of cocoa was in the Amazon wild forests of South America. The leading cocoa producers of the world are Ghana, Nigeria, Ivory Coast, Brazil and Cameroon.

As per the report of International Cocoa Organization (ICCO) in the year 2018-'19, the world cocoa bean production has reached to 4.8 Million Tonnes. Out of this, nearly 76% of cocoa beans was produced from West Africa, Cote d'Ivoire, Ghana, Nigeria and Cameroon. Cote d'Ivoire is the largest producer in the world followed by Ghana and Indonesia. The contribution of India to the global production is about 0.50% compared to other major producing countries, still India has much potential to increase the area and production. Cocoa cultivation is gaining more importance in the southern parts of India. In India, it is mainly cultivated in Karnataka, Kerala, Andhra Pradesh and Tamil Nadu mainly as an intercrop with arecanut and coconut. Cocoa is cultivated in India in an area of 94008 hectares with total production of 23981 tonnes (DCCD statistics 2018-'19). Andhra Pradesh stands first with 35% area (32,949 ha), 40% production (9615 tonnes) and it stands first in the case of productivity (750 Kg/Ha) also. Kerala comes

second with 35% of production (8507 tonnes) followed by Karnataka which shares 15% and Tamil Nadu which shares 10% out of total production. Today India earns more foreign exchange through the export of cocoa products in the global cocoa market. For the last ten years it was observed that there was a steady increase in demand of cocoa and the rate of increase in demand is higher than the increase in cocoa production (DCCD Cochin).

Cocoa is one of the most emerging plantation crops in the state of Kerala. The cocoa cultivation in Kerala was started in the year 1980. The area under cocoa in Kerala was about 10,700 hectares in 2010-'11 and it has spread out to 16,590 hectares in 2017-'18 (DCCD statistics 2017-'18). In the year 2018-'19, Kerala ranked second in area of cultivation of cocoa and Andhra Pradesh ranked first. With respect to the production, Kerala shared 35% of the total production in the year 2018-'19 followed by Karnataka and Tamil Nadu (DCCD statistics 2018-'19). The highest productivity was witnessed in Andhra Pradesh followed by Kerala with 850 kg/ha. Research activities are to be made to increase the area, production and productivity to meet the growing demand from foreign countries.

Hence a study was designed with an overall objective of exploiting different statistical tools on cocoa research activities so that an efficient production model can be generated which would provide an insight to the cocoa farmer's income in Kerala. This study would help in future in bringing proper policy and decision making to achieve self- sufficiency in this sector.

In planning and decision-making processes, prediction of future events is very critical and forecasting can help in making rational decisions (Armstrong,2001). Quantitative forecasting methods make use of historical data and a forecasting model to extrapolate past and current behaviour into the future. Time series forecasting under quantitative forecasting methods consists of time domain approach and frequency domain approach. The dependence of adjacent observations is an inherent feature of time series and in time series forecasting we try to quantify this dependence. This

necessitates the development of stochastic and dynamic models for time series data for forecasting purposes.

The most frequently used important time series models are ARMA, ARIMA, ARIMAX, SARIMA etc. These models have been proved very useful in forecasting yearly and monthly changes in area, production and productivity of crops. The ARIMAX model is a multiple regression model with one or more autoregressive (AR) terms and one or more moving average (MA) terms together with other explanatory variables. It can be used for data which is stationary or non-stationary. Seasonal Autoregressive Integrated Moving Average (SARIMA) model is an extension of ARIMA model that specially supports univariate time series data with a seasonal component. The SARIMA model is proved to be the best model with replacement of ARIMA models when the time series data contain seasonality and can be used for prediction with good accuracy. Exponential smoothing method is used for smoothing discrete time series data to forecast the future values. Simple exponential smoothing model is the most widely used method which is used to forecast the data that has no clear trend or seasonal pattern. It involves a single parameter called alpha (α) which is a smoothing factor or smoothing coefficient. Holt's exponential smoothing model is used for forecasting the time series data that has both trend and seasonality. It involves two parameters, alpha (α) to control the smoothing factor for the level and beta (β) to control the decay of the influence of the change in trend.

When we deal with perennial crops, the repeated measurements on yield is to be taken care of while analysing the results. The General Linear Model (GLM) Repeated Measures procedure can be employed to model the values of multiple dependent scale variables measured at multiple time periods. The significant difference between factors eliminating the effect of time can be studied using this procedure.

Crop production is affected biophysically by meteorological variables including rising temperatures, changing rainfall etc. An attempt to study the correlation of crop yield and climatic variables and their cumulative effect and regression of crop yield on

climatic variables may be helpful for identifying the most important variables that affect the crop production and also the incidence of several diseases of crop due to the vagaries of climate.

Farmers are really facing great difficulty to make decisions which require them to come up with realistic estimates of future yields of their crops. It is highly essential according to their point of view to decide which varieties of the crop they have to select or how much fertiliser they have to purchase or what would be the total expenditure etc. To give an idea about the yield and the percentage of infected pods out of the total we can make use of probability distributions which will guide us to acquire a robust knowledge about the internal structure of the data and the corresponding probabilities with respect to each class interval of the realised observations. This information can be productively made use of for future planning.

An empirical analysis to identify the factors perceived by farmers which influenced production of cocoa in actual field conditions can be validated by conducting a survey among cocoa farmers and collecting information on their demographic details, cultivation and management practices, expenditure incurred, production details, constraints they have faced etc. through pre-tested structured questionnaire. Since we have to deal with a number of independent and dependent variables in the yield production model, a path analysis which is a special case of structural equation modelling can be effectively made use of to address such complex system of pathways leading to the ultimate income of cocoa farmers. Probit analysis can be effectively made use of to identify the factors which lead to some important decision-making process in cultivation practices like usage of plant protection measures, pesticide use etc. Coefficient of concordance is a simple but an efficient tool to examine the overall agreement among respondents to rank a list of statements according to their preference and the same can be utilised to list out the constraints faced by the farmers in cocoa production.

In the present study, time series data on area, production and productivity of cocoa for the period from 1980 – 2017 with respect to Kerala state, the cocoa production details of 100 selected hybrids of cocoa trees having same age collected from Cocoa research Centre, College of Horticulture, Vellanikkara, KAU and primary data pertaining to cultivation and management practices together with demographic details from 100 farmers engaged in cocoa cultivation in Iritty Panchayat of Kannur district and Veliyamattom panchayat of Idukky district who have interactions with the Cocoa project “Mondelez International Ltd., Ernakulam” have been made use of to make a multiphase analysis of cocoa production in Kerala with the following objectives.

- To predict the area, production and productivity of cocoa in Kerala
- To study the impact of weather factors on yield
- To assess the yield gap
- To delineate the factors influencing farmer’s decisions on cultivation practices and to develop yield prediction models through structural equation modelling



REVIEW OF LITERATURE

Chapter 2

Review of Literature

A critical review of literature is required to provide evidence to support our research findings. It helps the researcher to identify the methodologies used in past studies on the same or similar topics. In line with the objectives, the review of literature is presented below as following sections:

2.1 Time series forecasting

2.1.1 ARIMA and Exponential Smoothing model

2.1.2 ARIMAX model

2.1.3 SARIMA model

2.2 General linear model

2.3 Probability distribution

2.4 Impact of climatic variables on crop yield

2.5 Assessing yield gap

2.6 Structural equation modelling

2.7 Probit regression model

2.8 Coefficient of concordance

2.1 Time series forecasting

2.1.1 ARIMA and Exponential smoothing model

Tahir and Habib (2013) used four models for trend analysis which were linear, quadratic, exponential and S-curve trend models to estimate the trend in area and production for maize in Pakistan. The accuracy measures such as Mean absolute percentage error (MAPE), Mean absolute deviation (MAD) and Mean squared deviation (MSD) were used to find the best fitted model among the four models. Since quadratic trend model had smaller values of all these measures it was considered as good fitted model with minimum forecasting errors. This model was chosen as the best model for forecasting.

Amin *et al.* (2014) selected time series models in order to forecast the wheat production in Pakistan. On the basis of model selection criteria ARIMA (1,2,2) had lowest AIC when compared to different time series models. Through this model it was predicted that wheat production of Pakistan would become 26623.5 thousand tons in 2020 and would become double in 2060.

Tripathi *et al.* (2014) forecasted the area, production and productivity of rice in Odisha by using the ARIMA model. ARIMA model was considered as the best model than any other model when it was necessary to forecast the yield and acreage before the crop harvest. As it was confirmed from the skewness and kurtosis that the time series data was non-normally distributed, so the non-parametric test - Mann-Kendall test was chosen as more suitable to detect trend. Based on the forecasting and validation results, it could be concluded that ARIMA model could be successfully used for forecast studies of rice.

Ankrah *et al.* (2015) projected weighted vector error correction model as a best statistical technique in forecasting cocoa production among many candidate models. They found that weighted ranking procedure was suitable for accurate forecasting of cocoa production in Ghana and even explained that the annual production

variability could be estimated by the use of predicted value of weighted vector error correction model.

Rajan *et al.* (2015) demonstrated trend analysis models namely linear, quadratic and cubic to study the trend of area, production and productivity of cotton in Tamil Nadu. Among three models, the cubic regression model was declared to be the best fitted model since it was having highest R^2 with respect to area, production and productivity. The future estimation for cotton was done using this selected model.

Karadas *et al.* (2017) determined three exponential smoothing time series methods such as Holt, Brown and Damped trend for forecasting the production of oil seed crops viz; sesamum, sunflower and soybean in Turkey. Holt method was used to study the trend in time series with two parameters alpha and beta as smoothing coefficients. Brown's linear exponential smoothing was another method used to evaluate the increase or decrease of trends in time series. The damped trend method was used for better forecasting. Holt exponential smoothing model was selected as a better fit method since it showed the lowest normalized BIC and greatest value for stationary R^2 when compared to other methods.

Saranyadevi and Mohideen (2017) identified presence of trend in the paddy production data. For smoothing of data, the common techniques like simple exponential, Brown exponential and Damped exponential smoothing models were used and forecasting of paddy production in Tamil Nadu was made. Among the various exponential smoothing methods, Holt's winter smoothing was promoted to be a better model based on model selection criteria and ARIMA (0,1,1) was found to be a better model to forecast the paddy production.

Hemavathi and Prabakaran (2018) applied time series methods to forecast area, production and productivity of rice in Thanjavur district of Tamil Nadu. Box Jenkin ARIMA model was fitted to the data for the period from 1990-'91 to 2014-'15. AIC and SBC were the criterion for the model selection. ARIMA (0,1,2) for area of rice, ARIMA (0,1,1) for production and ARIMA (0,1,1) for productivity had lowest

AIC and SBC values. The results from the forecast showed that by the year 2020 the area, production and productivity would be about 158.15 hectare, 637.05 thousand tons and 3.79 thousand kg per ha.

Rathod and Mishra (2018) developed a hybrid model to forecast the yield of mango and banana in Karnataka. Hybrid forecasting models as made by combining ARIMA with ANN and ARIMA with Nonlinear Support Vector Regression (NLSVR) models were found to overcome the problem of linear and non-linear components contained in the time series. The hybrid models ARIMA-TDNN (Timely delay neural network) and ARIMA-NLSVR showed better performance when compared to single models viz, ARIMA, TDNN and NLSVR.

2.1.2 ARIMAX model

Paul *et al.* (2013) identified five models at five different growth stages of wheat crop to forecast the wheat yield in Kanpur district of Uttar Pradesh. To develop ARIMAX model, the annual wheat yield data of Kanpur district from 1972 to 2011 was used. The data from 1972 to 2007 was used for model building and the remaining data from 2008 to 2011 was used for validation of the model. Similarly, the weather data at various stages of wheat crop, since from CRI stage to dough stage during the same period was obtained. The weather variables were taken as exogenous variables to build the ARIMAX model. On the basis of minimum Akaike information criterion (AIC) and Schwartz-Bayesian criterion (BIC) values, best ARIMAX model was determined. To forecast the wheat yield, the five models at five stages of wheat crop i.e. CRI stage, tillering stage, anthesis stage, milk stage and dough stage were developed. The forecasted values (in Quintals/Hectare) for the year 2012 were computed as 32.34, 32.33, 32.07, 33.28 and 34.71. From the above information it could be concluded that the ARIMAX model can be used for forecasting of crop yield based on weather data.

Sanjeev and Urmil (2016) made a research in predicting the yield of sugarcane in Karnal, Ambala and Kurukshetra districts of Haryana. The ARIMA model with exogenous variables as input series makes an ARIMAX model. The weather data

obtained over the growth period of crop was used as input series along with the sugarcane yield for forming an ARIMAX model. The time series yield data from the year 1966 to 2009 were used for training set and the remaining data i.e. from 2010 to 2014 were used for validation of the model. Since the ARIMA model was not enough to explain the overall explanatory power of a model, the ARIMAX model was established. The predictive performance of the model was diagnosed based on the values of mean absolute percentage error (MAPE), RMSE and RD% etc. The model ARIMAX (0,1,1) for Karnal, ARIMAX (0,1,1) for Ambala and ARIMAX (1,1,0) for Kurukshetra gave better forecasted results of sugarcane yield for the years 2010 to 2014.

Singh *et al.* (2018) made a weather-based forecasting related to the genotypes of wheat in Varanasi region of India. The time series data of weather which included monthly average maximum temperature ($^{\circ}\text{C}$), monthly average minimum temperature ($^{\circ}\text{C}$), monthly average rainfall (mm) and solar radiation (MJ/m^2) for the years 1985 to 2016 were collected from India Meteorological Department, New Delhi (India). The forecasting of climate change over Varanasi region of India were estimated by Multiple Linear Regression (MLR), Autoregressive Integrated Moving Average (ARIMA) model, Autoregressive integrated moving average with exogenous variable (ARIMAX) model and Artificial Neural Network (ANN). From the results it was outlined that among all other models ARIMAX model (2, 0, 2) was the best fitted model for forecasting of production with highest R - square value of 37.4 %.

2.1.3 SARIMA model

Mwanga *et al.* (2017) forecasted the quarterly yield of sugarcane in Kenya. A seasonal ARIMA model had been fitted to predict the quarterly yield of sugarcane. The quarterly data was obtained from the year 1973 to 2014. A SARIMA (2,1,2) (2,0,3) model was chosen as the best model. Seasonal ARIMA model includes the additional seasonal component in the ARIMA model, it was denoted by ARIMA (p,d,q)(P,D,Q)_s where (p,d,q) was the trend component and (P,D,Q) was seasonal component. Among

the SARIMA models the SARIMA (2,1,2) (2,0,1) was chosen as the best model based on the minimum value of AIC and BIC. The results obtained from the fitted model showed that there was a drop in quarterly yield from the year 2016 to 2019 and there was rise in yield from 2024 to 2029.

Unnikrishnan *et al.* (2018) conducted a study to develop SARIMA model to predict the weather parameters of Thrissur district, Kerala for the next six years. The data on weather parameters such as maximum temperature ($^{\circ}\text{C}$), minimum temperature ($^{\circ}\text{C}$), humidity, total rainfall (mm), number of rainy days and wind speed (km/hr) were collected from 2012-2017 from the Department of Agricultural Meteorology, College of Horticulture, Kerala Agricultural University, Vellanikkara. Based on the selection criteria such as coefficient of determination (R^2), Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) the best model was chosen. SARIMA (0, 0, 2) (0, 1, 1)₁₂ for highest maximum temperature and for minimum temperature, SARIMA (0, 0, 0) (0, 1, 0)₁₂ for average monthly rainfall, SARIMA (1, 0, 0) (0, 1, 1)₁₂ for monthly rainy days, SARIMA (0, 1, 1) (0, 1, 1)₁₂ for average wind Speed, and SARIMA (2, 0, 11) (0, 0, 0)₁₂ for average humidity were chosen as the best model to forecast the weather parameters in future.

Alsharif *et al.* (2019) applied the SARIMA model to forecast the daily and monthly solar radiation. The research was conducted in Seoul, South Korea by taking an hourly solar radiation data for about 37 years (1981-2017) from Korean meteorological administration. From the results it was obtained that ARIMA (1,1,2) model with RMSE value 104.26 and R^2 value 68% was used to forecast daily solar radiation whereas seasonal ARIMA (4,1,1) model of lag 12 for both auto-regressive and moving average parts with RMSE value of 33.18 and R^2 value of 79% was used to explain the monthly solar radiation.

2.2 General Linear Model

Hoepfner and Dukes (2012) used mixed effects models with a randomized block, split-plot, repeated measures ANOVA design and restricted maximum likelihood (REML) estimation for all biomass and species diversity response variables and described responses of an old-field herbaceous community to a factorial combination of four levels of warming (up to 4 °C) and three precipitation regimes (drought, ambient and rain addition) over 2 years. In the drought treatment, warming suppressed total production, shoot production, and species richness. The study proved that according to warming or precipitation change the herbaceous component of old-field communities would not provide significant increase in forage production or a negative feedback to climate change later this century.

Roth *et al.* (2015) carried out a field trial on a 15 year old *Miscanthus* stand, subject to nitrogen fertilizer treatments of 0, 63 and 125 kg-N ha⁻¹, measuring N₂O emissions, as well as crop yield over a full year. To test the effects of treatment and time a repeated measures two-way ANOVA followed by a Bonferroni post-test was performed on N₂O flux, WFPS, soil temperature, soil nitrate concentration and soil ammonium concentration data. N₂O emission intensity (N₂O emissions calculated as a function of above-ground biomass) was significantly affected by fertilizer application, with values of 52.2 and 59.4 g N₂O-N t⁻¹ observed at 63 and 125 kg-N ha⁻¹, respectively, compared to 31.3 g N₂O-N t⁻¹ in the zero fertilizer control.

Shukor *et al.* (2015) studied growth and physiological response of Jack seedlings to over-top- filling treatment by imposing different levels of soil over – top – filling ie, 10,20 and 30 cm. Soil was mounted above the normal collar and covered. Growth and physiological characteristics were assessed and repeated measures analysis was used to analyse the differences among times and treatments. The results showed various patterns of morphological growth throughout the experiment.

2.3 Probability distribution

Wilcox (2005) focused on the utilization of crop yield distribution information for better on-farm decision making. The crop yield probability distributions were extensively used in farm support system. The mean, median, mode, skewness and kurtosis were the important statistics to be estimated in crop yield probability distribution. Since crop yields were dependent on environmental factors, rarely it follows normality. The rational decisions made by farm decision makers should include the likely situations they had faced, since from the starting with the true mid-point yield to the variance and skewness of crop yields. The variance and standard deviation were used as measure of data variability. The best estimates of future yield were obtained if the farm decision maker considered location, individual management and field history. Farm decision makers generated the crop yield distributions by entering the field history records into the spread sheet specific to their farm.

Kulshrestha *et al.* (2007) forecasted the weekly rainfall probabilities of Anand state of Gujarat, India. The Gamma distribution model (GDM) and Artificial neural network were used to predict the weekly rainfall probabilities by using the 48 years of rainfall data series. The actual probabilities for the weekly rainfall data was computed using MATLAB function, then the probabilities obtained by GDM were compared with actual probabilities. The probabilities computed by ANN were significant to probabilities found by GDM and GDM probabilities were significant to the actual probabilities. Therefore, it was concluded that probabilities by ANN were also significant to actual probabilities.

Bhagat and Patil (2014) used three probability distribution functions to estimate a reference crop evapotranspiration at different probability level for Solapur district of Maharashtra. The probability distributions were identified to be log normal, Gumbel and Weibull's probability distribution functions. The standard Penman-Monteith method was used to collect the weekly reference crop evapotranspiration for the period 1977-2007 and tested at 10 percent to 90 percent probability levels. The chi-

square test was applied to test the fitness of probability distribution and the log normal distribution was considered to be the best fit for maximum weeks, which showed the lowest values of chi-square at 5 percent level of significance.

Dutt *et al.* (2016) studied the spatial pattern of occurrence of insect pest (Green stink bug) on pigeon pea crop and fitted a negative binomial distribution. The parameters of Negative binomial distribution were estimated by the Method of Moments (MM) and Method of proportion of Zeroth cell (MPZC). The Negative binomial distribution is mainly used in the aggregation (or) clustering behavior of plants, animals (or) insect population. This empirical distribution provided an excellent model when the distribution had variance larger than the mean. From the distribution with parameter p and r it was revealed that the spray of insecticides during the first week of December was more effective to keep the crop free from Green stink bug.

Subudhi *et al.* (2019) worked on the probability analysis of annual, seasonal and monthly rainfall data of Rayagada district of Odisha. The study was undertaken to overcome the low yield of crop due to improper crop planning. The rainfall data was collected for 17 years from 2001 to 2017. The different values of data were then subjected to various probability distribution functions namely, normal, log-normal (2-parameter), log-normal (3-parameter), gamma, generalized extreme value, Weibull, generalized pareto distribution, log-Pearson type-III and Gumbel distribution. The various parameters like mean, standard deviation, RMSE values were obtained and noted for different distributions. The best fitted distribution for different months, seasons and annual was determined. From the study it was concluded that the rainfall during June to September was less than 1000 mm and cropping pattern like paddy was taken followed by mustard. However, kharif rain could be harvested and it could be reused for another rabi crop by using sprinkler or drip irrigation.

2.4 Impact of climatic variables on crop yield

Ajayi *et al.* (2010) established that annual rainfall had an inverse relationship with the cocoa yield. They found that rainfall had a constraining effect on the cocoa yield over some area of Ondo state Nigeria. It was observed from graphical tools that as the range of the rainfall increased the yield of cocoa decreased.

Lawal and Ommonona (2014) evaluated the annual measurement of three climatic parameters viz; rainfall, temperature and RH about three decades 1980-2011 and the yield of cocoa over those periods in Nigeria. The data was exposed to inferential statistics and regression analysis using STATA. It was concluded that the rainfall showed negative correlation, meanwhile the temperature and RH showed positive correlation against the yield of cocoa. From the experiment it was inferred that the rainfall declined the yield of cocoa, whereas the temperature and RH improved the physiological processes in cocoa.

Manikandan *et al.* (2014) studied the relationship between cocoa yield fluctuations and weather variations in Vellanikkara, Thrissur. Karl Pearson correlation was applied to know the relationship between the weather parameters and cocoa yield. It was found that, only maximum and minimum temperature had significant effect on cocoa yield. The study revealed that high maximum temperature from January to middle of march along with high rainfall during the rainy season assessed to be detrimental to obtain better yield of cocoa.

Chizari *et al.* (2017) urged that the autoregressive distributed lag (ARDL) co-integration approach was a better analytical approach to know the impact of climate change on the yield of cocoa in Kenya, which helped to estimate, forecast, and stimulate the levels of cocoa production based on climate changes. They found that the production trend was positive and the increase in temperature and rainfall would lead to about 6.06% rise in annual yield.

Sitienei *et al.* (2017) found a multiple linear model to forecast the effect of climatic variables on the yield of tea crop using the climatic variables such as maximum temperature, minimum temperature and precipitation in Kenya. The study found the statistical relationship between the climatic variables through scatter diagrams, correlation analysis and trend analysis. The output obtained from the regression model was verified through contingency tables. The results obtained from the verification of the model showed that 70% of the model forecasts were having high degree of accuracy.

Wiah and Ankrah (2017) investigated the effect of four major climatic variables viz; maximum temperature, minimum temperature, precipitation and number of rainy days on cocoa yield in Ghana. Vector Auto regressive (VAR) model was used to check the vigorous influence of climate change on cocoa yield. The maximum temperature was found to be the climatic parameter with the largest number of significant cross correlation on yield preceded by the minimum temperature. Granger causality tests gave F- statistic for maximum temperature (4.12), minimum temperature (3.04) and precipitation (0.87) which showed significant effect on yield, whereas the F- statistic for number of rainy days (0.47) was insignificant. Therefore, it was suggested that agriculture sector should provide new cocoa seedlings that would have resistance to higher temperature and low precipitation to maintain high yields of cocoa in Ghana.

Sujatha *et al.* (2018) used the data for weather variables such as rainfall, maximum temperature, minimum temperature, sunshine hours, pan evaporation and RH for the period from 1970-2012 at Vittal, Karnataka and using SPSS, linear correlations and regressions were developed to assess the quantitative relationship between weather variables and cocoa yield. A multiple regression analysis was performed using weather variables. The results showed that the regression analysis of monthly variables like RH, maximum temperature, sunshine hours and rainfall explained more yield variability of cocoa, whereas minimum temperature and evaporation explained less yield variability.

2.5 Assessing the Yield gap

Job (2006) assessed the yield gap of rice in Alappuzha by using three stage random sampling scheme. The Frontier production function was employed to estimate the maximum feasible yield (MFY) and yield gap. The rice yield gap in Alappuzha was found to be 1588 kg/ha with an MFY of 5447 kg and actual yield of 3859 kg/ha.

Verma *et al.* (2012) applied FLDs (Front line demonstrations) to assess the yield gap on mustard crop. The FLD was conducted by KVK in the adopted villages of Faziabad district to quantify the average technology gap and extension gap. The average technology gap was found to be 4.40 percent and the extension gap ranged between 2.45 q to 4.72 q/ha.

Aneani and Frimpong (2013) performed an yield gap analysis of cocoa in Ghana. It was done by subtracting achieved average yield from the yield potential of cocoa. It was performed by a cross sectional survey of selected districts by adopting multistage sample technique for interviews. It was concluded that farmer's yield gap was smaller compared to experimental yield gap as well as model yield gap. The experimental yield gap was found to be 1553.4 kg/ha which indicated 82.1% of the experimental yield potential where as farmers yield gap was 1537.2 kg/ha which accounted for 82% of the farmer yield potential.

Pushpa and Srivastava (2014) conducted a study to assess the gap between the current and potential yields of major crops namely wheat, rice and sugarcane in eastern region of Uttar Pradesh. The yield gap differed from 20.01 to 53.85 %, 15.56 to 30.10% and 5.8 to 28.89% with the average gap of 28.26 %, 20.93% and 17.5% for rice, wheat and sugarcane crops, respectively in the irrigated region of Uttar Pradesh. The yield gap in percentage for paddy based on overall farm size (marginal, small, medium and large) was found to be 28.82%, while the yield gap in percentage for wheat crop based on overall farm size (marginal, small, medium and large) was found to be 20.93% whereas the yield gap in percentage for sugarcane crop based on overall farm size (marginal, small, medium and large) was found to be 17.5%.

Elum and Sekar (2015) examined the yield gap in seed cotton in Tamil Nadu. The survey was conducted using both the purposive and multistage random sampling. The determinants of yield gap were assessed through multiple regression model. The results from the study showed that the yield gap in BT cotton was significantly higher than that of conventional cotton and also the potash gap had significant but negative effect on yield, whereas nitrogen had significant and positive effect.

Singh *et al.* (2015) applied crop simulation model to examine the potential yield in eastern and north eastern regions of India. The CERES- rice model was used to determine the potential yields of rice for about 21 years. The daily weather data, the district wise rice yield data, the soil information data and the crop genetic data were used as input data for the model. From the simulation analysis it was concluded that the yield gap was large in different districts of eastern and north eastern India, which suggested that the rice yield could be increased up to 11 to 22 percent in different districts with better management practices. Thus, for better decision making to improve the response use efficiency CERES-Rice model could be used.

2.6 Structural Equation Modelling

Sefriadi *et al.* (2013) identified Structural Equation Model (SEM) to carry out the path analysis for cocoa production in west Sumatra, Indonesia. Based on the views of the cocoa farmers regarding the constraints they faced during the cocoa production which affected their incomes an SEM model was developed. Path analysis was a method to interpret the correlation between the variables in a linear causal model. Path analysis was a way to approach the SEM where it was visually represented in the form of path diagram involving complex variables. The goodness of fit for the model was tested by looking into the values of Root mean squared error of approximation (RMSEA), comparative fit index (CFI) and Tucker Lewis index (TLI). The results showed the value of CFI to be 0.941 (>0.9), TLI to be 0.928 (>0.9) and RMSEA to be 0.070 (<0.08). It was proved that the model was absolutely fit.

Shadfar and Malekmohammadi (2013) developed a model to construct the state intervention policies in rice production development in Iran. To examine those policies, Structural Equation Model (SEM) and Confirmatory Factor Analysis (CFA) was applied. A theoretical model was obtained by using SEM and then the validity and reliability of that model was done by CFA. The proposed model was first tested for the GOF (Goodness of Fit) indices and then the validity and reliability was checked. The GOF statistic was SRMS (Standardized Root Mean Square Residual), the value obtained was 0.064 (< 0.08) and RMSEA (Root Mean Squared Error of Approximation) value was 0.087 which declared that the model was adequately fit.

Monaganta *et al.* (2018) conducted a survey to study the factors that influence interdependence of cocoa farmers in central Sulawesi province. The custom random sampling technique was used to get the sample from 380 respondents. The sample data was examined through SEM (Structural equation modeling). The study was mainly aimed to improve the competence, capacity and interdependence of farmers.

Utami *et al.* (2018) proposed the Structural Equation Model (SEM) as a primary method to identify the key factors that affect the productivity of small holder cocoa farming in Indonesia. The model of the cocoa production was demonstrated by using SEM. By using the Confirmatory Factor Analysis (CFA) the variables and the factors that were strongly correlated to the productivity of cocoa were identified. The results of the Goodness of Fit (GOF) revealed that the RMSEA value for the proposed model was 0.076 (< 0.08), which showed that the model was a best fit model. The best fit model was used for further Confirmatory Factor Analysis (CFA), from the results obtained it was found that the construct variables Natural resource capital and Economic capital showed the higher standard loading factors value (> 0.2) and hence concluded that those two variables were the dominant variables that affected the cocoa production.

2.7 Probit regression model

Rahman (2008) used a bivariate probit analysis to identify the socio-economic factors with regard to the decision of farmers to crop choices in Bangladesh. For the diversified cropping system, the dependent variable took the value 1 and 0 otherwise. The value 1 refers to the farmer growing modern rice and 0 otherwise. The study was done based on adoption of diversified cropping system or modern rice technology. From the model it was observed that availability of irrigation was an important determinant to adopt modern rice technology, whereas the variables like farmers education, farming experience, farm asset as well as the share of non-agricultural income were found to be significant to adopt the diversified cropping system.

Raguindin and Vera (2011) investigated the adoption of WST (water saving technologies) such as controlled irrigation, direct seeding, land levelling and aerobic system by the rice farmers in Philippine. The significant variables that influenced the WST were education, experience in rice farming, family income of the farmers and size of manpower involved in farming. However, age and debt were negatively correlated to adopt controlled irrigation. The multivariate probit model was considered to be an efficient model which would be a generalization of the Probit model used to estimate several correlated binary outcomes jointly, since the four WST were correlated. The adoption rates of different water saving technologies were found to be 17.4%, 14.7% and 16.2% for direct seeding, land levelling and aerobic rice system respectively. The controlled irrigation had a high adoption rate over 50%.

Samal *et al.* (2011) made a research regarding the spread of modern rice varieties in different water regimes in the rainfed coastal Orissa. It was found that coverage of modern varieties was 37 percent in medium land and 11 percent in lowlands. The study revealed that the wider spread of modern varieties depended on development of new varieties that were adopted to adverse agro-climatic conditions and it was found that once the new varieties were developed, the irrigation and land reform facilitated the faster spread of modern rice varieties in coastal Orissa. A multivariate

probit model was applied to study the factors affecting adoption of modern varieties. The model was used by applying maximum likelihood method to estimate the coefficients. The dependent variable assumed the value 1 if modern varieties were adopted and 0 if not adopted.

Duniya and Adinah (2015) recommended the cotton farmers of Nigeria to access the formal source of credit instead of informal source from the study through the use of probit regression model. From the model it was observed that the factors such as formal education, off-farm income, household size, farm size and farming experience significantly influenced the credit availability to farmers. Based on the study conducted it was guided that extension agents should educate farmers and create awareness regarding importance of higher level of formal education.

Anang (2016) applied probit model to study the decisions of cocoa farmers to adopt fertilizers in Bekwai district of Ghana as the dependent variable which was binary in nature. The dependent variable in the probit model assumed the value 1 for the adoption and 0 for the non- adoption of fertilizers. Age, house hold size, farm size, extension contact, farm income, access to mass spraying were considered to be the independent variables among which the farmer's age, farm size and farm income were the significant factors that affect fertilizer adoption decision, however the remaining were insignificant in adoption of fertilizers. Thus, the result showed that farmer's age, farm size and farm income were the critical determinants of decision making and it was decided that the fertilizers should be subsidised by the government to promote adoption and the extension service should be improved so that the farmers get information on improved production practices.

Denkyirah *et al.* (2016) determined the factors which helped the farmers to access credit by checking the significance of different independent variables using probit model. The study was conducted in the upper east region of Ghana, where the rice farmers invested the credit for non-agricultural activities which hindered the adoption of technologies introduced in the region. The probit model showed that age,

marital status, membership of farm-based organisation, extension visit, record keeping and farm income were the significant variables that helped to access the credit, whereas the age and farm income showed negative impact to access the credit. From the study it was revealed that rice farmers should be advised to utilise the credit for agricultural activities to increase the productivity and the extension department should guide the farmers to maintain the record that positively influenced the farmers to attain the credit.

Kehinde and Adeyemo (2017) made a detailed study regarding the dis-adoption of improved technologies by the farmers in cocoa farming system of south eastern Nigeria. The improved technologies were improved seed varieties, fertilizers, recommended spacing, recommended mixed cropping and pesticides. They used the probit regression model to evaluate the factors affecting dis-adoption of improved technologies. The factors identified were membership of an association, years of formal education, access to credit, farm size, household size, gender and contact with extension agent. Through the analysis the significant variables were taken into consideration and efforts were made to reduce the dis-adoption of improved technologies in cocoa based farming system.

Chandio and Yuansheng (2018) adopted probit regression model and identified the factors affecting adoption of improved rice varieties by smallholder farmers in Northern Sindh, Pakistan. The results obtained from the probit econometric model showed that education level, farming experience, soil quality, market information, farm machinery ownership and extension contact had significant and positive effect on the adoption of improved rice variety whereas age had significantly negative effect.

Hambisa (2018) studied the determinants that helped access to formal credit to coffee farmers in Bodji Dirmeji district of west wollega, Ethiopia. The probit model depended on the nature of dependent variable which assumed the value 1 for access to formal credit and 0 otherwise. Through the maximum likelihood estimation of probit regression model it was showed that education level of the household, sex, family size,

extension contact frequencies, perception of group leading were significantly affecting the coffee farmers to access formal credit.

Shee *et al.* (2019) examined the determinants of post-harvest losses at each post-harvest stage of maize and sweet potato value chains for small holder farmers which was an important pathway to food and nutrition security in sub-Saharan Africa. Since the dependent variable was found to be ordered and categorical in nature, the use of ordinary least square and multinomial logit/probit types models were not appropriate, instead an ordered probit model was mostly widely used in empirical econometric applications.

