

**GRADIENT BOOSTING-BASED RESPONSE SURFACE METHODOLOGY
WITH APPLICATION IN MODELLING POTATO POST-HARVEST
LOSSES**

Erick Cheruiyot Kirui

**A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of
Doctor of Philosophy in Statistics of Murang'a University of Technology**

August, 2026

DECLARATION

Student Declaration

I hereby declare that this thesis is my original work and, to the best of my knowledge, has not been presented with a degree award at this university or any other university. I understand what plagiarism is and I am aware of the university's policy on this. I understand that any false claim regarding this work shall result in disciplinary action in accordance with the University Anti-Plagiarism Policy

Erick Cheruiyot Kirui
AS500/5772/2022

Date

Supervisor's Approval

The undersigned certify that they have read and hereby recommended for acceptance of Murang'a University of Technology a thesis entitled, **Gradient Boosting-Based Response Surface Methodology with Application in Modelling Potato Post-Harvest Losses**

Dr. Ayubu Anapapa
Department of Mathematics and Actuarial Science
Murang'a University of Technology, Kenya

Date

Dr. John Mutuguta
Department of Mathematics and Actuarial Science
Murang'a University of Technology, Kenya

Date

DEDICATION

This thesis is dedicated to my family who have always loved and supported me, encouraged and inspired me, and sacrificed for me in the process of my academic life. To my dear mother Catherine, thank you for your constant prayers, direction, and unwavering faith in me. I've always felt inspired by you and your perseverance. To my dear wife, Flossy, thanks for your patience, understanding, and never-ending support through the many challenges and demands of this study. You have been a huge inspiration to me. To my precious daughter Natasha, your joy, love and presence have given me the motivation and strength to persevere and strive for excellence. I also pay tribute to my siblings and extended family for their encouragement, support and belief in my abilities. Whether you've written a small or large portion of this book, your help has been incredibly important to me in reaching this point. It is a testament to the love, sacrifices and support you have all given me.

ACKNOWLEDGEMENT

I wish to express my sincere gratitude to the Management of Murang'a University of Technology for the sponsorship and support that made this study possible. I also owe a debt of gratitude to the School of Pure and Applied Sciences and the Department of Mathematics and Actuarial Science for their enabling academic environment; the support provided in terms of resources and constant encouragement during my research journey. I extend my heartfelt appreciation to my supervisors, Dr. Anapapa and Dr. Mutuguta, for their invaluable guidance, constructive criticism, patience, and unwavering support during this study. Their expertise, mentorship, and dedication greatly contributed to the successful completion of this work.

I also thank everyone in the School of Pure and Applied Sciences, and the Department of Mathematics and Actuarial Science in particular, for all the support, encouragement and professional help they rendered to me while pursuing my studies. I would like to express special gratitude to my dear mother Catherine, whose sacrifice, prayer, encouragement and unconditional love have been an enduring source of strength and inspiration. I would like to express my gratitude to my wife Flossy, for her understanding, patience and constant support during the challenging times of this research. I also would like to acknowledge my daughter Natasha who was a source of joy, motivation and determination on this journey. Lastly, I would like to express my gratitude to all who, through their efforts in any aspect, have made this study a success. Thank you for your support and encouragement.

ABSTRACT

Modelling the complex nonlinear systems is one of the most challenging problems in statistics and the conventional response surface models are generally not sufficient in capturing the intricate nonlinear relationship and higher order interaction and advanced machine learning models are highly predictive but have generally not been developed to have statistical interpretability and optimization properties. Therefore, there is no single modelling tool that allows for accurate predictions, interpretable inferences and response optimization of nonlinear data. To overcome this methodological limitation, this study introduced a novel Gradient Boosting Based Response Surface Methodology (GB-RSM) framework, combining the optimization and interpretability properties of Response Surface Methodology (RSM) with the nonlinear learning ability of Gradient Boosting. Specifically, the study aimed at identifying important storage factors that affect potato post-harvest losses using Response Surface Methodology (RSM); Modeling the potato post-harvest losses by a Gradient Boosting model to model nonlinear relationship; Developing a hybrid Gradient Boosting-Response Surface Methodology (GB-RSM) model for better nonlinear modelling; and comparing the performance of the developed model with the standalone RSM and Gradient Boosting models. A mixed method research design was used that combined experimental simulation and a cross-sectional survey. A four factor rotatable Central Composite Design (CCD) experiment was simulated to develop and test the proposed modelling framework and fieldwork was conducted among 628 potato value-chain actors who were sampled using stratified sampling in Nakuru and Nyandarua Counties, which were purposively selected as they are among the major potato growing counties in Kenya and intensive production activities as well as a high number of post-harvest handling activities exists, providing an appropriate setting for the evaluation of the proposed methodology. Data preprocessing was done using K-Nearest Neighbor imputation and normalization. A second order response surface model was fitted to the residuals and a Gradient Boosting model trained on the residuals to create a hybrid GB-RSM model to capture nonlinear structures not accounted for by the polynomial approximation. Furthermore, a multilevel beta regression technique was used to find factors influencing proportional post-harvest losses and Principal Component Analysis (PCA) was used to build a Composite Loss Index (CLI) for integrated modelling. The GB-RSM model was the best among the three models in terms of predictive performance ($RMSE = 0.0896$ and $R^2 = 0.993$) and more model interpretable and optimized with response capability as compared to the other two models. Storage methods, storage duration, market distance, chemical use, transport method and extension training were identified as factors that significantly affect post-harvest losses, but formal education was not significant. The integrated modelling also confirmed that storage duration, potato variety, relative humidity and temperature were the most important factors influencing the Composite Loss Index, and that the curing time did not have a significant effect in the range of the experiment. The study demonstrates that the proposed GB-RSM framework can be a powerful and precise and interpretable method for modelling complex nonlinear systems, while utilizing the complementary strengths of classical statistical modelling and machine learning. The framework is suggested to be implemented for the modelling and optimization of problems with strong nonlinearity in a variety of application areas and to be extended to other machine learning algorithms, high-dimensional data and adaptive experimental designs in the future.

TABLE OF CONTENTS

DECLARATION.....	ii
DEDICATION.....	iii
ACKNOWLEDGEMENT.....	iv
ABSTRACT.....	v
LIST OF TABLES	ix
LIST OF FIGURES	x
ACRONYMS AND ABBREVIATIONS.....	xi
CHAPTER ONE: INTRODUCTION	1
1.1 Background of Study.....	1
1.1.1 Potato Post-Harvest Losses.....	6
1.1.2 Gradient Boosting.....	12
1.1.3 Response Surface Methodology	13
1.2 Statement of the Problem	16
1.3 Research Objectives	17
1.3.1 General Objective	17
1.3.2 Specific Objectives	17
1.4 Hypotheses	17
1.5 Significance of the Study	17
1.6 Limitations of the Study	18
1.7 Contributions of the Thesis	19
1.8 Organization of the Thesis	20
CHAPTER TWO: LITERATURE REVIEW.....	22
2.1 Introduction	22
2.2 Post-Harvest Management	22
2.3 Poisson Regression in Agricultural and Post-Harvest Loss Modeling	29
2.4 Negative Binomial Regression.....	30
2.5 Beta Regression for Proportional Outcomes	32
2.6 Multilevel Beta Regression	34
2.7 Machine Learning Application.....	38
2.8 Factors Causing Post-Harvest Losses	39
2.9 Varietal differences in post-harvest loss	41
2.10 Composite Indices and Multivariate Loss Measurements.....	42
2.11 Response Surface Methodology.....	43
2.12 Gradient Boosting	46
2.13 Integrated Hybrid Approach.....	51
2.14 Chapter Summary.....	53
CHAPTER THREE: MATERIALS AND METHODS	54

3.1 Introduction	54
3.2 Sampling Design and Data Collection	55
3.3 Model Evaluation	56
3.4 Ethical Consideration	57
3.5 Chapter Summary	57
CHAPTER FOUR: DEVELOPMENT AND EVALUATION OF HYBRID	
RSM-GBM MODEL.....	58
4.1 Introduction	58
4.2 Study Design	58
4.3 Response Surface Methodology	59
4.4 Gradient Boosting	59
4.5 Model Evaluation	63
4.6 Experimental Design and Data Simulation	63
4.6.1 Regression Model Results	64
4.6.2 Model Performance Summary	69
4.6.3 Model Interpretation	71
4.6.4 Residual and Diagnostic Analysis	71
4.6.5 Residual Diagnostics of the RSM Model	73
4.6.6 3D Surface with Contour Plots	75
4.7 Chapter Summary	79
CHAPTER FIVE: MULTILEVEL BETA REGRESSION ANALYSIS.....	81
5.1 Introduction	81
5.2 Study Design	81
5.3 Definition of variables.....	82
5.4 Statistical modeling framework.....	86
5.4.1 Multilevel Negative Binomial Model	87
5.4.2 Likelihood Function.....	88
5.4.3 Laplace Transformation.....	89
5.4.4 Testing Random Effects	89
5.4.5 Interclass Correlation Coefficients (ICC)	90
5.4.6 Model Diagnostics	90
5.5 Multilevel Beta Regression Model	90
5.5.1 Beta Distribution.....	91
5.5.2 Marginal Likelihood and Estimation	92
5.5.3 Interpretation of random effects	94
5.6 Results and Interpretation of Multilevel Beta Regression Model	94
5.6.1 Model Selection and Fit statistics	95
5.6.2 Intraclass Correlation Coefficients (ICC)	95
5.7 Poisson Regression Results	96

5.8 Multilevel Beta Regression Results	102
5.9 Independent variables.....	108
5.9.1 Storage method and storage duration	108
5.9.2 Extension Services and Education level	110
5.9.3 Market Distance and Mode of Transport.....	112
5.9.4 Wald Tests	114
5.9.5 County Level Random Effects.....	115
5.9.6 Model Diagnostics	116
5.10 Chapter Summary.....	119
CHAPTER SIX: APPLICATION OF HYBRID RSM-GBM MODEL	
USING CLI.....	120
6.1 Introduction	120
6.2 Experimental Design	121
6.3 Variables and Response Measurements	122
6.4 Composite Loss Index	123
6.5 Integrated hybrid framework.....	125
6.6 Statistical Analysis	126
6.7 Results and Discussions	127
6.7.1 Principal Component Analysis	127
6.7.2 RSM Model Estimation and Diagnostics	129
6.7.3 Response Surface Analysis.....	139
6.7.4 Residual GBM Residual Importance	143
6.7.5 Hybrid Model Actual vs Predicted	144
6.7.6 Hybrid Model Residuals Diagnostics	145
6.8 Chapter Summary.....	150
CHAPTER SEVEN: CONCLUSION AND RECOMMENDATIONS.....	151
7.1 Conclusion.....	151
7.2 Recommendations	154
7.2.1 Recommendations for Policy.....	154
7.2.2 Recommendations for Future Work	156
REFERENCES.....	157
APPENDICES	171