Adhikari *et al.* (2020) worked on probit regression model and identified the factors that impacted on the decision to use herbicides by the farmers in wheat production in Nepal. The factors such as education, membership in organisations, migration of household members and wheat cultivated area were identified as the significant factors that influenced the decision of farmers to use herbicides. The dependent variables were dichotomous in nature and took the value 1 for adopter and 0 for non-adopters.

2.8 Coefficient of Concordance

Anang *et al.* (2011) evaluated the benefits and constraints faced by the cocoa farmers in Bibiani Anhwiaso-Bekwai district of Ghana. The constraints were ranked and applied Kendall's coefficient of concordance (W) to assess the degree of agreement among the farmers. The value of the coefficient W was 0.46. In regard with the constraints, pests and diseases were the highest followed by other constraints. Therefore, it was instructed that the Cocoa Diseases and Pests Control Exercise Committee (CoDAPEC) should monitor to ensure that cocoa farms would be properly sprayed to control pests and diseases.

Donkoh and Awuni (2011) studied the perception of farmers on the most important farm management practices that increased the output or income of low land rice production in Northern Korea. The ranking of the variables was done through Kendall's coefficient of concordance. The order of ranking was done in such a way that the most important variable had the smallest sum of ranks, while the least important variable had the greatest sum of ranks. The degree of agreement among the respondents was estimated by the value of W. The Kendall's coefficient of concordance was 51% at 1% level of significance, which showed that there was 51% of agreement among the respondents over the ranking of farm management.

Codjoe *et al.* (2013) measured the agreement among the ranks allotted by the cocoa farmers which identified the constraints that influenced the efficient functioning of cocoa-based Agriculture Knowledge and Information System (AKIS). The Kendall's coefficient of concordance was applied to detect the range of disagreements and agreements among the respondents. The respondents ranked fifteen (15) constraints in a descending order and the coefficient of concordance was observed to be 0.227 with 14 degrees of freedom. The results from the study revealed that inadequate interaction with researchers and extension agents had a mean rank of 4.57 which represents the highest-ranking order.

Abbeam *et al.* (2014) verified the constraints that inhibited the consumers about the consumption of local rice in the Tamale metropolis, Ghana. The Kendall's coefficient of concordance was a measure used to detect the agreements among several judges (respondents) who ranked some of the constraint's preference for local rice. The test of significance for the Kendall's coefficient of concordance was done by using chi-square statistics, which revealed that calculated chi-square was greater than the chi-square critical, so that the null hypothesis was rejected. From the statistical analysis it was observed that the poor packing of local rice was ranked as the most prominent constraint that inhibited the consumer's preference for local rice with mean rank of 2.32 followed by other constraints. Therefore, it was found that the local rice processors should be worked at improving the packaging to make it competitive in the market.

Amedi (2014) used the Kendall's coefficient of concordance to study the agronomic constraints among the rice farmers under the MiDA in the Hohoe municipality. The Kendall's coefficient (W) 0.61 showed that there was 61% agreement among the farmer's constraints ranked. In the system of ranking the constraint with highest rank was used to be the least important and the constraint with lowest rank assumed to be the most important. The rice farmers ranked the factors such as poor climatic condition, high incidence of pests, poor yield, high cost of farm inputs and problem of poor milling equipment to be the topmost five constraints.

Nirmala (2015) made a situation analysis of hybrid rice seed production in Telangana and Andhra Pradesh. The hybrid rice seeds producers ranked the factors associated with SWOT (strength, weakness, opportunities and threats) by the method of Kendall's coefficient of concordance (W). The coefficient of concordance was estimated to be 0.61 with 7 degrees of freedom and the value of chi-square was found to be 512.41 which was greater than the critical value (14.067) and hence they rejected the null hypothesis that there was no agreement among the sample seed producers.

Tanko (2017) investigated the challenges of shea butter producers in the northern region of Ghana. The dealers faced the challenges with respect to selling and processing of shea butter and they used the Kendall's coefficient of concordance and ranked the challenges. It was evident from the study that the poor-quality input was the most important challenge with mean rank of 1.02 and the other most challenging factor of processing and marketing was out dated equipment for processing shea butter, high cost of input, poor transformation network *etc.* The Kendall's coefficient (0.728) showed greater agreement among respondents in the ranking of challenges.



MATERIALS AND METHODS

CHAPTER 3

MATERIALS AND METHODS

In this chapter, a brief description of materials and statistical methods employed to analyse the data pertaining to various objectives of the study are discussed.

The study was conducted in three phases. The first phase made use of time series data pertaining to area, production and productivity of cocoa in Kerala for the years 1980 to 2017. The second phase made use of data on monthly yield and infected pods from 100 selected cocoa hybrids of Cocoa Research Centre, Kerala Agricultural University for the analysis. In the third phase an empirical analysis was done by collecting primary data on demographic characters, cocoa cultivation and management practices, constraints faced in cocoa farming and important factors leading to cocoa yield and ultimately to the net income of farmers. The 100 farmers who have contacts with the cocoa project “Mondelez International Limited” were selected from Iritty panchayat of Kannur district and Veliyamattom panchayat of Idukky district.

The contents of the chapter is outlined as follows

3.1 Box and whisker plot

3.2 Time series Forecasting

3.2.1 ARIMA model and Exponential smoothing model

3.2.2 SARIMA model

3.2.3 ARIMAX model

3.3 General Linear Model

3.4 Probability distribution

3.5 Impact of Climatic variables on cocoa yield

3.6 Structural Equation Modelling

3.7 Probit regression model

3.8 Coefficient of concordance

3.9 Assessing the Yield gap

3.1 Box and whisker plot

A box plot (box and whisker plot) is a diagram through which the distribution of a data set can be represented. It represents the summary of five measures of the data. Five measures of the data are the minimum value, the first quartile, median, the third quartile, and maximum value of the data.

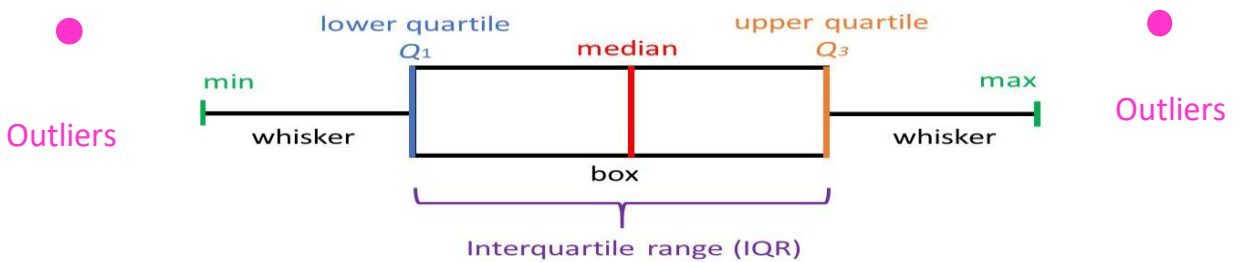


Fig. 3.1 Box plot

Here in the box diagram the box is plotted along the first quartile and the third quartile. The median is defined by the Vertical line in the box. At both the quartiles the whiskers pass and defines the maximum value and the minimum value. It also tells about the outliers and their values. The boxplot is used to know whether the data is normally distributed, variability of the data and whether the data is positively or negatively skewed. The boxplot ensures more advantageous when compared to histogram and density plot. It occupies less space which is useful in comparing distributions among many groups and datasets.

Q1 – the first quartile (it is the middle value between the lowest number and the median in the dataset)

Q2 – the second quartile (represents the median in the dataset).

Q3 – the third quartile (it is the average between the median and the maximum value in the dataset).

IQR – Interquartile range (it is the range between quartile 1 and quartile 3).

Whiskers (drawn in black).

Outliers (drawn in pink circles).

Maximum value = Quartile 3 + 1.5* Interquartile range.

Minimum value = Quartile 1 -1.5* Interquartile range.

3.2 Time series forecasting

3.2.1 ARIMA and Exponential smoothing model

The yearly data on area, production and productivity of cocoa in Kerala was obtained from the ‘Directorate of Cashew nut & Cocoa Development board (DCCD)’ located at Cochin, Kerala. To forecast the area, production and productivity of cocoa in Kerala univariate time series data for the years 1980-2017 were analysed. The data for the years 1980-2011 were used for building the model (training period). After validation of the model for the remaining years a suitable model was fitted for the whole data to forecast the area, production and productivity of cocoa in Kerala for the next 5 years from 2018-2022. The work was undertaken by applying ARIMA (Auto Regressive Integrated Moving Average), Simple exponential smoothing and Holt’s exponential smoothing models were also fitted using the SPSS statistical software package.

Forecasting is mainly required in business, Industry, government and in many institutions for making policies and future planning. There are several different ways of forecasting; the selection of the method depends on the intent and the importance of prediction, as well as the expense of methods involved. The frequently used time series forecasting method is the Box and Jenkins Auto Regressive Integrated Moving Average (ARIMA) model. In a Box and Jenkins ARIMA models, the univariate time series data

is the data where the observations are collected at sequence of point of time and the predictions are made based on the previous values. The important characteristics of this data is the dependency of the successive observations. The values in the observed data series, Y_t , is regarded as the attainment of a stochastic process $\{Y_t\}$, which is a family of random variables $\{Y_t, t \in T\}$, where $T = \{0, \pm 1, \pm 2, \dots\}$.

Box and Jenkins researched ARIMA models extensively throughout 1968, and their names were also used synonymously with the ARIMA method for time series forecasting. The stochastic model for the time series data is used to forecast the future values. There is either a stationary or non-stationary stochastic process and most of the time series are non-stationary.

The major steps in Box-Jenkins forecasting model are as follows.

Identifying the model

Estimation of parameters

Diagnostic check and Forecasting

Stationarity of a Time Series process

If a TS is generated with a constant mean then it is called as stationary data with its autocorrelation function and variance basically constant over time.

It is used for auxiliary regression parameter

$$\Delta_1 y_t = \rho y_{t-1} + \alpha_1 \Delta_1 y_{t-1} + \varepsilon_t$$

Where Δ_1 denotes the differencing operator i.e. $\Delta_1 y_t = y_t - y_{t-1}$

The appropriate null hypothesis is $\rho = 0$, which means that the series is non-stationary.

Whereas, the alternative hypothesis is $\rho < 0$, which says that the series is stationary.

The differencing of the data is usually done until the ACF shows a few significant autocorrelations with an interpretable pattern.

Autocorrelation functions

Autocorrelation is a measure of correlation between the observation Y_t at point of time t , with the observation Y_{t-p} which lags at p periods from the current observation Y_t . The correlation between the two observations (Y_t, Y_{t-p}) is given by

$$r_p = \frac{\sum_{t=1}^{n-p} (Y_t - \bar{Y})(Y_{t+p} - \bar{Y})}{\sum_{t=1}^n (Y_t - \bar{Y})^2}$$

The value of r_p lies between -1 to +1. It was also found that the maximum number of useful r_p are calculated as $N/4$, where N is the number of periods from which the information on y_t is available.

Partial autocorrelation

Partial autocorrelation is a statistical method used in forecasting. It is a measure of degree of association between the two values y_t and y_{t-p} in a time series data when the y -effects at other time lags 1, 2, 3, ..., $p-1$ are removed.

Pankratz (1983). mentioned that the Autocorrelation function (ACF) and partial autocorrelation function (PACF) are calculated for different models taken for different values of orders of autoregressive and moving average components i.e. p and q . Hence, a correlogram is plotted for a given TS data by plotting the sample ACFs against the lags and compared with the theoretical ACF/PACFs in order to find the appropriate match and selecting one or more ARIMA models.

The general features of theoretical ACFs and PACFs are outlined in the table (here the word ‘spike’ represents the line at different lags in the plot with length equal to magnitude of autocorrelations)

Model	ACFs	PACFs
AR (p)	Spikes decay to zero with exponential pattern	Spikes cutoff to lag p
MA (q)	Spikes cutoff after lag q	Spikes decay to zero exponential pattern
ARMA (p, q)	Spikes decay to zero with exponential pattern	Spikes decay to zero exponential pattern

Description of ARIMA models

Autoregressive (AR) Model

An auto regressive model is a stochastic model which is widely used in the representation of series which occurs practically. Here the current value y_t depends on the finite and linear addition of previous values of the process with \mathcal{E}_t . The values of the time series are equally spaced at time intervals $t, t-1, t-2, \dots$ by $y_t, y_{t-1}, y_{t-2}, \dots$, then y_t can be written by the following equation:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \mathcal{E}_t$$

The autoregressive operator of order p is expressed by

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p$$

where B is the backshift operator such that $B y_t = y_{t-1}$, the autoregressive model can be written as

$$\varphi(B)y_t = \mathcal{E}_t$$

Moving Average (MA) Model

A moving average model is a MA(q) model that has a great practical application where the value y_t is the weighted moving average of the lag values of error terms. The regression equation is represented by

$$y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

The moving average operator of order q is represented by

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

where B is the backshift operator such that $By_t = y_{t-1}$, the moving average model can be written as

$$y_t = \theta(B) \varepsilon_t$$

Autoregressive Moving Average (ARMA) Model

To forecast the future values with great accuracy by fitting the actual time series data it is most important to combine the autoregressive and moving average model which leads to ARMA model.

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \dots + \varphi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q}$$

or

$$\varphi(B) y_t = \theta(B) \varepsilon_t$$

An ARMA model is the addition of autoregressive and moving average model which is denoted by ARMA (p, q) model. It is used for forecasting only stationary time series data.

Autoregressive Integrated Moving Average (ARIMA) Model

An ARIMA model is the extension of ARMA model which applies differencing into the model. ARIMA model is more advantageous over ARMA model which is extensively used for fitting the non-stationary time series data. The simple example of reducing a non-stationary process into a stationary one after differencing is Random Walk. A process $\{y_t\}$ follows an Integrated ARMA model, denoted by.

ARIMA (p, d, q) , if $\nabla^d y_t = (1 - B)^d \varepsilon_t$ is ARMA (p, q) . The model is written as

$$\varphi(B) (1 - B)^d y_t = \theta(B) \varepsilon_t$$

Where $\varepsilon_t \sim WN(0, \sigma^2)$, WN indicating White Noise. The integration parameter d is a non-negative integer. When $d = 0$, ARIMA $(p, d, q) \equiv$ ARMA (p, q) .

The technique of ARIMA follows three steps i.e. selecting the model, estimation of model parameters and diagnosis checking of fitted model. In the first step the ARIMA is selected tentatively. The parameters of the selected model are assessed in the second step and the accuracy of the model is tested in the third step i.e. diagnostic check. All the three steps are repeated when the model is found inadequate until best satisfactory model is obtained for the time series data. Box *et al.* (2011) has explained the different aspects of this approach. The analysis of fitting the ARIMA model is done by the standard software packages, like SAS, SPSS, R and EViews.

Exponential smoothing methods

It is an effective method of forecasting which can be used as a substitute to replace the most common Box-Jenkins ARIMA model. Exponential smoothing is a tool for the estimation of univariate time series data. Here in this technique the predictions are made from the past values which are weighted averages. The name exponential smoothing technique reveals that the weights decline exponentially as the values get older. It means that the recent values get more weights than the past values.

The main concept behind this technique is that the recent values get high significance in a time series. When the values get older the significance for those values get decline exponentially.

Types of Exponential smoothing methods

1. The method with single parameter: Single Exponential smoothing (also called Simple Exponential smoothing)

2. The method with two parameters: Double Exponential smoothing

- i. Brown's Liner Method with single parameter
- ii. Holt's Linear Method with two parameters

3. The method with three parameters: Winter's exponential smoothing model

- i. Winter's Multiplicative Method
- ii. Winter's Additive Method

Simple Exponential Smoothing

This method is the simplest among exponential smoothing methods so it is named as simple exponential smoothing (SES). This approach is appropriate for predicting data without any specific trend and the data which do not show any seasonal behaviour. In this approach all the predicted future values are said to be equal to the last observation of the series,

$$y_{T+h/T} = y_T$$

for $h=1, 2, \dots$

In this approach it is assumed that most of the information is provided by the recent values rather than the past observations. This can be explained by the weighted average where more weights are given to the last observations. From the average method the future forecasted values are simple average of the observed data.

$$y_{T+h/T} = \frac{1}{T} \sum_{t=1}^T y_t$$

$h=1, 2, \dots$ From the above equation it is described that all the observations are given equal importance which gives equal weights during the forecasting.

We expect something more between the extreme values of the observed data. It is important to provide the higher weights to the recent observations than to the observations which are far behind. This is the concept regarding the simple exponential technique. Forecasts are calculated based on the weighted averages where the weights get decline exponentially and the older observations get smaller weights.

$$Y_{T+h/T} = \alpha Y_T + \alpha(1-\alpha) Y_{T-1} + \alpha(1-\alpha)^2 Y_{T-2} + \dots$$

α is the smoothing coefficient which ranges between 0 to 1. It determines the rate at which the weights decrease. From the above equation it is confirmed that the one-step forward forecast for the time $T+1$ is the weighted average of all the observations in the series Y_1, \dots, Y_t .

Holt's Exponential smoothing model

Holt's winter exponential smoothing model is an extension of simple exponential smoothing model. Here in this model it is considered for two parameters so it is called Holt's two parameter model (or) Holt's double exponential model. Holt's (1957) developed this model to forecast the time series data with the trend. In this method the forecasting equation depends on the trend and level of the time series.

$$\text{Level equation } L_t = \alpha Y_t + (1-\alpha) [L_{t-1} + T_{t-1}]$$

$$\text{Trend equation } T_t = \gamma [L_t - L_{t-1}] + (1-\gamma) T_{t-1}$$

$$\text{Forecast equation } F_{t+1} = L_t + k T_t$$

Where L_t refers to the level estimate of the time series at time t and T_t denotes the trend estimate of the time series at time t , and α is the smoothing coefficient of the level equation which ranges between 0 to 1. γ is the smoothing coefficient of the level equation which ranges between 0 to 1.

3.2.2 SARIMA model

The monthly cocoa yield data of 100 plants for the period from 2003 to 2017 obtained from Cocoa Research Centre, Vellanikkara, Thrissur, Kerala were made use of to fit a SARIMA model for the prediction of monthly cocoa yield. Since the time series data obtained was seasonal in nature with repeating cycles the simple ARIMA modeler was not suitable where it only holds for the time series data that do not contain any seasonal component.

Seasonality is the seasonal behaviour of the time series that occurs regularly after every P periods, where P refers to the time period over which the pattern repeats when the time series data contains the seasonal component then we use the forecasting model called Seasonal ARIMA (also called SARIMA) model.

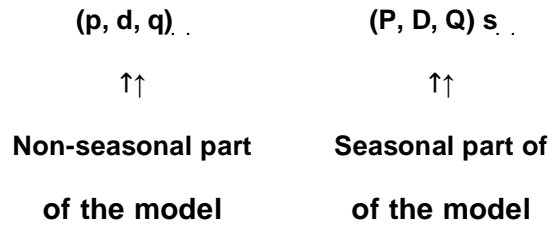
The SARIMA model is designed for two types of data. For the monthly data $S = 12$ and for the quarterly data $S = 4$, where S refers to the time periods in a year.

The time series component in a SARIMA model contains the AR and MA terms which predict the value Y_t based on the past values as well as the error terms at times with certain lags which are multiples of S (span of seasonality)

For example,

- In a first order seasonal auto regressive model for a monthly data with $S = 12$ will use Y_{t-12} lag values to predict the value Y_t
- Similarly, a second order seasonal auto regressive model would use Y_{t-12} and Y_{t-24} to predict the value Y_t .

The seasonal ARIMA model will be divided into two components, which are Non-seasonal component and seasonal component.



The Non-seasonal component also known as trend component with three parameters denoted by p, d, q.

p - non-seasonal autoregression order.

d - non-seasonal difference order.

q - non-seasonal moving average order.

The Seasonal component contains four elements denoted by P, D, Q and s.

P - Seasonal AR order.

D - Seasonal difference order.

Q - Seasonal MA order.

s - span of seasonality

3.2.3 ARIMAX Model

ARIMAX is an extension of ARIMA with exogenous variables which add the explanatory value of the model. Theoretically, ARIMAX is a mix up of regression and ARIMA modelling. If an ARIMA model is unfit to explain the overall explanatory power then the ARIMAX model is used to forecast.

When the ARIMA model includes other time series as input variables, the model is referred to as ARIMAX. In addition to past values of the response series and past errors, the current and past values of exogenous variables are also taken to model the response series.

$$y_i = \beta x_i + \sum_{j=1}^p \phi_j y_{i-j} + \varepsilon_i + \sum_{j=1}^q \theta_j \varepsilon_{i-j}$$

Apart from the Production data of cocoa in Kerala for the period from 1980 to 2017, area under cocoa for the same period was considered as exogeneous variable for building the ARIMAX model. Taking the production as dependent variable (endogenous variable) and area as independent variable (exogenous variable) the ARIMAX model was evaluated. The analysis was done using the ‘SPSS’ statistical software package.

3.3 General Linear Model

An empirical analysis was done to know the effect of different time periods on cocoa yield. The monthly cocoa yield data was collected for 100 trees of same age from the Cocoa Research Centre, Vellanikkara, Thrissur, Kerala. The repeated measurements on total number of cocoa pods from the same 100 trees over different time periods were made on a monthly basis for the period from 2003 to 2017. The analysis was done by using General Linear Model (GLM) Repeated Measures one-way ANOVA to study the effect of time over cocoa yield and thereby to estimate the interaction effect of time with respect to low and high yielding groups of trees. The analysis of variance and regression methods cannot be used to solve the repeated measures data because the data will not satisfy the prescribed assumptions.

In GLM repeated measures mainly three kind of effects are analysed

i) The effects of between-subject or GROUP effect

ii) The effects of within-subject or TIME effect

iii) The interaction effect of both the main effects or GROUP*TIME interaction effect

General linear model (GLM) repeated measures ANOVA is a statistical procedure used to compute the values of dependent, or criterion variable, measured as correlated and non-independent data at different time periods. The variables that are independent in GLM can be categorical and continuous. The key effects of within and between the subjects and their interaction, also the effects of covariates including the interaction of covariates and between subject factors are estimated in GLM repeated measures analysis.

Analysis of Variance (ANOVA) is a powerful statistical test which is used to compare the mean values obtained during an experiment from various conditions or groups. There are many different types of ANOVA and here One-Way Repeated-Measures ANOVA is used.

3.3.1 Test of Hypothesis of ANOVA in GLM

In the ANOVA of GLM repeated measures test, an assumption is made regarding the hypothesis to be tested for any differences between related population means.

The null hypothesis (H_0) tells that there is no significant difference between the related populations

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_m$$

In which μ = population mean and m = number of related groups.

The alternative hypothesis (H_1) tells that there is a significant difference between the related population means

H_1 : there is a significant difference between at least two means

3.3.2 ANOVA in GLM

The ANOVA in GLM under repeated measures is a special case of between-subjects ANOVA. The total variability in the between-subjects ANOVA is divided into two components i.e. variability between the groups (SS_b) and variability within the groups (SS_w), as described below

Partitioning of variability in a repeated measures ANOVA

The error variability (SS_{error}) in this ANOVA is nothing but the variability within the groups (SS_w). The F-statistic is estimated by the ratio between the mean sum of squares for between-groups (MS_b) and within-groups (MS_w) with the relevant degrees of freedom.

3.3.3 Multivariate test to compute F value

- Pillai's trace is a method to calculate the F value. The value of this statistic ranges between 0 to 1. The growing values of this statistic shows effects that add more value to the model
- Hotelling's trace is a multivariate test calculated by the sum of diagonal values of a test matrix where the diagonal elements are the eigen values. The value of this test is always positive. Usually the value of Hotelling's trace is greater when compared to Pillai's trace. When the eigen values are small it becomes equal to Pillai's trace.
- The other Multivariate test to compute F value is the Wilks' Lambda. The value of this statistic ranges between 0 to 1. It measures the effect of level of each independent variable that contributes to the model. Here the value 0 refers to total discrimination and 1 refers to no discrimination.
- Roy's largest root is one more test calculated from the test matrix. It is the eigen value of test matrix with the maximum value. Either it is less or sometimes equal to Hotelling's trace. The value of this test is always positive.

3.3.4 Partial eta-squared is a measure of effect size. It determines the amount of effect of independent variable on the dependent variable. It tells about the magnitude of the effect.

It is calculated as the ratio of the sum of squares of the effect to the sum of squares of error term associated with that effect in an ANOVA.

It is denoted as $\eta_{partial}^2$

$$\eta_{partial}^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

Where:

SS_{effect} = the sum of squares of the effect

SS_{error} = the sum of squares of error term associated with that effect

3.3.5 Sphericity test

One of the assumptions of ANOVA in GLM repeated measures is the sphericity test, sphericity is the homogeneity of variance of differences between all combinations of related time points. The SPSS software tests for sphericity using Mauchly's Test for sphericity. According to the rule, the sphericity is assumed if the Sig. > 0.05.

3.4 Probability distribution

This study was undertaken to know the pattern of distribution of the number of infected pods of cocoa. The data of infected pods was collected for the 100 cocoa plants from the Cocoa Research Centre, Vellanikkara, Thrissur, Kerala on a monthly basis from the year 2003 to 2017. The probability distribution fitted was the geometric distribution for the number of infected pods. The distribution for the data was fitted by using the statistical software "EasyFit 5.5".

The distribution was fitted to the data by considering the frequency of infected pods. A frequency distribution is the frequency of outcomes obtained in a tabulated form. A probability distribution defines the possibility of each outcome of a random experiment or event. Probability distribution is a function that describes the possible values of a random variable with their associated probabilities. A set of random variables with their corresponding probabilities will form a probability distribution.

3.4.1 Probability Distribution

Based on the measurement of values two types of probability distribution are defined

1. Discrete distribution of probability
2. Continuous distribution of probability

In a Discrete probability distribution, random variable will take only a discrete and finite set of values. The probability distribution function formed by discrete random variables is called as a probability mass function. Uniform distribution, Bernoulli distribution, Binomial distribution, negative binomial distribution, Geometric distribution, Hypergeometric distribution, Multinomial distribution etc. come under this category.

3.4.2 Negative Binomial Distribution

The negative binomial distribution can be used to improve the fit of a Poisson model. The Poisson distribution is applied in rare and random events. The number of events in a given time will follow a Poisson distribution if the events are independent and occur at random. When these assumptions don't hold, a negative binomial will give a compatible fit to the data because it have an extra parameter and the Poisson distribution is a limiting case of negative binomial distribution. The bacterial clustering, insect death, number of infected pods etc. tends to negative binomial distribution.

A negative binomial distribution (or Pascal) is a discrete distribution which is comprised of sequence of independent trials. The experiment consists of x repeated trials which are independent with two feasible outcomes either success or failure. Let p be the probability of success and $(1 - p)$ be the probability of failure. The experiment continues until r successes are observed, where r is specified in advance. The outcome of one event does not affect the outcome of other event since the trials are independent.

The negative binomial distribution comprising of n Bernoulli trials.

- i) all the trials are independent
- ii) the trials remain constant with probability of success 'p'

Let us define a probability distribution function $f(x; r, p)$ comprised of $x+r$ trials, where x is the number of failures preceding the r th success and last trial would be a success with probability p . So, we get $r-1$ success from the remaining $(x+r-1)$ trials and its probability is obtained by

$$\binom{x+r-1}{r-1} p^{r-1} q^x$$

Therefore, by compound probability theorem, the above equation is written by the product of probability p , i.e.,

$$\binom{x+r-1}{r-1} p^{r-1} q^x p = \binom{x+r-1}{r-1} p^r q^x$$

Thus, the probability mass function of a negative binomial distribution with random variable X is given by

$$p(x) = P(X=x) = \binom{x+r-1}{r-1} p^r q^x; x = 0, 1, 2, \dots$$

$$= 0, \text{ otherwise}$$

Also

$$\sum_{x=0}^{\infty} p(x) = p^r \sum_{x=0}^{\infty} \binom{-r}{x} (-q)^x = p^r * (1 - q)^{-r} = 1$$

Therefore $p(x)$ represents the probability function and the discrete variable which follows this probability function is called the negative binomial variate.

3.4.3 Geometric distribution

The negative binomial distribution reduces to geometric distribution to get a single success from x repeated trials. In negative binomial distribution when we consider the number success (r) is equal to 1 it leads to geometric distribution.

If we take $r = 1$, we have

$$p(x) = q^x p ; x = 0, 1, 2, \dots$$

Which is the probability mass function of geometric distribution.

Hence the distribution of negative binomial is regarded as the generalization of geometric distribution

Goodness of fit test

The software “Easy Fit 5.5” calculates the Goodness of fit (GOF) statistics used to identify the distribution that best fits to the data. Here in this software it allows to display the ranks to some distribution in a tabulated form. The Goodness of fit test is used to check the agreement of sample data with theoretical distribution. These test provides the best distribution that fits well to the the original data.

The Goodness of fit tests adopted in Easy Fit software are:

- Kolmogorov-Smirnov
- Anderson-Darling
- Chi-Square

3.4.4 Kolmogorov-Smirnov test

Kolmogorov-Smirnov test is a non-parametric test which measures the largest distance between the empirical distribution function $F_{data}(x)$ and the theoretical function $F_0(x)$, measured in vertical direction.

The test statistic is denoted by D

Where, $D = \sup |F_0(x) - F_{data}(x)|$

$F_0(x)$ = the hypothesized distribution

$F_{data}(x)$ = the empirical distribution function of observed data

Test of Hypothesis

- H_0 : the null hypothesis says that the distribution fits well to the actual data
- H_1 : the alternative hypothesis that the specified distribution does not fit well to the actual data

The decision to reject or accept the null hypothesis is made based on the test statistic D. If the test statistic D is greater than the critical value at certain level of significance α (0.01, 0.05 etc.) then the null hypothesis is rejected. If the test statistic D is less than the critical value at certain level of significance α then the null hypothesis is accepted.

P- value

The P-value is another criterion to reject or accept the null hypothesis with respect to some fixed values of α . The P-value is tested against the α value of significance. The null hypothesis is accepted for all the α values less than the P-value. For example, the null hypothesis is rejected if the significance levels α (0.05 and 0.1) is greater than the $P=0.025$ value. Similarly, the null hypothesis will be accepted if the significance levels α (i.e. 0.01 and 0.02) is less than the $P=0.025$ value.

3.4.5 Anderson-Darling test

The Anderson-Darling test is another alternative test to identify the distribution that appropriately fits to the data. In this test the critical value does not rely on the distribution to be tested, it is free from the distribution. It makes use of some specified distribution to estimate the critical value.

The test statistic of A-D test is denoted by A^2

Test of Hypothesis

- H_0 : the null hypothesis says that the distribution fits well to the actual data
- H_1 : the alternative hypothesis that the specified distribution does not fits well to the actual data

when the value of test statistic, A^2 becomes larger than the critical value, then hypothesis of the specified distribution is rejected at some level of significance (α).

3.5 Impact of climatic variables on cocoa yield

An attempt to understand the impact of parameters of weather on cocoa yield recorded in terms of number of cocoa pods have been made. The monthly cocoa pod yield data have been collected for 100 trees for the years 2003 to 2017 from Cocoa Research Centre, Kerala Agricultural university, Vellanikkara. The Weather data was obtained for the period 2002 to 2017 from the Department of Agricultural Meteorology, College of Horticulture, Kerala Agricultural University, Vellanikkara.

The weather data comprised of daily records on maximum temperature ($^{\circ}\text{C}$), minimum temperature ($^{\circ}\text{C}$) Relative humidity (%) (RH1 and RH2), sunshine hours (Hrs), wind speed (Km/hr), rainfall(mm) and number of rainy days.

3.5.1 Correlation

Correlation determines the degree of agreement between the variables. It was introduced by Karl Pearson. The Pearson correlation coefficient is denoted by 'r'. It explains the association between the two variables which estimates the magnitude and direction of variables. The value of 'r' ranges between -1 to +1. If the value of 'r' is equal 1 which means that there is a perfect positive correlation. It means that when value of one variable X increases there is also increase in the other variable Y. When the value of 'r' is negative then there is an inverse relationship between the two variables. Here when the value of one variable X increases then there is decrease in the value of the other variable Y.

3.5.2 Linear Regression

Linear regression is a mathematical expression which explains the functional relationship between the dependent variable and independent variables which are linearly associated. The variable which is dependent is also called as response variable and the independent variable is called as predictor variable. It estimates regression coefficients of the linear equation that explains the change in the value of dependent variable corresponding to a unit change in the independent variable.

The Data considerations for Linear Regression

The dependent and independent variables should be quantitative. Categorical variables are to be recorded as binary (dummy) variables.

Assumptions:

- i) The distribution of the response variable should obey normality.
- ii) There should be linearity between the response variable and predictor variable.
- iii) The observed variables should be independent.
- iv) The variance of residual is same for any value of independent variable

Simple linear regression model is a mathematical model in which the model contains one dependent variable Y which depends on one independent variable X.

- The variable Y is also known as response or endogenous variable.
- The variable X is also known as predictor or exogenous variable

If the regression model is built with more than one independent variable then it is named as multiple linear regression model. Here the dependent variable depends on more than one predictor variable.

3.5.3 Simple linear regression

The mathematical expression of simple linear regression model is

$$y = \beta_0 + \beta_1 x + \varepsilon$$

The model contains one dependent variable and one independent variable. From the above regression model, the variable y refers to the dependent variable and x refers to the independent variable. The expression β_0 is the y intercept and β_1 is the regression coefficient or slope of the regression equation which measures the amount of change in the dependent variable for a unit change in the independent variable. Where ε is the residual effect.

3.5.4 Multiple linear regression

Multiple linear regression (MLR) is a regression equation which contains one dependent variable which is influenced by more than one independent variable. The model is expressed as below

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon$$

From the model it can be seen that there are several independent variables along with their respective regression coefficients including the intercept and error term (ε). The parameters of the regression equation are calculated by the method of ordinary least-squares (OLS). The above Regression model is given by

Y_i = i^{th} dependent variable, where $i = 1, 2, \dots, n$

X_i = i^{th} independent variable

β_0 = constant term (y-intercept)

β_p = regression coefficients of the respective independent variable

ε = error term (residual)

3.5.5 Stepwise Regression

Stepwise regression is one step forward than the multiple regressions. It is built by either adding (or) removing the variables in a step-by-step manner. The addition or deletion of variables is mainly based on the calculation of F-tests (or) T-tests for each of the variables. In this process the variables which are more significant are retained in the model and the variables which are least significant are removed from the model. In this regression technique all the variables are not taken into account simultaneously but it provides a parsimonious model.