LIST OF TABLES

Table 4.1: RSM Output	65
Table 4.2: MLE Model Performance	68
Table 4.3: Model Summary	69
Table 5.1: Types of Losses	83
Table 5.2: Storage Methods of Individual Respondents	84
Table 5.3: Market Access for respondents	86
Table 5.4: Model Comparison Statistics	95
Table 5.5: Intraclass Correlation Coefficient	96
Table 5.6: Multilevel Negative Binomial Fixed Random Effects	98
Table 5.7: Multilevel Beta Regression Coefficient Random Effects	104
Table 5.8: Type III Chi-Square Tests. Multilevel Beta Regression	115
Table 6.1: CCD Factor Levels for the Experiment	122
Table 6.2: PCA Loadings	128
Table 6.3: PCA Variance	129
Table 6.4: RSM Model Fit Statistics	130
Table 6.5: Model Performance on CLI	131
Table 6.6: Predictor Effects on CLI	132
Table 6.7: Predictor Effects on CLI with Effect Sizes	135
Table 6.8: Residual GBM Variable Importance	143

LIST OF FIGURES

Figure 4.1: Residual Density Analysis	71
Figure 4.2: RSM Residuals	73
Figure 4.3: Response Surface.....	75
Figure 4.4: Predicted vs Observed Values for RSM, GBM and Hybrid GB-RSM Models	76
Figure 4.5: RSM Contour Plots	77
Figure 5.1: Multilevel Negative Binomial model	97
Figure 5.2: Multilevel Beta Regression	103
Figure 5.3: Proportional loss by method of storage	109
Figure 5.4: Proportional loss by Duration of storage.....	110
Figure 5.5: Model-predicted proportional loss by extension training.....	111
Figure 5.6: Proportional loss predicted by education Level	111
Figure 5.7: Proportional loss modeled through market distance category.....	112
Figure 5.8: Model-predicted proportional loss by mode of transportation	113
Figure 5.9: Proportional loss by duration of storage.....	114
Figure 5.10: Proportional loss predicted based chemical use	114
Figure 5.11: Multilevel random intercepts at the county level of the BRM	116
Figure 5.12: Dharma Residual Diagnostics Multilevel Negative Binomial Model	117
Figure 5.13: Multilevel Beta Regression Model, Dharma Measure of Residual Diagnostics.....	118
Figure 6.1: 3D Response Surface.....	140
Figure 6.2: Interactive Contour Plot	142
Figure 6.3: Hybrid Model: Predicted vs Actual CLI on the test set.....	144
Figure 6.4: Hybrid Model: Residuals vs Predicted CLI on the test set.....	145
Figure 6.5: Interact Q-Q Plot Hybrid Model Residuals	146
Figure 6.6: Normal Q-Q Plot of Hybrid Model Residues.....	147
Figure 6.7: RSM Four-Panel Diagnostic Plots.....	148

ACRONYMS AND ABBREVIATIONS

ANOVA	Analysis of Variance
ASDS	Agricultural Sector Development Strategy
BTR	Beta regression model
CCD	Central composite design
CGIAR	Consultative Group for International Agricultural Research
CIP	International Potato Center
CIPC	Chlorpropham
CLI	Component Loss Index
FAO	Food and Agriculture Organization
GBM	Gradient Boosting Model
ILRI	International Livestock Research Institute
KALRO	Kenya Agricultural and Livestock Research Organization
KEPHIS	Kenya Plant Health Inspectorate Services
KIRDI	Kenya Industrial Research and Development Institute
KS	Kolmogorov-Smirnov
NRTC	National Root and Tuber Crops Policy
PCA	Principal Component Analysis
RSM	Response Surface Methodology
SPS	Seed Potato Strategy (SPS)
SVM	Support Vector Machines
WHO	World Health Organization

CHAPTER ONE INTRODUCTION

1.1 Background of Study

Africa has struggled to feed its rapidly increasing populace for too long, bearing in mind that food is the most fundamental human right. It is estimated that nearly 700 million people in Africa are facing starvation and food insecurity, out of which about 300 million people are facing catastrophic levels of food insecurity (Akhtar Gul et al., 2022). The weather patterns and climate change have radically increased food insecurity in Africa, exacerbating an issue that has worsened in recent years. The World Bank warns that Africa is heading from a health crisis into a food crisis, and the United Nations is predicting that the rate of chronic malnutrition over the years could double, triggering a humanitarian food crisis (Spencer et al., 2020). It is estimated that millions of poor people in developing countries across the world, where Africa has a lot of developing countries, have a high risk of dying from hunger. A study by the African Union estimated that approximately 1.4 million people in Northern Kenya are facing starvation and need urgent food supply (Njogu, 2022).

African countries have quickly become more reliant on externally sourced food at a time when international markets are experiencing unprecedented changes. However, an agricultural organization which has a station in Kenya named the Consultative Group for International Agricultural Research (CGIAR), based in the International Livestock Research Institute (ILRI) Nairobi, estimates that 25-40% of staple food, mainly maize potatoes and rice produced in sub-Saharan Africa is lost through post-harvest losses. The organization estimates that maize, rice and potatoes are staple food to more than 500 million people, and the loss translates to over \$ 250 billion (Zsögön et al., 2022). For the past few years, climate change, especially rain patterns and other weather factors such as drought season, has entirely inflicted more pain across Africa in staple food production, and this, in turn, has led to hunger since production has either declined or the food produced is lost during harvesting and storage.

Africa enjoys at least two rainy seasons per year, but despite that, when it rains or shines, there are many deaths associated with hunger and malnutrition in Africa. This overdependence on rain for agricultural production is here to stay, and it is also estimated that food production in Africa may decrease by 30% - 40% by the year 2030.

Hence, Africa finds it difficult to realize its self-reliance in feeding its population and alleviating deaths due to hunger or malnutrition (Sirba et al., 2021). One of the ways to solve and mitigate this and help the continent is by maximizing and utilizing what we already have or are expecting and minimizing the losses. Novel strategies that directly impact the use of existing weather patterns and help reduce losses from farms could prevail well with the vision of ensuring that Africa as a continent is self-sufficient.

Agriculture is the staple of economic growth in most developing nations, especially in sub-Saharan Africa. The industry is a significant contributor to gross domestic product (GDP) and has high percentage workforce in rural regions and supports household food security and income source. The Food and Agriculture Organization of the United Nation's reports that agriculture contributes around 23% of GDP in sub-Saharan Africa and provides more than 60% of labor force (Olakiumide, 2021). Although there have been tremendous improvements in agricultural production and policy support, there is still a high percentage of food produce that goes to waste before it reaches final consumers. These losses, which are also referred to as post-harvest losses (PHL), are also one of the most urgent problems that food systems in developing economies face.

Post-harvest losses refer to quantifiable quantitative and qualitative losses in food quantity, quality, or nutritional value that takes place between point of harvest and the point of consumption. This loss is experienced at different points of the agricultural value chain such as during harvesting, handling, storage, processing, transportation and marketing. It has been estimated that developing countries suffer post-harvest losses of between 20 and 40 percent of all agricultural production each year (Olakiumide, 2021), (Zsögön et al., 2022). In the case of perishable commodities like fruits, vegetables and root crops, these numbers may be even greater. The impacts are extensive, including loss of income in smallholder farmers, high consumer prices, food insecurity and unwarranted environmental impacts due to the wastages of land, water and energy (Mburu et al., 2023; Mugo et al., 2021).

The potato crop in Kenya forms a very essential part of the national food system as both staple food and a very important source of income to millions of smallholder farming households. The crop is mainly grown in highlands of the country such as the

Rift Valley, Central and East, that offer the cool weather and fertile soils that support growth of tubers (Mwakidoshi, 2021). Current statistics show that Kenya yields about 2.5 million metric tons of potato every year, which makes it one of the top potato producing nations in the sub-Saharan region (Mburu et al., 2023).

Although production of potatoes is significant in Kenya, the crop is associated with some of the highest post-harvest losses in the agricultural industry. Investigations in different potato-producing areas have recorded losses of between 20 and more than 45% of the amount harvested, depending on production zone, storage facilities, market access and value chain management practices (Devaux et al., 2021). The cause of these losses is a multidimensional interaction of factors occurring at various levels of the value chain. Mechanical damage at the farm level occurs because of poor farming methods and lack of proper on-farm storage of tubers to humidity, temperature changes, and pests. The improper containers and the road infrastructure have led to bruising and physical deterioration during the transportation process. Poor market facilities, poor sanitation and extended holding periods at wholesale and retail markets also enhance the rate at which quality is deteriorated.

Economic implications of post-harvest losses in the potato value chain are very high. Losses leads to revenue lost, which lowers the household savings and investment level, which continue to perpetuate poverty cycles among smallholder farmers. Post-harvest losses decrease the effective food supply without a proportional decrease in the resources spent on production, which is a food system inefficiency that compromises national food security objectives (Bendinelli et al., 2020). It is thus important to understand factors that cause such losses so that specific interventions can be developed to enhance efficiency of value chain.

The post-harvest loss data is difficult to statistically analyze in several aspects. Post-harvest losses are often given as proportions or rates i.e. the percentage of total produce is lost at a particular stage, and these measures lie between zero and one. These proportional results are not in line with the assumptions of the traditional linear regression models, which assume the presence of a continuous unbounded response variable with constant error variance (Chen & Yang, 2024). Beta regression was specifically developed to model continuous proportional observations on the open

interval (0, 1) using a logit link function and can deal with the skewness and heteroscedasticity inherent in proportional data (Geissinger et al., 2022).

Another complexity of post-harvest loss data is the hierarchical nature of data collection process. The counties in which farmers and traders operate are characterized by similar environmental, infrastructural, and socioeconomic factors. It cannot, therefore, be considered that observations across the same county or even the same stage of the value chain to be statistically independent. Such clustering and hierarchical dependence are ignored, with the result of underestimating the standard errors, overstating Type I error rates, and making unreliable statistical inference (Geissinger et al., 2022; Zhang & Lai, 2024). Multilevel modeling (also known as hierarchical linear modeling or mixed-effects modeling) offers a conceptualized statistical approach to the analysis of data that has a nested design by modeling variation at multiple levels simultaneously.

This research addresses these methodological challenges by developing and applying multilevel beta regression model to analyze determinants and spatial heterogeneity of post-harvest potato losses across the potato farm-market value chain in Kenya. It is a very powerful methodology that has an excellent theoretical basis applied to approximate the determinants of proportional losses, quantify the between-county, and between-stage variation, and generate policy-relevant information regarding how to minimize post-harvest losses in the Kenyan potato industry by integrating the flexibility of the beta distribution plus the hierarchical nature of multilevel modeling.

In this study, the use of a multilevel beta regression model is a significant methodological extension with a special focus in the analysis of post-harvest losses. Although the multilevel negative binomial model best represents the absolute quantity of losses, it fails to capture the variance on the scale of different value chain actors that process different amounts of potatoes. The study focuses on the modeling of the proportion of losses expressed as the ratio of the quantity lost to the total quantity produced. This method offers a more significant and comparable performance measure because it indicates the effectiveness of managing practices and not only the size of losses. The beta regression model can be used for examining proportions and this enables a greater understanding of the occurrence of losses when compared to the

degree of activity among farmers, transporters, wholesalers, and retailers (Endalew et al., 2023; Qange et al., 2024).

The response variable can thus be built as a continuous proportion with a range of zero to one hence beta regression is the best modeling framework. The beta regression model is specifically used with fractional data unlike the linear regression which assumes normally distributed results, the model guarantees that the predicted value does not go beyond the admissible range. A significant benefit in this regard is the ability and flexibility of the beta distribution to accommodate skewness and heterogeneity found in agricultural data. The model is parametric with a mean and a precision parameter where the mean is the predicted proportion of the losses and the parameter of accuracy is the dispersion level. The beta regression model has two parameters: the mean parameter is the expected proportion of losses, and the precision parameter controls the spread of observations from the mean.

This study proposed to come up with new strategies that have a direct impact on lowering post-harvest losses of staple food to maximize and use what we can produce optimally and model conditions that would inform the farmer to adapt to the expected weather changes so that they can increase agricultural production to feed the alarming hunger-stricken Africans and avert humanitarian food crisis. Post-harvest losses account for about 25 to 40% of production losses, meaning that every person in sub-Saharan Africa loses between 150 – 200 kg of food produced yearly (Ali et al., 2021). However, post-harvest losses remain a significant problem, primarily because of inadequate handling and storage procedures. These losses worsen economic instability and food insecurity and influence smallholder farmers' livelihoods. As a result, employing efficient storage procedures is imperative to reduce potatoes' post-harvest losses in Kenya. There is a need to identify factors that could reduce food losses, increase food security, and make Kenya self-reliant in feeding her populace.

Gradient boosting algorithms such as XGBoost and LightGBM together with CatBoost present native and preprocessing solutions to effectively manage missing values. The designed algorithms operate competently with missing data while creating their tree structures. During training XGBoost and LightGBM choose optimal split direction (left or right) for missing values when the direction minimizes the loss function. The model assigns missing values dynamically and strategically through

pattern-based data analysis to optimize performance. CatBoost adds missing values as an independent category during categorical feature processing thus directly integrating missingness conditions into the modeling processing.

Both preprocessing methods and built-in algorithm functions provide solutions to handle missing data. The mean-imputation approach together with median or mode techniques are typical methods for filling numerical variables whereas mode assignment as a placeholder works best for missing categorical data. Much advanced analysis techniques work together with K-Nearest Neighbors (KNN) imputation to use similarity measures between data points for missing value prediction while multivariate imputation uses feature relationship-based statistical techniques to forecast missing entries. When many data points are missing this approach becomes essential for retaining model precision since it enables better prediction despite missing values.

The analysis of missingness data often involves producing features from it. Analysis models benefit from binary indicator values that show missing data patterns. A standard replacement value such as -1 or another unique number can fill in for missing values to mark data points. The algorithm can now detect missingness as a meaningful data sign rather than automatically discarding them. Gradient boosting algorithms eliminate columns from training when all values within the column are missing. When partial data values are missing gradient boosting algorithms employ their built-in data treatment and preparatory methods to prevent performance degradation. Gradient boosting models transform a weakness into a strength because they merge different strategies to process missing data which in turn produces extra predictive insights.

1.1.1 Potato Post-Harvest Losses

Potato (*Solanum Tuberosum*) is the mainly cultivated non-cereal crop in the world and the fourth most important crop after maize, wheat and rice in terms of production and consumption. Its global production is estimated to be over 350 billion metric tons annually (Devaux et al., 2021). Potatoes are a critical food and cash crop for smallholder farmers in Kenya with substantial incomes. It is the third staple food after maize and rice and is the second most cultivated crop after maize in the highland parts of Kenya (Mugo et al., 2021). It is mainly sold to small-scale retailers as fresh produce from the farm, who then sell it to households as foodstuffs. However, its production

is affected by various factors ranging from production, harvesting, transportation, storage, farmers' practices, and weather conditions. Other factors affecting its production are improper land preparation, poor pest and disease control, and poor timing. Also, smallholder farmers do not use certified and quality seeds, which may compromise the quality and yields produced.

Nearly a third of food produced for commercial and domestic use is estimated to be lost (Mugo et al., 2021). In Africa, this is not the case, as most potato growers are small-scale farmers who cannot afford sophisticated storage facilities, and this causes an inconsistent supply of tubers in the market during the dry season. After harvesting potato tubers, the objective is to maintain their quality and shelf life during storage. To ensure this is achieved, it is essential to delay dormancy break while minimizing sprouting, weight loss and sweetening of potatoes, which occurs as the storage duration increases. To boost dormancy while delaying wastage through sprouting, greening, and sweetening, optimal storage conditions must be maintained.

Potatoes play a critical role as a food and cash crop for a significant number of people around the world. Preventing or reducing post-harvest losses ensures food availability and, in turn, increases food security worldwide (Mugo et al., 2021). Potato growers, suppliers, and consumers in the potato value chain bear significant financial losses due to post-harvest losses. By improving the structure of the storage system, farmers stand a better chance to increase income levels and make a profit, hence improving their living standards. Potato losses after harvest can be addressed in a way that impacts the elimination and reduction of waste, thus saving resources used in production, such as water and soil. The more farmers can sell their post-harvested agricultural produce, the higher their income and, therefore, the lower the poverty levels. Also, potatoes provide consumers with basic nutrients since they are rich in essential nutrients such as carbohydrates, vitamins and minerals. Reducing post-harvest losses allows farmers to increase their capacity to supply the markets effectively. Therefore, this research aims to improve food storage and reduce post-harvest loss of potatoes to enhance food security and economic productivity and address several universal issues, including food waste, sustainability, and climate change (Mburu et al., 2023).

Post-harvest losses are among agriculture's significant challenges, affecting food security, economic stability, and the environment (Bendinelli et al., 2020). These

losses occur at different stages, from harvest to final consumption, such as poor storage methods, inefficient handling, long transportation time, and poor processing methods. Overcoming post-harvest losses is vital for making the future food supply chain more sustainable, mainly due to population increase and climate change. However, the complexity of exploiting PHL data is difficult due to the inherent nature of the data, such as nonlinearity and missing values (Kaguongo, 2014).

The quality management of crops during post-harvest is highly dependent on optimal storage conditions. The most suitable storage conditions of potatoes are the temperature of 4-10°C and relative humidity (RH) of 90-95 % (Khorramifar et al., 2023). It is, however, noted that tubers are large with high humidity levels, high respiratory rate, thin skin, and delicate flesh and delicate texture thus turning them into easily affected during storage. The adverse effects of post-harvest storage of this crop on fresh as well as processed sweet potato crops include moisture loss, alteration in nutrient composition or soft rot due to pests and diseases that may seriously impact on the quality and reproduction. Long-term storage of the potato tuber crop relates to the quantitative transformations associated with the processes of transpiration, respiration, and sprouting, which are characterized by the weight loss of the initial tuber weight. During storage, factors like relative humidity and temperature mainly determine the transpiration, respiration and sprouting processes. Natural losses of the potato tubers could also be due to meteorological conditions during the vegetation season as well as genetic characteristics of varieties and mechanical damages to tubers that is caused during harvesting and post-harvest process. Low temperature storage is a common approach that is adopted to preserve the freshness of food.

The low-temperature environment can slow down the physiological and biochemical reaction speed of fruits and vegetables, retard the senescence of fruits and vegetables, and slow down the multiplication rate of microorganisms and even sterilize them, to extend the fresh period of fruits and vegetables. The lower the temperature, the more the starch is transformed into sugar, and as the temperature rises the sugar is converted back to the starch. The process of aging in the potato is a similar one but this time it is irreversible. Thus, one should bear in mind that the average long-term preservation temperature of potatoes is between 5 and 6 degrees Celsius. To keep the quality of the potatoes over a long period of time it is important to store them under specific storage

conditions, which helps reduce the loss of both quality and the product itself, and ideally keeping it in the most optimal temperature of 4-7 °C to be able to consume them (Sáenz-Baños et al., 2022).

Owing to the ever-increasing cost of living, malnutrition and food insecurity, there is a need to ensure that the food produced is not lost. Also, due to the emerging climatic changes, competition for natural resources such as water and soil must be fully optimized to ensure a consistent food supply. To ensure this consistent supply of potato tubers throughout the year to meet harvesting demands, post-harvest handling, management, storage, and transportation of tubers are vital. Mechanical injuries during harvest, storage and transportation are known to reduce the lifespan of tubers (Jakubowski & Królczyk, 2020). Temperature and humidity control are the primary ways to preserve the shelf life of fresh tubers, fruits and vegetables. Low temperature is desirable in potato storage since high temperatures increase tuber respiration rate and lower their storage life (Firdous, 2020). The storage temperature of fresh potatoes depends on the potatoes' maturity stage and or cultivar. Levaj et al., (2023) have also shown that the temperature and variety significantly impact on the potato quality features such as color, texture, activation energy, total phenol content and weight loss during storage time. Potatoes lose their essential vitamins when exposed to high temperatures and low humidity and stored for a long time.

In sub-Saharan Africa, the Consultative Group for International Agricultural Research (CGIAR), which conducts operations under the International Livestock Research Institute (ILRI) in Nairobi, estimates that between 25% and 40% of staple foods grown in the region, mostly maize, rice, and potatoes, are lost during post-harvest processes. These losses happen across a multifaceted value chain that includes farm-level storage, transportation, wholesale, and retail and they have massive economic implications. The approximate financial cost of these losses is more than USD 250 billion per year, which affects a food system that feeds over 500 million people (Ndaka et al., 2012). To small-holder farmers who in any case operate on lean economic margins, post-harvest losses would not be a logistical inconvenience but an existential threat to the household income and livelihood security.

Kenya provides an instructive case study on post-harvest losses in potato subsector. The country is one of the potential producers of potatoes in sub-Saharan Africa where

the crop is a critical source of income and nutrition to more than 2.5 million farming households located in the highlands agroecological zones of central, western and Rift Valley areas (Andati et al., 2023; Gikundi et al., 2023; Macharia & Mutungi, 2015). Although this is important, post-harvest losses in the Kenyan potato industry are estimated to range between 30 and 50% of total production, which reflects the combination of inadequate storage infrastructure, poor handling practices, limited access to cold chain technology and over reliance on ambient temperature storage systems which is mismatched with the physiological demands of potato tubers (Ndaka et al., 2012).

The complexity of the processes involved (deterioration) is compounded by the multifactorial and interactive nature of the deterioration processes involved. The post-harvest quality of potatoes is simultaneously controlled by both physical and environmental factors such as temperatures and relative humidity in storage, exposure to light, mechanical damage during harvesting and handling, length of storage, and the use of curing treatments as well as by the inherent genetic characteristics of the variety under storage (Githumbi et al., 2024; Qange et al., 2024). Post harvest deterioration is not a linear function of any single factor but is caused by a combination of factors and the interaction between them. All these interactions influence the rate and degree of the three main causes of deterioration: fresh weight loss from transpiration and respiration, sprouting after dormancy, and rotting caused by fungi and bacteria. Learning and measuring these interactions are a basic requirement that must be met before the development of evidence-based, variety-specific, and context-appropriate storage management recommendations may be formulated to benefit Kenyan potato farmers. The main objective of this research was to apply the integrated response surface methodology with gradient boosting in modeling potato post-harvest losses.