Steps in performing the stepwise regression are:

Doing Backward step: This method is followed when there are sufficient number of variables. In this method some of the variables are removed from the available variables in each step as the model get progress. The removing of the variable is mainly done based on the calculation of F-value. The variable with lowest F-value is removed from the model at each step. This process is done in two steps,

- At first a test statistic is calculated for each variable in the model and then it is squared to obtain the F-values for the variables
- The variable with least F-value is removed from the model

Doing Forward step: this step is quite contrast with the Backward step. It deals with the adding of variables into the model. This step is carried out when there are large number variables available. The process of adding of variables is done by calculating the test statistic for each of the variables which is not present in the model. Then the variable with more F-value is added into the model.

3.5.6 Durbin Watson test

In the regression analysis it is assumed that the residuals are independent. In order to check this, there is a test called the Durbin-Watson (DW) test. It measures the autocorrelation between the error terms over a progressive period of time. This test is derived by James Durbin and Geoffrey Watson. This test is preferred

only for time series data and it cannot be used for cross-sectional data since it depends on the sequence of data points.

The hypothesis of Durbin-Watson test is:

H_0 = no first order autocorrelation

H_1 = first order autocorrelation exists

The test statistic is calculated by the following formula

$$DW = \frac{\sum_{t=2}^T (e_t - e_{t-1})^2}{\sum_{t=1}^T e_t^2}$$

The D-W test statistic value ranges between 0 to 4.

- If the test statistic is equal to 2 then there is no autocorrelation.
- If the test statistic is above 0 and below 2 then there is a positive autocorrelation.
- If the test statistic above 2 and below 4 then there is a negative autocorrelation

The standard rule that usually followed is that, if the value of statistic lies between 1.5 and 2.5 then we declare that it has no autocorrelation.

3.6 Structural Equation Modelling

Sample and study regions:

The aim of this study was to conduct a path analysis using the structural equation model (SEM) and to develop a model based on the critical factors of cocoa production that influence the income of cocoa famers of Kerala state. The primary data for this study was collected by directly interacting farmers engaged in cocoa cultivation in Kerala. A pre-tested structured questionnaire was used to collect data on demographic details of the farmers, cocoa cultivation and management practices,

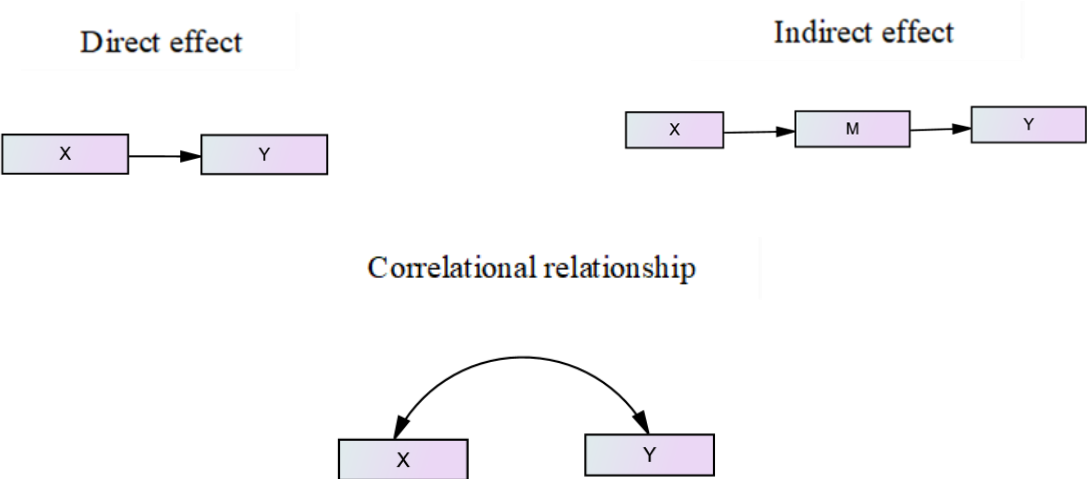
production, constraints faced etc. The respondents involved in this study were 100 small holder cocoa farmers who have contacts with the cocoa project “Mondelez International Ltd, Ernakulam”. The Veliyamattom Panchayat of Idukky district and Iritty Panchayat of Kannur district were two regions of Kerala where the selected farmers were engaged in cocoa cultivation.

Methodology

SEM is a structural model that describes the interrelations between the dependent and independent variables. Where the variable can be discrete (or) continuous. SEM is a Very complex method which has got a multidimensional structure. It allows calculation of modification indices which help researcher to develop best fit model to the data.

Path analysis is a special case of SEM. Path analysis is a statistical technique used to examine the strength of direct and indirect relationships among variables. It is an extension of multiple regression. SEM is an advanced regression analysis that analyses more than two causal models identified by researchers. It examines how the independent variables are statistically related to a dependent variable.

3.6.1 SEM model components



3.6.2 Different steps in conducting SEM

- 1) Collecting the research articles and reviews to assist in building a model
- 2) Defining a model
- 3) Access the model identification
- 4) Determine the measures for all the variables chosen in the model
- 5) Necessary data collection
- 6) Constructing the path model
- 7) Assigning the model fit

Goodness of fit indices – GFI, AGFI, NFI, CFI should be > 0.9 (Hair *et al.* 2010)

Badness of fit indices – RMSEA < 0.08 (Hair *et al.* 2010)

Model comparison – AIC, BIC and CFI

The SEM model is developed using the “Amos” software package. SEM is a diagrammatic model built by several observed variables and latent variables. In this path diagram the observed variables are included in a rectangles or squares. The unobserved variables are included in circles or ellipses. The direct effect of one variable on the other variable is shown by a single headed arrow. The covariance between the two independent variables is connected by the curve with two headed arrows.

In this study the results are drawn by applying the structural equation modelling to examine the interdependence of factors related to demographic details, cocoa cultivation and production data provided by the selected cocoa farmers of Kerala state.

3.6.3 Model fit summary of SEM

It is important to check whether the obtained SEM model adequately fits the data. The accuracy of the model is tested based on the goodness of fit test. The root mean squared error of approximation (RMSEA), comparative fit index (CFI) and Tucker Lewis index (TLI) are the statistical measures that tests the Goodness of fit of the model. The value of RMSEA alters when there is a change in the degrees of freedom and sample size. For the high value of degrees of freedom with large sample size the RMSEA value gets reduces. When the RMSEA is zero then the model is the best fit and if the value is ≤ 0.08 the model is defined to be a good fit. There is some other goodness of fit indices which are CFI and TLI values that checks the fitness of the base model compared with the hypothesized model. The suggested value of CFI should be > 0.90 which lies between the 0 to 1.

Model specification

In SEM there are two kinds of variables. One is the Observed or manifest variable that can be measured and the other is latent or unobserved variable. The SEM model is represented diagrammatically where the variables which are observed are included in the squares or rectangles and the latent variables are included in the circles or ellipses.

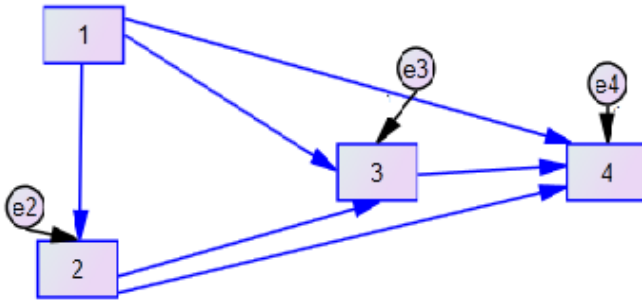
Since SEM contains the dependent and independent variables, the relationship among them is interpreted by regression equations as,

$$y_i = \alpha_i + \mathbf{B}\mathbf{X} + \varepsilon_i$$

Where y_i indicates the i^{th} dependent variable, α_i is the i^{th} intercept in a regression equation, \mathbf{X} denotes the vector of independent variables, whereas \mathbf{B} refers to the regression coefficients vector of corresponding variables in \mathbf{X} and ε_i represents random error associated with i^{th} dependent variable.

3.6.4 Calculating Path Coefficients

In Path analysis we work with the path coefficients showing the direct effect of one variable on the other variable. There is a standard form of denoting the variable i.e. in terms of X scores. Let us consider a path diagram as follows



Here we have

$$X_1 = \varepsilon_1 \quad (1)$$

$$X_2 = p_{21}X_1 + \varepsilon_2 \quad (2)$$

$$X_3 = p_{31}X_1 + p_{32}X_2 + \varepsilon_3 \quad (3)$$

$$X_4 = p_{41}X_1 + p_{42}X_2 + p_{43}X_3 + \varepsilon_4 \quad (4)$$

From the model it can be seen that the variable X_1 is not explained by any other variable except the external cause which is unobserved. in the model.

The second variable X_2 which is directly affected by the first variable X_1 and some external causes (or) error ε_2 . The above equations are in accordance with the path diagram. In each equation the variable X is associated only with the effect not by the indirect effect. For example, the variable X_3 is not associated by the indirect effect P_{21} .

Observed correlations are used to calculate the path coefficients

$$r_{12} = \frac{1}{N} \sum X_1 X_2$$

$$r_{12} = \frac{1}{N} \sum X_1 (P_{21} X_1 + \varepsilon_2)$$

$$r_{12} = P_{21} \frac{\sum X_1 X_2}{N} + \frac{\sum X_1 \varepsilon_2}{N}$$

= path coefficient x the variance of $X_1 (=1)$ + correlation between X_1 and $\varepsilon_2 (=0$, because of one of the assumptions in path analysis)

$$= P_{21}$$

Therefore $r_{12} = P_{21}$.

Thus, the effect of independent variable on the dependent variable in a path diagram is connected by a single headed arrow and this effect is measured in terms of path coefficient. When the dependent variable is associated with a single independent variable then the path coefficient will be equal to the correlation coefficients.

When we consider the variable X_3 , it is influenced by two variables X_1 and X_2 . The paths between the three variables X_1 , X_2 and X_3 is mainly computed by the correlations between them. But the error terms due to some external causes which are not correlated can be certainly left out.

$$r_{13} = \frac{1}{N} \sum X_1 X_3$$

$$r_{13} = \frac{1}{N} \sum X_1 (P_{31} X_1 + P_{32} X_2)$$

$$r_{13} = P_{31} \frac{\sum X_1^2}{N} + P_{32} \frac{\sum X_1 X_2}{N}$$

$$r_{13} = P_{31} + P_{32} r_{12}$$

Now we have r_{12} and r_{13} but P_{31} and P_{32} are not known. In this case r_{23} can be used to estimate the other path coefficients by producing a system of simultaneous equations.

$$r_{23} = \frac{1}{N} \sum X_2 X_3$$

$$r_{23} = \frac{1}{N} \sum X_2 (P_{31} X_1 + P_{32} X_2)$$

$$r_{23} = P_{31} \frac{\sum X_1 X_2}{N} + P_{32} \frac{\sum X_2^2}{N}$$

$$r_{23} = P_{31} r_{12} + P_{32}$$

Hence, it produces two equations

$$r_{13} - P_{32} r_{12} = P_{31}$$

$$r_{23} = P_{31} r_{12} + P_{32}$$

The first equation is subtracted from $P_{32}r_{12}$ from both sides

$$r_{13} = P_{31} + P_{32} r_{12}$$

$$r_{23} = (r_{13} - P_{32} r_{12}) r_{12} + P_{32}$$

$$P_{32} = \frac{r_{23} - r_{13}r_{12}}{1 - r_{12}^2}$$

This formula shows that the path coefficients are same as that of the standard regression weights which are also called as beta weights. The coefficients are calculated effectively with the correlation between the three variables

Hence,

$$r_{13} = \beta_{31.2} + \beta_{32.1} r_{12}$$

$$r_{23} = \beta_{31.2} r_{12} + \beta_{32.1}$$

We can notice that the path coefficients and beta weights are similar

It was already estimated that the first path coefficient is equal to the correlation coefficient which is also equal to beta weights since the variable X_2 had a single path from X_1 so it requires one regression equation to find the path coefficient. But now the variable X_4 is connected by three paths from three variables (X_1, X_2, X_3). So now it requires three simultaneous equations to estimate the path coefficients which are not known.

$$r_{14} = \frac{1}{N} \sum X_1(P_{41}X_1 + P_{42}X_2 + P_{43}X_3)$$

$$r_{14} = P_{41} \frac{\sum X_1^2}{N} + P_{42} \frac{\sum X_1X_2}{N} + P_{43} \frac{\sum X_1X_3}{N}$$

The remaining coefficients are reduced to

$$r_{24} = P_{41} r_{12} + P_{42} + P_{43} r_{23}$$

$$r_{34} = P_{41} r_{13} + P_{42} r_{23} + P_{43}$$

The path coefficients are resulted from several multiple regression equation. For example, the variable X_4 is influenced by three independent variables X_1, X_2 and X_3 treated as independent variables. In this case we have three simultaneous regression equations to estimate the 3 path coefficients P_{41}, P_{42} and p_{43} . Again, when we consider the variable X_3 as the dependent variable influenced by X_2 and X_1 , we expect two multiple regression equation to estimate the path coefficients P_{31} and P_{32} . Finally, for the variable X_1 it requires a simple regression equation to estimate a single path

coefficient P_{21} which is equal to the correlation coefficient r_{12} . Then the path coefficients are estimated by the system of simultaneous regression equations.

In general, the correlation coefficient can be reduced into 4 parts:

1. Direct Effect (DE) of independent variable X on dependent variable Y
2. Indirect Effect (IE) of one variable on the other variable through intermediate variable
3. Unanalysed (U) effect from correlated exogenous variables
4. Spurious (S) effect caused from a third variable

In order to better understand all kinds of effects in a path analysis, the correlations are decomposed.

We have $r_{12} = P_{21}$. Since r_{12} is due to a single path, it indicates a direct effect.

$$\begin{aligned} r_{13} &= P_{31} + P_{32}P_{21} \\ &= \text{DE} + \text{IE (direct effect + indirect effect)} \end{aligned}$$

Significance and goodness of fit: OLS and maximum likelihood methods are used to predict the path coefficient. Statistical software such as AMOS, M-Plus, SAS, LISREL, etc. are software that calculates the path coefficient and goodness of fit statistics automatically.

In SEM, the shared variance among exogenous variables can be accounted through by drawing covariances connecting those variables before estimating various effects in the model.

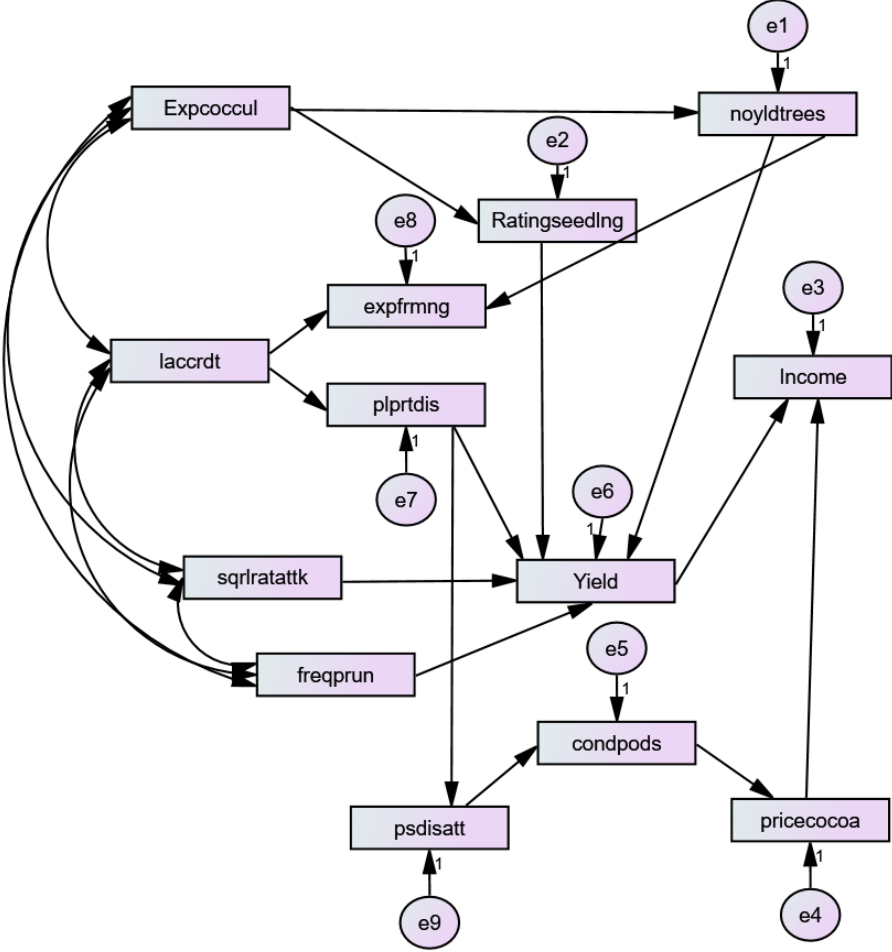


Fig. 3.2 Base model of cocoa production

Based on the critical factors of cocoa production that influence the income of cocoa famers of Kerala state a SEM model was developed and illustrated in Fig.3.2. The figure depicts the opinion of cocoa farmers regarding the constraints they are facing and other important factors that are influencing the income of cocoa famers. It can be seen that all the variables are represented by rectangles since they are observed variables. But the errors are unobserved or latent variables and are drawn in circles. The model shows the different pathways where the independent variables leads to dependent variables.

Cocoa yield is an important variable where most of the variable acts as independent variables to it. Farmer's income is the ultimate factor where the pathways get end up. The whole model speaks about the pathways that runs from the cultivation and management practices during cocoa production, constraints faced during the production and ends up with the economic condition of cocoa farmers.

The model drawn in Fig.3.2 can be illustrated by 8 structural equations. Equations (1), (2), (7) describe the factors contributing to the yield of cocoa. Equations (3), (4), (6) explain the factors contributing to the income of cocoa farmers. Equation (5) describes the factors affecting the quality of cocoa pods. Equation (8) deals with expenditure on cocoa production.

$$\text{Ratingseedling} = \alpha_1 + \beta_1 \text{Expcoccul} + \varepsilon_1 \quad (1)$$

$$\text{noyldtrees} = \alpha_2 + \beta_2 \text{Expcoccul} + \varepsilon_2 \quad (2)$$

$$\text{Income} = \alpha_3 + \beta_{31} \text{yield} + \beta_{32} \text{pricecocoa} + \varepsilon_3 \quad (3)$$

$$\text{pricecocoa} = \alpha_4 + \beta_4 \text{condpods} + \varepsilon_4 \quad (4)$$

$$\text{condpods} = \alpha_5 + \beta_5 \text{psdisatt} + \varepsilon_5 \quad (5)$$

$$\text{yield} = \alpha_6 + \beta_{16} \text{noyldtrees} + \beta_{26} \text{Ratingseedling} + \beta_{36} \text{plprtdis} + \beta_{46} \text{sqrlratattk} \quad (6)$$

$$+ \beta_{56} \text{freqprun} + \varepsilon_6$$

$$\text{plprtdis} = \alpha_7 + \beta_7 \text{laccrdt} + \varepsilon_7 \quad (7)$$

$$\text{expfrmng} = \alpha_8 + \beta_8 \text{laccrdt} + \varepsilon_8 \quad (8)$$

In the above structural equations,

α_i refers to the intercept corresponding to the i^{th} dependent variable

β_{ij} refers to the path coefficient corresponding to the effects of j^{th} independent variable on the i^{th} dependent variable

ε_i refers to the error associated in predicting the i^{th} dependent variable

The model contains 13 measured variables consisting of 9 dependent and 4 independent variables and some intervening variables. The dependent variable in an equation can also act as independent variable for some other equation. For example, in the equation (5) the condpods (condition of pods) acts as dependent variable. Whereas, in equation (4) it acts as independent variable. The noyldtrees (number of yielding trees) in equation (2) acts as dependent variable. Whereas, in equation (6) it becomes the independent variable. The SEM model allows to determine the effect of sqrlratattk (squirrel, rat, civet etc. attack) and freqprun (frequency of pruning) on the income of cocoa farmers through the variable yield (cocoa yield) which acts as intermediate variable in the model. Thus, SEM helps us to determine the simultaneous effect of several interrelated variables on the income of cocoa farmers which is the economic status of the farmers.

Table: 3.1 Description of variables in the SEM model of cocoa production

Variables	Description	Measures
Noyldtrees	Number of yielding trees	Number
Income	Income of cocoa farmers	In Rupees
pricecocoa	Price of cocoa	In Rupees
yield	Number of cocoa pods	Kilogram
expfrmng	Expenditure on cocoa farming	In Rupees
condpods	Condition of harvested cocoa pods	1= Partially ripe, 2= Mix of partially and fully ripe, 3= Fully ripe
plprtdis	Plant protection & disease management	1=Never, 2=Occasionally, 3=Regularly
psdisatt	Pest and disease attack	1= mild, 2= moderate, 3 = severe
sqrлатatk	Attack from squirrel, rat, civet etc	1= mild, 2= moderate, 3 = severe
freqprun	Frequency of pruning on cocoa trees per year	1= once, 2= twice, 3= thrice
laccrdt	Lack of access to credit	2= yes, 1= no
Ratingseedling	Rating of quality of cocoa seedlings	1=ordinary, 2= good, 3=very good
Expcoccul	Experience in cocoa cultivation	No. of years

3.7 Probit regression model

This study was conducted to assess the factors influencing farmer's decision to use plant protection measures using Probit regression model. The cross-sectional data was collected from 100 cocoa farmers of Kerala.

In statistics, a Probit regression model is a regression equation where the response variable is categorical or dichotomous which can take two possible values, for example presence/absence or success/failure etc.

This model is used to test the probability that an entity with a specific characteristic will belong to any one of the categories. This model is also called as binary response model if the dependent_variable is classified into two classes. Here the endogenous variable Y is dichotomous and takes binary outcomes which is denoted by 1 and 0.

In this method the dependent variable is dichotomous and the Probit model is expressed in terms of a cumulative standard normal distribution function $\Phi(\cdot)$.

The Probit model assumes that the values 0 and 1 for the dependent variable Y_i are observed. To determine the value of Y_i there is a latent variable or unobserved continuous variable Y_i^* . This model determines the probability whether it belongs to category 0 (or) 1. Let Y_i be the dependent variable influenced by several independent variables X_i and the model is written as

$$\Pr(Y_i = 1/x_i) = F(\beta'x_i) = \Phi(\beta'x_i) \quad (1)$$

where \Pr is the probability, Y_i denotes the binary variable representing adoption of plant protection measures and Φ is the cumulative standard normal distribution function (CDF). β be the coefficient vector which is unknown.

An assumption is made that the unobserved variable Y_i^* can be in the form

$$Y_i^* = \beta_0 + \sum_{n=1}^N \beta_n x_{ni} + \varepsilon_i \quad (2)$$

and

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Where X_i represents a vector of independent variables, u_i represents the random error term, N is the total sample size, and β is the coefficient vector where the coefficients are calculated by maximum likelihood method and the total sample size is given by N .

In probit regression model the decision to check the most significant variable that influence the dependent variable is done by estimating the marginal effects for each of the independent variables. These marginal effects for each of the coefficients are considered as highly informative.

To estimate the marginal effect, we differentiate equation (1) with respect to x_i

$$\frac{dY_i}{dx_i} = \phi(\beta' x_i) \beta_i \quad (4)$$

Where ϕ represents the probability density function of the standard normal distribution.

The actual probit model is determined by:

$$y_i = \beta_0 + \sum_{n=1}^8 \beta_n x_{ni} + \varepsilon_i \quad (5)$$

Where Y_i = Decision to make use of plant protection measures (=1 if farmers adopt plant protection measures, 0 otherwise); X_1 = age; X_2 = education; X_3 = occupation; X_4 = family size; X_5 = landholding size; X_6 = experience in cocoa cultivation; X_7 = membership in organisations ; X_8 = frequency of extension contact.

3.8 Coefficient of concordance

This study was to identify the important factors which influence the cocoa production and ultimately the farmer's income. The Kendall's coefficient of concordance is the most preferred method to study the farmer's perception to identify the important factors using primary data collected from farmers with respect to cocoa cultivation and production. The respondents involved in this study were 100 small holder cocoa farmers of the Veliyamattom Panchayat of Idukky district and Iritty Panchayat of Kannur district of Kerala where the farmers were engaged in cocoa cultivation.

The Kendall's W is a non-parametric test for K related samples that measures the agreement among the group of farmers (judges) who rank the statements to identify the most important factors influencing cocoa production. The value of W ranges between 0 and 1. It is mathematically expressed as the ratio of variance of sum of ranks to the maximum variance of the ranks. The concept behind this technique is to look for the agreement among the respondents in ranking the factors. This was done by finding the sum of ranks for each of the factors and finding the variability for this sum of ranks.

Kendall's coefficient of concordance (W) uses the χ^2 statistic for testing the significance. If Kendall's W is 1, then there is a perfect agreement among the respondents and if the value comes to 0, then we say that there is no agreement in ranking the factors by the respondents. The value in between 0 and 1 gives the amount of degree of agreement among the respondents.

The Kendall's coefficient of concordance (W) is given by the relation

$$W = \frac{12 S}{K^2 (n^3 - n) - K^T}$$

where,

W denotes the Kendall's coefficient of concordance

k denotes number of respondents

n refers to the number of factors to be ranked

T refers to the correction factor for the tied ranks

S refers to the sum of squares statistic over the row sum of ranks (R_i).

The statistic sum of square (S) is defined by:

$$S = \sum_{i=1}^n [R_i - (\sum R_i / n)]^2$$

where R_i is the sums of rank for each row and $\sum R_i / n$ is the average value of sum of ranks.

The correction factor for tied ranks (T) is given by:

$$T = \sum_{k=1}^m (t_k^3 - t_k)$$

Where t_k is the number of tied ranks in m groups of ties

The chi-square statistic is used as a test of significance of the Kendall's coefficient of concordance in this method:

$$\chi^2 = k (n-1) W$$

Where n refers to the number of factors to be ranked, k denotes the number of judges and W is the index for Kendall's coefficient of concordance.

Based on the value of chi-square statistic the decision to reject (or) accept the null hypothesis is decided. The null hypothesis says that the rankings of the respondents are unrelated. The alternate hypothesis says that the rankings of the respondents are related.

The thumb rule is defined as, if the value of χ^2 statistic is greater than the critical value then the null hypothesis (the k rankings are unrelated) is rejected and the alternate hypothesis (there is agreement among the judges) is accepted.

3.9 Assessing the Yield gap

An attempt was done to assess the national yield gap of cocoa. Aneani and Frimpong (2013) defined the yield gap as the difference between the estimated national average yield and the potential yield (maximum yield) attained during the research trials in research stations. In this study the yield gap is the difference between the experimental yield potential and estimated national average yield of cocoa. Yield potential for a crop is estimated when the crop is grown under a favourable environment with sufficient nutrients and suitable moisture without attack of any pest and diseases (Gommès, 2006; Lobell *et al.*, 2009).

The experimental yield potential for cocoa was obtained from an on-station trial in which the crop was grown with adequate water and nutrients without any biotic stress and effectively controlling the pest and disease attacks. The national average yield of cocoa was collected from the yield data given by current socio-economic survey (Aneani *et al.*, 2007).



RESULTS AND DISCUSSION

CHAPTER 4

RESULTS AND DISCUSSION

Aiming at the objectives of the study titled “Multiphase analysis of cocoa production in Kerala”, the data generated in different aspects were subjected to statistical methodologies explained in chapter 3. In this chapter, the salient findings of the research are discussed in detail under different headings.

4.1 Trend analysis of area, production and productivity of cocoa in Kerala

Table 4.1: Descriptive statistics for the time series data of cocoa cultivation in Kerala from 1980 to 2017

Parameters	Minimum	Maximum	Mean	Std.Deviation	Variance	Skewness	Kurtosis	CV
Area (hectares)	6907	23506	12476.47	4096.25	16779268	1.072	0.828	32.83%
Production (tonnes)	1461	7507	5185.63	1356.98	1841418	-0.749	0.367	26.16%
Productivity (tonnes per hectare)	0.08	0.64	0.45	0.14	0.2	-1.193	1.274	31.11%

From Table 4.1 it can be observed that the maximum area of cocoa in Kerala was 23506 ha in the year 1980. Though the area was more, the production was less in the beginning years. The highest production was 7507 tonnes obtained in the year 2017. It indicated that there was an increasing trend in the production. Kerala was having highest productivity of cocoa when compared to other states and the highest productivity was 0.64 tonnes per hectare in the year 1994. The values of variance for the area, production and productivity emphasized that the data was widely scattered. Coefficient of variation which is a measure of consistency was 32.83% for area, 26.16% for production and 31.11% for productivity. The values of skewness and kurtosis showed that the given time series data of area, production and productivity of cocoa was non-normal in nature. Area under cocoa showed a positive skewness, the distribution was having a long right tale resulted from concentration of more number of values on the left side of the probability density curve and few high values pulling the mean towards the right to make the mean more than the mode and median. But in the case of production and

productivity, concentration of more number of values were towards the right half of the density curve and few smaller values pulling the mean towards the left making it smaller than median and mode.

The area under cocoa of Kerala for the years 1980 to 2017 was used for estimating trend. From Fig. 4.1, it can be inferred that the graph showed a quadratic trend. The quadratic regression model had significant coefficients of the time variables (β_1) and (β_2) with a high value of R^2 equal to 0.93. The area of cocoa in the beginning of the graph showed a decreasing trend up to 1994, and then it again showed an upward trend. The maximum area was found during the year 1980.

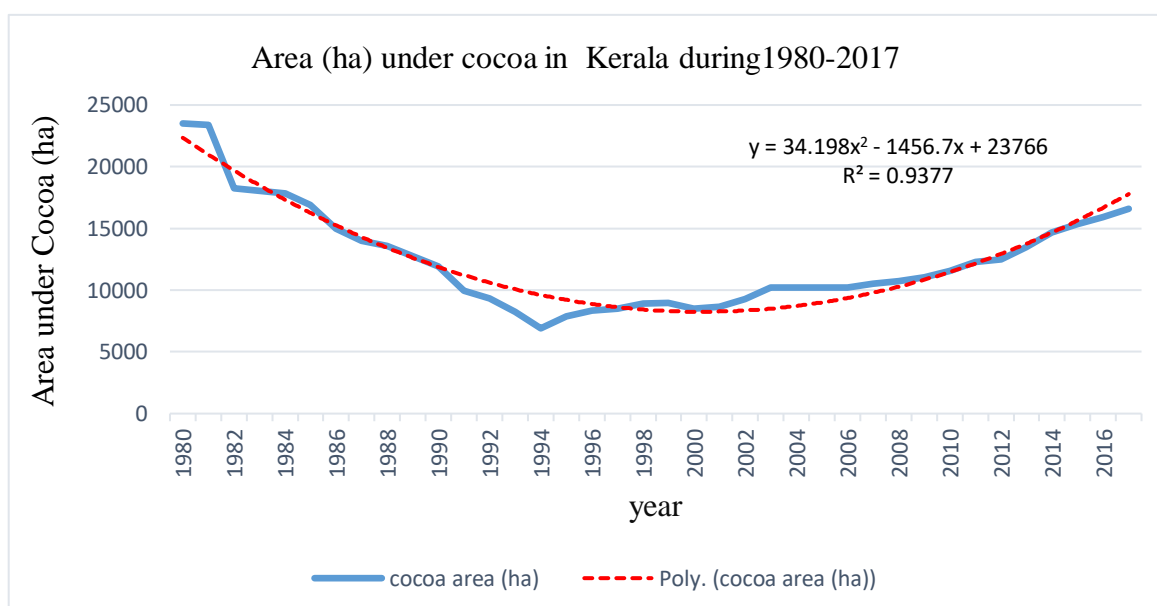


Fig. 4.1 Area under cocoa in Kerala during 1980-2017

Table 4.2: Quadratic regression model with estimated parameters for area under cocoa of Kerala from 1980-2017

Equation	Model summary					Parameter Estimates		
	R square	F	df1	df2	Sig	constant	β_1	β_2
Quadratic	0.938	263.294	2	35	.000	23766.02	-1456.70	34.198

From Fig. 4.2 it is observed that the graph on production of cocoa in Kerala for the years from 1980 to 2017 showed a stochastic linear trend. High variation could be seen in production. The production was just 3020 tonnes in the beginning of the year 1980 and it was decreased to 1461 tonnes in 1982. By the end of the year 2017, the production reached 7507 tonnes. From Fig.4.2, the linear regression model showed that the coefficient of the time variable β_1 was significant with R^2 value equal to 0.445. Therefore, the linear regression model was significant and the graph showed an increasing trend in the time series data on cocoa production in Kerala.

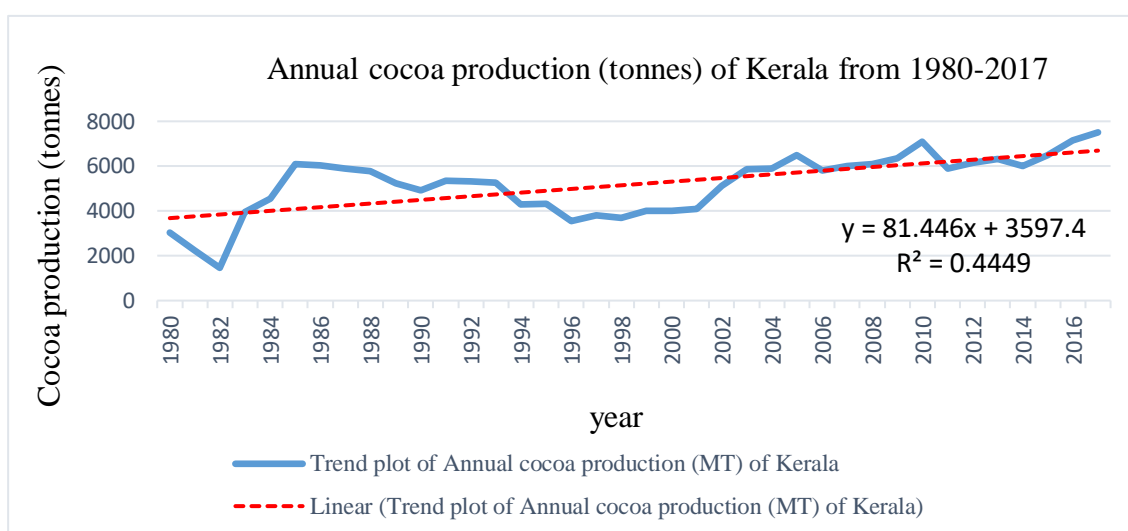


Fig 4.2: Annual cocoa production of Kerala from 1980 -2017

Table 4.3: Regression model to test the presence of linear trend in cocoa production of Kerala from 1980-2017

Equation	Model summary					Parameter Estimates	
	R square	F	df1	df2	Sig	constant	β_1
Linear	0.445	28.85	1	36	.000	3597.43	81.44

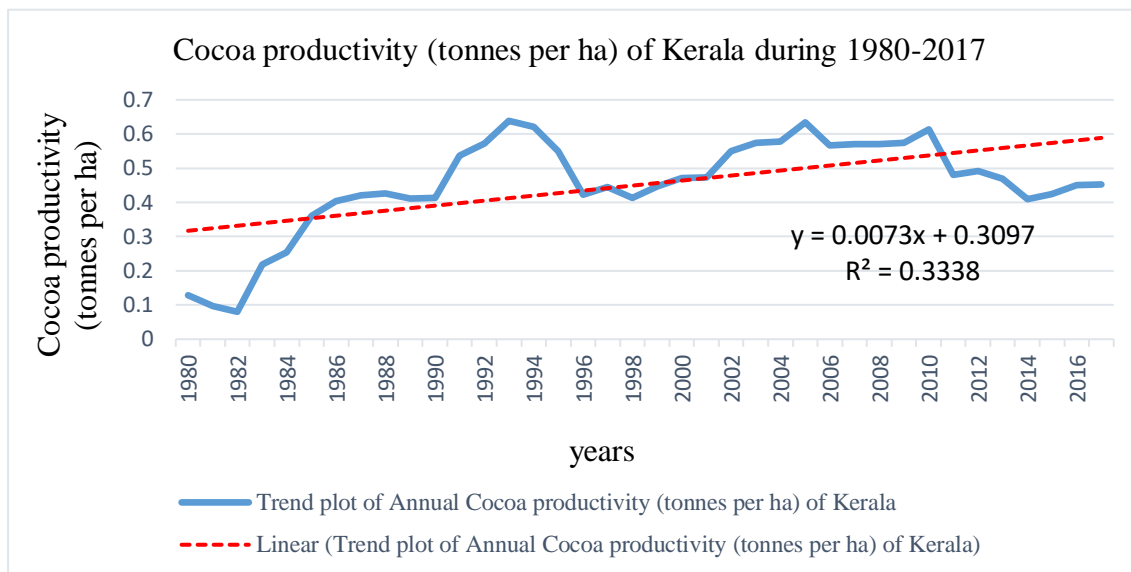


Fig 4.3 Cocoa productivity of Kerala during 1980 - 2017

Fig 4.3 showed that productivity was very low in the starting of the year in 1982, which was just 0.08 tonnes/ha. The productivity then increased to 0.63 tonnes/ha in 1993 and reached 0.61 tonnes/ha in 2010. By the end of 2017-18 it has dropped down to 0.45 tonnes/ha

Table 4.4: Linear regression model to test the presence of trend in cocoa productivity of Kerala from 1980-2017

Equation	Model summary					Parameter Estimates	
	R square	F	df1	df2	Sig	constant	β_1
Linear	0.334	18.03	1	36	.000	0.310	0.007

The Augmented Dickey Fuller test makes use of the hypotheses for the test as:

- The null hypothesis :There is a unit root, $\alpha = 1$ (i.e. the data needs to be differenced to make it stationary)
- The hypothesis: The time series is stationary (or trend-stationary)

If the p-value obtained is less than the significance level (say 0.05) then the null hypothesis is rejected, assuming the presence of unit root, i.e. $\alpha=1$, where α is the coefficient of first lag on y, thereby, inferring that the series is stationary.

Table 4.5: Unit root test for area, production and productivity of cocoa in Kerala from 1980 to 2017

Variables	Level	P- Value (0.05)
Cocoa Area	Intercept and trend	0.71
Cocoa Production	Intercept and trend	0.53
Cocoa Productivity	Intercept and trend	0.29

From Table 4.5, the results of unit root test indicated that the data for area, production and productivity were non-stationary since the p value was more than the significance level (say 0.05) and the null hypothesis was accepted, assuming that there was presence of unit root i.e. $\alpha =1$.

Time Series Forecasting

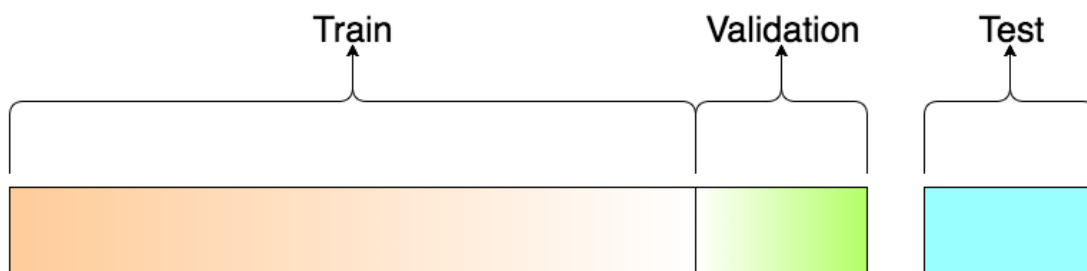
Time series forecasting is the use of statistical methods to predict future behaviour based on historical data. A time series is a sequence of data points recorded through time. Thus, when dealing with time series data, ‘order’ matters. Specifically, values in a time series express a dependency on time. Usually, time series data have two important properties.

- Data is measured sequentially and equally spaced in time.
- Each time unit has at most one data measurement.

In addition, when doing time series forecasting, we usually have two goals.

- First, we want to identify patterns that explain the behaviour of the time series.
- Second, we want to use these patterns to forecast (predict) new values.

In any time series forecasting, first the data is divided into training dataset and validation data set. The actual dataset that we use to train the model is called the training data set. The model sees and learns from this data. The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model parameters is called validation dataset. An unbiased evaluation of a final model fit can be done using test data set.



Before finalising the forecasting models for cocoa in Kerala for the period from 1980 to 2017, the time series data of area, production and productivity for the year 1980 to 2011 was used to train the models and the remaining 6 years data was used for validation of the models. The model building was done using SPSS20 statistical software. Based on the validation, suitable model based on the whole data was selected and used for forecasting the future values.

4.1.1 Forecasting of area under cocoa

The data for the years from 1980 to 2011 was taken as a training period with respect to area under cocoa for fitting a forecast model. The expert modeller in SPSS selected the Holt’s exponential smoothing model as the best model to forecast the area under cocoa in Kerala. After validation of the model, using the data for the period from 2012 to 2017 forecasts for five years were made using the Holt’s exponential smoothing model. The results are outlined in Table 4.7, which showed high value of R square along with other values of RMSE, MAPE, MAE and BIC.

The parameters of the exponential smoothing coefficients of the Holt’s model for area under cocoa are summarized in Table 4.6. The coefficients of the model were $\alpha = 0.711$

and $\gamma = 0.316$, where α was significant at 1% level and γ was significant at 10% level. Thus, the Holt's exponential smoothing model could be used to forecast the area under cocoa in Kerala for the next 5 years from 2018 to 2022.

Table 4.6: Parameters of the exponential smoothing coefficients of the Holt's exponential smoothing model for area under cocoa in Kerala for the period from 1980 to 2011

	Estimate	SE	T	Sig.
Alpha (Level)	0.711	0.169	4.218	0.00
Gamma (Trend)	0.316	0.165	1.91	0.06

Table 4.7: Accuracy measures of Holt's exponential smoothing model

Fit statistic	Holt's model
R ²	0.943
RMSE	1041.71
MAPE	5.813
MAE	687.36
BIC	14.11

Table 4.8 Validation of predicted area under cocoa in Kerala using Holt's model for 2012-2017

Year	2012	2013	2014	2015	2016	2017
Actual area (ha)	12483	13483	14650	15344	15894	16594
Forecasted area for validation (ha)	12550.6	12964.1	13377.6	13791.2	14204.7	14618.3

Table 4.8 gives a perusal of the validation of the model developed to predict area under cocoa by a comparison of the actual and forecasted values of area under cocoa in Kerala for the years 2012 to 2017.

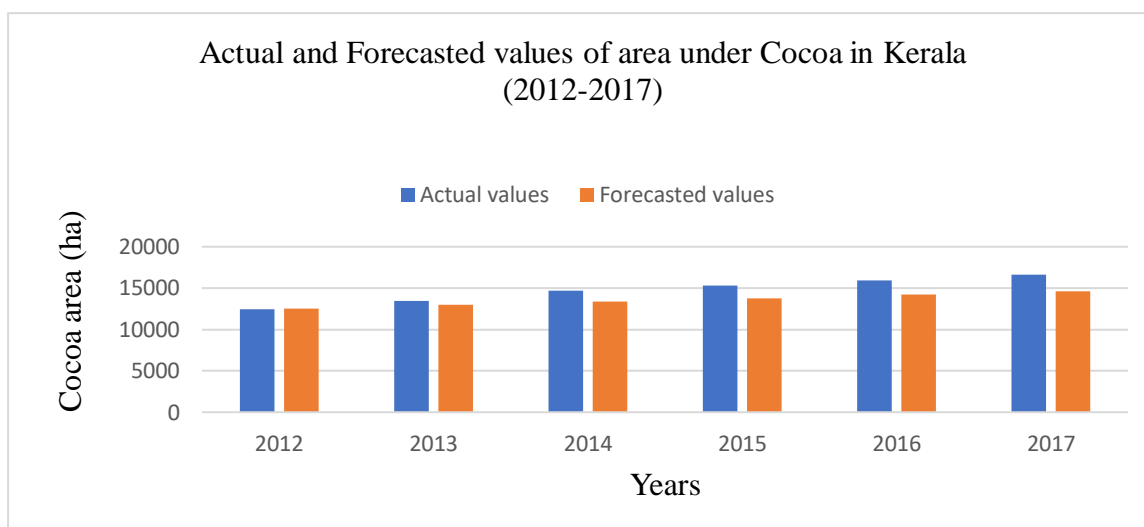


Fig 4.4 Validation of forecasted values of area under cocoa in Kerala for the years 2012 to 2017

The time series data of area under cocoa for the period from 1980-2017 was exposed to ARIMA models also. ARIMA (0,2,2) was identified to forecast the values for the next 5 years from 2018-2022. Using the accuracy and the efficiency measures of forecasting such as R^2 , MAE, MAPE, RMSE and BIC the best model was identified.

Holt's exponential smoothing was chosen as the best model which showed the highest R^2 and the smallest RMSE, MAPE, MAE and BIC than in the case of ARIMA model (Table 4.9).

Table 4.9: comparison of accuracy measures of ARIMA (0,2,2) and Holt's exponential model

Fit statistic	ARIMA (0,2,2)	Holt's Exp. smoothing
R^2	0.88	0.94
RMSE	1168.78	966.15
MAPE	5.56	5.21
MAE	660.4	627.97
BIC	14.52	13.93

The parameter estimates will be provided for the selected model. The parameters of the exponential smoothing coefficients of the Holt's model on area under cocoa are summarized in Table 4.10. The coefficients of the model were $\alpha = 0.72$ and $\gamma = 0.325$, in which α was significant at 1% level.

Table 4.10: Parameters of the Holt's exponential smoothing model to forecast area under cocoa in Kerala for the period from 1980 to 2017

	Estimate	SE	t	Sig
Alpha (Level)	0.72	0.155	4.658	0.00
Gamma (Trend)	0.325	0.154	2.10	0.42

Holt's exponential smoothing model

(Level of the series at time 't', coefficient $\alpha = 0.72$) $L_t = \alpha Y_t + (1-\alpha) [L_{t-1} + T_{t-1}]$

(Trend of the series at time 't', coefficient $\gamma = 0.325$) $T_t = \gamma [L_t - L_{t-1}] + (1-\gamma) T_{t-1}$

Forecast for k step ahead $F_{t+1} = L_t + k T_t$

Level equation $L_t = 0.72 Y_t + (1-0.72) [L_{t-1} + T_{t-1}]$

$$= 0.72 Y_t + 0.28 [L_{t-1} + T_{t-1}]$$

Trend equation $T_t = 0.325 [L_t - L_{t-1}] + (1- 0.325) T_{t-1}$

$$= 0.325 [L_t - L_{t-1}] + 0.675 T_{t-1}$$

Forecast equation $F_{t+1} = L_t + k T_t$

Where $k = 1, 2, 3, 4, 5$. (forecasting for the period from 2018 to 2022)

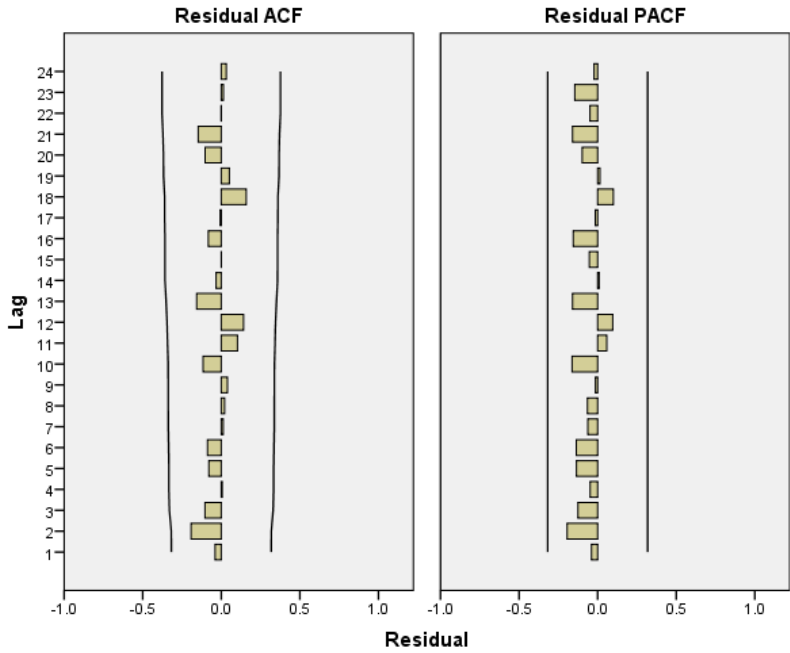


Fig 4.5: Residuals of ACF and PACF plots of Holt's exponential smoothing model

From Fig. 4.5, it can be seen that all the residuals in the ACF and PACF plots were within the confidence limits or the residuals were pretty close to white noise

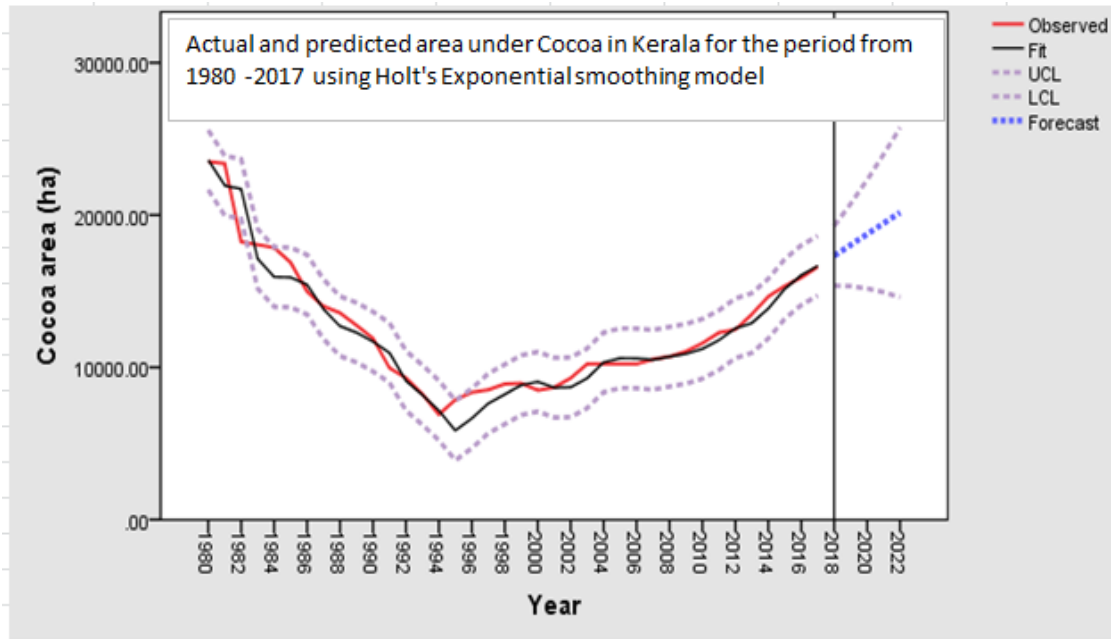


Fig 4.6: Actual and forecasted area under cocoa in Kerala

In Fig.4.6, the actual and forecasted values of the area under cocoa in Kerala are depicted. The closeness of the two curves showed the validity of the model developed to forecast the area under cocoa.

Forecasts for area under cocoa for the years from 2018 to 2022 are given in Table 4.11. An increasing trend for the area can be visualised i.e. from 17326.16 ha in 2018 to 20171.79 ha in 2022.

Table 4.11: Forecasted values of area under cocoa (ha) in Kerala for the period 2018-2022

Year	2018	2019	2020	2021	2022
Area (ha)	17326.16	18037.56	18748.97	19460.38	20171.79

4.1.2 Forecasting of cocoa production

The time series data of cocoa production for the period from 1980 to 2011 was used for training the model and validation of the model was done using the next 6 years data. In SPSS the expert modeller selected the Simple exponential smoothing model as the best model to forecast the production of cocoa for the years 2018 – 2022.

The parameters of the Simple exponential smoothing model for prediction of cocoa production in Kerala is summarized in Table 4.12. The coefficient in the model was observed as $\alpha = 1$, which was significant. Thus, the simple exponential smoothing model could be used to forecast the cocoa production in Kerala and forecasts have been made for 5 years from 2018 to 2022.

Table 4.12: Parameters of Simple exponential model of cocoa production for the years 1980 to 2011

	Parameter Estimate	SE	t	Probability(p)
Alpha (Level)	1	0.174	5.748	0.00

Table 4.13, shows statistical measures viz; R square, RMSE, MAPE, MAE and BIC associated with the Simple exponential smoothing model

Table 4.13: Accuracy measures of Simple exponential smoothing model

Fit statistic	Simple exponential model
R ²	0.680
RMSE	731.063
MAPE	11.673
MAE	493.151
BIC	13.29

Table 4.14: Validation of cocoa production for the period 2012-2017 using Simple exponential model

Year	2012	2013	2014	2015	2016	2017
Actual production (tonnes)	6136	6320	6000	6500	7150	7507
Forecasts of cocoa production (tonnes) for validation	6136	6136	6136	6136	6136	6136

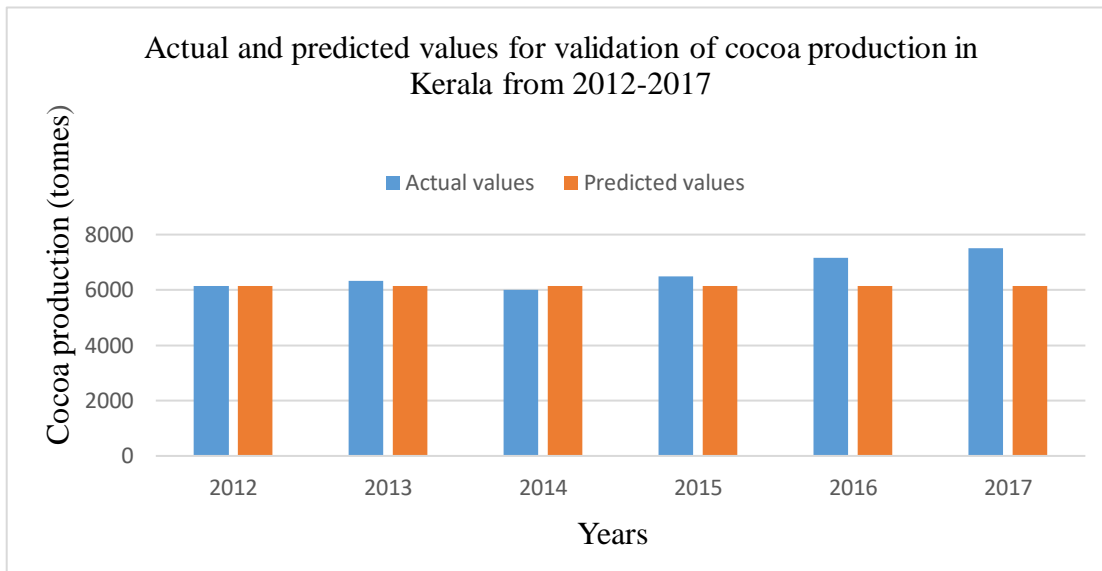


Fig 4.7: Validation of cocoa production for the period 2012-2017 using Simple exponential model

The time series data of cocoa production for the period 1980-2017 was exposed to ARIMA models also and it has resulted in identifying ARIMA (0,1,1) as the best for forecasting the values for the next years from 2018-2022. Both models were compared based on the accuracy measures of forecasting such as the R², MAE, MAPE, RMSE and BIC and the best model was determined.

The results obtained are entered in Table 4.15. ARIMA (0,1,1) model was obtained as the best model which had the highest R² and the smallest RMSE, MAPE, MAE and BIC to predict the cocoa production in Kerala.

Table 4.15: Comparison of accuracy measures of ARIMA (0,1,1) and Simple exponential model

Fit statistic	ARIMA (0,1,1)	Simple Exp. model
R ²	0.72	0.67
RMSE	717	741
MAPE	11.41	11.9
MAE	487	501
BIC	13.44	13.366

The parameter estimates of the selected model will be provided in Table 4.16.

Table 4.16: Parameters of the model ARIMA (0,1,1)

Variable	Parameters	Estimate	SE	t	probability
Cocoa Production	Constant	106.95	250.86	0.426	0.675
	Difference	1			
	MA Lag 1	-0.002	0.172	-0.012	0.991

ARIMA (0,1,1) model

$$Y_t = \mu + Y_{t-1} - \theta_1 \epsilon_{t-1}$$

$$= 106.95 + Y_{t-1} - (-0.002) \epsilon_{t-1}$$

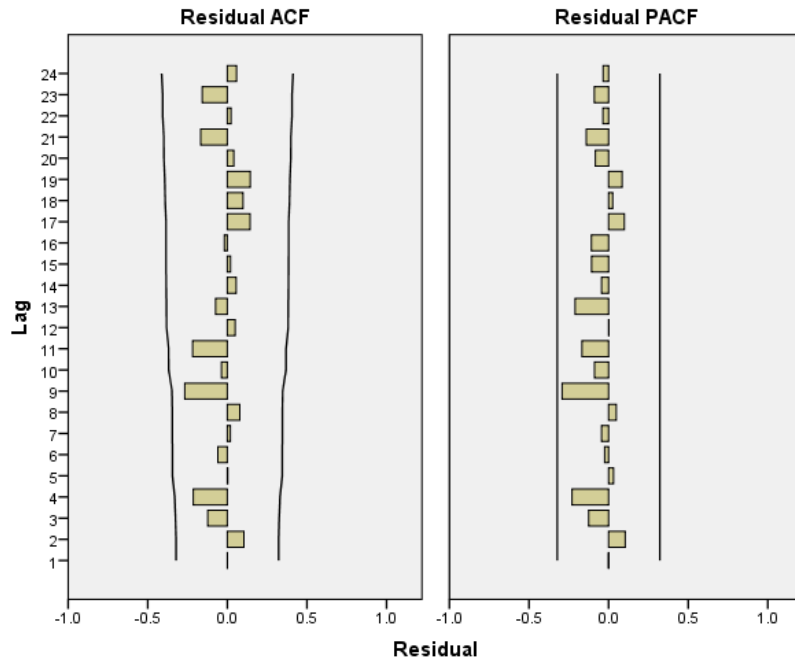


Fig 4.8: Residual plots of ACF and PACF of ARIMA (0,1,1) model

From Fig. 4.8, it can be seen that all the residuals in the ACF and PACF plots were within the confidence limits and the residuals were almost white noise

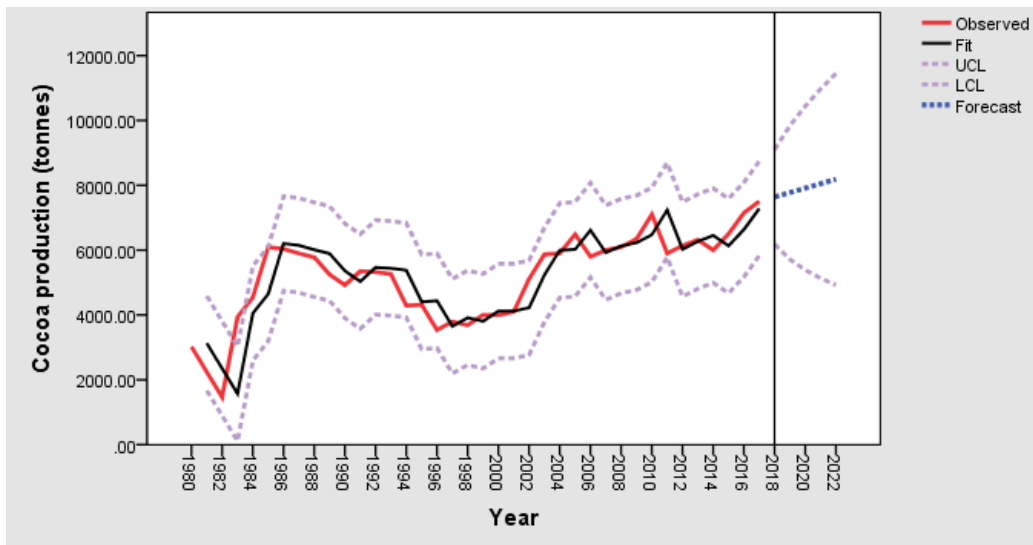


Fig. 4.9: Actual and forecasted production of cocoa in Kerala

Table 4.17: Forecasted values of cocoa production (tonnes) in Kerala for the period 2018-2022

Year	2018	2019	2020	2021	2022
cocoa Production (tonnes)	7642.22	7777.71	7913.91	8050.83	8188.46

Fig. 4.9 depicts that the observed series of cocoa production moves close to the forecasted values. Forecasts for cocoa production in Kerala for the years from 2018-2022 are given in table 4.17. The graph showed an increasing trend for the production of cocoa for the period from 2018-2022 by an amount of 7642.22 to 8188.46 tonnes.

4.1.3 Forecasting of cocoa productivity

Similarly, we use the time series data of cocoa productivity of Kerala for the period from 1980 to 2011 as a training period and validation of the model was done using the data for the next six years from 2012 to 2017. The expert modeller in SPSS selected the Simple exponential smoothing model as the best model to forecast the productivity of cocoa in Kerala.

The parameters of the Simple exponential smoothing model for predicting cocoa productivity are summarized in Table 4.18. The coefficient of the model was observed as $\alpha = 1$, which was significant. Thus, the simple exponential smoothing model could be used to forecast the cocoa productivity in Kerala.

Table 4.18: Parameters of the Simple exponential smoothing model for cocoa productivity in Kerala

	Estimate	SE	t	Probability
Alpha (Level)	1	0.194	5.148	0.00

Table 4.19: Statistical measures of Simple exponential smoothing model for prediction of cocoa productivity

Fit statistic	Simple exponential model
R ²	0.844
RMSE	0.061
MAPE	11.455
MAE	0.043
BIC	-5.491

Table 4.20: Validation of cocoa productivity in Kerala for the period 2012-2017 using Simple exponential smoothing model

years	2012	2013	2014	2015	2016	2017
Actual cocoa productivity (tonnes per ha)	0.49	0.46	0.41	0.42	0.45	0.45
Forecasted cocoa productivity (tonnes per ha)	0.48	0.48	0.48	0.48	0.48	0.48

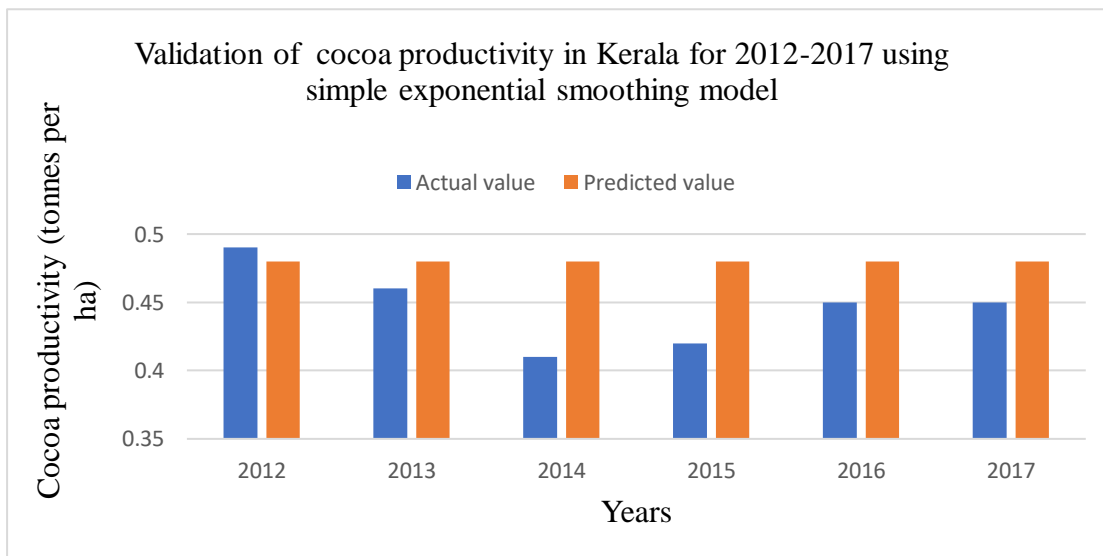


Fig 4.10: Validation of cocoa productivity in Kerala for 2012 to 2017 using simple exponential smoothing model

The time series data of cocoa productivity for the period 1980-2017 was exposed to ARIMA models also. ARIMA (0,1,1) was identified as the best for forecasting. Both models were compared based on the accuracy measures of forecasting such as the R², MAE, MAPE, RMSE and BIC and the best model was determined.

The results obtained are outlined in Table 4.21, the simple exponential smoothing model was obtained as the best model which had the highest R² and the smallest RMSE, MAPE, MAE and BIC values

Table 4.21: Comparison of accuracy measures of ARIMA (0,1,1) and Simple exponential model for cocoa productivity in Kerala

Fit statistic	ARIMA (0,1,1)	Simple Exp. model
R ²	0.829	0.838
RMSE	0.056	0.057
MAPE	12.037	10.47
MAE	0.041	0.040
BIC	-5.461	-5.637

The parameters of the Simple exponential smoothing model for cocoa productivity are summarized in Table 4.22. The coefficient of the model was observed as $\alpha = 1$.

Table 4.22: Parameters of Simple exponential smoothing model for cocoa productivity in Kerala

	Estimate	SE	t	Sig
Alpha (Level)	1	0.164	6.110	0.00

Simple exponential smoothing model

(Level of the series at time 't') $L_t = \alpha Y_t + (1-\alpha) L_{t-1}$

Forecast for k step ahead $F_t(k) = L_t$
 $= Y_t$ (since $\alpha = 1$)

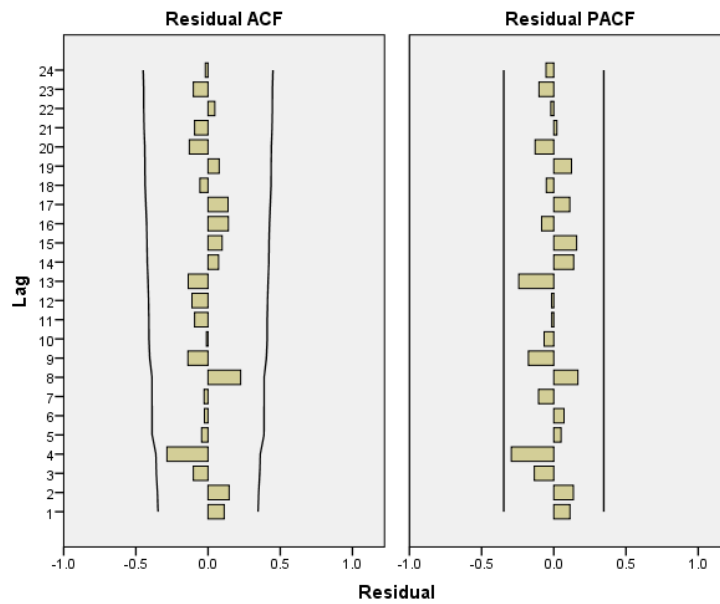


Fig 4.11: Residual plots of ACF and PACF of Simple exponential smoothing model for cocoa productivity in Kerala

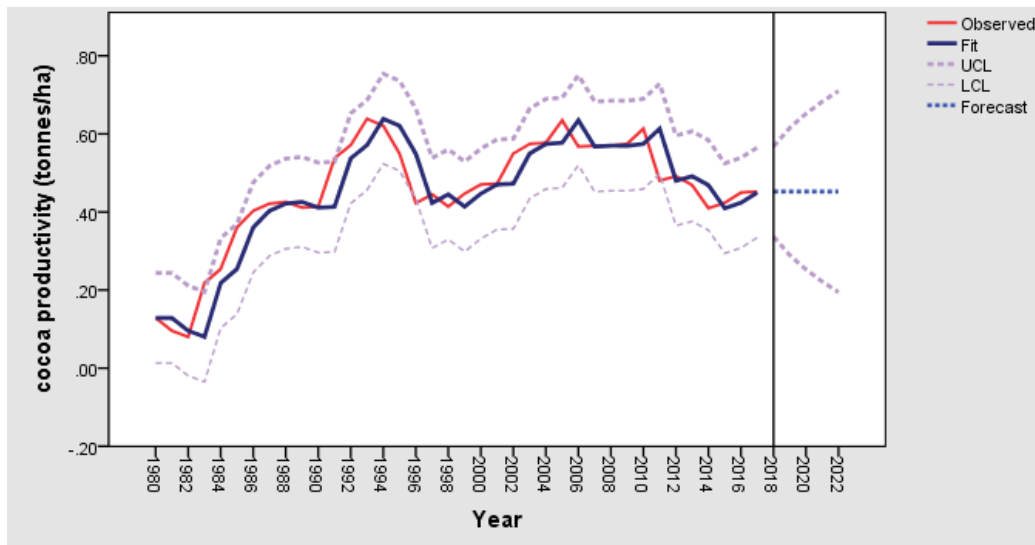


Fig 4.12: Forecasting of cocoa productivity (tonnes/ha) in Kerala using simple exponential model

From Fig. 4.12, it can be inferred that the two series, actual versus predicted values of cocoa productivity move together very closely depicting the efficiency of the model developed.

Table 4.23: Forecasted values of cocoa productivity (tonnes/ha) in Kerala using simple exponential smoothing model

Year	2018	2019	2020	2021	2022
Forecasted Productivity of cocoa(tonnes/ha)	0.45	0.45	0.45	0.45	0.45

Forecasts for cocoa productivity for the years 2018-2022 are given in Table 4.23. The graph showed a constant trend for the productivity. The forecasted productivity for all the years from 2018-2022 came out to be 0.45 tonnes/ha. From the production data of cocoa in Kerala an increasing trend could be noticed. So, the constant values of productivity revealed that area under cocoa cultivation was also getting increased which maintain the productivity at a constant level.