According to Mburu et al., (2023), the education level of farmers, access to extension services and credit were significant variables for improving the production of potatoes. Also, adopting modern farming techniques such as crop rotation, use of certified seeds, and proper timing enhances the production of potatoes, hence alleviating farmers' living standards. To guarantee high profits for potato growers, the government of Kenya, through state agencies and departments, must revitalize the potato sector by implementing policies such as the Agricultural Sector Development Strategy (ASDS)

of 2010-2020, the Seed Potato Strategy (SPS) of 2009 and the National Root and Tuber Crops (NRTC) Policy of 2010 as well as implementing the legal directives aimed at addressing the packaging of seeds.

Consequently, regulatory bodies such as the Kenya Plant Health Inspectorate Services (KEPHIS) mandated seed certification, inspection and analysis of soil, water, agricultural produce, and farm inputs like fertilizer and pesticides must step up to ensure farmers access quality seed varieties. On the other hand, the Kenya Industrial Research and Development Institute (KIRDI), which is mandated to carry out research and development which includes reducing post-harvest losses and adopting appropriate processing and storage technologies, needs to up its game since a good percentage of the produced potatoes is lost during and after harvesting. They suggested that KIRDI should work in tandem with the International Potato Center (CIP), which is an international research institute responsible for developing and distributing new and improved potato varieties and technologies aimed at improving yields and market access to farmers (Ayieko et al.,2021).

Potato post-harvest losses continue to be one of the key challenges facing food security, agricultural sustainability and farmer livelihoods in Kenya and other developing countries. Potatoes are one of the most significant food and cash crops in Kenya; they play a significant role in household income, nutrition and national food supply. Despite the increasing production of potatoes due to rising urbanization, population growth, and market demand, substantial quantities of harvested potatoes are lost before reaching consumers. These losses occur at different stages of the farm-to-market value chain including harvesting, handling, transportation, storage, wholesaling, and retailing. Poor storage structure, poor transportation system, mechanical injury due to handling, pest attack, unfavorable environmental conditions and failure to use improved post-harvest technologies are blamed for the losses. In the potato producing areas of Kenya post-harvest losses have been reported to be between 20% and 45% based on storage conditions, market accessibility and handling practices. These losses affect food availability, profitability for smallholder farmers, inefficiency in markets, and food insecurity and economic instability in the agricultural sector.

1.1.2 Gradient Boosting

Gradient Boosting is a machine learning model developed to produce a final prediction by learning and combining several weak models or decision trees (Upadhyay et al., 2021). It is based on an algorithm in which a new learner whose error is based on the previous learner's error is trained successively. This differs from other methods like those of the Bagging approach which comprises of Random Forests and parallel ensemble learning making Gradient Boosting appropriate for big data sets. The essence of performing Gradient Boosting is to learn a specific type of loss function, such as mean square error in regression or log loss in classification by using the gradient descent approach. In each of these, the algorithm identifies a solution that reduces the loss for the remaining errors between the current ensemble and the new model. Gradient Boosting thus focuses on such residuals to actualize this goal of improving the predictive capacity of the model in regions of poor modeling by the current models (Otchere et al., 2022).

Gradient Boosting is an ensemble learning algorithm in machine learning where weak models are created successively with a reference to the previous model by minimizing the error. It usually focuses on the identification of nonlinear associations between variables by continuously improving and integrating relatively poor performers known as decision trees. The technique improves the model's capability by applying regularization, cross validation, and hyperparameters optimization. However, many of the machine learning methods are sensitive to missing values (Sibindi et al., 2023).

Gradient Boosting's versatility can be illustrated by various working combinations of loss functions, including the more stable variants like Huber loss that enhances the model's performance on the data with a high level of outliers. However, techniques such as shrinkage that help in reducing the contribution of a tree and early termination of the learning procedure that tends to occur when the model starts to overfit have been implemented to handle this problem. The ability to work with high-dimensional data makes Gradient Boosting the best model in modern-day predictive modeling (Trizoglou et al., 2021).

According to (Zhang & Jánošík, (2024), gradient boosting and deep learning frameworks have also been introduced primarily in the hybrid forms. These strategies utilize gradient boosting to capture important features, data transformation, and neural

networks for applications that require high-level features. Such innovations explain the progressive character of gradient-boosting algorithms in a context of solving problems in classical and modern approaches to machine learning. Another application of Gradient Boosting is in natural language processing (NLP). Due to the use of categorical feature handling and reliable tree-based model, CatBoost has been used in applications of sentiment analysis, spam detection, document classification. For instance, the big data sentiment analysis has been enhanced by using gradient boosting for ranking to counter the overfitting despite big data containing noisy data and the lack of one or more parameters. These applications prove how general the algorithm is when addressing various problems and its applicability in numerous disciplines (Bacanin et al., 2022).

The nonlinearity of the models and random missing data have emerged to be a challenge in data science in various fields such as agriculture, healthcare or finance since such data can be incomplete. Conventional methods of statistical analysis may not work well in the case of a nonlinear system or when some of the values of a sample are missing, which leads to poor prediction. The efficient development of advanced methodologies for such tasks is hence very significant (Vieira et al., 2024). The purpose of this research is to present a methodological solution which combines Gradient boosting; a machine learning technique characterized by outstanding results while processing nonlinear data sets with Response Surface Methodology.

1.1.3 Response Surface Methodology

Response surface methodology (RSM) is a statistical method of analyzing problems with the goal of finding the best level of factors that affects the response variable of interest (Bayuo et al., 2020). It is used to maximize or minimize the response variable when more than one is involved and is used by industries, social sciences and engineering to eliminate process. This method was invented by Box and Wilson in 1951, and it has been refined over time (Bezerra et al., 2008). The main aim of RSM is to identify the best function that approximates the true response surface. It is usually represented graphically for visualization using contour plots.

RSM is mainly used for process optimization and studying relationships among several input variables to a given response variable. RSM constructs response surfaces mathematical models showing the variation in the output as a function of input

variables and is thus very valuable in determining the optimum operating conditions. RSM, usually employed in industrial and experimental design, has been very effective where nonlinear interactions between variables need to be determined and optimized. However, with its dependence on linear regression models, RSM is the least adapted for such nonlinearity found in incomplete data. This framework has been developed to improve the understanding of the system performance by incorporating the imputation techniques with gradient boosting. The hybrid model is adequate and efficient for finding good prediction for nonlinear problems and there is missing or incomplete data.

The integration of gradient boosting with RSM results in the design of an accurate and efficient model for solving nonlinear mathematical problems when data has missing values or incomplete. Gradient boosting is a type of ensemble learning where decision trees are built sequentially, with each tree focusing on minimizing residuals errors made by the previous tree. Since gradient boosting involves building a model in stages, it can find intricate, nonlinear interactions between the features and the target variable, which is why the algorithm can model data better than linear regression or some decision trees. Due to the ability to operate flexibly based on distributed data, it is applied to predict the further behavior of consumers and identify potential cases of fraud.

Agricultural data is always considered incompletely observed through surveys and may include missing values due to sampling and data collection shortcomings and restricted access to data in agronomic databases particularly in the developing world (Ahmad et al., 2024). With reference to this technique, the farmers, supply chain managers, and policymakers are in a better position in terms of decision-making processes regarding handling and storing harvested crops. Furthermore, the model's capability of dealing with missing data makes it appropriate, especially when data are poorly collected in developing countries and other parts of the globe.

This model can be applied in various industries, including the health and finance industries. It has also been beneficial for e-commerce sites since it gives reliable forecasts of the outcome of specific product promotions. For instance, gradient boosting at Amazon to present its products or to anticipate the client's choice (Shahbazi et al., 2020). Such model factors as prior purchase history, history of use,

and evaluations about the product are incorporated to provide a shopping experience that meets the consumers' needs while adding to customer confidence.

The cross-validation and the hyperparameter tuning help enhance the quality of the model by tuning the parameters to avert overfitting (Montesinos López et al., 2022). Therefore, the use of gradient boosting together with RSM is an improvement on traditional statistical modelling. It enables the inclusion of nonlinearities into the model while providing an unprecedented means of managing the interaction between the variables using imputation methods to address cases of missing values. This results in a relatively flexible and versatile model that can be easily employed to solve different kinds of problems, starting with improving farming practices in agriculture and ending up with increased efficiency of treatments in medicine. Unlike many traditional models, gradient boosting can handle missing data through surrogate splits that precisely produce outcomes through an incomplete dataset. RSM enhances this capability through the imputed or missing data in the model to have reasonable interaction with the blank or imputed cases (Trabelsi et al., 2023).

This approach also has the advantage of efficient hyperparameter optimization. Gradient Boosting models are sensitive to the parameters, including the number of trees, learning rate and tree depth (Zhang & Haghani, 2015). This is made easier in RSM by ensuring hyperparameters are incorporated and systematically fine-tuned to give precision and the best output. Furthermore, the gradient boosting -RSM hybrid model can perform better, particularly when finding a solution for high-dimensional problems (Alzu'bi et al., 2015). Datasets in domains like genomics or health care, which are likely to have missing values, are challenging not to overfit using conventional methods. Gradient Boosting Machine naturally performs the feature selection as the model is trained, and RSM can be applied further to fine-tune the important variables and their interplay. Several studies in comparative literature indicate that the gradient boosting-RSM model outperforms other machine learning models, especially in nonlinear interaction and imbalanced datasets (Alzu'bi et al., 2015). Gradient Boosting is more effective than machine learning models such as Random Forests, SVMs, and Logistic Regression, with significantly higher dimensional, complex and nonlinear data.

This study proposed a new technique for integrating the gradient boosting model with the RSM. It combines the working of the gradient-boosting decision trees concerning the prediction part and the stringent RSM classifier on optimization part to help improve the model for the analysis of post-harvest potato losses. These limitations and changing trends explain why nonlinear relations and missing value in PHL modelling is crucial. These include non-linearities in the historical agricultural data since most are interrelated such as temperature, humidity and different crop varieties (Misra et al., 2022).

1.2 Statement of the Problem

Modeling complex nonlinear systems is still a methodological problem, since many statistical methods are based on several assumptions that do not adequately capture the complex nonlinear structure and higher-order interactions between variables. Response Surface Methodology (RSM) is well suited for experimental design, optimization and interpretation of factor effects, but suffers from being restricted using second-order polynomial approximations for highly nonlinear response surfaces. On the other hand, gradient boosting machines (GBM) are powerful tools that can model complex nonlinear relationships and interactions with high predictive accuracy; however, they do not offer the statistical interpretability and optimization necessary for inference and decision making. Existing methods, however, provide a compromise between interpretability and predictive power, and create a gap in nonlinear statistical models. Even though there are complementary strengths in RSM and GBM, few studies have addressed how to combine them into a single modelling framework that will enhance the prediction, maintain interpretability, and allow optimization. It is aimed to overcome this gap by creating a gradient boosting-based response surface methodology (GB-RSM) framework for modelling complex nonlinear systems and using potato post-harvest loss data as a case study to test the proposed methodology.

1.3 Research Objectives

1.3.1 General Objective

To develop a hybrid model integrating gradient boosting and response surface methodology to model potato post-harvest loss data.