The area, production and productivity of cocoa in Kerala for the period from 2018 to 2022 were forecasted with high accuracy. For forecasting of area under cocoa, the ARIMA (0,2,2) and Holt's exponential smoothing model were compared. Based on the performance evaluation measures it was found that Holt's model was having high R^2 value (0.94) and low RMSE, MAE, MAPE and BIC values. So, Holt's exponential smoothing model was selected as the best model to forecast area under cocoa. The forecast of area under cocoa showed an increasing trend for the 5 years from 2018 to 2022. In case of forecasting of cocoa production, ARIMA (0,1,1) and simple exponential smoothing model were compared. Based on the selection criteria measures the ARIMA (0,1,1) was found to be an appropriate model with high R^2 value (0.72) and low RMSE, MAE and BIC values. In the forecasted period, the cocoa production increased from 7642.22 tonnes in 2018 to 8188.46 tonnes in 2022. If it is realized it would be a boon to the chocolate manufacturing companies of Kerala. Similarly, for forecasting the cocoa productivity also the ARIMA (0,1,1) model was compared with simple exponential smoothing model and found that simple exponential smoothing model performed better with higher R^2 value (0.831). The forecasted value of

productivity was 0.45 tonnes/ha which remained constant for all the 5 years from 2018 to 2022. Even though there was an increasing trend for the production, the trend of future productivity remained constant. It might be because of the increase in cultivated area. Therefore, it is important to take necessary steps to increase the cocoa productivity by advising good cultivation practices and high yielding varieties of cocoa to the farmers.

4.1.4 ARIMAX model

For the purpose of crop yield forecasting, the autoregressive integrated moving average (ARIMA) was the most widely used model in past. But this model cannot incorporate exogenous variables. Hence, Autoregressive Integrated Moving Average with Exogenous variables (ARIMAX) is preferred over ARIMA in order to forecast the crop yield more accurately. To perform ARIMAX, first the independent variable has been modelled and forecasted up to 2022 using ARIMA technique. Then these forecasted values were used as independent variable in the ARIMAX model.

The data for the period from 1980 to 2011 was used for training the ARIMAX model and the next six years data was used for validation of the model to forecast the cocoa production for the years ahead. The expert modeller in ‘SPSS’ software identified ARIMAX (0,1,0) as the best to forecast the cocoa production for the next five years.

In Table 4.24, the parameters of the ARIMAX (0,1,0) model for cocoa production in Kerala are summarized.

Table 4.24: Parameters of ARIMAX (0,1,0) model

		Estimate	SE	t	Sig
Production	Constant	-234.512	441.89	-0.531	0.600
	Difference	1			
Area	Numerator Lag 0	0.028	0.036	0.778	0.443

Table 4.25: Accuracy measures of ARIMAX model

Fit statistic	ARIMAX (0,1,0)
R square	0.660
RMSE	753.832
MAPE	13.21
MAE	527.211
BIC	13.469

The results obtained from the software is outlined in Table 4.25 which provides the accuracy measures of the model. The value of R square showed that 66% of the variation in cocoa production can be explained through the ARIMAX model by taking the area under cocoa as exogenous variable.

Table 4.26, gives the comparison of the actual values and predicted values of cocoa production for years from 2012 to 2017 for validation

Table 4.26: comparison of actual and predicted values of cocoa production for the years 2012 to 2017 in Kerala

Actual Values of cocoa production in tonnes	6136	6320	6000	6500	7150	7507
Forecasted values of cocoa Production in tonnes for validation	6015.34	6158.71	6334.79	6530.32	6741.26	6971.82

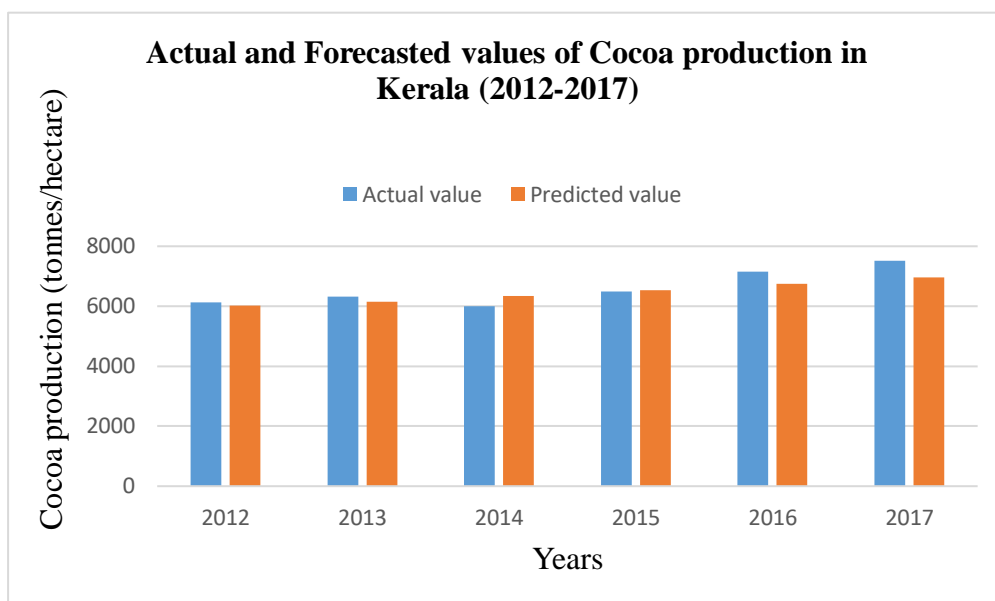


Fig 4.13: Validation of cocoa production in Kerala for 2012 to 2017 using ARIMAX (0,1,0) model

After validation of the model the data pertaining to area and production of cocoa for the period from 1980 to 2017 was used for forecasting the cocoa production for the years 2018-2022. The time series data of cocoa production for the period 1980-2017 was exposed to expert modeller in SPSS software after converting the figures to natural logarithm to reduce the variability in whole data series of area and production. ARIMAX (0,1,0) was chosen as the best model to forecast the cocoa production.

In Table 4.27, the parameters of the ARIMAX (0,1,0) model for cocoa production in Kerala are summarized.

Table 4.27: The parameter estimates of ARIMAX (0,1,0) model for cocoa production (tonnes) in Kerala

Variable	Parameters	Estimate	SE	t	sig
Production model Area	Difference	1			
	Numerator Lag 0	0.539	0.212	2.538	0.017
	Lag 1	-0.421	0.189	-2.224	0.035
	Difference	2			

$$Y_t = Y_{t-1} + 0.539 X_t - 0.421 X_{t-1} + \epsilon_t \quad [\text{ARIMAX (0,1,0) model}]$$

Table 4.28.: Statistical measures of ARIMAX model for cocoa production (tonnes) in Kerala

Fit statistic	ARIMAX (0,1,0)
R ²	0.837
RMSE	0.090
MAPE	0.823
MAE	0.07
BIC	-4.58

Table 4.28 provides the statistical measures for the selected ARIMAX model. The value of R² showed that 83.7% of variation in cocoa production can be explained through the ARIMAX model taking area under cocoa as the independent variable.

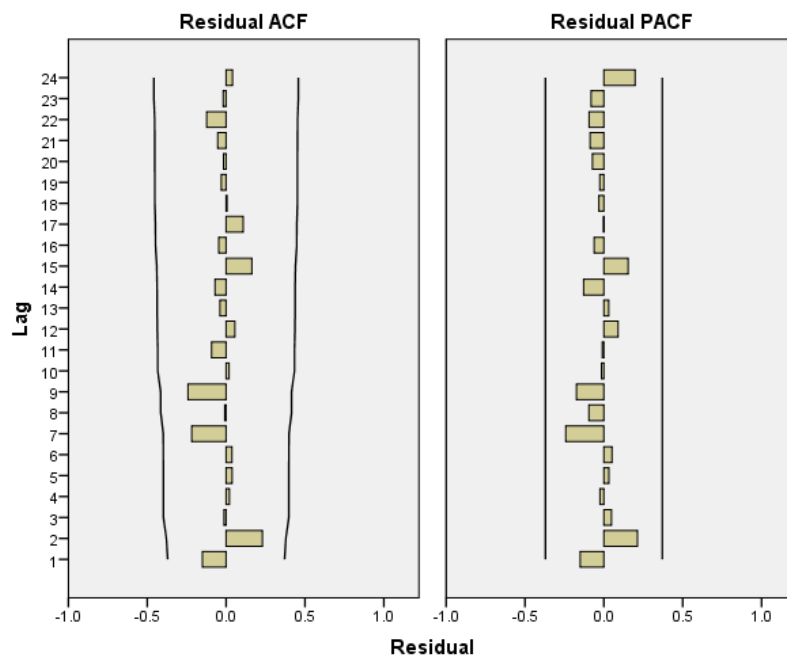


Fig. 4.14: Residual plots of ACF and PACF in the ARIMAX model

From Fig. 4.14, it can be seen that all the residuals in the ACF and PACF plots were within the confidence limits and so the residuals were almost white noise

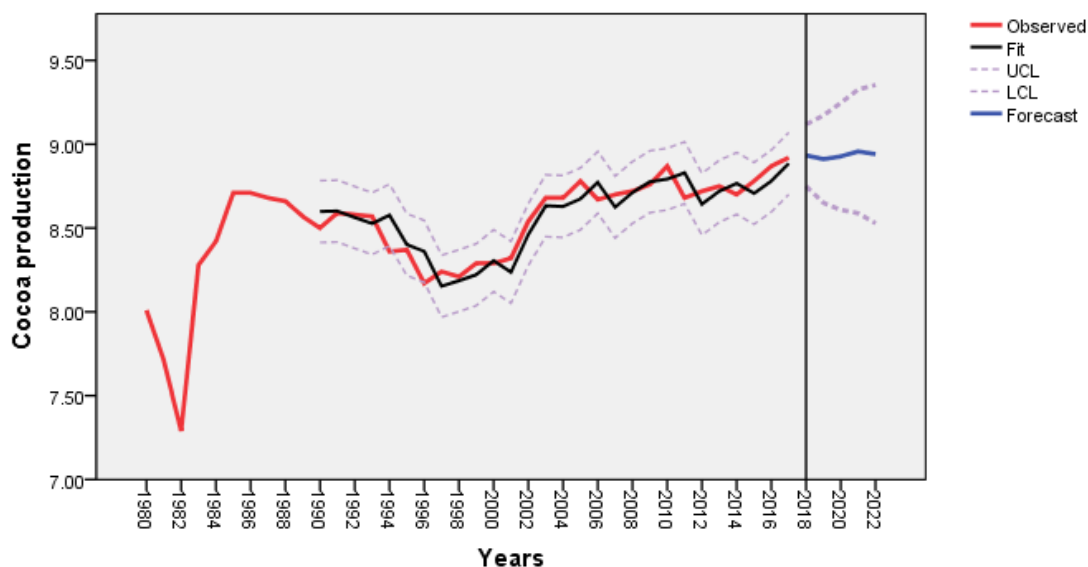


Fig 4.15: Forecasting of cocoa production of Kerala through ARIMAX (0,1,0) model

Forecasts for cocoa production for the years 2018-2022 period are given in Table 4.29. The graph showed an increasing trend for the production of 2018-2022 period, from 7628.27 to 8113.25 tonnes.

Table 4.29: Forecasted values of cocoa production (tonnes) in Kerala for the period 2018-2022

Year	2018	2019	2020	2021	2022
Production (tonnes in natural log values)	8.93	8.91	8.93	8.96	8.94
Production (tonnes in original values)	7555.265	7405.661	7555.265	7785.357	7631.197

From the results it was assessed that compared to the traditional ARIMA model, the Autoregressive Integrated Moving Average with exogenous inputs i.e. (ARIMAX) model was more powerful in prediction of cocoa production. The ARIMAX (0,1,0) model by considering the area under cocoa in Kerala for the period from 1980 to 2017 as exogenous input to forecast the cocoa production of Kerala state for the year 2018 to 2022 resulted in forecasted values which showed an increasing trend in cocoa production from 7555.27 tonnes to 7631.20 tonnes for the years from 2018 to 2022. The time series data on area under cocoa as input variable helps in improving the cocoa yield forecasts. The ARIMAX models performed well to get the short-term forecasts of

cocoa production in Kerala. and excelled the ARIMA models in forecasting the cocoa yield well in advance of the crop harvest for Kerala state.

4.2 Performance Evaluation of cocoa hybrids in Cocoa Research Centre, KAU

The performance of cocoa hybrids maintained at Cocoa Research Centre was evaluated by collecting the monthly yield and infected number of pods of 100 selected hybrids from the Centre. The number of cocoa pods for all the months from each of 100 cocoa plants for the period from 2003 to 2017 was obtained from Cocoa Research Centre, Vellanikkara, Thrissur, Kerala. The total number of cocoa pods for all the 100 plants for 12 months in each year was computed. Then the average number of cocoa pods were calculated. This gave the average number cocoa pods for the 15 years with 12 months in each year providing 180 set of data points, a univariate time series data which was seasonal in nature.

Table 4.30: Descriptive Statistics for the average monthly cocoa yield from 100 plants for the period from 2003 to 2017

Months Years	Months											
	Jan	Feb	Mar	April	May	Jun	July	Aug	Sept	Oct	Nov	Dec
2003	4.9	4.21	4.13	2.09	1.98	2.17	2.65	4.23	7.79	13.7	14.69	15.11
2004	8.76	5.5	4.7	2.21	2.08	2.41	2.73	4.43	7.96	12.15	14.42	17.25
2005	6.02	4.95	4.68	2.5	2.45	2.77	3	3.9	6.56	18.42	20.27	19.2
2006	4.82	4.46	4.08	1.7	1.85	2.58	3.33	5.06	7.03	9.32	17.57	13.41
2007	6.85	6.04	5.89	3.12	3.13	3.43	4.27	5.03	8.21	22.36	20.48	17.42
2008	5.15	4.88	4.05	3.29	3.32	3.62	4.05	4.54	6.91	17.49	18.29	15.47
2009	5.54	5.07	4.64	4.07	3.82	3.56	4.05	4.7	8.6	18.22	17.82	13.64
2010	5.01	3.93	3.33	3.96	3.65	3.7	3.79	4.2	8.88	20.42	17.76	15.32
2011	6.22	3.88	1.8	4.1	3.01	4.06	4.1	5.89	13.36	19.24	19.22	14.23
2012	5.39	4.21	2.93	4.61	3.42	3.76	4.21	4.68	7.75	17.56	18.32	14.33
2013	5.37	3.96	2.71	4.36	3.46	3.59	3.17	5.3	14.8	18.16	19.19	13.74
2014	6.67	4.01	2.8	4.36	3.09	4.4	3.85	5.48	8.94	20.61	20.56	10.62
2015	6.84	3.68	2.65	3.79	3.56	3.19	3.78	4.5	11.67	19.47	17.7	11.47
2016	6.53	3.17	2.87	3.51	2.74	3.02	3.56	4.51	9.18	21.76	19.06	14.14
2017	5.86	3.6	3.08	3.56	3.06	3.94	3.38	3.78	5.49	21.7	16.68	13.04
Mean	6.00	4.37	3.62	3.42	2.97	3.35	3.59	4.68	8.88	18.04	18.14	14.56
SD	1.04	0.78	1.08	0.91	0.63	0.64	0.53	0.58	2.55	3.70	1.85	2.22
CV (%)	17.33	17.78	29.79	26.73	21.06	19.17	14.74	12.44	28.72	20.52	10.17	15.27
Skewness	1.30	0.69	0.41	-0.62	-0.64	-0.38	-0.51	0.48	1.25	-1.19	-0.73	0.35
Kurtosis	2.34	0.04	-0.21	-0.81	-0.74	-0.64	-0.94	-0.09	1.14	0.99	0.26	0.39

In Table 4.30, the descriptive statistics for average monthly cocoa yield of 100 plants obtained from Cocoa Research Centre, Vellanikkara, is depicted. The mean number of cocoa pods was highest in the month of November followed by October, December and September. The least average number of pods was produced in the month of May followed by June and April. The value of standard deviation was highest for the month of October which indicated that the data was widely spread from the mean value. This was followed by the months of September and December. The CV was lowest in the month of November which showed that the consistency was highest during November with respect to average monthly yield over the years. Thus, November came first with respect to average monthly cocoa yield with highest consistency. The values of skewness and kurtosis showed that the average number of cocoa pods from 100 plants was not normally distributed in different months.

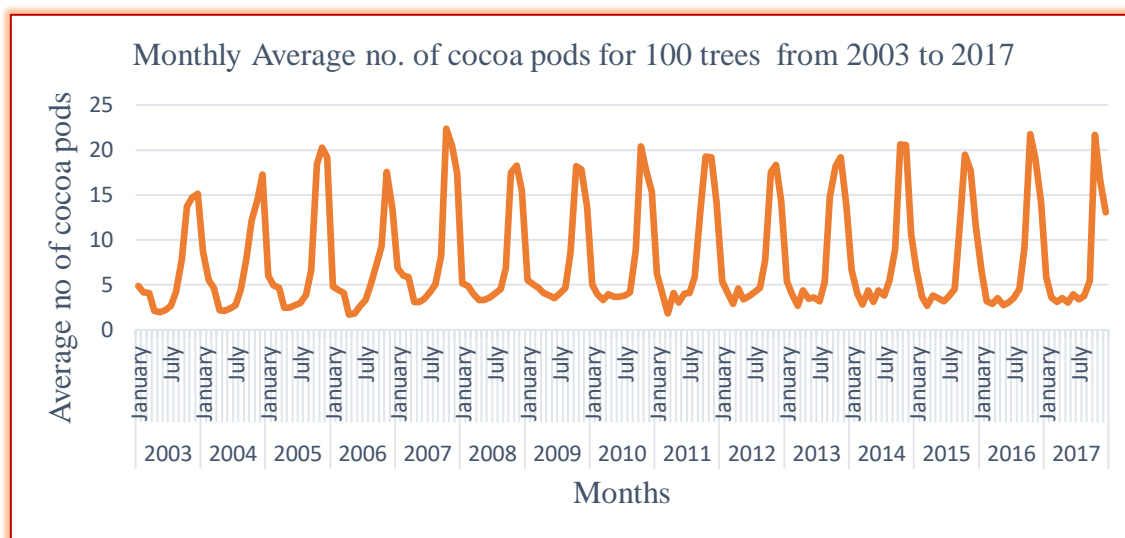


Fig 4.16: Monthly average no. of cocoa pods for 100 trees from 2003 to 2017

From Fig. 4.16, it can be understood that seasonality was present in the average monthly cocoa yield data from 100 plants. The seasonal component changes slowly as shown in figure that exhibited similar pattern after every consecutive years. The regular pattern

of changes in average monthly cocoa yield data that repeats again after every 12 months represented the seasonality.

4.2.1 SARIMA MODEL

To capture the seasonality present in the monthly yield data of the cocoa hybrids, an attempt was made to fit SARIMA model to the monthly yield data. The data for the years 2003 to 2016 were used for training the SARIMA Model and validation of the model was done using the monthly data for the year 2017. The expert modeller in SPSS selected the SARIMA (1, 0, 0) (1, 1, 0)₁₂ as the model type to forecast the yield for the next 12 months, where the model contained two elements, trend element (1,0,0) and seasonal element (1,1,0).

Table 4.31: Parameter estimates of SARIMA (1,0,0)(1,1,0)₁₂ model

Models	Estimate	Significance
AR Lag 1	0.314	0.000
AR, Seasonal Lag 1	-0.543	0.000
Seasonal Difference	1	

From Table 4.31, it can be seen that a monthly Seasonal ARIMA (1,0,0)(1,1,0)₁₂ model is divided into two parts viz; the non-seasonal part and the seasonal part. The non-seasonal part (1,0,0) was the first-order autoregressive model and the series was stationary with significant autocorrelation. The seasonal element (1,1,0) was the seasonal part of the model with first-order seasonal auto regression which was significant and first-order seasonal difference.

From Table 4.32, it can be followed that the model SARIMA (1, 0, 0) (1, 1, 0)₁₂ used had high value of R² (0.923) which confirmed that the model possessed good explanatory power which was suitable for forecasting the cocoa yield for the next 12 months.

The results obtained has been outlined in Table 4.32 which shows the highest value of R square with lowest value of RMSE, MAPE, MAE and BIC.

Table 4.32: Statistical measures for SARIMA (1,0,0)(1,1,0)₁₂ model

Fit statistic	SARIMA (1,0,0)(1,1,0) ₁₂
R ²	0.923
RMSE	1.648
MAPE	13.543
MAE	1.023
BIC	1.063

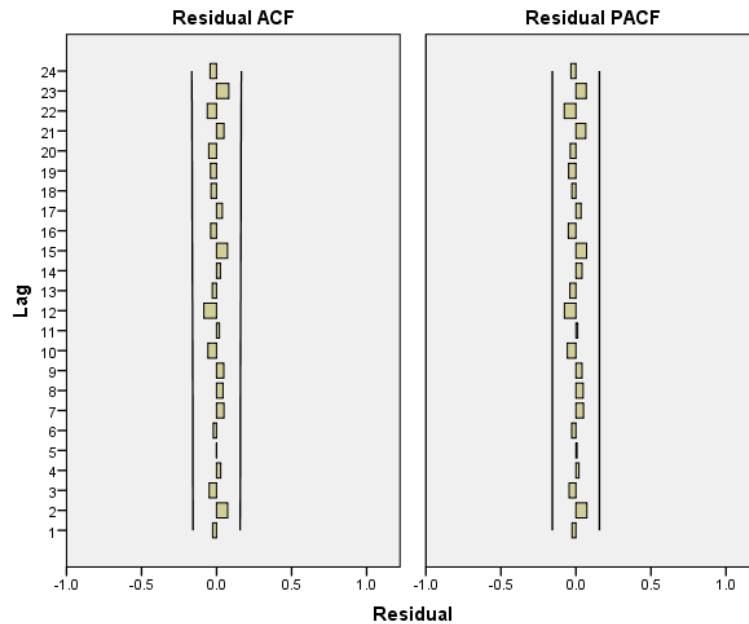


Fig 4.17: Residual plots of ACF and PACF of SARIMA(1,0,0) (1,1,0)₁₂ model

From Fig. 4.17, it can be seen that all the residuals in the ACF and PACF plots were within the confidence limits and thus the residuals could be considered as white noise.

Table 4.33: Validation of the SARIMA model for average monthly cocoa yield for the next 12 months for the year 2017

Model	Jan 2017	Feb 2017	Mar 2017	Apr 2017	May 2017	Jun 2017	July 2017	Aug 2017	Sep 2017	Oct 2017	Nov 2017	Dec 2017
Actual Values (Av. no. of cocoa pods)	5.86	3.60	3.08	3.56	3.06	3.94	3.38	3.78	5.49	21.70	16.68	13.04
Forecasts (Av. no. of cocoa pods)	7.36	3.58	2.82	3.73	3.22	3.17	3.74	4.58	10.64	20.84	18.63	12.84

Table 4.33, gives the comparison of validated values with the actual values of average number of cocoa pods for 12 months in the year 2017. It can be observed that the validated values were very close to the actual values. Thus the SARIMA (1,0,0)(1,1,0)₁₂ model is suitable for forecasting the average monthly cocoa yield.

The complete data on average monthly cocoa yield from 2003 to 2017 was used to forecast the monthly yield for the next year 2018 applying the model SARIMA (1, 0, 0) (1, 1, 0)₁₂ to forecast the yield for the next 12 months.

Table 4.34: Parameter estimates of SARIMA (1,0,0)(1,1,0)₁₂ model

Models	Estimate	Significance
AR Lag 1	0.286	0.000
AR, Seasonal Lag 1	-0.512	0.000
Seasonal Difference	1	

SARIMA (1,0,0)(1,1,0)₁₂ model

$$(1-\phi_1B)(1-\Phi_1B^{12})(1-B^{12})Y_t = C$$

Where, $(1-\phi_1B)$ – Non-seasonal AR (1) with $\phi_1 = 0.286$

$(1-\Phi_1B^{12})$ – Seasonal AR (1) with $\Phi_1 = -0.512$

$(1-B^{12})$ – Seasonal difference

From Table 4.34, it is obvious that monthly Seasonal ARIMA $(1,0,0)(1,1,0)_{12}$ model was divided into two parts viz; the trend element and the seasonal element. The trend elements $(1,0,0)$ was the non-seasonal part of the model which was the first-order autoregressive model and the series was stationary and with significant autocorrelation. The seasonal element $(1,1,0)$ was the seasonal part of the model with first-order seasonal autoregressive terms which was significant and first-order seasonal difference.

The obtained results is outlined in Table 4.35 which shows the high value of R square with lowest value of RMSE, MAPE, MAE and BIC.

Table 4.35: Statistical measures for SARIMA $(1,0,0)(1,1,0)_{12}$ model

Fit statistic	SARIMA $(1,0,0)(1,1,0)_{12}$
R^2	0.920
RMSE	1.679
MAPE	13.959
MAE	1.044
BIC	1.098

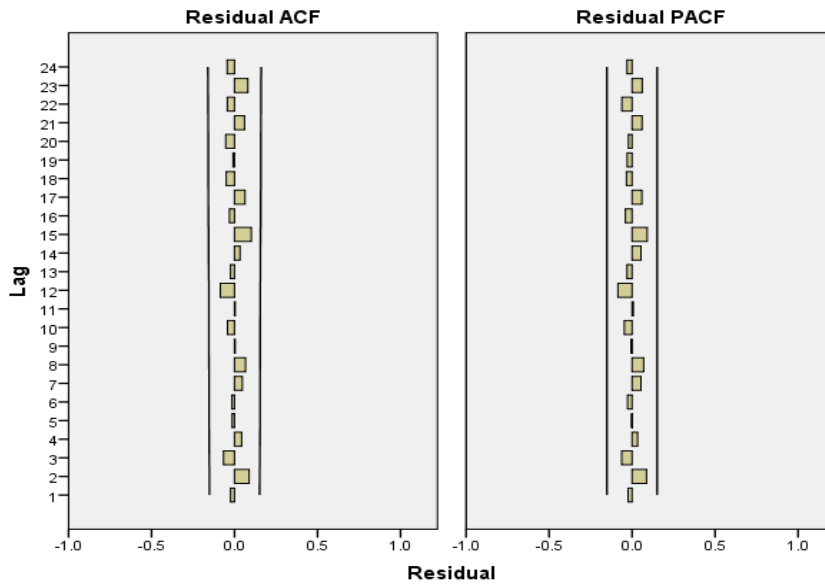


Fig 4.18: Residual plots of ACF and PACF of SARIMA $(1,0,0) (1,1,0)_{12}$ model

From Fig. 4.18, it can be seen that all the residuals in the ACF and PACF plots were within the confidence limits.

Table 4.36: Forecasted values of average monthly cocoa yield for the next 12 months for the year 2018

Month and year	Jan 2018	Feb 2018	Mar 2018	Apr 2018	May 2018	Jun 2018	July 2018	Aug 2018	Sep 2018	Oct 2018	Nov 2018	Dec 2018
Forecast (Av. No. of cocoa pods)	6.34	3.44	3.03	3.60	2.94	3.50	3.53	4.21	7.27	22.12	18.18	13.84

From Table 4.36, it is apparent that the SARIMA (1,0,0)(1,1,0)₁₂ model could be effectively used to generate the forecasted values for average number of cocoa pods for all the months in the year 2018. The number of pods has been increased especially in the months of September, October, November and December compared to previous months.

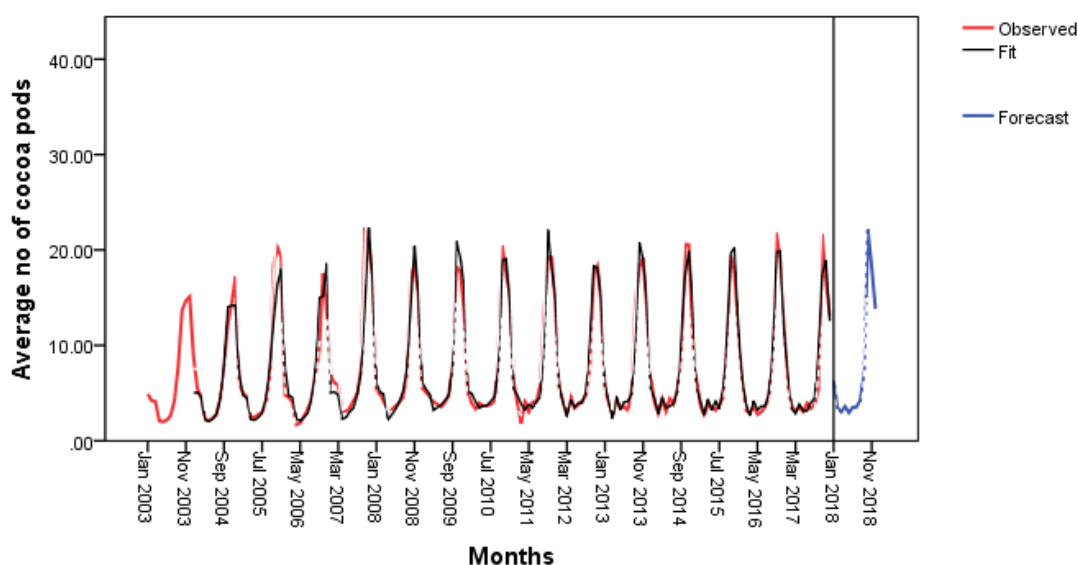


Fig 4.19: A twelve month forecast of average cocoa yield using SARIMA (1,0,0)(1,1,0)₁₂ for the year 2018

From Fig. 4.19, it is revealed that the SARIMA(1,0,0)(1,1,0)₁₂ gave the average monthly forecasted values of number of cocoa pods for all the months of the year 2018 with high level of accuracy

The main purpose of the above attempt was to fit a model that could be used to forecast average monthly cocoa yield of cocoa trees from Cocoa Research Centre, vellanikkara, Thrissur, Kerala. The study explored that the Seasonal ARIMA model could be very effectively used to forecast the average monthly cocoa yield. At first the cocoa yield data collected on monthly basis from 2003-2016 was used for training the model and SARIMA (1,0,0)(1,1,0)₁₂ model came out to be the best fit after validation to forecast the yield for the next 12 months for the year 2018. From the results it was discovered that the model SARIMA (1, 0, 0) (1, 1, 0)₁₂ used showed relatively high value of R² (0.92) with low values of RMSE, MAPE, MAE and BIC values. The model predicted the average monthly cocoa yield in 2018 for the months as September (7.27 average no.of cocoa pods), October (22.12 average no. of cocoa pods), November (18.18 average no. of cocoa pods) and December (13.84 average no. of cocoa pods) . Thus, it is concluded that in order to deal with the time series data containing seasonality, the SARIMA models are proved to be the best model with replacement of ARIMA models. Seasonal ARIMA model can predict the yield with good accuracy and can be used in all sectors.

4.2.2 General Linear Model

An insight to the total yearly cocoa yield of 100 cocoa hybrids cultivated in the Cocoa Research Centre, KAU can be depicted through a Box plot.

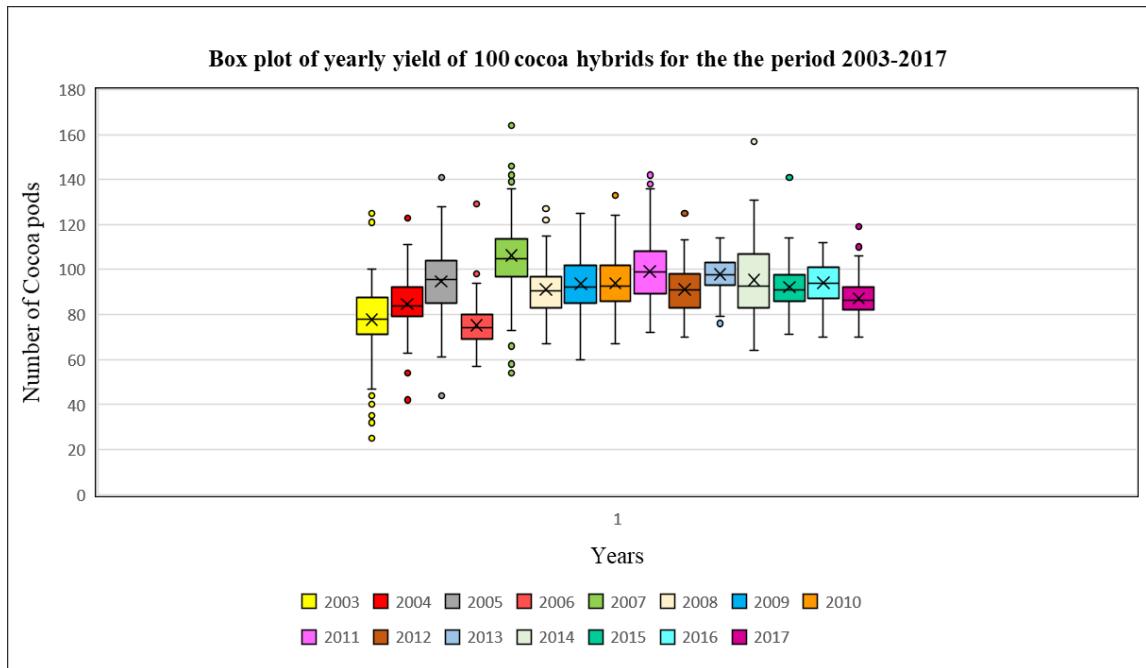


Fig. 4.20 Box plot for total number of cocoa pods of 100 cocoa hybrids for 2003 -2017

The box plot was drawn to compare the yearly yield of cocoa for 100 trees for the 15 time periods.

The boxplot was built on the maximum value, minimum value, first quartile (Q1), median (Q2), and third quartile (Q3) of the data.

From the Figure 4.20 it is depicted that the peak yield was observed in the year 2007 with highest number of cocoa pods followed by 2011. The lowest yield was observed in the year 2006.

By looking at the shape of the box plot it can be explained that the data set is normally distributed or Skewed. The box plot for the year 2003, 2004, 2006, 2009, 2010 and

2014 showed that the median was closer to the lower quartile and the mean > median. It means that the data set was positively skewed. While, the box plot for the year 2005, 2008, 2011 and 2012 showed the median closer to the upper quartile and the mean < median. It means that the data was negatively skewed. Whereas, the box plot for the year 2007 and 2015 showed that the median got coincided with the mean and the data set was normally distributed.