1.3.2 Specific Objectives

1. To develop gradient boosting -based Response Surface Methodology framework for modeling complex nonlinear relationships between independent and dependent variables.
2. To evaluate the statistical properties of the hybrid GB-RSM in comparison with standalone RSM and GBM.
3. To apply a multilevel beta regression model to identify the determinants and variability of potato post-harvest losses in the farm-market value chain.
4. To apply the integrated hybrid model in modeling potato post-harvest loss data to identify the most significant predictor variables

1.4 Hypotheses

H₀₁: Integrated gradient boosting -based Response Surface Methodology framework does not significantly improve the modeling of non-linear relationship between dependent and independent variables.

H₀₂: There is no significant difference in predictive accuracy between the integrated hybrid model and standalone gradient boosting and RSM.

H₀₃: There is no difference in model fit, adequacy and prediction performance between Poisson, Negative binomial and Multilevel Beta regression in modeling potato PHL.

H₀₄: The hybrid model integrating gradient boosting with RSM does not perform better than standalone methods like RSM and GBM.

1.5 Significance of the Study

Potatoes serve as a fundamental agricultural commodity for a substantial global population. Potatoes are among the leading income-generating crops for smallholder farmers in many regions globally. Reducing post-harvest losses guarantees food

supply stability and enhances global food security. Farmers and stakeholders in the potato supply chain suffer significant economic losses due to post-harvest losses. Through the optimization of storage systems, farmers have the potential to enhance their profitability and make a substantial contribution to the financial sustainability of their enterprises. The decrease in food waste by minimizing post-harvest losses in potatoes can help save resources used in the production of potatoes. Minimizing post-harvest losses enhances the marketing efforts of agricultural products thus increasing incomes and subsequently eradicating poverty. Potatoes are rich in nutrition value as they contain some of the most important nutrients such as carbohydrates, vitamins and minerals. Implementation of appropriate storage techniques is critical in enhancing the nutritional quality of potatoes thus ensuring that consumers are availed with appropriate nutritious food. Through reducing the losses that occur after harvest, farmers can increase their capacity to supply the markets better. Thus, the proper storage of food and potatoes and reducing the rate of post-harvest loss ensures food security and economic stability by solving more generalized problems such as food waste, food sustainability, climate change among others. This research can offer valuable insights to farmers, consumers, and global efforts towards achieving sustainable development goals for food and agriculture. The findings help in the identification of appropriate post-harvest handling practices for potatoes to assist the farmers and vendors in the marketing phase to enhance food safety and reduce post-harvest losses.

1.6 Limitations of the Study

In this study, a gradient boosting-based response surface methodology (GB-RSM) framework was developed and tested on potato post-harvest loss data for the purpose of modelling complex nonlinear systems. While the proposed framework showed better prediction accuracy and retained the interpretability and optimization aspect of Response Surface Methodology, certain limitations must be recognized. The model has been developed and validated with data obtained from two key potato growing counties in Kenya. The conditions of these counties are appropriate for testing the proposed methodology, but the results may not be representative of production systems that have a significantly different climatic, agronomic, or socioeconomic profile. Therefore, the model's generalizability to other geographical regions should be treated with some caution and should be validated externally.

Second, the experimental design used in the study was a four-factor rotatable Central Composite Design (CCD), where the four selected factors are believed to account for sufficient variance in the response. The design explained well the main effects, interaction and quadratic effects, but other possible factors not accounted for in this study might also affect the nonlinear response behavior. The presented model could be further refined with the inclusion of other predictors in future applications. Thirdly, the hybrid approach was tested by combining only Response Surface Methodology and Gradient Boosting. Other advanced machine learning algorithms, such as Extreme Gradient Boosting (XGBoost), LightGBM, CatBoost, Gaussian Process Regression, and deep neural networks, were not considered. The study therefore does not prove that the proposed hybrid framework outperforms any of the other possible nonlinear modelling methods. Predictive accuracy, model fit, and optimization performance were the main considerations examined when assessing the proposed methodology. This study did not examine the theoretical statistical properties of the hybrid estimator such as consistency, asymptotic behavior, statistical efficiency, and convergence characteristics. These properties would be explored in the proposed framework for a richer theory and would be the future direction of methodology studies.

1.7 Contributions of the Thesis

This thesis presents a hybrid approach, Gradient Boosting Based Response Surface Methodology (GB-RSM), which combines the interpretability and optimization properties of RSM with the nonlinear prediction properties of Gradient Boosting. The framework models the systematic linear, quadratic, and interaction effects using RSM, and the residual nonlinear patterns that are not accounted for using Gradient Boosting. In addition, the thesis presents an application of the multilevel beta regression for modelling potato post-harvest losses as a proportion, considering the hierarchical nature of the data. Furthermore, a Composite Loss Index (CLI) is generated through Principal Component Analysis to give an overall assessment of weight loss, sprouting and rotting. Empirically, the study quantifies significant factors involved in potato post-harvest losses and shows that hybrid GB-RSM method can be used to provide better performance in predicting losses with useful interpretation for optimization of the potato response.

1.8 Organization of the Thesis

This thesis consists of seven chapters, each one of them discusses a different aspect of modelling nonlinear data using the Gradient Boosting Based Response Surface Methodology (GB-RSM) framework. Chapter One presents the introduction to the study. It gives the background of potato post-harvest losses, problem statement, research objectives, hypotheses, significance of the study, scope, and limitations. It also provides the rationale for integrating Response Surface Methodology and Gradient Boosting for modelling complex nonlinear relationships. Chapter Two presents the literature review. It discusses potato post-harvest losses and their determinants, varietal differences, composite loss measures, RSM, gradient boosting, and hybrid modelling approaches. Finally, the methodological and empirical gaps which justify the study are identified.

Chapter Three outlines materials and methods used in the study. It presents research design, sampling procedures, data collection, experimental design, variables and response measurements, data preprocessing, and statistical techniques. The chapter also describes the procedures used for Principal Component Analysis, multilevel beta regression, RSM, gradient boosting, and the development and evaluation of the hybrid model. Chapter four describes the development and evaluation of the proposed GB-RSM framework based on simulated Central Composite Design data. First a second-order Response Surface Model is fitted to collect the linear, quadratic and interaction effects between the predictors. Residuals from the RSM model are then fed into the Gradient Boosting model to capture the patterns that were not explained by the polynomial model. The hybrid GB-RSM model is then evaluated against the RSM and Gradient Boosting models individually to examine the predictive performance and modelling power of the hybrid model.

Chapter five applies multilevel beta regression to potato post-harvest losses in the farm-market value chain. The chapter tests for a proportionality of postharvest losses and addresses the hierarchical nature of the data. It reviews the impact of storage practices, storage time, market distance, transport, chemical application, extension education, and other factors on relative post-harvest losses. Chapter six applies the hybrid framework of RSM-GBM and employs the Composite Loss Index (CLI). Principal Component Analysis is used to reduce weight loss, sprouting and rotting to

an integrated measure of post-harvest deterioration. The chapter then fits and compares the RSM, Gradient Boosting and hybrid RSM-GBM models and uses the response surface and variable importance analysis to determine the important factors affecting the overall post-harvest losses. Chapter Seven presents the conclusions and recommendations of the study. It presents a summary of the main findings in the context of the research objectives, identifies the contribution made by the proposed GB-RSM framework, and offers practical and policy recommendations to mitigate potato post-harvest losses. The chapter also discusses the limitations of the study and suggests areas for further research, including the application of the hybrid framework to other nonlinear problems and machine-learning algorithms.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents a review of literature relevant to the modelling of potato post-harvest losses and the development of the proposed hybrid Gradient Boosting-Based Response Surface Methodology (GB-RSM) framework. The review provides the theoretical and empirical foundation for the study by examining the nature and causes of potato post-harvest losses, factors influencing losses along the farm-market value chain, and differences in loss susceptibility among potato varieties. It further examines approaches for combining multiple loss indicators into a composite measure.

The chapter also examines statistical and machine learning methods applied to Agri-data modelling, particularly Response Surface Methodology (RSM), Gradient Boosting and multilevel beta regression. The advantages and disadvantages of these methods are explored within the context of the potato post-harvest loss data, which is nonlinear, proportional, and hierarchical. Finally, the chapter summarizes the existing hybrid modelling approaches that use both statistical and machine-learning methods and highlights a few gaps in the literature that require further research and development of the proposed framework for hybrid modelling (GB-RSM). The literature reviewed in the preceding chapters is thus used to inform the choice of methodology in the following chapters of the thesis.

2.2 Post-Harvest Management

Post-harvest losses in agriculture pose a major problem globally leading to food insecurity and economic losses and waste of resources. Specifically, potatoes, as one of the most popular vegetables worldwide, are prone to post-harvest losses which can be considered considerable. Such losses occur at different points of the potato supply chain, including during harvesting, storage, transport and in the market. Based on FAO estimates, approximately one-third of all the foodstuff produced by humans is wasted (Olakiumide, 2021). Factors that contribute to these post-harvest losses are numerous and include; mechanical damage during harvest, inappropriate storage conditions, the environment, including humidity and temperature, as well as ineffective handling during transportation. Besides the disruption to food supply, these losses also have economic consequences of farming, particularly in those regions where potatoes are

widely cultivated. Potatoes are susceptible, which makes storage and transportation a special challenge.

Different studies have demonstrated that proper storage conditions such as regulated temperatures and moisture can substantially minimize post-harvest losses in potatoes. Along with this, better handling procedures that include reducing mechanical damage during harvesting and transportation have proved to be vital in keeping potato quality at good levels during post-harvest processes. poor handling and the absence of such infrastructure in some parts of the globe is a major factor of huge losses experienced across the world (Bisht & Singh, 2024).

According to Hameed et al., (2018), advanced investment in breeding and technology improvements in gene engineering have improved potato production in the recent past. Current research on post-harvest losses and quality improvement of seed potatoes reveal that approximately 30-40% of potatoes lose their quality after harvest. Although much work has been done in post-harvest management of potatoes to improve quality and storage life, according to some estimates, post-harvest losses of potatoes are still very high globally. Like other perishable food supply systems, the potato supply system employs many players within the food supply chain system using many production and supply techniques (Degebasa, 2020).

The effects of transportation and multi-supply chain handling on the quality of fresh potatoes have not been studied extensively or understood. It is thus essential to understand the rate at which quality starts dwindling and include the data in central control of the potato supply chain that ensures customer satisfaction and income improvement for the farmers. Decontamination employing different agents is critical to addressing microbial load on the surface (Levaj et al., 2023). It minimizes post-harvest spoilage and guarantees the product's compliance with legal microbial quality requirements. Potato products need to be treated with chemicals that do not harm the environment and can also improve the quality and storage duration of those products, as chlorine, which is usually used in disinfection, has adverse effects on the environment and human health (Zhu et al., 2024).

Post-harvest losses (PHL) are quantifiable both in quantity and quality losses in agricultural produce between the time of harvest and the time of final consumption.

The concept includes not only physical losses the weight or volume of produce decreases but also the qualitative losses that decrease the nutritional value, marketability, or consumer safety (Levaj et al., 2023). Post harvest losses are caused by a multifaceted interaction between biological, mechanical, environmental, and socioeconomic factors that act at various nodes of the agricultural value chain.

Structural shortcomings in agricultural infrastructure, storage technology, transportation networks, and market systems are the main causes of disproportionately high post-harvest losses in developing countries, compared with high-income countries (Bisht & Singh, 2024). According to the estimates made by Food Agricultural Organization (2019), about a third of all food that is produced worldwide to be consumed by humans is lost or wasted, with post-harvest losses within this category making a significant part of this data (Mugo et al., 2021; Olakiumide, 2021). In sub-Saharan Africa, the total losses in form of cereals at the farm and storage levels have been estimated to cost the smallholder farmers a total of US 4 billion each year (Ariong et al., 2023; Yonas et al., 2021). Due to the high moisture content and the biological sensitivity of perishable crops like fruits, vegetables, and root tubers, their loss rates are relatively high compared to cereals.

The value chain of perishable crops suffers losses because of numerous factors. Microbial deterioration, physiological aging, and pest and disease infestation are biological determinants. Mechanical factors include bruising, abrasion and physical damage that occur due to the harvesting tools, handling practices and vehicles used in transportation. Biochemical deterioration is accelerated by environmental factors like high temperatures, changes in humidity, bad ventilation, etc. The socioeconomic environment, access to credit, lack of information, poor infrastructure, and ineffective institutional response limit access to loss-reducing technologies and practices among smallholder farmers and traders (Mohd Ali et al., 2022; Njogu, 2022).

Value chain analysis offers a practical conceptual framework of establishing where exactly the cause of loss arises and what promotes losses. Value chain approach follows the path of the agricultural produce starting with production through transportation, distribution to ultimate consumption identifying the value adding activities, actors and the relationship in each stage (Ariong et al., 2023). When used to analyze the post-harvest losses, this framework can allow researchers to break down

total losses by stage, ascribe losses to stage-specific causal variables and leverage points through which interventions can be targeted. Other studies that use the value chain framework to determine the post-harvest losses in horticulture in sub-Saharan Africa are by (Mburu et al., 2023) and they all highlight the multi-stage character of the losses and the need to focus on systemic, not isolated, interventions.

The amount of post-harvest losses differs significantly depending on crop, geographic location, stage of value chain and the technological and institutional environment within which production and marketing activities are conducted. Potatoes, in particular, a high-moisture and perishable tuber crop, which is prone to mechanical damage, disease, and sprouting, can be especially devastating in drivers of post-harvest losses when the conditions used in storage and handling are inadequate (Bisht & Singh, 2024). Quantification, understanding, and ultimately reducing these losses demand not only solid empirical data, but also proper techniques of analysis-areas where this research would be aiming to make its contributions.

The meaning of the term "quality" is relative to the consumer, the perspective they take, and how they propose to use it. Quality can be described either as the absence of defects or as the extent of the best standard of a particular item. Consumer-oriented product qualities focus on those attributes that affect consumer acceptability and purchase of farm products (Petrescu et al., 2020). Most post-harvest research focuses on the fresh commodities' product attributes, which some standardized testing may measure. Fresh market potatoes are characterized by physical, nutritional, chemical, and safety attributes (Angeli et al., 2020). Customers who consume potatoes are susceptible to certain qualities provided that they include size, shape, taste and color. The quality attributes of fresh potatoes include texture, skin color, size, form, and weight (Bedini et al., 2024). It has been observed that the genetic makeup of the potato cultivar, production technology, post-harvest handling and storage technology influence potato processing and end-use quality parameters. Indeed, throughout the maturity of the fresh potato, it experiences minor changes in shape and size. Typically, the potato tuber decreases in size at the final stage of its development due to respiration and reduction in water content (Gunathilake et al., 2021). The alteration of tuber shape and size results from the degradation of pectin, cellulose and lignin in the cell wall.

As a result, post-harvest handling should minimize water loss and avoid undesirable respiration rates or loss of firmness in fruit. Controlling humidity and temperature are the key methods to prolong the shelf-life of fresh tubers, fruits and vegetables. Low temperature is preferred in potato storage because higher temperatures accelerate tuber respiration and reduce their shelf-life (Firdous, 2020). The storage temperature of fresh potatoes depends on the maturity stage and cultivar of the potatoes. Levaj et al. (2023) demonstrated that the temperature, variety and duration of storage significantly affect the degradation of potato quality, including factors such as color, texture, activation energy, total phenolic content, and weight loss. High temperature, low relative humidity, and lengthy storage conditions result in the degradation of essential elements in potatoes, notably vitamins.

In Kenya, potato production and consumption have been rising, with the crop ranking third in consumption behind maize and rice (Andati et al., 2023). The major potato-growing regions in Kenya are Nyandarua, Bomet, Meru, Uasin Gishu, Narok, Nyeri, Nakuru, Kiambu, West Pokot, and El-Keiyo Marakwet counties (Onditi et al., 2022). Most potatoes grown in Kenya are first sold as fresh products on the local market before being processed into various cuisines at the home or commercial level (Musita et al., 2019). Additionally, processed potato goods like chips, crisps, and French fries are highly sought after by urban consumers and make up many of the menu items in major restaurants and hotels. This means that for many Kenyan consumers, potatoes constitute a significant portion of their diet.

Post-harvest management of potato is essential to help maintain the nutrient quality of the crop and reduce postharvest losses. However, post-harvest handling and storage greatly affects the safety standards of potatoes to be consumed by humans. This is because potatoes contain a family of steroidal toxic secondary metabolites known as glycoalkaloids (GAs) and are naturally toxic compounds that exist in every part of the potato and are produced to protect the plant from parasites and diseases due to their antimicrobial insecticide, and fungicide properties (Schrenk et al., 2020). The substances may be hazardous to humans when taken internally in large volumes. A tyrosine kinase inhibitor can lead to toxicity with potential life-threatening issues such as impaired neuronal function and gastrointestinal problems which if not properly managed, may lead to death. These stress variables of post-harvest handling such as

light; storage temperatures and physical injuries/bruising affect the levels of these toxins. The stress factors connected to post-harvest handling, such as exposure to light, storage temperatures and injuries or bruising of fruits, noticeably raised the levels of these toxins. Marketing and distribution form another key challenge to the potato value chain in Kenya, and considerable losses have been observed in this phase of post-harvest handling.

Musita et al., (2019) estimated that about 25-45 % of the potatoes available in Kenyan markets are green or damaged. Some traders expose the potatoes to adverse temperature, light and other stress factors that enhance the synthesis of glycoalkaloids (GAs) in the tubers. Besides raising alarm over food losses, it poses health risk for all consumers of potatoes, given that the sale of these substandard, contaminated or greened potatoes keeps putting them through these toxins. By addressing the post-harvest handling issues among potato dealers, food losses can be avoided, and food safety can be promoted, thus safeguarding consumer health. Storage conditions such as air ventilation, relative humidity and temperature levels are vital in maintaining the quality of tubers. Post-harvest losses of potatoes during storage occur due to weight loss, greening, sprouting and rotting. This weight loss is caused by respiration, where sugar and starch are converted into carbon dioxide and water and moisture loss due to evaporation (Grudzińska et al., 2022).

Martín et al., (2022) argued that if the quality of the seed is compromised, good storage, proper harvesting techniques, handling and storage will not guarantee the longevity of tubers. Immature harvesting of tubers and mechanical injuries alleviate the loss of edible tubers and the quality of seeds for the next planting season. High temperature and poor aeration during storage increase sprouting and rotting, reducing edible quality. They argued that temperature and humidity control help lower sprouting and spoilage rates, increasing the potatoes' shelf-life. Also, implementing effective pest and disease control measures ensuring high quality and disease-free potatoes are well stored maintains the potato tubers for a more extended period than disease-infested tubers that are susceptible to spoilage.

Among the notable methods of storing potatoes and preserving their quality include sprout suppressants, pit storage and cold rooms. Potatoes stored in pits exhibit high temperatures, which accelerate the sprouting rates. To preserve the quality of tubers

stored in pits, approved sprout inhibitors such as Chloropropham (CIPC) are used to increase the shelf life and maintain its quality. It is worth noting that these inhibitors have, in recent years, been banned by the European Union Commission in all European Union countries due to rising health and environmental concerns. These chemicals have yet to be adopted in Kenya since smallholder farmers cannot afford them. Instead, they sell their produce immediately after they identify a potential buyer and store the remaining tubers in dark rooms and pits since storing in dark rooms prevents the formation of chlorophyll and subsequently reduces sprouting and spoilage (Gikundi et al., 2023)

In a study to determine the cause of post-harvest loss in sweet potatoes, (Wu et al., 2023) established that temperature is a critical factor in maintaining the quality of the harvested roots. They argued that storing and transporting potatoes at a reduced temperature will likely maximize their shelf life due to reduced respiration rate and deterioration of the edible quality. Consequently, storing sweet potatoes under lower temperatures causes chilling characterized by discoloration, decay, and surface pitting. Subjecting them to high temperatures also accelerates sprouting and pathological and physiological activities that destroy their quality. They found that post-harvest losses occur during harvesting, grading and sorting, transportation, packaging, storage, marketing and processing. To curb this wastage and guarantee food security, a comprehensive approach geared towards post-harvest handling and management of roots and tubers includes devising new harvesting technologies, storing, packaging and husbandry to develop new varieties resistant to pests and diseases.

According to Komen et al., (2017), fast-food outlets such as Chicken Inn, Kentucky Fried Chicken, and Pizza Inn, among other outlets, have established their branches in Kenya, and a consistent supply of good quality tubers is critical. However, due to losses during harvest, storage, and transportation, our local markets need help adequately and consistently to meet their demands, necessitating the external sourcing of tubers. Therefore, the findings of this study help educate farmers on best practices for potato storage and post-harvest handling and management of tubers as they await better prices.

2.3 Poisson Regression in Agricultural and Post-Harvest Loss Modeling

The Poisson regression is the model of the analysis of count data and has been used extensively in the agricultural sciences to model discrete count outcomes, including the number of pests, the occurrence of a disease and the number of units of produce lost during storage or handling (Ma et al., 2025). The model was built on the assumption that the response variable is a Poisson variable with mean equal to the variance and a log linking function. Poisson regression has been applied to estimate the number of damaged units at each stage of the supply chain, which serves as a statistically based baseline model instead of more general alternatives.

The count data regression models were thoroughly covered by (Swartz et al., 2022) who showed how the model can be applied to various disciplines such as epidemiology, ecology, and agricultural economics. He pointed out that Poisson regression works in theory, but in practice nearly all counts data in social and agricultural sciences are over dispersed. This particularly applies to the data of post-harvest losses since the losses are dependent on the biological (sprouting, rotting, microbial activity), physical shocks in transit, which add extra variability. When the overdispersion of the Poisson regression is ignored, the standard errors are understated and the test statistics are inflated, as well as the effects of predictors are spuriously insignificant (Khan & Hasan, 2022; Saha et al., 2026).

Poisson regression has been employed as a baseline model of agricultural production and loss outcomes in several empirical studies in sub-Saharan Africa before moving on to more sophisticated count models. (Wubetie, 2024) used Poisson models on food security data in Ethiopia and concluded that overdispersion greatly underestimated standard errors, which supports the importance in assessing whether to infer the Equi dispersion assumption is met. Likewise, in an experiment to determine the extent of crop damage by pests in smallholder farms in Tanzania, (Busungu, 2025) established that, when the underlying distribution was heavy-tailed, Poisson regression systematically underestimated the prevalence of losses. The results are directly applicable to the Kenyan potato scenario in which distribution of losses must be right skewed with most of the actors having moderate losses with small percentage with catastrophic losses.