The variability in the data set is best explained by the interquartile range (IQR). The dispersion in the data is the extent to which a distribution is stretched or squeezed. The interquartile range (IQR) in the boxplot showing the 50% of the data points can be calculated by subtracting first quartile from the third quartile. When the data is skewed, we prefer interquartile range as the best to measure the variability than the standard deviation. The boxplot for 2011 and 2014 was highly stretched which showed that the data points were highly variable or scattered. The boxplot for the year 2013, 2015 and 2017 was compressed which showed that the distribution of data points was squeezed or concentrated. While, the boxplot for the remaining years was normally distributed.

There are outliers present beyond the upper whisker and the lower whisker. These outliers are considered as the abnormal values that affect the data set. With the help of this box plot the outliers can be identified and discarded from the data set. From the figure it is observed that most of the years contained outliers beyond the upper whisker and the lower whisker but only the box plot for the year 2009 and 2016 did not contain any outliers.

Repeated Measures ANOVA

The repeated measure design is also called within subject design since the comparisons are made multiple times (repeated) within the same subject. The subjects are the cocoa plants on which the observations are taken. In this study the dependent variables that is the within subject measures are the total yearly yield of each of the 100 cocoa plants repeatedly measured over 15 time levels (t_1, t_2, \dots, t_{15}) from 2003 to 2017. When the subjects are in groups that are independent of one another, then the group is called

between subject factor. In the study the 100 cocoa hybrids were divided into 5 groups by first arranging the yield data in 2003 in a sequence with respect to ascending order of magnitude. Each group contained 20 cocoa plants and all the 5 groups were considered as the between subject factors. The general linear model repeated measures ANOVA helps us to compare the 5 groups eliminating the effect of time.

Table 4.37 and Table 4.38 show the Within and Between subject factors

Table: 4.37 The dependent variables in GLM

Time (years)	Dependent variable
2003	t1
2004	t2
2005	t3
2006	t4
2007	t5
2008	t6
2009	t7
2010	t8
2011	t9
2012	t10
2013	t11
2014	t12
2015	t13
2016	t14
2017	t15

Table: 4.38 Between subject factors

Factors	Number of cocoa plants in each group (Independent variables)
F ₁	20
F ₂	20
F ₃	20
F ₄	20
F ₅	20

The analysis for repeated measures was done by using the statistical software package “SPSS”.

Table 4.39: The accession ids of cocoa hybrids included in the GLM repeated measures analysis

Sl.No.	Factor1	Factor2	Factor3	Factor4	Factor5
1	33.9	5.8	2.3	12.27	6.4
2	20.3	11.3	5.1	2.9	10.12
3	23.21	5.2	7.1	5.12	3.1
4	22.22	5.11	9.12	10.13	4.1
5	20.16	8.2	2.1	12.1	8.6
6	11.22	12.6	2.6	6.2	8.11
7	1.11	4.4	2.7	1.3	5.3
8	3.8	4.6	4.1	2.13	12.4
9	4.2	5.9	4.5	7.6	1.1
10	10.7	5.14	8.25	10.24	2.4
11	1.1	12.25	10.6	9.25	9.24
12	3.7	1.4	12.12	10.4	10.16
13	3.1	1.5	4.13	4.9	10.26
14	12.5	3.5	5.13	6.1	1.8
15	5.1	7.24	6.5	8.3	4.3
16	10.1	8.26	3.2	10.23	10.2
17	6.1	1.7	3.6	11.8	4.8
18	1.12	2.5	8.5	2.1	10.18
19	2.2	5.6	10.5	2.12	1.9
20	2.8	12.1	12.2	10.14	10.19

Table 4.40: Mean and S.D of yearly production of 5 groups (Factors) of Cocoa hybrids used for GLM

M denotes mean and S.D denotes standard deviation

Factor	Time	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
	F ₁	M	54.4	79.05	83.45	74.65	100.35	89.1	97.5	96.05	99.9	96.95	96.5	99.9	94.95	91.45
	S.D	13.05	19.4	18.29	15.58	26.56	14.34	13.04	13.17	18.96	11.29	6.83	22.4	15.4	9.17	10.74
F ₂	M	72.6	87.5	95.15	75.5	107.15	88.2	93	90.55	97.7	93.2	99.4	91.5	91.55	93.55	86.3
	S.D	2.15	9.84	9.14	6.99	12.81	10.06	15.64	10.96	13.14	9.32	6.81	14.7	8.6	5.84	7.82
F ₃	M	78.5	83.75	95.15	73.6	110.15	94.95	93.25	93.6	98.9	91.35	95.95	91.5	92.55	96.4	88.05
	S.D	1.19	10.09	14.46	7.06	13.16	13.54	16.21	14.25	11.78	8.52	8.34	8.81	8.92	9.58	5.93
F ₄	M	85.4	86.6	98.3	74.8	108.05	92.7	89.25	96.95	102.15	85.85	97.5	97.6	92.3	95	84.3
	S.D	2.64	12.21	14.65	4.37	16.94	14.57	12.20	8.7	9.78	6.77	7.89	11.36	7.46	10.36	6.72
F ₅	M	97.3	86.1	101.55	77.45	105.45	90.35	95.65	92.6	96.9	88.5	99.7	96.45	90.15	93.85	86.8
	S.D	9.3	9.95	15.67	9.64	12.04	9.02	11.94	13.21	11.77	6.78	7.02	14.4	6.38	10.09	8.34
TOTAL	M	77.6	84.6	94.72	75.21	106.23	91.06	93.73	93.95	99.11	91.17	97.81	95.39	92.3	94.05	87.17
	S.D	16.01	12.94	15.70	9.42	17.15	12.51	13.92	12.2	13.31	9.35	7.41	15.14	9.82	9.12	8.17

Table 4.40 gives the set of values of mean and standard deviation for 15 time levels (dependent variable) from 2003 to 2017 which were subjected to 5 factors (F1 to F5) formed by groups of cocoa hybrids with homogeneous yield within group for comparison of variation of yield at different time levels. The Factor 3 at 5th time level (for the year 2007) showed the highest mean value of yearly cocoa pod yield of 20 cocoa trees with the mean value of 110.15. Among the mean value for all the years, the year 2007 (5th time level) showed the highest number of yearly cocoa pod yield with value of 106.23 average number of cocoa pods for all the 100 cocoa trees. Irrespective of all the groups (Factors) this result hold true as seen from Table 4.40. The peak average number of pods for different groups at the 5th time interval were 100.35 for F1, 107.15 for F2, 110.15 for F3, 108.05 for F4, 105.45 for F5 with the least S.D for F5=12.04 and with an overall average yield coinciding at 106.23. The value of standard deviation provides the variation of yearly cocoa yield over different time periods. It was found that in the year 2007 (5th time level) the highest value, 17.15 as the value of standard deviation which indicated that the value of cocoa yield spread over wide range from the mean value in that time period. It could be inferred that the peak yearly yield could be obtained during the 5th year after the first harvest from plants irrespective of the hybrids under consideration.

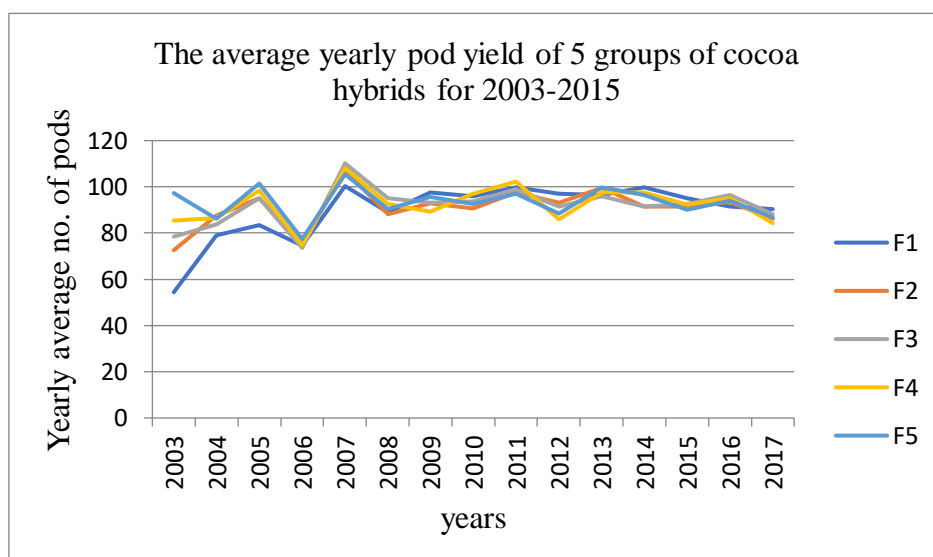


Fig.4.21 The average yearly pod yield of 5 groups of cocoa hybrids for 2003 -2015

From Fig.4.21, it can be observed that yield of F₁ was significantly lower than F₅ in the beginning year but subsequently the difference in yield between different groups get reduced. Thus, the significant effect of time on cocoa yield could be visualised. The peak yield was attained during the year 2007, i.e. during the fifth year from the start of harvest of the plants. A biennial tendency could also be evidenced in the average yearly pod yield as the time (years) progressed. The percentage reduction in yield (average number of pods) at the 15th time period with respect to the peak period (5th period) for each group were F1(9.92%), F2(19.46%), F3(20.06%), F4(21.98%) and F5(17.69%). From this the significant interaction effect of time with factors comprised of different hybrids was proved. Even though the yielding capacity of the plants in the first group were lowest in the beginning year of harvest, the plants could maintain the average yield moderately well when compared to the yield of the peak harvest. Based on these observations, important decisions can be made to plan whether a plant is to be culled or maintained in the farm after several years.

Table 4.41: Multivariate Tests to compute F values in GLM

Effect	Value	F	Hypothesis df	Error df	Sig.
Time					
Pillai's Trace	.935	84.09	14	82	0.00
Wilks' Lambda	.065	84.09	14	82	0.00
Hotelling's Trace	14.35	84.09	14	82	0.00
Roy's Largest Root	14.35	84.09	14	82	0.00
Time * Factor					
Pillai's Trace	1.17	2.52	56	340	0.00
Wilks' Lambda	.130	3.96	56	321.13	0.00
Hotelling's Trace	4.60	6.61	56	322	0.00
Roy's Largest Root	4.16	25.28	14	85	0.00

The Multivariate Tests in Table 4.41 shows the result of GLM repeated measures one-way ANOVA. SPSS estimates four different statistics to calculate the F value for MANOVA. From Table 4.41 we have to look at the Time effect and Time*Factor interaction effect and the corresponding parameters of Wilks' Lambda in a row. The

value of the Wilks' Lambda was observed to be 0.065 for the Time effect and 0.13 for Time*Factor interaction effect which were significant at 1% level. Therefore, it can be concluded that the Time effect and Time*Factor interaction effect were significant.

Table 4.42: Mauchly's Test of Sphericity in GLM

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Time	0.011	408.762	104	0.00	0.557	0.637	0.071

From Table 4.42 it can be observed that the the p value was < 0.05 which revealed that the assumptions of sphericity was violated.

The violation of sphericity is nothing but a loss of power (i.e. an increased probability of Type II error) and a test statistic F value calculated could not be compared to tabulated value of F distribution.

Since the data violated the assumption of sphericity, another method called Greenhouse-Geisser which made an adjustment to the degrees of freedom of the repeated measures ANOVA was used.

Table 4.43: Tests of within-subject effects

source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Sphericity Assumed	88198.6	14	6299.9	46.008	0.00	0.327
	Greenhouse-Geisser	88198.6	7.79	11310.29	46.008	0.00	0.327
	Huynh-Feldt	88198.6	8.92	9888.15	46.008	0.00	0.327
	Lower-bound	88198.6	1	88198.6	46.008	0.00	0.327
Time * Factor	Sphericity Assumed	29795.4	56	532.06	3.892	0.00	0.141
	Greenhouse-Geisser	29795.4	31.19	955.21	3.892	0.00	0.141
	Huynh-Feldt	29795.4	35.67	835.10	3.892	0.00	0.141
	Lower-bound	29795.4	4	7448.85	3.892	0.006	0.141
Error (Time)	Sphericity Assumed	181803.33	1330	136.69			
	Greenhouse-Geisser	181803.33	740.81	245.40			
	Huynh-Feldt	181803.33	847.36	214.55			
	Lower-bound	181803.33	95	1913.71			

Table 4.43 gives the Tests of within-subjects effects, which tells weather there was an overall significant difference between the means at the different time points. It gives the F value and its associated significance level and effect size (Partial Eta Squared) for the Time and Time*Factor interaction. As the data violated the assumption of sphericity, the values for Greenhouse-Geisser had been made use of. From the results it could be stated that, using an ANOVA with repeated measures with a Greenhouse-Geisser correction to the degrees of freedom, there was an overall significant difference between the means of cocoa yield at different time points evidenced from both Time effect ($F_{(7.798, 740.818)}=46.088$, $p < 0.05$) with Partial Eta squared value of 0.327 and Time*Factor interaction effect ($F_{(31.192, 740.818)}=3.892$, $p < 0.05$) with Partial Eta squared value of 0.141. Thus 32.7% of the variability in average yearly yield could be explained by the main effect of time keeping all other variables fixed. The factor*time interaction seemed to be significant and it could be established that the different factors viz; F1, F2, F3, F4 and F5 comprising of low yielding to high yielding group of trees have significantly different effect with the time variable with partial eta squared equal to 0.141. Thus 14.1% of the variation in yield could be attributed to the factor*time interaction keeping all other variables fixed.

Table 4.44: Tests of within-subjects contrasts

Source	Time	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Time	Linear	10602.60	1	10602.60	40.15	0.00	0.297
	Quadratic	23280.54	1	23280.54	267.63	0.00	0.738
	Cubic	178.40	1	178.40	1.42	0.23	0.015
Time * Factor	Linear	10876.12	4	2719.03	10.29	0.00	0.302
	Quadratic	6722.17	4	1680.54	19.32	0.00	0.449
	Cubic	2196.84	4	549.21	4.38	0.003	0.156
Error (Time)	Linear	25087.21	95	264..07			
	Quadratic	8263.55	95	86.98			
	Cubic	11906.41	95	125.33			

In repeated measures the contrast variables are the linear combinations of the responses over time period for individual cocoa trees. A set of contrast variables were used to check trends over time and to make comparisons between times in repeated measures data. Table 4.44, provides orthogonal polynomials which represented linear, quadratic, cubic, etc., trends over time. From the table it can be observed that the linear and quadratic models were good fit. As the partial eta squared for the quadratic model was 0.738 for the time effect and 0.449 for time*factor effect which was higher than the effect size due to linear models, it could be inferred that quadratic model was the best to account for the time effect as there was no further improvement in the value of eta square for the cubic model. This result was also similar to the yearly trend of cocoa in Kerala in which case it could be well fitted using a quadratic model.

Table 4.45: Tests of Between-Subjects Effects

Source		df	Mean Square	F	Sig.	Partial Eta Squared
Intercept	12588588.15	1	12588588.15	64512.22	0.000	0.999
Factor	2369.7	4	592.42	3.03	0.021	0.113
Error	18537.81	95	195.13			

Table 4.45 shows that the value of significance for the factor was 0.021 indicating that there was a significant difference between the factors. Since the factors were significantly different a post hoc test was performed to identify the treatments which were significantly different.

Table 4.46: Multiple comparison of means of different factors

(I) Factor	(J) Factor	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1.00	2.00	-1.2233	1.14057	1.000	-4.5014	2.0547
	3.00	-2.2033	1.14057	.564	-5.4814	1.0747
	4.00	-2.8100	1.14057	.156	-6.0881	.4681
	5.00	-3.6133*	1.14057	.021	-6.8914	-.3353
2.00	1.00	1.2233	1.14057	1.000	-2.0547	4.5014
	3.00	-.9800	1.14057	1.000	-4.2581	2.2981
	4.00	-1.5867	1.14057	1.000	-4.8647	1.6914
	5.00	-2.3900	1.14057	.388	-5.6681	.8881
3.00	1.00	2.2033	1.14057	.564	-1.0747	5.4814
	2.00	.9800	1.14057	1.000	-2.2981	4.2581
	4.00	-.6067	1.14057	1.000	-3.8847	2.6714
	5.00	-1.4100	1.14057	1.000	-4.6881	1.8681
4.00	1.00	2.8100	1.14057	.156	-.4681	6.0881
	2.00	1.5867	1.14057	1.000	-1.6914	4.8647
	3.00	.6067	1.14057	1.000	-2.6714	3.8847
	5.00	-.8033	1.14057	1.000	-4.0814	2.4747
5.00	1.00	3.6133*	1.14057	.021	.3353	6.8914
	2.00	2.3900	1.14057	.388	-.8881	5.6681
	3.00	1.4100	1.14057	1.000	-1.8681	4.6881
	4.00	.8033	1.14057	1.000	-2.4747	4.0814

Table 4.46 gives level of significance for the differences between the individual Factors. The column "Mean Difference (I-J)" gives the value of differences between the means of the Factors. It can be observed from the table that only Factor 1 and Factor 5 had the sig. value < 0.05, which indicated that Factor 1 was significantly different from Factor 5 only. All other factors were on par with Factor 1. Thus, it can be inferred

that there was an overall significant difference over 15 time periods (years) between Factor1 and Factor5.

The aim of this study was to analyse the repeated measures data obtained from 100 cocoa trees for the period from 2003 to 2017 by forming 5 factors, grouping trees with homogeneous yields and estimating the between factor effect over time. General Linear Model (GLM) was used to analyse the repeated measures data. The Multivariate tests displayed the results that both the model effects viz; Time and Time*Factor interaction effects were significant for all the four tests. From the table of tests of within-subject effects it was found that the significance value of Greenhouse-Geisser test for both the Time and Time*Factor interaction was 0.00 ($p < 0.05$) which explained that the mean value of yearly cocoa pods for different groups over different time periods from 2003 to 2017 were significantly different. The value of partial eta square 0.327 for time effect indicated that 32.7% of variation in mean value of yearly cocoa pods at different time points was contributed by main effect of Time and 14.1% of variation was contributed by Time*Factor interaction keeping all other variables fixed.

From the table of tests of between-subjects effects it was found that there was an overall significant difference between the factors and the multiple comparison test obtained by Bonferroni post hoc test revealed that the significant difference occurred only between the Factor 1 and Factor 5. The mean value of yearly cocoa pods obtained for 15-time intervals from the cocoa hybrids which came under Factor 1 and Factor 5 showed significantly different mean value of yearly cocoa pods.

4.2.3 Probability distribution

The frequency of monthly infected pods observed for 100 cocoa hybrids were subjected for fitting a probability distribution in Easyfit software. 'EasyFit' is a data analysis software allowing to fit probability distributions to sample data and to select the best model. A perusal of percentage number of monthly infected cocoa pods is given as box plot in Fig 4.22.

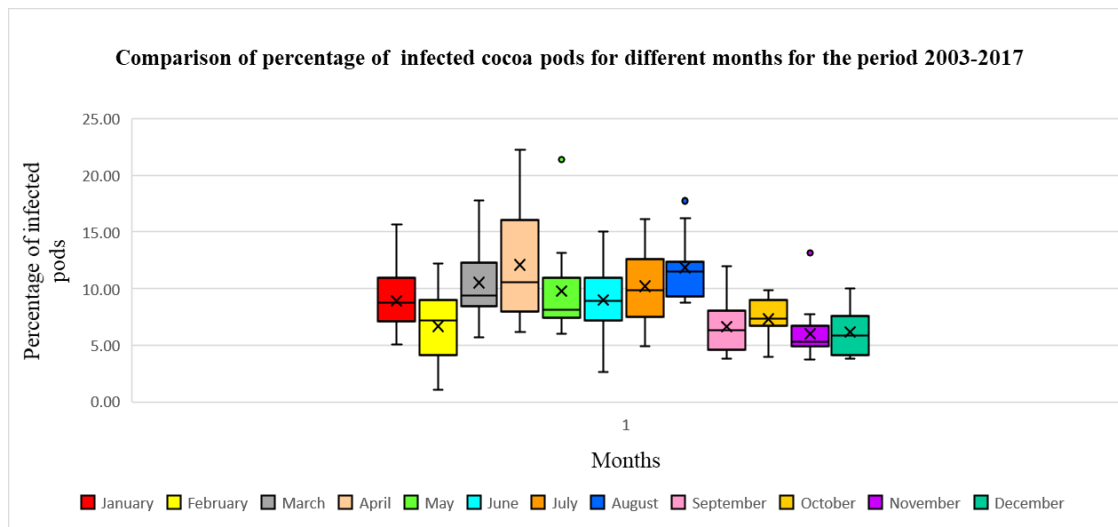


Fig. 4.22 Box plot for percentage of infected Cocoa pods

The boxplot was drawn based on the monthly data of percentage of infected cocoa pods out of the total pods for the period from 2003 to 2017. The monthly data of number of cocoa infected pods was recorded from the 100 cocoa hybrids.

The percentage of infected cocoa pods is calculated by the formula;

$$\text{Percentage of infected cocoa pods} = \frac{\text{number of infected cocoa pods}}{\text{total number of cocoa pods}} \times 100$$

From the boxplot (Fig.4.22) it can be observed that the distribution of the data for most of the months was not symmetric. The median was not in the middle of the boxplot instead it was closer to the bottom of the box or closer to the top of the box. The mean in the boxplot for the month January, March, April, May, July, August, September, November and December was above the median and the whisker was shorter on the lower end of the box. Thus, the distribution of infected pods for those months were positively skewed. Whereas, the mean in the boxplot for the month February and October was below the median and the whisker was shorter on the upper end of the box and the distribution for those months were negatively skewed. While the boxplot for the month of June was coinciding with the median line and the length of the whisker

seemed to be balanced at both the ends and hence the percentage of distribution of infected cocoa pods was symmetric.

The maximum value of percentage of infected pod was observed in the month of August followed by that of April and July. The minimum value of percentage of infected pods was observed in the month of November.

The variability in the data set is best explained by the interquartile range (IQR). The dispersion in the data is the extent to which a distribution is stretched or squeezed. The interquartile range (IQR) in the boxplot showing the middle 50% of the data points can be calculated by subtracting first quartile from the third quartile. When the data is skewed, we prefer interquartile range as the best to measure the variability than the standard deviation. The boxplot for the month of April was highly stretched which showed that the data points were highly variable or scattered. The boxplot for the month of November seemed to be compressed which showed that the distribution of data points was squeezed or concentrated.

There are extreme outliers present beyond the upper whisker and the lower whisker. The months of May, August and November months showed outliers beyond the upper whisker.

Goodness of Fit test

The goodness of fit (GOF) tests measure the compatibility of frequency of infected pod data with a theoretical probability distribution function. In other words, these tests show how well the distribution is fitted to the input data. 'EasyFit' calculates the GOF statistics for each of the fitted distributions and it ranks the fitted distributions. This test is conducted to decide which distribution describes the data in a best way.

Table 4.47: Goodness of fit test

Distribution	Kolmogorov Smirnov		Anderson Darling	
	Statistic	Rank	Statistic	Rank
Geometric	0.192	1	1.411	1
D. Uniform	0.294	2	19.463	3
Poisson	0.339	3	15.7	2

From Table 4.47, it was observed that ‘EasyFit’ enabled to allot the rank for the fitted distribution with respect to the two goodness of fit tests viz; Kolmogorov Smirnov and Anderson Darling test. The values of statistic for the two goodness of fit test was produced for fitted distributions. The Geometric distribution was given rank 1 in both the test conveying that Geometric distribution was the best fit to the frequency of monthly infected pod data. From the results of goodness fit test, geometric distribution could suitably fit to the data with the parameter p (probability of success) equal to 0.192.

Hypothesis Testing

The null and the alternative hypotheses were:

- H_0 : the data follow the specified distribution;
- H_A : the data do not follow the specified distribution.

Table 4.48: Summary of Kolmogorov-Smirnov test

Kolmogorov-Smirnov test					
Distribution	Geometric				
Sample Size	43				
Statistic	0.19283				
P-Value	0.07106				
Rank	1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.159	0.182	0.202	0.226	0.243
Reject?	Yes	Yes	No	No	No

The D value for the Kolmogorov-Smirnov test was 0.192. The null hypothesis was rejected at significance level ($\alpha = 0.2$ and 0.1) where the D value was greater than the critical value. While the null hypothesis was accepted at significance level ($\alpha = 0.05$, 0.02 and 0.01) where the test statistic D was smaller than the critical values. Therefore, from the Kolmogorov-Smirnov test the frequency of monthly infected cocoa pod data followed Geometric distribution at significance level ($\alpha = 0.05$, 0.02 and 0.01)

Based on the P value, the null hypothesis was evaluated. It was calculated based on the test statistic. The null hypothesis was accepted at significance level ($\alpha = 0.05$, 0.02 and 0.01) less than the P value.

Table 4.49: Summary of Anderson Darling test

Anderson Darling test					
Distribution	Geometric				
Sample Size	43				
Statistic	1.41				
Rank	1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.37	1.92	2.50	3.28	3.90
Reject?	Yes	No	No	No	No

Unlike Kolmogorov-Smirnov test the critical values of Anderson Darling test statistic depends on the specific distribution going to be tested. Here the critical value depends on sample size which was calculated by using the approximation formula. The hypothesis was rejected at significance level ($\alpha = 0.2$) where the statistic A^2 (1.41) was greater than the critical value. While it was accepted at significance level ($\alpha = 0.1$, 0.05 , 0.02 and 0.01) where the statistic A^2 (1.41) was greater than the critical value.

Whereas, when we look at the Goodness of fit test for other distributions (Discrete uniform and Poisson) mentioned in Table 4.47, the null hypothesis was rejected at all level of significance ensuring that Geometric distribution was best fit to the data.

Geometric distribution

From the results it was stated that the distribution of infected pods data followed Geometric distribution with parameter p (probability of success) = 0.19. The geometric distribution is a special case of negative binomial distribution. In negative binomial distribution when we consider the number success (r) equal to 1 it leads to geometric distribution.

The probability mass function of a negative binomial distribution is given by

$$p(x) = P(X=x) = \binom{x+r-1}{r-1} p^r q^x; x = 0, 1, 2, \dots$$

If we take $r = 1$, we have

$$p(x) = q^x p; x = 0, 1, 2, \dots \text{ (probability function of geometric distribution)}$$

Table 4.50: Parameters of Geometric distribution

Parameters of Geometric distribution	
Descriptive measures	Values
Mean ($\frac{1-p}{p}$)	4.18
Mode	0
Variance ($\frac{1-p}{p^2}$)	21.709
Standard deviation	4.65
CV (%)	1.11
Skewness ($\frac{2-p}{\sqrt{1-p}}$)	2.01
Kurtosis ($6 + \frac{p^2}{1-p}$)	6.04
PDF	$(1-p)^x p$

The properties of the distribution showed that the parameter p (the probability of success) was 0.19. The mean of the distribution was 4.18 (approximately equal to 4) which was the monthly expected number of infected pods and it was expected to vary by about 4 (which was the standard deviation). The value of Skewness and Kurtosis showed that the Geometric distribution was positively skewed and leptokurtic

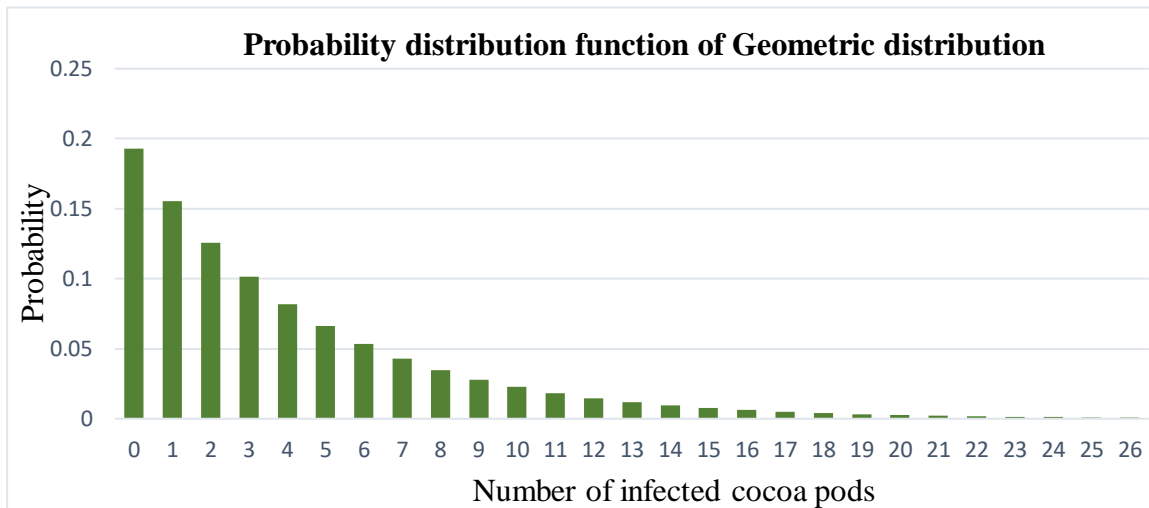


Fig. 4.23 Geometric probability distribution function of monthly infected cocoa pods

The probability distribution function of Geometric distribution fitted on the data of monthly infected cocoa pods described the probability of getting x number of infected pods in a month with the parameter p . Here the p value was 0.19.

From Fig. 4.23, it can be observed that the probability of getting zero number of monthly infected pods was 0.19. The probability of getting one infected pod was 0.16. As the number of monthly infected pods increased then the corresponding probability of its occurrence get decreased as it could be witnessed from the graph that the height of bars erected on the x axis decreased sequentially. Thus it could be inferred that the probability of getting more number of infected pods in a month is very small and the probability goes on reducing as the number of infected pods is increasing.

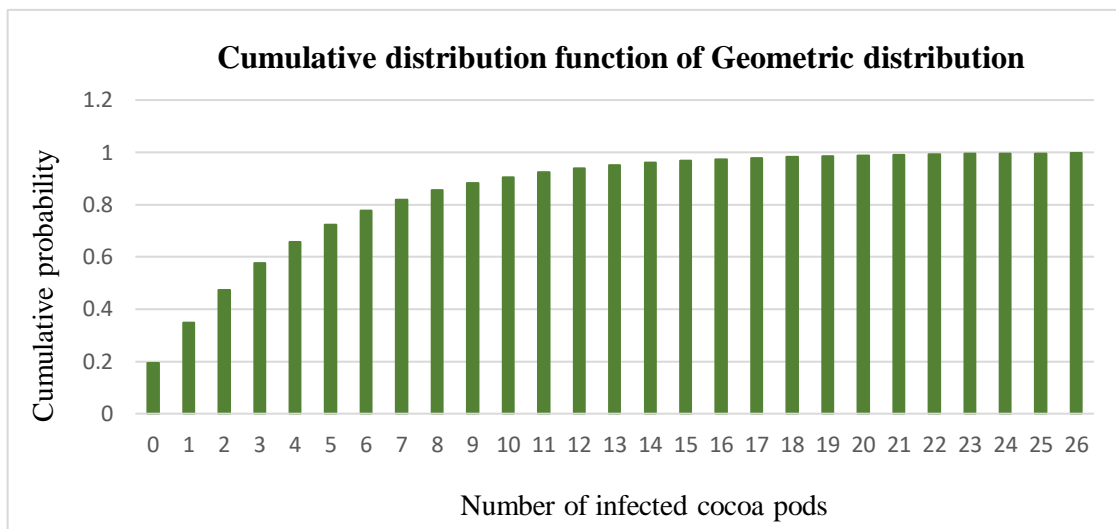


Fig. 4.24: Cumulative distribution function of Geometric distribution

The cumulative distribution function is the probability that the variable takes the value less than or equal to x . That is

$$F_X(x) = P(X \leq x) \text{ for all } x \in \mathbb{R}$$

where ,

$$\lim_{x \rightarrow -\infty} F(x) = 0 ; \lim_{x \rightarrow \infty} F(x) = 1$$

From the figure the cumulative distribution function is interpreted as

The first bar on the X axis is represented as $F_X(0) = P(X \leq 0) = 0.19$

The second bar on the X axis is represented as $F_X(1) = P(X \leq 1) = P(X=0) + P(X=1) = 0.35$

and so on. As x increased the height of the bar increased. From the graph it could be observed that for a certain value of x , the probability became 1 which was the highest and continued as such for further values of x .

The study examined the fitting of probability distribution on the frequency of number of monthly infected cocoa pods. It was revealed that the Geometric distribution was the good approximation for the monthly infected pod data taken for 100 trees. The properties of the distribution showed that the parameter p was 0.19. The mean of the distribution was 4.18 (approximately equal to 4) which was the expected number of monthly infected pods and it was expected to vary by about 4 pods (which was the standard deviation). The coefficients of asymmetry and kurtosis would read about the shape of the distribution. The realised geometric distribution was positively skewed and right tailed with skewness equal to 2.01 and with kurtosis 6.04 leading to leptokurtic distribution. The PDF and CDF of Geometric distribution showed that the information on monthly infected cocoa pod data was well fitted with the distribution and based on the goodness of fit test it was proved that the Geometric distribution was ranked 1 for the input data. Therefore, this distribution is useful for further studies on infected pods of cocoa and related areas.

4.2.4 Impact of Climatic variables on cocoa yield

The increasing number of infected cocoa pods during heavy rain period and extreme hot seasons prompt a researcher to study the impact of weather variables on cocoa yield. The five-month cumulative average for each of the weather variables viz; maximum and minimum temperature, RH1 and RH2, sunshine hours, wind speed, total rainfall and total number of rainy days were first computed. The monthly cocoa yield (number of pods per tree) from January 2003 to December 2017 were taken and the correlation coefficient was worked out using the previous five months cumulative weather variables. After identifying the significant weather variables, regression equations were tried to predict monthly cocoa yield at least one month before harvest. Multiple linear regression as well as stepwise regression were tried using the statistical software package "SPSS". Apart from this the correlation coefficients of current month's weather variables with monthly average cocoa yield for 100 cocoa trees have also been worked out. Current month's weather variables were not included for fitting regression as it won't serve the purpose of forecasts well in advance.

Table 4.51: Correlation of previous five months cumulative weather variables with average monthly cocoa yield of 100 plants

weather variables	correlation of av. monthly cocoa yield with weather variables
Maximum temperature (°C)	-.759**
Minimum temperature (°C)	-.114
RH1 %	.740**
RH2 %	.802**
Wind speed (Km/hr)	-.526**
Sunshine hour (Hrs)	-.798**
Rainfall (mm)	.762**
Rainy day (days)	.814**

** denotes significance at 1% level

From Table 4.51, it can be observed that the weather parameters such as maximum temperature, RH1, RH2, wind speed, sunshine hours, rainfall and number of rainy days showed significant relationship towards the cocoa pod production but the minimum temperature was insignificant with respect to the cocoa pod yield. RH1, RH2, total rainfall and number of rainy days were positively correlated with the cocoa yield. However, there was an inverse relationship showed by maximum temperature, wind speed and sunshine hours with the cocoa yield. The number of rainy days was highly correlated with the cocoa pod yield and the correlation was 0.81. This showed the importance of distributed rainfall over number of days rather than heavy rainfall for increasing the production of cocoa pods.

Since the minimum temperature was insignificant towards the pod production, it was not considered for fitting regression

Multiple Linear Regression

A multiple linear regression equation was fitted taking current monthly average yield of 100 cocoa plants from Cocoa Research Centre, Kerala Agricultural University, Vellanikkara raised under homogenous environmental and climatic conditions and under homogeneous cultivation practices on previous five months cumulative weather variables. The impact of weather variables was estimated taking average of previous five months weather variables as the independent variables viz; maximum temperature, RH1, RH2, sunshine hours, wind velocity, total rainfall and total number of rainy days. The regression analysis using 7 weather variables resulted in a forecasting equation with adjusted $R^2 = 72\%$.

$$\begin{aligned} \text{Cocoa yield} = & 100.96 + (-1.944) \text{ Max temp.} + (-0.261) \text{ RH1} + (-0.261) \text{ RH2} + (-0.744) \\ & \text{Wind speed} + (0.080) \text{ Sunshine hour} + (-0.012) \text{ Total rainfall} + (1.326) \\ & \text{No. of rainy days} \end{aligned}$$

Thus, it was inferred that 72% of the variation in average monthly cocoa yield could be explained through unit changes in the cumulative weather variables used as input variables in the regression equation. But the independent variables used for fitting the regression seemed to have high degree of multicollinearity. Hence a stepwise regression was attempted to extract a suitable prediction equation for cocoa yield using climatic variables.

In most of the cases the problem with the multiple regression is the multicollinearity among the independent variables of the regression model. Multicollinearity is the occurrence of intercorrelation between two or more predictor variables. Large number of independent variables in the multiple regression leads to multicollinearity. When multicollinearity exists the information's through one variable may be masked by the presence of other independent variables which have high inter correlations. So, an attempt was made to build a regression model which could provide complete information about the predictor variables.

The step wise regression was chosen to build a parsimonious forecast equation for average cocoa production using significant average weather variables of the past five months.

Table 4.52: The variables entered or removed in doing the Step wise regression

Models	Variables Entered	Prob. level	Variables Removed	Prob. level
1	Number of rainy days	0.00	Maximum temperature RH1 RH2 Windspeed Sunshine hour Rainfall	0.00 0.051 0.323 0.516 0.225 0.059
2	Number of rainy days Maximum temperature	0.00 0.00	RH1 RH2 Windspeed Sunshine hour Rainfall	0.248 0.105 0.124 0.966 0.059

There were two models (1 and 2) formed by removing the remaining non-significant variables. In model 1, the number of rainy days was chosen as the best predictor variable, which resulted in a value of adjusted R^2 of 66%.

In model 2, the number of rainy days and maximum temperature were chosen as the best predictors for which the adjusted R^2 was 69%.

Table 4.53: Coefficients of the Step wise regression for Model1 and Model 2

Model	Unstandardized coefficients	standardized coefficients	t value	sig
	B	Beta		
1. Constant	0.336		0.721	0.472
No. of rainy days	0.778	0.814	18.730	0.00
2. Constant	41.921		4.390	0.00
No. of rainy days	0.550	0.575	8.371	0.00
Maximum temperature	-1.229	-0.300	-4.359	0.00

Table 4.53 gives the coefficients of the predictor variables for Model 1 and Model 2 which had contributed significantly towards the production of cocoa pod yield. Out of the results obtained in Table 4.53, two step wise regression models could be outlined

Model 1: $Y = 0.336 + 0.778 \text{ number of rainy days}$ (Adjusted $R^2 = 0.66$)

Model 2: $Y = 41.921 + 0.550 \text{ number of rainy days} - 1.229 \text{ maximum temperature}$ (Adjusted $R^2 = 0.69$)

Table 4.54: Model summary of the Step wise regression

Model	R square	Adjusted R square
1	0.663	0.662
2	0.696	0.693

The R^2 value for the above stepwise regression equation reflected the proportion of variation in the dependent variable that can be explained by the model's input.

According to Model 1, 66% of variation in cocoa pod yield could be explained by the single weather parameter, number of rainy days having positive effect on yield confirming the importance of distributed rainfall rather than heavy rainfall obtained for previous months of harvest.

The R^2 value for Model 2 showed that 69% of the variation in cocoa pod yield could be explained by the weather parameters viz; total number of rainy days and cumulative average maximum temperature.

From the correlation and regression studies using weather variables it could be concluded that the two important weather variables cumulated for five months before harvest of cocoa were number of rainy days which had a significant positive effect and maximum temperature which had a significant negative effect on yield.

Table:4.55 Correlation of current month's weather variables with average monthly cocoa yield of 100 plants

weather variable	correlation of av. monthly cocoa yield with weather variables
Maximum temperature (°C)	-.178*
Minimum temperature (°C)	-.375**
RH1 %	-.141
RH2 %	-.010
Wind speed (Km/hr)	0.135
Sunshine hour (Hrs)	0.082
Rainfall (mm)	.762**
Rainy day (days)	-.202**

In contrast to the correlation results with respect to the previous five months average weather variables, increasing number of rainy days during the month of harvest was detrimental to the cocoa pod yield even though the rainfall had some positive effect. The increase in minimum temperature during the month of harvest also had significant negative effect.

The main aim of the correlation and regression analysis was to develop the best regression model to predict the yield of cocoa atleast one month in advance of harvest. The monthly cocoa yield data was recorded in terms of number of cocoa pods from 100 cocoa trees for a period from 2003 to 2017. Pooled weather data for the previous five months and average cocoa pod yield for the current month were used for fitting the regression model and thereby to identify the significant variables that were mainly contributing to the prediction of cocoa yield. To check for the significance of weather parameters correlation analysis was adopted, which identified all the weather parameters which were significant. Except the minimum temperature, all other previous

months cumulated weather variables considered were found to have significant effect towards the cocoa yield.

By using the significant weather variables, a suitable Multiple Linear Regression model was developed with adjusted value of $R^2 = 0.72$. But since the model showed drawbacks of multicollinearity, a step wise regression model was also attempted to develop a parsimonious forecast model which resulted in extracting two competent models to predict the cocoa pod yield effectively.

In Model 1 with adjusted $R^2 = 0.66$, the number of rainy days was identified as the best predictor variable with positive relationship with yield which showed the importance of distributed rainfall rather than accumulated heavy rainfall for cocoa pod yield.

In Model 2, the number of rainy days and maximum temperature were determined as the best predictors significantly affecting the cocoa pod yield for which the adjusted R^2 was 0.69

Therefore, cumulative previous five month's average maximum temperature which had negative correlation and total number of rainy days for previous five months of harvest which had positive correlation with yield could be considered as the most important weather factors that significantly contributed in predicting the average number of cocoa pods per tree one month in advance of harvest at Cocoa Research Centre, Vellanikkara, Thrissur, Kerala.

4.3 Empirical approach to identify the important factors perceived by farmers in cocoa production

4.3.1 Structural Equation Modelling

An empirical analysis to identify the factors perceived by farmers to influence their cocoa production was done taking a sample of 100 farmers. The survey was done in Iritty Panchayat of Kannur district and Veliyamattom Panchayat of Idukki district of Kerala. In connection with a training programme on cocoa cultivation jointly arranged by the Cocoa project “Mondelez International limited, Ernakulam” and Cocoa Research Centre, College of Horticulture, Vellanikkara, KAU these farmers were gathered at their respective places and primary data on their demographic details, cocoa cultivation and management practices, production constraints etc. were collected through a structured questionnaire. The data collected were analysed and sufficient variables were identified that maximises the cocoa yield and there by the income of farmers.

Table 4.56: Demographic characteristics of the Respondents

Variables		Frequency	Percentage
Gender	Male	68	68
	Female	32	32
Age	≤ 35 yrs	5	5
	36 – 50 yrs	26	26
	> 50yrs	69	69
Educational status-	Illiterate	1	1
	Primary school	26	26
	High school	46	46
	Intermediate/+2	16	16
	Graduate	8	8
	Post graduate	3	3
Occupational status	n=1	25	25
	2	64	64
	3	11	11
	n denotes the no. of engagements leading to income		
Land holding size (Area)	< 1 acres 1	17	17
	1-3 acres 2	61	61
	3-4 acres 3	10	10
	4-5 acres 4	5	5
	> 5 acres 5	7	7
Experience in cocoa cultivation upto	3 yrs	24	
	4 yrs	27	24
	5 yrs	18	27
	6 yrs	20	18
	above 6 yrs	11	20
Organisational membership (no. of organisations)	0	18	18
	1	39	39
	2	25	25
	above 2	18	18
Extension contact	- yes	45	45
	No	55	55
Trainings received	- yes	48	48
	No	52	52
Family Type	Nuclear -1	76	76
	Joint -2	24	24

The base model using SEM was developed using the variables viz; Experience in cocoa cultivation, lack of credit, plant protection measures, expenditure on farming, rating of quality of seedlings, squirrel, rat and civet attack, frequency of pruning, pest and disease attack, condition of pods at the time of harvest, price of cocoa, number of yielding trees, yield and Income

Table 4.57: Model fit summary of base model of SEM on cocoa production

Indices	Value	Suggested value
Chi-square value	184.113	-
DF	57	-
CFI	0.913	>0.90
TLI	0.880	>0.90
RMSEA	0.150	<0.08

Based on the goodness of fit the base model was tested for how better it fitted the data. The value of CFI, TLI and RMSEA were 0.913, 0.880 and 0.150 respectively and the value of chi-square was 184.113. Though the value of CFI was > 0.90, the value of TLI and RMSEA did not reach the expected value and also the value of chi-square was high. Hence the base model led to worse fit to the data. This was due to the fact that some of the variables were insignificant in some of the paths. Therefore, it was necessary to improve the goodness of fit and the model needed to be modified by removing insignificant variables and eliminating paths from the model or by building some more paths in the model. The other tests to improve the model was based on the modification indices test which enabled us to add or remove the paths to improve the model fit.

After the modification of the base model, a better model was resulted which showed an improvement in parameters of the model and was considered as a good SEM model for cocoa production. The generated final model is illustrated in Fig. 4.25.

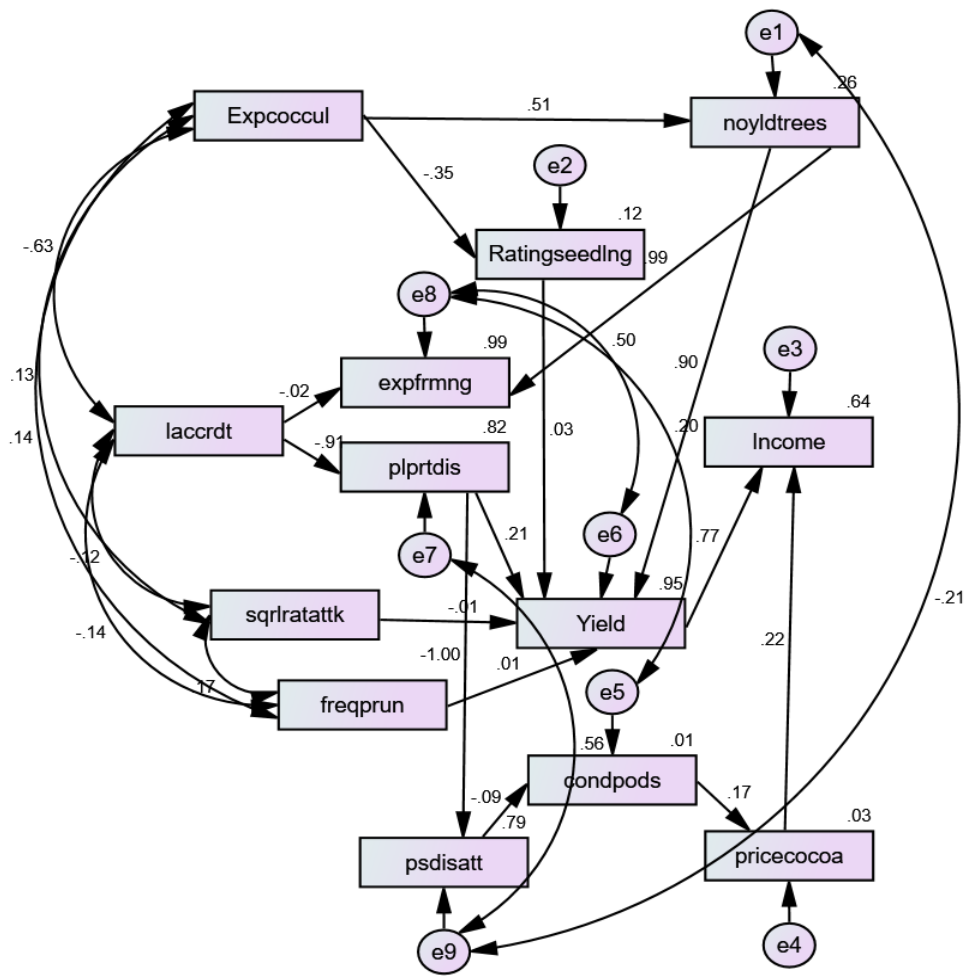


Fig 4.25: Final SEM model based on standardized coefficients for cocoa production

- Ratingseedling - Rating of quality of cocoa seedlings yield - number of cocoa pods
- Expcoccul - Experience in cocoa cultivation Income - Income of cocoa farmers
- sqrlratattk - Attack from squirrel, rat, civet etc. psdisatt - Pest and disease attack
- plprtdis - Plant protection & disease management lacrrdt - Lack of access to credit
- expfrmng - Expenditure on cocoa farming year pricecocoa - Price of cocoa
- noyldtrees - Number of yielding trees
- condpods - Condition of harvested cocoa pods
- freqprun - Frequency of pruning on cocoa trees/year

After modification, the final model was generated and illustrated in figure 4.25. The numbers near the arrows are standardised coefficients between the variables, the R^2 value for the dependent variable is shown above its rectangle in the diagram.

Table 4.58: Coefficients of variables in Structural Equation Model (SEM)

Variables			Unstandardized Co-efficients (B)	S.E of B	Standardized Co-efficients (Beta)	P-value
plprtdis	<---	laccrdt	-1.642	0.77	-0.906	***
psdisanatt	<---	plprtdis	-1.028	0.052	-1.003	***
Ratingseedlng	<---	Expcoccul	-0.202	0.055	-0.349	***
noyldtrees	<---	Expcoccul	28.691	4.808	507	***
condpods	<---	psdisanatt	-0.076	0.085	-0.089	0.371
yield	<---	noyldtrees	0.917	0.025	0.896	***
pricecocoa	<---	condpods	1.285	0.761	0.167	0.091
yield	<---	sqrlratatk	-0.929	3.188	-0.006	0.771
yield	<---	Ratingseedlng	2.943	2.075	0.029	0.156
yield	<---	freqprun	0.750	2.277	0.007	0.742
yield	<---	plprtdis	19.650	2.235	0.211	***
Income	<---	pricecocoa	187.855	52.656	0.216	***
Income	<---	yield	48.365	3.819	0.767	***
expfrmng	<---	laccrdt	-96.156	53.348	-0.015	0.071
expfrmng	<---	noyldtrees	37.711	0.329	0.991	***

*** = 000

Model coefficients and their significance

We use unstandardized coefficients to know the amount of change in the dependent variable due to a unit change in the independent variable. Unstandardized coefficients are the raw coefficients produced by the data in a regression analysis. Standardised coefficients also can be obtained as path coefficients.

From Table 4.58, Unstandardised coefficient of lacrdt (lack of credit) on plprtdis (plant protection & disease management) was -1.642 which represented the partial effect of lack of credit, holding the other path variables as constant. The estimated negative sign implied that lack of credit had negative impact on acquiring the plant protection measures. Since the farmers were facing the problem of access to credit and it would have led to lack of capital and they might be reluctant to apply improved quality fertilizers and plant protection measures which in turn resulted in reduced cocoa yield. The purchasing power of plant protection aids by the farmers decreased by 1.64 for every unit decrease in the access to credit and the coefficient value was significant at 5% level. Also it could be observed that when there was unit increase in lack of credit it has got a significant negative impact (-96.156 as the regression coefficient) on the farmer's willingness to expend more on farming and when the number of yielding trees increased the expenditure also would get increased (37.711 as the regression coefficient) and a small holder farmer may not be able to cope up with those situations when there was no access to credit.

Unstandardised coefficient of plprtdis (plant protection and disease management) on psdisanatt (pest and disease attack) was -1.028 which represented the partial effect of plant protection and disease management measures, holding the other path variables as constant. The estimated negative sign implied that, the attack of pest and disease on cocoa plants would decrease by 1.02 for every unit increase in plant protection and disease management measures and this coefficient value was significant at 5% level.

Similarly the Unstandardised coefficient of Expcoccul (experience in cocoa cultivation) on rating of quality of seedlings was -0.202, Unstandardised coefficient of psdisanatt (pest and disease attack) on condpods (condition of pods) was -0.076 and Unstandardised coefficient of sqrlratatk (squirrel,rat,civet etc. attack) on yield of cocoa was -0.929 which represented the negative effect of one variable on the other variable.

The Unstandardised coefficient of Expcoccul (experience in cocoa cultivation) on noyldtrees (number of yielding trees) was 28.69 which showed that as the number of years of experience on cocoa cultivation increased it would naturally increase the age of the trees and number of yielding trees also get increased. Unstandardised coefficient of number of yielding trees on yield of cocoa was 0.9. It is evident that the yield would increase according to the increase in number of yielding trees. Unstandardised coefficient of condition of pods on price of cocoa was 1.28 which showed that for a unit increase in level of condition of pods the price would increase by 1.28 unit. Unstandardised coefficient of rating of seedling on yield of cocoa was 2.94, Unstandardised coefficient of frequency of pruning on yield of cocoa was 0.75. When the level of plant protection and disease management measures increased by one unit it had a significant positive effect on yield of cocoa by a unit of 19.65. For a unit increase in the level of price of cocoa there would be an increase of 187.8 unit in farmer's income. For a unit increase in the yield of cocoa there would be an increase of 48.36 unit increase in farmer's income.

The selected model resulted in an R^2 value of 0.95 for yield and 0.64 for income from cocoa cultivation. Hence 95 percentage variation in cocoa yield and there by 64 percentage variation in the income generated from it could be explained by the final model.

The regression of pest and disease attack on cocoa pod, condition of pods on price of cocoa, rodent attack on cocoa yield, rating of quality of seedling on cocoa yield, frequency of pruning on yield and lack of credit on farming expenditure were having P value > 0.05 which implied that those variables were insignificant at 5% level of

significance even though they were expected to produce significant effect. But those variables were still retained in the model because those variables were also treated as dependent and independent variables in some other pathways where it was significant and also acted as intervening variables between the independent and dependent variables. So, removal of those variables from the model would affect the other significant variables and it would not make sense from the theoretical aspect of view.

Modification of the fitted model using modification indices

Table 4.59: Covariances between the independent variables and between the error terms in Structural Equation Model (SEM)

Variables / errors	Estimate
laccrdt <--> Expcoccul	-.444
Expcoccul <--> sqrlratatk	.102
Expcoccul <--> freqprun	.152
laccrdt <--> freqprun	-.051
laccrdt <--> sqrlratatk	-.033
sqrlratatk <--> freqprun	.070
e9 <--> e7	.089
e1 <--> e9	-6.143
e8 <--> e6	2525.221
e8 <--> e5	41.118

When the parameters of the modified model was compared with the base model there was an improvement in the model. Based on the modification indices test, the modification of the model was done which provided a satisfactory fit and was used for further analysis and interpretation. A larger chi-square value for a model indicated that the model was not a good fit. The modification indices (also called LaGrange multiplier or Score test) is an estimate of the amount by which the chi-square will be reduced where there will be as many modification indices as imposed restrictions. It enables us to add or remove the paths to improve the model fit. The error terms which showed the high modification indices values in the model were joined by two head arrow

(covariance) and repeated the same process until modification indices were on par where all the error terms were cleared with modification indices value.

Covariances between the independent variables or between the errors of dependent variables determines the non-causal connections between the respective variables. Standardized residual covariances are much like modification indices. Modification indices identify the discrepancies between the proposed and estimated models. Hence, we use modification indices for the covariances.

From Table 4.59 it can be observed that the covariances between the errors of dependent variables (e9 and e7, e1 and e9, e8 and e6, e8 and e5) drawn based on the values of modification indices has improved the model.

For the purpose of testing the model fit, null hypothesis and alternative hypothesis were framed as

Null hypothesis: The hypothesized model had a good fit.

Alternate hypothesis: The hypothesized model did not have a good fit.

Table 4.60: Model fit summary of SEM

Indices	Value	Suggested value
Chi-square value	108.027	-
DF	53	-
CFI	0.962	>0.90
TLI	0.944	>0.90
RMSEA	0.102	<0.08

From Table 4.60, it can be seen that the value of CFI, TLI and RMSEA are 0.962, 0.944 and 0.102 respectively. The values indicated that the model was satisfactorily fitted. The value of CFI and TLI satisfied the suggested value which was > 0.90 but the value of RMSEA was > 0.08.

The value of RMSEA is calculated as

$$\text{RMSEA} = \frac{\sqrt{(X^2 - df)}}{\sqrt{df (N - 1)}}$$

Where

X^2 – Chi-square value (108.027)

df – degrees of freedom (53)

N – sample size (100)

The value of RMSEA is sensitive to degree of freedom and sample size. When the degrees of freedom is higher and the sample size is larger, the value of RMSEA is smaller. It produces better result when the sample size is adequately large. When the sample size is large, the term $[1/ (n-1)]$ gets closer to zero. Since the sample size is small and due to the low degrees of freedom the value of RMSEA obtained was 0.102. But still the model was considered to be productive because the estimate for the other two parameters supported better fit of the model. MacCallum et al, 1996 suggested that the RMSEA less than 0.08 show an absolutely good fit model. Some important variables like details of drying and fermentation of cocoa beans had not been included in the model explained in Fig.4.25 as most of the farmers were not having the facilities for fermentation or drying up of cocoa beans which in turn have decreased their income.

Hence, the Null hypothesis was accepted that the hypothesized model had a good fit.

A brief discussion has been made to illustrate the application of structural equation modelling to study the causal relationship of the critical factors of cocoa production that influenced the income of cocoa famers of Kerala state based on the views of the cocoa farmers of Veliyamattom of Idukky district and Iritty of Kannur district of Kerala. The SEM model was developed to study the interdependence of factors related to demographic details, cocoa cultivation and management practices and production data and to evaluate how these variables influenced the income of cocoa farmers.

From the SEM model, 8 structural equations were generated and the results showed that price of cocoa was the most influencing variable on the income of cocoa farmers. With the increase of price of cocoa by 1 unit, the income of the cocoa farmers would increase by 187.85 units, leading to the conclusion that fixing the price of cocoa beans was an important factor. When the cocoa beans are sold at good price it stabilizes the economic condition of the farmers, even the government and other manufacturing companies should involve in fixing the remunerative prices. The price of cocoa beans get stimulated if it is sold as dried or in fermented form. But most of the farmers lacked drying and fermentation facilities. The second most influencing variable was the yield which would increase the income of the farmers by 48.36 units which was undoubtedly true that increase in the cocoa production would directly affect the income of farmers. For increasing the yield there were several influencing factors such as good quality seedlings, plant protection and disease management measures, protection from attack of squirrel, rat, civet etc. which would indirectly affect the condition of pods as the farmers would make an early harvest of cocoa pods otherwise because of the threat of rodent attacks. The damage caused by rodents and pest and disease attack should be taken care to increase the yield.

It is quite natural that lack of access to credit would lead to lack of capital and it was a critical factor that indirectly affected the application of fertilizers and plant protection measures adequately. The farmers would be reluctant to expend more on plant protection measures if they lack capital. The farmers were denied to access the credit since they lacked information on how to get the credit. The farmers who just started and were new to the cocoa farming were facing the problem of credit accessibility. The experienced farmers expressed the importance of quality planting materials. The farmers expect that government should take initiatives to provide them better quality seedlings so that they can ensure better produce from it. Some farmers were even collecting seedlings from other local farmers based on the high yielding performance of cocoa trees with them, but since in the case of cocoa, nothing could be assured about the next generation of hybrids, care should be taken to assess the quality of seedlings when they are procured. Pest and disease attack would affect the condition of pods and

it was having a negative effect on quality of cocoa beans. Management of pest and disease is an important mechanism where the non-infected pods can be used further for fermentation purpose and the quality of cocoa beans can be improved.

In short, the Structural Equation Model (SEM) is an important statistical frame work which can be used to represent complex relationships between the observed and unobserved (latent) variables in a diagrammatic path way and can be effectively employed to study the factors affecting yield of cocoa and there by the income of farmers. It is an advanced method of regression analysis which solves systems of several linear equations simultaneously. The ordinary multiple linear regression analysis has got several limitations as multiple dependent or outcome variables are not permitted, mediating variables cannot be included in the same single model as predictors, each predictor is assumed to be measured without error, the error or residual variable is the only latent variable permitted in the model, multicollinearity among the predictors may hinder result interpretation etc. Amos can fit models that are not subject to theses limitations. Thus, SEM has become more popular in recent times to study the interdependence of variables involved in different crop production programs also.

4.3.2 Assessing the Yield gap

Yield gap is the difference between potential yield and national average yield. The Potential yield is defined as the yield when the crop variety or hybrid was grown under controlled conditions without growth limitations of water and nutrients and without pests and diseases problems. In this study the potential yield of cocoa is estimated in terms of dry bean weight per tree per year. The experimental yield potential of cocoa is 4 kg dry bean weight per tree per year and the national average yield is 2.5 kg dry bean weight per tree per year. Thus, the yield gap of cocoa is obtained by subtracting national average yield from the potential yield.

Table 4.61: Estimation of Yield gap of cocoa in Kerala

Estimated values	Yield potential (Kg/tree/year)	National average yield (Kg/tree/year)
Estimated cocoa yield (dry bean weight in Kg per tree per year)	4	2.5
Yield gap (dry bean weight in Kg per tree per year)	1.5	-
Percentage Yield gap to potential (%)	37.5 %	-

Table 4.61 provides the yield gap and their proportion to yield potentials. The results showed that the national yield gap is 1.5 Kg dry beans per tree per year, accounting for 37.5% of yield gap to the cocoa yield potential.

The study revealed that the national yield gap of cocoa is 1.5 Kg dry beans per tree per year, accounting a gap of 37.5% to reach the cocoa yield potential which need some great attention.

4.3.3 Probit regression model

A probit regression model was performed to identify the maximum likelihood estimates of parameters for decision making by cocoa farmers in Kerala to adopt plant protection measures or not

Table 4.62: Probit model on decision making to make use of Plant protection measures in cocoa cultivation

Variables	Coefficient	Std.Error	P value	Marginal effect
Constant	-10.82	2.66	<0.0001 ***	
Age	0.97	0.42	0.021 **	0.150
Education	0.08	0.21	0.706	0.012
Occupation	0.05	0.38	0.889	0.008
Family size	0.21	0.47	0.656	0.032
Land holding size	0.79	0.34	0.02 **	0.123
Experience in cocoa cultivation	1.26	0.29	<0.0001 ***	0.194
Membership in Organisations	0.59	0.21	0.0054 ***	0.092
Frequency of Extension contact	0.89	0.46	0.051 *	0.134

***, ** and * refers to statistical significance at 1%, 5% and 10% level.

Table 4.63: Diagnostic measures of the Probit regression model

Log likelihood	-28.35
McFadden R-squared	0.547
Adjusted R-squared	0.404
Schwarz criterion	98.163
S.D. dependent variable	0.468
Akaike criterion	74.716
Hannan-Quinn	84.205

In Table 4.63 the diagnostic measures showed an excellent fit of the model developed. The percentage of the cases correctly classified was 85%.

From Table 4.62 it can be assessed that the factors that had significant positive influence on the farmer's decision to adopt the plant protection measures were Age, Land holding size, Experience in cocoa cultivation, membership in organisations and frequency of Extension contact whose marginal effects were statistically significant.

The age of the farmers was positively related to the adoption of plant protection measures and it was significant at 5% level. A unit increase in the level of age improved the probability of adoption of plant protection measures by 0.150 keeping all the other variables at a constant level. It might be resulted through the knowledge gained by the farmers from their experience and in the present study there were adequate number of farmers with age group greater than 50.

Land holding size had a positive relationship with the decision-making process of adoption of plant protection measures which was significant at 5% level. The result implied that an increase in the level of land holding size by one unit would increase the probability of adoption of plant protection measures by 0.123 keeping all other variables constant. It is quite natural that the farmers with huge land size would be

financially sound and might not be facing any problem of lack of capital and would be ready to expend adequate amount on plant protection measures. Also, the farmers with large area of cultivation might be more conscious about the recommended plant protection measures.

Experience in cocoa cultivation exhibited a positive relationship with the decision making of adoption of plant protection measures which was significant at 1% level. The result showed that an increase in the level of experience of cocoa farmers by one unit would increase the probability of adoption of plant protection measures by 0.194 keeping all other variables constant. When the farmers became experienced, they were most likely to use the plant protection measures resulted through their own experience and by the interaction with other farmers.

Organisational membership score had a positive effect on the decision making for adoption of plant protection measures which was significant at 1% level. Increase in this score by one unit would increase the probability of adoption of plant protection measures by 0.092. It means that as the farmers get attached to some organisations like Krishibhavan, farmer's club, co-operative society, Banks SHGs etc. either as a member or an office bearer regularly they would get more chances to interact with experienced farmers and officials and got trained themselves to know the importance of using plant protection measures.

Frequency of Extension personnel contact showed a positive relationship with the decision making of adoption of plant protection measures which was significant at 10% level. This indicated that the extension personnel has got an important role in imparting basic knowledge about the cultivation practices of crops to increase the productivity. The marginal effect of this factor was found to be 0.134.

The study identified the significant factors that influenced the cocoa farmers to adopt the plant protection measures. It revealed the marginal effect of a factor that would influence the cocoa farmers to make decision to adopt plant protection measures. The results showed that age, land holding size, experience in cocoa cultivation, membership

in various organisations and extension contact were the significant factors that influenced the cocoa farmers to adopt the plant protection measures.

As plant protection measure is an important management practice which would help to reduce pest and disease attack, ultimately leading to enhanced yield of the crop, these results would give a better insight to improve the application of those measures. The study discussed some of the important factors that influenced the cocoa farmers of Kerala to adopt the plant protection measures. Thus, the Probit analysis helps the researcher to identify the probability that an entity with a particular characteristic would fall into either the class of adopters or non-adopters of plant protection measures.

4.3.4 Major factors influencing the cocoa production as perceived by cocoa farmers

To study the important factors which influenced the cocoa production and income generated through it, the farmers were asked to rank statements listed under different heads according to their importance by first giving rank1 to the most important, rank 2 to the second most important etc.

The hypotheses in the study were stated as follows:

H0: There was no agreement among the cocoa farmers in ranking the statements

H1: There was agreement among the cocoa farmers in ranking the statements

The major heads and the statements selected as important by cocoa farmers were

1. Production and labour related

PC1 - Procurement of superior planting materials

2. Control of Pest and disease attack

PS1 - Black pod disease

PS2 - attack of rat, squirrel, civet etc.

3. Financial constraints

FC1 - Difficulty in securing working capital

FC2 - Insufficient financial assistance from financial institutions

4. Marketing constraints

MC1- Low price for the produce

MC2- Lack of fermentation facility

MC3 -Lack of drying facility

5. Information and publicity

IF1 - Knowledge about cultivation practices

IF2 - Sufficient training and demonstration

Table 4.64: Ranks of different factors perceived by farmers

Factor	Mean Rank	Rank
PC1	2.06	1 st
PS1	3.06	2 nd
PS2	3.82	3 rd
PS3	4.59	4 th
FC1	4.82	5 th
FC2	6.65	6 th
MC1	6.88	7 th
MC2	7.65	8 th
MC3	8.41	9 th
IF1	8.65	10 th
IF2	9.41	11 th

Table: 4.65 Test statistic of Kendall's coefficient of concordance

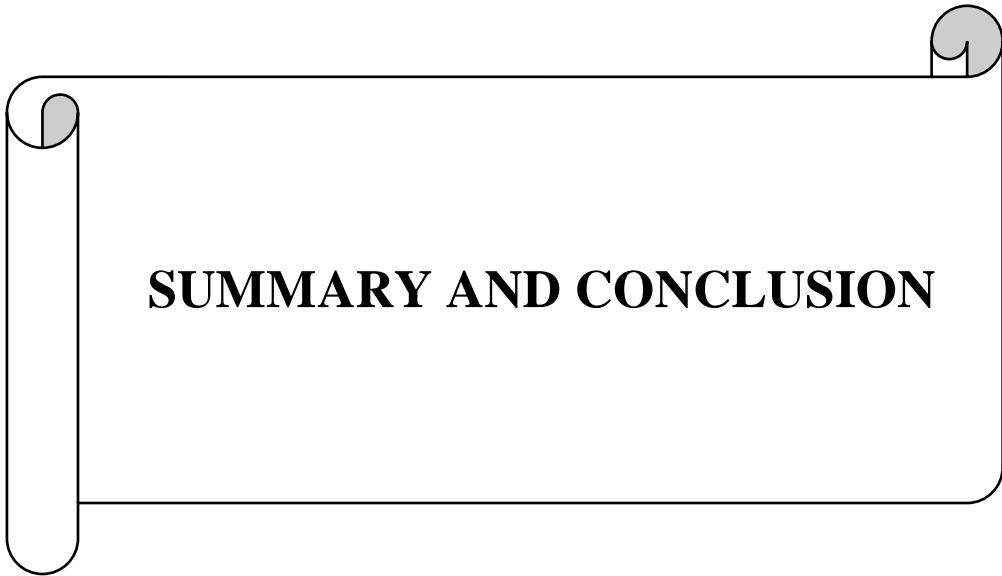
Sample size (N)	100
Kendall's W	0.552
Chi-Square	551.62
Degrees of freedom	10
Asymptotic significance	0.00

The value of Kendall's coefficient of concordance obtained in the analysis was 0.552 with a chi-square value of 551.62 which was significant at 1% level of significance. Hence there was high degree of agreement among the cocoa farmers in ranking the statements to identify most important factors which influenced cocoa production leading to their net income. The statements were ranked according to their importance as (1) procurement of superior planting materials (2) The second most important factor ranked by the farmers was protection of crops from pest and disease attack. The high level of yield loss in cocoa cultivation was due to Black pod disease (PS1), attack by rat, squirrel and civet (PS2) and attack by Tea mosquito bugs (PS3). The third ranked statement was related to financial constraint which included difficulty in securing working capital and insufficient financial assistance from financial institutions. Next was the marketing constraints such as low price for the produce (MC1), lack of fermentation facility (MC2) and lack of drying facility (MC3). Next factor was related to acquiring sufficient knowledge about cultivation practices and getting sufficient training and demonstration.