Poisson regression, although limited, is a very important reference model in multilevel modelling systems. (Almasi et al., 2016) extended Poisson model to a multi-level context and demonstrated how a two-tier Poisson model (with random effects at the cluster level) could account for within-group correlation and still have ratio of incidence rates that are interpretable. The deviance, Pearson Chi-square goodness-of-fit, statistics of Poisson regression are diagnostic standards used to compare negative binomial and zero-inflated alternatives based on likelihood ratio tests and information criterion (AIC and BIC). To estimate the extent of overdispersion in the count-based loss outcome after harvest, Poisson regression is estimated as a baseline model to be used in the current study prior to the estimation of the negative binomial specification.

2.4 Negative Binomial Regression

The most common extension of the Poisson regression to the over dispersed count data is negative binomial (NB) regression model. It is defined as a Poisson-Gamma mixture, but the Poisson rate parameter itself is a Gamma-distributed random variable, with an extra dispersion parameter, which relaxes the Equi dispersion condition (Morris & Sellers, 2022). The NB model has found broad application in health sciences, ecology, and agricultural economics to model counts with more variability than is possible with the Poisson model. Negative binomial regression has been chosen as the model of choice in agricultural post-harvest literature due to its ability to both explain overdispersion and provide rate ratios on the covariate effects that are easy to interpret.

In their original study on count data econometrics, (Almasi et al., 2016; Khan & Hasan, 2022) developed the theoretical basis of NB regression, differentiating between the NB_1 specification (variance is proportional to the mean) and the NB_2 specification (variance is quadratically proportional to the mean). Negative binomial 2 (NB_2) parameterization also abbreviated $NB(\mu, \theta)$ when $var(Y) = \mu + \frac{\mu^2}{\theta}$ is by far the most used in practice since it can have dispersion that increases with the mean, which is empirically true of most agricultural and ecological data. The NB_2 model is a convergence Poisson as $\theta \rightarrow \infty$, giving a natural nested model structure, making possible likelihood ratio tests of the overdispersion hypothesis.

(Alhassan et al., 2025; Nohamba et al., 2022) used negative binomial regression on smallholder farm data in sub-Saharan Africa and discovered a significant overdispersion of the number of losses, which is evidence that NB is applicable, as opposed to Poisson regression. Their experiment involving a variety of crops and various nations revealed that storage technology, market accessibility, and agro-ecological zone were major predictors of the number of losses and the NB regression offered more accurate confidence intervals than the Poisson regression. A systematic review of post-harvest loss assessment approaches in sub-Saharan Africa by (Bassey, 2018; Ndaka et al., 2012; Nohamba et al., 2022) suggested that descriptive methods of post-harvest loss assessment were statistically inadequate and a count regression model, such as NB regression, was necessary to model discrete loss outcomes in staple crop value chains.

A more recent methodological advancement that is especially applicable to the present study is the use of multilevel or mixed-effects negative binomial models (Almasi et al., 2016). Multilevel NB models are direct extensions of the fixed-effects NB model in that random effects are included at the higher-level grouping unit, which could be a county or a stage of a value chain, thus explaining the clustering of observations within groups. The two-level NB model was parameterized using a random intercept, which revealed that the conditional mean is decomposed into a fixed effect and a normally distributed random effect (Zhou & Gilmour, 2024). This method produces cluster-specific prediction and can separate the total variance into within group and between groups with intraclass correlation coefficients.

(Ariong et al., 2023) carried out one of the first systematic loss analyses in Uganda with descriptive statistics and ordinary least squares regression within the narrow scope of the study of potential measure of potato post-harvest losses in East Africa. Even though their work offered valuable empirical standards, it failed to take into consideration the hierarchical organization of data gathered in different production zones. (Githumbi et al., 2024) extended this to the Kenyan setting, using logistic regression on the dichotomous outcomes of loss, but did not model the number of absolute losses or proportional losses as continuous losses. This research fills this gap by estimating the multilevel NB regression on the outcome of the absolute count, with counties and value chain stages symbolizing the grouping of levels as the random

effects and hence the capacity to gain a better statistical inference along with a more substantive understanding of the determinants of loss throughout the entire potato value chain.

2.5 Beta Regression for Proportional Outcomes

According to (Chen & Yang, 2024), Ferrari and Cribari-Neto (2004) introduced beta regression to the statistical literature as a regression model specifically designed to estimate continuous response variables that are in the open unit interval $(0, 1)$. In contrast to linear regression which assumes the normally distributed, unbounded residuals, beta regression assumes that the response takes a beta shape that is bounded and may be of an extremely wide variety of shapes, such as symmetric, left-skewed, right-skewed, and bimodal based on its two shape parameters. (Ferrari & Cribari-Neto, 2004; Figueroa-Zuñiga et al., 2012) reparametrized the beta distribution using a mean parameter μ and a precision parameter ϕ and linked the mean to a linear predictor using the logit function to obtain a model with structural similarities to logistic regression with binary outcomes but a continuous proportion (Ferrari & Cribari-Neto, 2004).

Smithson and Verkuilen (2006) broadened the conceptual and applied scope of beta regression by showing through a series of simulation studies that beta regression performs significantly better than linear regression used with bounded proportional data (Abonazel et al., 2022). Their analysis demonstrated that the ordinary least squares (OLS) with proportion data gives biased coefficient estimates, inappropriate standard errors and predicted values that do not lie under the $(0, 1)$ interval that are removed when using beta regression. Smithson and Verkuilen (2006) further demonstrated the application of beta regression with empirical data, both in psychology and social science, and showed its applicability in other contexts than econometric use. They inspired the use of beta regression throughout the research that deals regularly with proportional results such as health sciences, environmental studies, education research, and agricultural economics.

The computational infrastructure of beta regression was also extended by Cribari-Neto and Zeileis (2010) who implemented the *betareg* package in R that offers maximum likelihood estimation, residual diagnostics, influence measures and model

misspecification tests (Cribari-Neto & Zeileis, 2010). Their study gave a comprehensive description of the score function and information matrix of the beta regression model and showed how to compute marginal effects, which are nonlinear functions of the estimates of the coefficients and the predicted means. This implementation allowed beta regression to become practically available to applied researchers and made it adoptable by agricultural researchers. (Fokkema & Zeileis, 2024) applied the *betareg* package to homoscedasticity in the precision parameter, through a variable dispersion sub-model, where the log of ϕ is modeled as a linear function of covariates, allowing extra flexibility in the context where the variability of proportional outcomes per se is modelled by predictor variables.

Beta regression has been applied to model various proportional results in the agricultural literature, such as allocation of land to crop, market participation rate and scores on dietary diversity (Anteneh & Endalew, 2023). A quasi-maximum likelihood estimator of the response model of proportional outcomes proposed by Papke and Wooldridge (1996) closely related to beta regression has been applied to agricultural economics in model estimation, such as intensity of fertilizer use and income shares (McDonald, 2009). Their estimator is strong to misspecification of the assumed distribution and yields consistent estimates provided the logit link function of the mean is properly specified. (Marina Pare, 2025; Misango et al., 2022) also illustrated that the Papke-Wooldridge method can be used in agricultural data analysis, and this method is superior to OLS and Tobit models in which the outcome is limited.

Specifically, the use of beta regression in post-harvest loss rates has been studied in a limited but expanding body of research. (Marina Pare, 2025) used fractional response models on post-harvest loss datasets that consisted of data on smallholder rice farmers in Tanzania and found that proportions of losses against harvested yield were much larger with the use of traditional methods of storage, with frequent location along tarmac roads, with limited access to extension services. It was observed that beta regression yielded more suitable models than OLS in all the specification tests, and estimation coefficients were significantly different in sign and magnitude compared to OLS estimation, especially of interaction terms (Bahrami et al., 2025). This observation highlights the significance of choosing the appropriate model of distribution in cases where the response variable is bounded.

In a review of post-harvest loss assessment and modelling methods of perishable crops in sub-Saharan Africa, (Qange et al., 2024) proposed beta regression to proportional loss results, stating that the conventional linear model failed to predict loss at the boundary levels, and underestimated it at the higher levels because of its inability to reflect the boundary limitations. They further noted that the heteroscedasticity of proportional agricultural data, with the largest variance at the intermediate levels of loss, hence naturally addressed by the precision parameter of the beta distribution, would under OLS be addressed by ad hoc variance-stabilizing transformations. Their analysis involved literature on Kenya, Uganda, Tanzania, Malawi and Ethiopia and showed that proportional value chain losses in potato, maize and horticulture models are invariably better explained by bounded distributional models.

2.6 Multilevel Beta Regression

Hierarchical linear modelling (alternatively known as multilevel modelling or mixed-effects modelling) was designed to analyze data whose structure is hierarchical or nested, with the observations at the lower level of the hierarchy clustered within the units of the higher level (Almasi et al., 2016; Mang et al., 2021; Zhang & Lai, 2024). They laid the theoretical background in their contribution toward the random-effects models of longitudinal data, and later by (Fokkema & Zeileis, 2024) to the generalized linear model framework. (Hoffman & Walters, 2026) has given a practical guide to multilevel modelling across fields in a comprehensive way that defines the framework that most applied researchers use to specify and interpret multilevel models. In agricultural studies, multilevel models have been used to explain the clustering of farmers by villages, by districts, and finally regions, a framework quite equivalent to the nesting of the value chain participants in counties in the present study.

The multilevel beta regression model of combination of beta regression with multilevel structure is a new methodological approach that considers both the proportionality of the response and the hierarchical dependence of observations at the same time. (Figueroa-Zuñiga et al., 2012) suggested a mixed-effects beta regression model which includes random effects at various levels in the logit model of the mean of the beta distribution and showed that it performs better than a basic beta regression in data sets where cluster membership explained non-trivial amount of variability in the proportional outcome. Their simulation experiment indicated that the omission of

group-level heterogeneity in beta regression inflated Type I error rates up to 30% and provided biased estimates of level 2 covariates that have a direct application in value chain data where county-level characteristics form a significant portion of post-harvest results.

(Ju et al., 2020) generalized the multilevel beta regression model to three level nested data designs and demonstrated the model using longitudinal data on educational achievement, where students were nested within classroom within a school. They highlighted that the appropriate specification of random effects at each level is the key to getting unbiased estimates of fixed effects and that the attribution of the variance across levels results in systematically misleading substantive inferences. Their method of estimation, adaptive Gauss-Hermite quadrature, has since been used in numerous statistical packages such as *glmmTMB* in R and the *meglm faml* of commands in Stata, and three-level beta regression models are now practically available to applied researchers. In the present research, the three-tier model of individuals within value chain steps within counties within individuals gives direct methodological support to the selected modelling strategy (Watson, 2024).