Farmers of Kerala showed much interest in the cultivation of cocoa since 1980. It has become an important source of livelihood for many of the small holder cocoa farmers. So, an attempt was made to find out the important factors identified by the farmers which would lead to enhanced production.

Results from the above analysis indicated that, procurement of superior planting materials was the most important factor and played a vital role in increasing the cocoa yield. According to the opinion of cocoa farmers government should take initiatives to supply superior quality seedlings. Regarding the pest and disease attack, the cocoa farmers are expecting high yield loss since the cocoa plant is more susceptible to pest

and disease right from the beginning of sowing of seeds and most of them have agreed that it was the second important factor leading to their income. It is known that the cocoa pod is the economic part of cocoa plant and most of the yield loss was occurred through the attack of black pod disease, tea mosquito bugs and the damage caused by some of the rodents. The pest and disease management is one of the most critical components to maintain the plant health. Many of the cocoa farmers are poor enough and facing financial problems and it was treated as a constraint. In order to overcome this, financial assistance to the farmers from the financial institutions is required which helps the farmers to access credit and secure working capital. Lack of availability of proper marketing facilities and lack of knowledge on post-harvest management are other important constraints of cocoa farmers. A stable market with good facilities will encourage the farmers towards the increase of cocoa production. Finally, the information and publicity on cocoa production was also important because of the fact that proper education and training about the cocoa cultivation practices would lead to better production. Thus, the study investigated the important factors perceived by the cocoa farmers of Kerala by considering the opinions of 100 respondents. Results showed that the value of Kendall's coefficient of concordance obtained in the analysis was 0.552 which stated that the respondents had high degree of agreement to rank the statements according to their importance.



CHAPTER 5

SUMMARY AND CONCLUSION

Over the years, cultivating cocoa has made several farmers financially independent and helped transform their livelihoods. A well-maintained cocoa intercrop farm doubles the income for the farmer if the recommended or best cultivation practices are adopted. Government of India has started a cocoa life program with an objective to create thriving cocoa growing communities. It also aims to support cocoa research at agricultural universities in India, communicating the benefits of growing cocoa to farmers, educating the farmers on good cultivation practices for cocoa and giving support on post-harvest operations like fermentation, drying and storage.

Cocoa is one of the estate commodities that play an important role in export earnings and employment opportunities. It is the main source of income for a number of small holder farmers. The importance of cocoa cultivation and the resulting income generation to the farmers motivate to implement new research areas on this crop. Apart from the trend in area, production and productivity of cocoa, the performance evaluation of cocoa in research stations as well as in farmers field conditions are equally important. The views of farmers to identify the factors which influence the production of cocoa which would ultimately lead to their gross income are also essential. In this context, a multiphase analysis of cocoa production in Kerala was made with the specific objectives such as (1) To predict the area, production and productivity of cocoa in Kerala (2) To study the impact of weather factors on yield (3) To assess the yield gap (4) To delineate the factors influencing farmer's decisions on cultivation practices and to develop yield prediction models through structural equation modelling.

To forecast the area, production and productivity of cocoa in Kerala, the respective time series data for 37 years from 1980 – 2017 were used. The maximum area under cocoa was in 1980 and the highest production was 7507 tonnes in 2017 showing an increasing trend. Kerala is having higher productivity of cocoa when compared to other states and the highest productivity was 0.64 tonnes/ hectare in 1994.

The area under cocoa showed a significant quadratic trend. The Holt's exponential smoothing model was chosen for prediction of area under cocoa with an adjusted $R^2 = 0.943$ depicting that 94.3% of the variation in area under cocoa could be captured by the model. The MAPE was 5.21. The forecasted area under cocoa came to be 20171.79 ha in 2022 in Kerala.

In the case of production of cocoa in Kerala, ARIMAX (0,1,0) model resulted as the best with an adjusted $R^2 = 0.837$ and MAPE= 0.823. For the year 2022, the predicted figure for cocoa production touches 7631.20 tonnes in Kerala.

The simple exponential smoothing model turned out to be the best to predict productivity of cocoa in Kerala with an adjusted $R^2 = 0.838$ and MAPE= 10.47. The predicted value of productivity of cocoa appeared to be 0.45 tonnes/ha in 2022. According to the statistics released by Directorate of Cashewnut and Cocoa development, the area under cocoa in Kerala was about 11044 ha in 2010-'11 and it has spread out to 15894 hectares in 2016-'17. India ranks eighteenth in cocoa production with majority of the cultivation in Kerala, Karnataka, Andhra Pradesh and Tamil Nadu. The overall productivity of the nation is still found to be low (0.2 MT) which remained same for the last few years.

Cocoa is the most important plantation crop grown in Kerala. Kerala ranked third in area of cultivation of cocoa and Tamil Nadu ranked first. Coming to the production Andhra Pradesh ranked second and then followed by Kerala. The productivity witnessed in Kerala was 0.45 tonnes/hectare but remained constant for several years. Hence efforts are needed to increase the area, production and productivity to meet the growing demand from foreign countries. India's potential to achieve self-sufficiency in cocoa is not exploited due to low involvement of capable farmers in cocoa cultivation.

Performance evaluation of cocoa trees maintained in the Cocoa Research Centre, KAU was made by collecting the monthly yield data in terms of number of cocoa pods and the monthly number of infected pods from 100 hybrids of same age.

The descriptive statistics showed that the average monthly yield was maximum during the month of November (18.14) with a S.D of 1.85 and C.V =10.17 and least in the month of May (2.97). The CV was lowest in the month of November which assured more consistent yield during this month.

Since the study dealt with a perennial crop where repeated measurements on yield were done on a monthly basis for the same tree, the general linear model repeated measures one-way ANOVA was performed which would help detection of significant difference between factors eliminating the time effect. The 100 trees were grouped into 5, with homogenous trees within a group by grouping the hybrids according to the yield in the ascending order of magnitude. After performing GLM repeated measures ANOVA a significant difference between the 1st (lowest yield group) and 5th group (highest yield) was noticed. A significant time x factor interaction was also found to exist. The peak average yearly yield was 106.23 number of pods attained during the fifth year of harvest. A biennial tendency was also found to exist with respect to yearly average cocoa yield.

A wide variation was realised in the average monthly yield data of 100 cocoa hybrids. So the average monthly yield data was subjected to SARIMA model building. The best model extracted was SARIMA (1,0,0)(1,1,0)₁₂ with an adjusted $R^2 = 0.92$ and MAPE =13.96 to predict the average monthly yield of cocoa for future months in the Cocoa Research Centre, KAU. Thus, it can be concluded that seasonal ARIMA models can be very effectively used to forecast the average monthly yield of cocoa.

An attempt was made to study the pattern of distribution of monthly infected pods of cocoa in the Research Centre. The total number of infected pods for the 100 trees were taken and subjected to fitting of probability distribution. The Geometric distribution proved to be the best to provide the probability distribution on the frequency of number of infected pods. The mean number of infected cocoa pods came out be 4 in a month.

The losses caused by black pod disease, Tea mosquito bugs etc. prompted to make a conclusion that those disease incidences were the outcome of external factors such as temperature, rainfall, relative humidity etc. So, a study was made to identify the significant contribution of climatic variables on cocoa yield. For this purpose, accumulated weather variables pertaining to previous five months of harvest were taken and correlations with yield were worked out. Average maximum temperature, average sunshine hours and average wind velocity were having significant negative correlation with yield whereas RH1, RH2, total rainfall and total number of rainy days were having significant positive correlation with yield. The results showed the importance of distributed rainfall over a number of days rather than accumulated heavy rainfall. When the correlation analysis was carried out using the current month's weather variables with average monthly cocoa yield, temperature and number of rainy days were having negative correlation which showed that during the month of harvest, for the ripened pods, too many number of rainy days and above average maximum temperature had significant negative effect.

The climatic variables put together may generate multicollinearity and hence a step wise regression was adopted to fit a regression model of cocoa yield on accumulated previous month's climatic variables and to predict the average monthly yield. A parsimonious regression equation with a single predictor variable viz; number of rainy days could explain 66% of the variation in yield and a regression equation with two variables viz; total number of rainy days and average maximum temperature as predictors could explain 69% of the variation in cocoa yield.

An empirical analysis to identify the factors perceived by farmers to influence their cocoa production was done taking a sample of 100 small holder farmers. The survey was done in Iritty Panchayat of Kannur district and Veliyamattom Panchayat of Idukki district of Kerala. Out of the total respondents 68% were males and 32% were females. Mainly the respondents belonged to the age group of above 50 years (69%) and 26 % were in between 36- 50 years and 5% less than or equal to 35 years. When educational qualification was considered, 1% was illiterate, 26% had primary school

level, 46% had high school level, 16% had intermediate level, 8% had graduate level and 3% had post graduate level education. Among the respondents, 76% belonged to nuclear type family and 24% belonged to joint type family. Land holding size was less than one acre for 17%, 1-3 acres for 61%, 3-4 acres for 10%, 4-5 acres for 5% and above 5 acres for 7% of the respondents. The farmers were having experience in cocoa cultivation as 3 years for 24%, 4 years for 27%, 5 years for 18%, 6 years for 20% and above 6 years for 11%. For 18% of the farmers, there was no membership in any organisation such as Krishibhavan, Farmer's club, Co-operative society, Banks, SHGs etc. There was association with at least one of these organisations to 39% respondents, 25% had association with 2 organisations and 18% had association with above 2 organisations. Out of the 100 farmers, 45 had contacts with extension personnel and the remaining 55 had no contact. Trainings in cocoa cultivation were received by 48 farmers whereas 52 didn't receive any training.

Based on the demographic characters of the farmers, cultivation practices, expenditure incurred, constraints faced etc. a path analysis was executed through structural equation modelling to develop cocoa yield prediction models and to identify the constraints faced by the farmers in cocoa cultivation which would ultimately lead to their income. Since a large number of exogenous and endogenous variables were to be simultaneously considered which had direct and indirect effects on yield and income of cocoa farmers, a large number of simultaneous regression equations were to be analysed and the path analysis through SEM resulted as an efficient tool to manage such situations.

From Structural equation modeling, 8 structural equations were generated and the results showed that price of cocoa was the most influencing variable on the income of cocoa farmers. With the increase of price of cocoa by 1 unit, the income of the cocoa farmers increased by 187.85 units, it means that fixing the price of cocoa beans is an important factor. When the cocoa beans are sold at good price it stabilizes the economic condition of the farmers, even the government and other manufacturing companies should involve in fixing the remunerative prices. The price of cocoa beans get stimulated if it is sold as dried or in fermented form. But most of the farmers lack drying and fermentation facilities. The second most influencing variable was the yield which

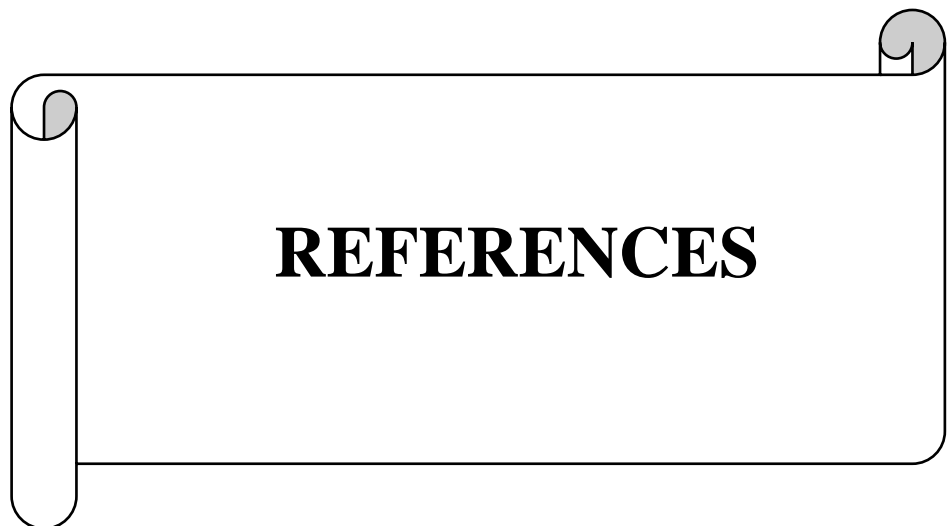
increase the income of the farmers by 48.36 units which is undoubtedly true that increase in the cocoa production will directly affect the income of farmers. For increasing the yield there are several influencing factors such as good quality seedlings, plant protection and disease management measures, protection from attack of squirrel, rat, civet etc. which would indirectly affect the condition of pods as the farmers would make an early harvest of cocoa pods otherwise because of the threat of rodent attack. The damage caused by these rodents and pest and disease attack should be taken care of to increase the yield.

Farmers of Kerala showed much interest in the cultivation of cocoa since 1980. It has become an important source of livelihood for many of the small holder cocoa farmers. So, the important factors perceived by cocoa farmers of Kerala are very important. Hence an attempt was made to identify the important factors leading to enhanced cocoa yield and ultimately to net income of farmers. Results from this study indicated that, procurement of superior planting materials was the most important factor. According to the opinion of cocoa farmers, government should take initiatives to supply quality seedlings so that any average farmer can own it. Regarding the pest and disease attack the cocoa farmers are expecting high yield loss since the cocoa plant is more susceptible to pest and disease right from the beginning of sowing of seeds and most of them have agreed that control of pest and disease attack as the second important factor. It is known that the cocoa pod is the economic part of cocoa plant and most of the yield loss was occurred through the attack of black pod disease, tea mosquito bugs and the damage caused by some of the rodents. The pest and disease management is one of the most critical component to maintain the plant health. Many of the cocoa farmers are poor enough and facing financial problems and it was treated as a constraint. In order to overcome this, financial assistance to the farmers from the financial institutions is required which helps the farmers to access credit and secure working capital. Lack of availability of proper marketing facilities and lack of knowledge on post-harvest management are other important constraints of cocoa farmers. A stable market with good facilities will encourage the farmers towards the increase of cocoa production. Finally, the information and publicity on cocoa production was also important because

of the fact that proper education and training about the cocoa cultivation practices would lead to better production. Kendall's coefficient of concordance obtained in the analysis was 0.552 which stated that the respondents had high degree of agreement to rank the factors according to their importance.

Probit analysis was done to identify the significant factors that influenced the cocoa farmers to adopt plant protection measures. It revealed the marginal effect of a factor that would influence the cocoa farmers to make decision to adopt plant protection measures. The results showed that age, land holding size, experience in cocoa cultivation, membership in various organisations and frequency of extension contact were the significant factors that influenced the cocoa farmers to adopt plant protection measures.

The results of yield gap analysis showed that the national yield gap of cocoa is 1.5 Kg dry beans per tree per year, accounting for 37.5% of yield gap to the cocoa yield potential.

A decorative scroll graphic with the word "REFERENCES" centered inside. The scroll is white with a black outline and has a grey shadow on the left side. The word "REFERENCES" is written in a bold, black, serif font.

REFERENCES

CHAPTER 6

REFERENCES

- Abbeam, G. D., Armed, M., and Baidoo, F. 2014. Determinants of consumer preference for local rice in Tamale metropolis, Ghana. *Int. J. Edu. Soc. Sci.* 1(2): 114-112.
- Adhikari, S. P., Ghimire, Y. N., Subedi, S., and Poudel, H. K. 2020. Decision to use herbicide in wheat production by the farm household in Nepal: A probit regression analysis. *J. Agric. Nat. Resour.* 3(1): 12-19.
- Ajayi, I. R., Afolabi M.O., Ogunbodede, E. F., and Sunday, A.G. 2010. Modeling rainfall as a constraining factor for cocoa yield in Ondo state. *Am. J. Sci. Ind. Res.* 1(2): 127-134.
- Alsharif, M., Younes, M., and Kim, J. 2019. Time series ARIMA model for prediction of daily and monthly average solar radiation: The case study of Seoul, South Korea. *Int. J. Symmetry* 11(2): 1-17.
- Amedi, M. 2014. Constraints among rice farmers under the MiDA agricultural credit programme in the Hohoe municipality. *Int. J. Novel Res. Mark. Manag. Econ.* 1(1): 1-9.
- Amin, M., Amanullah, M., and Akbar, A. 2014. Time series modeling for forecasting wheat production of Pakistan. *J. Anim. Plant Sci.* 25(4): 1444-1451.
- Anang, B. T. 2016. A Probit analysis of the determinants of fertilizer adoption by cocoa farmers in Ghana. *Asian J. Agric. Ext. Econ. Sociol.* 8(1): 1-8.

- Anang, B. T., Adusci, K., and Mintah, E. 2011. Farmer's assessment of benefits and constraints of Ghana's cocoa sector reform. *Curr. Res. J. Soc. Sci.* 3(4): 358-363.
- Aneani, F. and Frimpong, K. 2013. An analysis of yield gap and some factors of cocoa (Theobroma Cocoa) yields in Ghana. *Sustain. Agric. Res.* 2(4): 117-127.
- Aneani, F., Anchirinah, V. M., Asamoah, M., and Owusu-Ansah, F. 2007. *Baseline socio-economic and farm managements survey*. A Final Report for the Ghana Cocoa Farmers' Newspaper Project. New Tafo-Akim: Cocoa Research Institute of Ghana (CRIG).
- Ankrah, S., Peiris, B. L., and Thattil, R. O. 2015. Weighted modelling and forecasting of cocoa production in Ghana: A multivariate approach. *Trop. Agric. Res.* 26 (3): 569-578.
- Armstrong, J. S. (ed.). 2001. *Principles of Forecasting: A Handbook for Researchers and Practitioners* (1st Ed.). Springer Science and Business Media, New York, 843p.
- Bhagat, A. D. and Patil, M. A. 2014. Probability distribution functions of weekly reference crop evapotranspiration for Solapur district of Maharashtra. *Int. J. Agric. Eng.* 7(2): 399-401.
- Box, G. E., Jenkins, G. M., and Reinsel, G. C. 2011. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons publications, Hoboken, 784p.
- Chandio, A. A. and Yuansheng, J. 2018. Determinants of adoption of improved rice varieties in Northern Sindh, Pakistan. *Rice Sci. J.* 25(2): 103-110.
- Chizari, A., Mohamed, Z., Shamsudin, M., and Seng, K. W. K. 2017. The Effects of Climate Change Phenomena on Cocoa Production in Malaysia. *Int. J. Environ. Agric. Biotechnol.* 2(5): 2599-2604.

- Codjoe, F. N. Y., Brempong, S. A., and Boateng, D. O. 2013. Constraints facing cocoa – based agricultural knowledge and information system in Ghana: Perception of cocoa farmers in the eastern region of Ghana. *Am. J. Agric. Res.* 1(8): 1-11.
- DCCD [Directorate of Cashew nut and Cocoa Development]. 2017-18. DCCD home page [on line]. Availabe: <https://dccd.gov.in>. [09 July, 2020].
- DCCD [Directorate of Cashew nut and Cocoa Development]. 2018-19. DCCD home page [on line]. Availabe: <https://dccd.gov.in>. [09 July, 2020].
- Denkyirah, E. K., Adu, D. T., Aziz, A. A., Denkyirah, E. K., and Okoffo, E. D. 2016. Analysis of the factors influencing small holder rice farmer’s access to credit in the upper east region of Ghana. *Asian. J. Agric. Ext. Econ. Social.* 10(4): 1-11.
- Donkoh, S. A. and Awuni, J. A. 2011. Adoption of farm management practices in lowland rice production in Northern Ghana. *J. Agric. Biol. Sci.* 44(2): 424-439.
- Duniya, K. P. and Adinah, T. I. 2015. Probit analysis of cotton farmer’s accessibility to credit in Northern Guinea Savannah of Nigeria. *Asian J. Agric. Ext. Econ. Social.* 4(4): 296-301.
- Dutt, V., Sharma, H. L., and Das, S. B. 2016. A negative binomial distribution for describing pattern of green stink bug in pigeon pea crop after spraying insecticides. *J. Crop Weed* 12(1): 96-100.
- Elum, Z. A. and Sekar, C. 2015. An empirical study of yield gap in seed cotton production in Tamil Nadu state, India. *Indian J. Agric. Res.* 49(6): 549-553.
- Gommes, R. 2006. Non-parametric crop yield forecasting, a didactic case study for Zimbabwe. *Proceedings of the ISPRS Archives: Remote sensing support to crop yield forecast and area estimates*, Stresa, Italy, pp 79–84.
- Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2010. *Multivariate Data Analysis* (7th Ed). Prentice Hall, New Jersey, United States, 761p.

- Hambisa, E. N. 2018. Determinants of coffee producing farmer's access of formal credit in Bodji district of West Wollega Zone, North West Ethiopia. *Eur. J. Bus. Manag.* 10(34): 5-10.
- Hemavathi, M. and Prabakaran, K. 2018. ARIMA Model for Forecasting of Area, Production and Productivity of Rice and Its Growth Status in Thanjavur District of Tamil Nadu, India. *Int. J. Curr. Microbiol. App. Sci.* 7(2): 149-156.
- Hoepfner, S. S. and Dukes, J. S. 2012. Interactive responses of old-field plant growth and composition to warming and precipitation. *Glob. Change Biol.* 18: 1754-1768.
- ICCO [International Cocoa Organization]. 2018. Quarterly bulletin of cocoa statistics. ICCO home page [on line]. Available: <https://www.icco.org>. [09 July, 2020].
- Job, E. 2006. Yield gap of rice in Alappuzha district of Kerala. *J. Trop. Agric.* 44(1-2): 89-90.
- Karadas, K., Celik, S., Eyduran, E., and Hopoglu, S. 2017. Forecasting production of some oil seed crops in Turkey using exponential smoothing methods. *J. Anim. Plant Sci.* 27(5): 1719- 1729.
- Kehinde, A. D. and Adeyemo, R. 2017. A Probit analysis of factors affecting improved technologies dis-adoption in cocoa-based farming systems of South-Western Nigeria. *Int. J. Agric. Econ.* 2(2): 35-41.
- Kulshrestha, M. S., George, R. K., and Shekh, A. M. 2007. Weekly rainfall probability analysis by gamma distribution and artificial neural network. *J. Agrometeorol.* 9(2): 196-202.
- Lawal, J. O. and Omonona, B. T. 2014. The effects of rainfall and other weather parameters on cocoa production in Nigeria. *Comun. Sci. J.* 5(4): 518-523.

- Lobell, D. B., Cassman, K. G., and Field, C. B. 2009. Crop yield gaps: Their importance, magnitudes and causes. *Ann. Rev. Environ. Resour.* 34: 1-26.
- MacCallum, R. C., Browne, M. W., and Sugawara, H. M. 1996. Power analysis and determination of sample size for covariance structure modelling. *Psychol. Methods J.* 1(2): 130-49.
- Manikandan, N., Prasada Rao, G. S. L. H. V., and Parasannakumari, S. 2014. Influence of climate on yield of cocoa over Vellanikkara, Thrissur, Kerala. *J. Innov. Agric.* 1(1): 31-36.
- Monaganta, A., Sumardjo, Sadona, D., and Tjitropranoto, P. 2018. Influencing factors the interdependence of Cocoa farmers in Central Sulawesi province, Indonesia. *Int. J. Sci. Technol.* 8(1): 106-113.
- Mwanga, D., Ongala, J., and Orwa, G. 2017. Modeling sugarcane yields in the Kenya sugar industry: A SARIMA model forecasting approach. *Int. J. Stat App.* 7(6): 280-288.
- Nirmala, B. 2015. Hybrid Rice Seed Production in Telangana and Andhra Pradesh States Of India: A Situation Analysis. *Int. J. Agric. Sci.* 7(14): 883-886.
- Pankratz, A. 1983. *Forecasting with Univariate Box- Jenkins Models: Concepts and Cases.* John Wiley publications, New York, 576p.
- Paul, R. K., Prajneshu., and Gosh, H. 2013. Statistical modelling for forecasting of wheat yield based on weather variables. *J. Agric. Sci.* 83(2): 180-183.
- Pushpa, and Srivastava, S. K. 2014. Yield gap analysis and the determinants of yield gap major crops in eastern region of Uttar Pradesh. *Econ. Aff. J.* 59(4): 653-662.
- Raguindin, D. R. and Vera, E. A. A. 2011. Multivariate probit analysis on the factors influencing the adoption of water saving technologies by rice farmers in Sto. Domingo, Nueva Ecija. *Philipp. Stat. J.* 61(1): 109-121.

- Rahman, S. 2008. Determinants of crop choices by Bangladeshi farmers: A bivariate probit analysis. *Asian J. Agric. Dev.* 5(1): 29-41.
- Rajan, S. M., Palanivel, M., and Mohan, S. K. 2015. Forecasting of cotton area, production and productivity using trend analysis. *Int. J. Appl. Sci. Eng.* 3(12): 516-520.
- Rathod, S. and Mishra, G. C. 2018. Statistical models for forecasting mango and banana yield of Karnataka, India. *J. Agric. Sci. Technol.* 20: 803-816.
- Roth, B., John, M., Finnan, Jones, M. B., James, I., Burke, and Williams, M. L. 2015. Are the benefits of yield responses to nitrogen fertilizer application in the bioenergy crop *Miscanthus 3 giganteus* offset by increased soil emissions of nitrous oxide?. *Glob. Change Biol. Bioenergy.* 7(1): 145-152.
- Samal, P., Pandey, S., Kumar, G. A. K., and Barah, B. C. 2011. Rice ecosystems and factors affecting varietal adoption in rainfed Coastal Orissa: A multivariate probit analysis. *Agric. Econ. Res. J.* 24(2): 103-110.
- Sanjeev and Urmil, V. 2016. ARIMA versus ARIMAX modelling for sugarcane yield prediction in Haryana. *Int. J. Agric. Stat. Sci.* 12(2): 327-334.
- Saranyadevi, M. and Mohideen. A. K. 2017. A stochastic modelling for paddy production in Tamil Nadu. *Int. J. Stat. Appl. Math.* 2(5): 14-21.
- Sefriadi, H., Villano, R., Fleming, E., and Patrick, I. 2013. Production constraints and their causes in the cocoa industry in West Sumatra: From the farmer's perspective. *Int. J. Agric. Manag.* 3(1): 1-13.
- Sitienei, Betty, J., Shem, G., Juma, S. G., and Opere, E. 2017. On the use of regression models to predict tea crop yield responses to climate change: A case of Nandi East, Sub-County of Nandi County, Kenya. *Climate* 5(3): 54.

- Shadfar, S. and Malekmohammadi, I. 2013. Application of structural equation modelling (SEM) in restructuring state intervention strategies towards paddy production development. *Int. J. Acad. Res. Bus. Soc. Sci.* 3(12): 576-618.
- Shee, A., Mayanja, S., Simba, E., Stathers, T., Bechoff, A., and Bennett, B. 2019. Determinants of postharvest losses along smallholder producers maize and sweet potato value chains: an ordered Probit analysis. *J. Food Secur.* 11(5): 1101-1120.
- Shukor, N. N., Hamid, H. A., Abdu, A., and Ismail, M. K. 2015. Growth and physiological response of *Azadirachta excelsa* (Jack) jacobs seedlings to over-top-filling treatment. *Am. J. Plant Physiol.* 10: 1-24.
- Singh, M., Mishra, G. C., and Mall, R. K. 2018. Time series models for forecasting the impact of climate change on wheat production in Varanasi district, India. *Int. J. Curr. Microbiol. App. Sci.* 7(11): 2687-2696.
- Singh, P. K., Singh, K. K., Bhan, S. C., Baxla, A. K., Gupta, A., Balasubramanian, R., and Rathore, L. S. 2015. Potential yield and yield gap analysis of rice (*Oryza sativa*) in Eastern and North Eastern regions of India using CERES rice model. *J. Agrometeorol.* 17(2): 194-198.
- Subudhi, C. R., Jena, N., Suryavanshi, S., and Subudhi, R. 2019. Rainfall probability analysis for crop planning in Rayagada district of Odisha, India. *Int. J. Hydrol.* 3(6). 507-511.
- Sujatha, S., Bhat, R., and Apshara, E. S. 2018. Climate change, weather variability and associated impact on arecanut and cocoa in humid tropics of India. *Int. J. Innov. Hortic.* 7(1): 27-37.
- Tahir, A. and Habib, N. 2013. Forecasting of maize area and production in Pakistan. *ESci J. Crop Prod.* 2 (2): 44-48.

- Tanko, M. 2017. Profit efficiency and constraints analysis of shea butter industry: Northern region of Ghana. *Korean J. Agric. Sci.* 44(2): 424-439.
- Tripathi, R., Nayak, A. K., Raja, R., Shahid, M., Kumar, A., Mohanty, S., Panda, B. B., Lal, B., and Gautam, P. 2014. Forecasting rice productivity and production of Odisha, India, using Autoregressive Integrated Moving Average Models. *Adv. Agric. J.* 2014(4): 1-9.
- Unnikrishnan, T., Anilkumar, P., and Gopakumar, C. S. 2018. SARIMA models forecasting of weather parameters for Thrissure district. *Int. J. Stat. Appl. Math.* 3(1): 360-367.
- Utami, Y. E., Maarif, M. S., Fahima, I., and Suroso, A. I. 2018. Factors affecting productivity and welfare among Indonesian cocoa farmers. *J. Agric. Vet. Sci.* 11(9): 62-70.
- Verma, S., Verma, D. K., Giri, S. P., and Vats, A. S. 2012. Yield gap analysis in mustard crop through front line demonstrations in Faizabad district of Uttar Pradesh. *Int. J. Pharmacogn. Phytochem. Res.* 1(3): 79-83.
- Wiah, E. N. and Ankrah, S. T. 2017. Impact of climate change on cocoa yield in Ghana Using Vector Autoregressive Model. *Ghana J. Technol.* 1(2): 32-39.
- Wilcox, D. 2005. *Yield Probability Distribution Analysis: A Forgotten Tool in the Farm Decision Toolkit*. Manitoba Agricultural Services Corporation, Manitoba, 15p.

MULTIPHASE ANALYSIS OF COCOA PRODUCTION IN KERALA

By

SHIVAKUMAR M

(2018-19-005)

ABSTRACT OF THE THESIS

Submitted in partial fulfillment of the requirement for the degree of

Master of Science in Agricultural Statistics

Faculty of Agriculture

Kerala Agricultural University, Thrissur



DEPARTMENT OF AGRICULTURAL STATISTICS

COLLEGE OF HORTICULTURE

VELLANIKKARA, THRISSUR- 680656

KERALA, INDIA

2020

CHAPTER 7

ABSTRACT

Cocoa (*Theobroma cacao L.*) is a very important crop as it provides food, income, employment and resources for poverty reduction. It ensures livelihood for millions of small holder farmers and offers raw material for the multibillion global chocolate industries. Despite the fact that Kerala has enormous potential in terms of suitable agricultural land, cocoa has failed to become a significant crop. As its domestic production is not sufficient to meet the increased demand, the industry has to resort to substantial imports. So, a comprehensive study titled “Multiphase analysis of cocoa production in Kerala” has been made on different aspects of cocoa cultivation, management practices, production and the constraints faced by actual growers.

The trend analysis and forecasting of yearly area, production and productivity of cocoa in Kerala using advanced time series models employed on the data for the period from 1980-2017 revealed a distinct quadratic trend for the area under cocoa, having an increasing trend now and more or less linear stochastic trends for production and productivity. The Holt’s exponential smoothing model was identified as the best to predict yearly area under cocoa with an adjusted R^2 equal to 0.94. The yearly production of cocoa could be well modelled by ARIMA (0,1,1) with an adjusted $R^2 = 0.72$. By incorporating area under cocoa as an independent variable, ARIMAX (0,1,0) model could improve the R^2 to 0.84 to predict the yearly production of cocoa. The productivity of cocoa seemed to be constant for several years (0.45tonnes/ha) which was well predicted through the simple exponential smoothing model with an adjusted $R^2 = 0.84$.

Evaluation of the performance of 100 selected cocoa hybrids in the Cocoa Research Centre, College of Horticulture, KAU, Vellanikkara showed that the peak average monthly yield was in the month of November (18.14pods) followed by the yield in October (18.04) and December (14.56). A pattern of biennial tendency persisted for the yearly yields of the hybrids. The results of General linear model repeated measures ANOVA highlighted the existence of a significant Time x Factor interaction with a

partial eta squared equal to 0.14 where factor denotes different subgroups of cocoa hybrids with homogeneous yield. After the first harvest, the peak average yield was noticed during the fifth year irrespective of different low and high yielding groups.

The income from cocoa farming depends on healthy pods harvested. So, an attempt was also made to account for the frequency of number of infected pods from each tree and it could be well demonstrated by geometric distribution which is a special case of Negative binomial distribution. Owing to the fact that the infected pods might be the outcome of external factors like weather variables, the influence of those factors with cocoa yield was also investigated. A stepwise regression of yield on previous five month's accumulated weather variables resulted in a parsimonious prediction equation with total number of rainy days as the single regressor which could explain 66% of the variation in yield. The adjusted R^2 could be enhanced to 69% by incorporating maximum temperature as the second most important regressor. The wide variation realised in the average monthly yield of cocoa hybrids could be well captured through SARIMA (1,0,0) (1,1,0)₁₂ model with an adjusted $R^2 = 0.92$.

An empirical analysis to identify the factors perceived by farmers to influence their cocoa production and ultimately their income was performed taking a total sample of 100 farmers from Veliyamattom Panchayat of Idukky district and Iritty Panchayat of Kannur district. From a path analysis through structural equation modelling several linear regression equations could be generated simultaneously leading to prediction equations for cocoa yield and income. The final model iterated resulted in goodness of fit measures viz; comparative fit index = 0.96 and Tucker Lewis index = 0.94. Price of cocoa turned out to be the most prominent factor which contributed to the income of a cocoa farmer highlighting the importance of fixing the marketing price of cocoa. Second factor was yield per tree which was the outcome of good quality seedlings, efficient cultivation practices, plant protection and disease management measures, protection from rodent attacks etc. Importance of access to credit which would help to overcome the problems of lack of capital was emphasised. Financial problems such as

inability to get assistance from financial institutions, lack of proper marketing facilities including drying and fermentation facilities of cocoa beans also were noticed.

Probit analysis identified the factors viz; age of the farmers, land holding size, experience in cocoa cultivation, membership in organisations like Krishibhavan, farmer's club, Cooperative society, Banks, SHGs etc. and frequency of contact with extension personnel to be significant for decision making to implement plant protection measures which were inevitable for successful crop management and ultimately leading to the net income of farmers.

The yield gap analysis revealed that as against the potential yield (dry bean weight) of 4kg/tree/year, the national average yield from cocoa farmers was only 2.5 kg/tree/year resulting in a yield gap of 37.5% which need adequate attention.