Multilevel beta regression has been used to model proportional outcomes like the self-rated health, disease prevalence rate, and body fat percentage in the health and epidemiology literature (Almasi et al., 2016). (Santa-Ramírez et al., 2023) used a hierarchical beta regression model to estimate the spatial distribution of prevalence of malnutrition in municipalities in Colombia with the finding that about 35% of total variation in nutritional outcomes was explained by the individual municipality-level effects holding all others constant. Their analysis showed the policy applicability of the division of variance at different levels: at the ICC of over 0.2, interventions based on the infrastructure and food system governance of municipalities were shown to be less costly than the individual-level nutritional programs. Similarly, when a large proportion of variability in proportional losses can be attributed to the infrastructure and market governance of counties, then this would suggest that county-scale policy interventions, like road improvement or cold chain investment, would produce a larger overall reduction in the aggregate losses than actor specific training programs (Santa-Ramírez et al., 2023).

In the agri-food systems literature, few studies have used multilevel beta regression to model proportional loss or efficiency results (Yaremych & Preacher, 2024). (Ogheneruemu & Ifoga, 2023) used a multilevel fractional logit model to model post-harvest loss information of yam farmers in Nigeria, categorizing farmers at the village level and district level and discovered that the village-level storage infrastructure and district level market connectivity moderated the individual level impact of storage technology on proportional losses. Farmers were nested within villages and districts and the study showed that the effect of the storage technology level at the individual level had a moderating effect on the proportional losses based on the level of storage infrastructure at the village level and the level of market connectivity at the district level. In research on enhanced grain storage in Uganda, (Sunday et al., 2022) also discovered that the share of proportional loss reduction due to improved hermetic storage bags was greater in villages where higher loss rates at harvest are observed at baseline, indicating that contextual factors that are measured at the village level are significant moderators of technology effects.

The data of agricultural value chain is inherently multilevel in nature. Farmers, transporters, wholesalers and retailers are in common agroecological spaces, market structures and regulatory environments that produce correlated results in clusters, despite the individual-level attributes that they have controlled (Sunday et al., 2022). Regression models based on classical assumptions such as using all observations as independent variables underestimate standard errors of coefficient estimates and also over indicate the accuracy of statistical inference when such clustering occurs (Ferrari & Cribari-Neto, 2004; Geissinger et al., 2022). The extent of this issue is determined by the intraclass correlation coefficient (ICC) and the measure of the percentage of the overall variability of the outcome that can be attributed to the differences in clusters. Analytically, (Dong et al., 2026; Stürner et al., 2024) demonstrated that Type I error rates of over 20% could be obtained by even modest levels of ICC of 0.05 to 0.10 when regression analyses were done without consideration of clustering, which is regularly surpassed in agricultural survey data.

The key diagnostic instrument that justifies the multilevel modelling approach is intraclass correlation coefficient (Devine et al., 2024). An intraclass correlation coefficient (ICC) is the ratio of the variance of observations within clusters to the

overall variance, which reflects the similarity of scores within clusters and therefore the dependence of observations within clusters (Devine et al., 2024). (Williams et al., 2024) suggested that ICC of 0.05 should be considered enough to warrant a multilevel model, whereas (Nugent & Kleinman, 2021) contended that even a small ICC can result in significantly biased inference with large samples and many clusters. Other non-Gaussian multilevel sources compute ICC in a more complicated manner since the level 1 residual variance is conditioned by the link function and the distributional family. The latent variable interpretation of the binary outcome is suggested by (Devine et al., 2024) to use in the context of multilevel logistic and probit models, in which the latent level 1 variance is estimated as $\frac{\pi^2}{3} = 3.29$. The identical approximation is applied in the context of the multilevel beta regression, as described by (Williams et al., 2024), since the beta regression is compared on the logit scale, and the logit-normal distribution has a level 1 variance in the latent scale of $\frac{\pi^2}{3}$.

Information criteria are typically used to select a model in multilevel regression, with the Akaike Information Criterion and Bayesian Information Criterion being the most common information criteria, and likelihood ratio tests being used as supplementary information criteria (Harbecke et al., 2024; Kasali & Adeyemi, 2022). According to (Kasali & Adeyemi, 2022) the authoritative source on information-theoretic model selection was given by Burnham and Anderson in 2002, which demonstrated that AIC is asymptotically equivalent to leave-one-out cross-validation, and thus suited to predictive model selection, whilst BIC punishes overfitting of a model more severely and is better suited to identifying the true model among a set of candidate models. In the multilevel modelling context, Raudenbush and Bryk (2002) suggested restricted maximum likelihood to compare models that only differ in their random effects structure, and full maximum likelihood to compare models that differ in their fixed effects, a distinction that is significant in specifying the test statistic in likelihood ratio tests correctly (Zhang, 2025). The current study exhibits this difference by the Laplace approximation of the marginal likelihood to derive the maximum likelihood estimates of the multilevel NB and multilevel beta regression models.

2.7 Machine Learning Application

Machine learning has enhanced the agricultural sector by providing new approaches that improve productivity, resource employment, and sustainability. The first and most useful ML application in agriculture is precision farming. The sensor and satellite feed collected data can be analyzed with the help of techniques like decision trees and neural networks to gain insights for effective resource management, especially in the irrigation process.

Studies show that by applying machine learning in irrigation, one can cut water use by 25%, hence less water usage but equivalent production (Sharma et al., 2020). They found that machine learning can enhance water usage efficiency in irrigation by raising yields by up to 25%, and thus, reducing water consumption without affecting production. In the same way, the models for the application of fertilizer and pesticide through machine learning have reduced the input cost and the impact on the environment thereby improving the sustainability of farming. Response surface methodology coupled with gradient boosting machines is a robust tool in complex, nonlinear modeling between numbers of predictors and responses in experimental and industrial applications (Ahmad et al., 2024). Response Surface Methodology is the traditional method that has been utilized in optimization and modeling of relationships between variables in an experimental design. It usually utilizes the use of polynomial regression which is perfect in approximating continuous or smooth relationships between the predictors and the responses. Nonetheless, the models may not be sufficient to explain complex and non-linear relationships in the real-world systems, and it is here where machine learning techniques such as gradient boosting machines come into play since they outperform the models by being more flexible and accurate in establishing the relationships in the model. This integration can make response surface modeling more efficient, it is possible to make optimization more efficient, gain a better understanding of interactions of different predictors in complex systems (Karatas et al., 2020).

Gradient Boosting Models or GBM are an ensemble machine learning algorithm, which uses weak learners (usually decision trees) to produce a powerful predictive model (Friedman, 2001; Sibindi et al., 2023). One of the most important strengths of GBM is that it can capture, nonlinear complex relationships between responses and predictors, which may be beyond the capability of linear regression models such as

those used in RSM. It is especially effective in the processing of large sets of data reflecting complicated interactions and highly precise predictions can be obtained. After the basic idea of GBM arose several variations of this general method namely XGBoost, LightGBM, CatBoost, etc. emerged enhancing the original algorithm in speed, working with categorical variables, and preventing its overfitting. It has proved very successful in areas like finance, healthcare, and machine learning competitions since it performs better than the other modeling methods (Huber et al., 2022).

Another advantage of using machine learning is in monitoring crops and detecting diseases. The current techniques, particularly Convolutional Neural Networks (CNNs), are extensively used for image-based diagnosis of diseases (Agarwal et al., 2020). For instance, CNNs have been shown to diagnose crop diseases with the accuracy of over 90% thus making it easy to diagnose diseases at an early stage and therefore applying the right measures. It also reduces vulnerability to pests and diseases hence improving on the productivity of the farmland. Such a process minimizes occasions that require labor and gives farmers an opportunity to combat certain challenges efficiently and accurately thus making the farm resilient to hardships (Lu et al., 2021).

The contribution of machine learning is also employed in the prediction of yields. Machine learning can provide very accurate yield forecasts employing different input data on weather conditions, types of soil, and types of crops (Sharma et al., 2020). These are particularly important in decision making since farmers can change their conduct to reflect the conditions necessary that are likely to ensure less loss and higher yields. Further, machine learning increases the efficiency of supply chain management. Predictive models that define storage and transportation requirements help control post-harvest losses. At the same time, recommendation engines enhance marketing opportunities by matching farmers to potential customers, promoting fairer prices, and minimizing food wastage (Khorramifar et al., 2023).

2.8 Factors Causing Post-Harvest Losses

Temperature is the most important environmental factor affecting potato post-harvest loss. Temperature affects all three loss pathways: it controls transpiration water loss via its effect on vapor pressure deficit; it affects the meristematic activity responsible for sprout growth; and it affects growth of the fungal and bacterial pathogens

responsible for rotting (Visse-Mansiaux et al., 2021). For sprouting, meristematic activity and hence sprout production is severely inhibited at low temperatures (below 4°C), but proliferation occurs rapidly at higher temperatures (above 15°C). For rotting, pathogens are generally dormant at very low and very high storage temperatures, with an optimum at intermediate temperatures (10-15°C) for many potato storage pathogens. For weight loss, the effect is roughly linear and positive with temperature, as it affects transpiration. The non-linearity and pathway-specificity of the influence of temperature justifies the use of squared temperature terms in the RSM.

Relative humidity (RH) has a complex effect on potato post-harvest quality. At low RH (less than 70%), the vapor pressure gradient between the tuber flesh and the surrounding air is high, leading to excessive transpiration that results in weight loss, skin wrinkling, and softening of the tuber tissue (Jakubowski & Królczyk, 2020; Khanal et al., 2024). However, at high RH (above 90-95%), condensation occurs on the skin surface, and the surface becomes highly susceptible to germination and growth of fungal and bacterial pathogens, such as *Fusarium* dry rot and soft rot (Gálvez & Palmero, 2022). An intermediate RH interval of 85-90% is the recommended condition for potato storage to avoid both these risks. The quadratic dependence of loss on RH justifies the inclusion of a quadratic RH term in the RSM model.

Time is a natural variable that accumulates the three losses. Weight loss increases with time due to transpiration and respiration losses. Sprouting is limited by the variety-specific dormancy: once the dormancy breaks, sprouts will appear and grow, with the rate of sprouting then determined mainly by temperature and light conditions. The percentage of rotting increases with time as the risk of infection and disease development increases, especially when mechanical injuries and high humidity are present. (Xue et al., 2023) showed that ambient storage of potatoes in humid highland regions exceeds 15% weight loss after 90 days. The interaction between storage time and temperature is particularly pronounced as lower temperatures significantly prolong the safe storage time by inhibiting the three pathways of loss.

Light exposure during storage induces chlorophyll synthesis in potato tubers: this results in the familiar green discoloration of potato tubers that coincides with solanine (a toxic glycoalkaloid) build-up (Schrenk et al., 2020). In addition to food safety

concerns, light is a well-known photo stimulant to potato sprouting: phytochrome signaling pathways mediate light-induced activation of endogenous gibberellin synthesis, thus promoting or hastening the end of dormancy to initiate meristematic growth (Zhu et al., 2023). The light and temperature interaction is significant: at low temperatures, the sprouting-inducing effect of light exposure is greatly reduced, whereas at ambient or high temperatures, light and temperature interact to accelerate sprouting. Darkness is recommended for the prevention of photo-induced sprouting.

Physical injuries incurred during harvest, grading, sorting, packing and transport result in damage to the potato periderm (skin) that provides entry for fungal and bacterial pathogens that directly contribute to rotting loss (Zhu et al., 2023; Zhu et al., 2024). Besides its pathogenic impacts, mechanical damage activates wound-induced ethylene production and respiratory compensation responses in both damaged and undamaged tissues, which hasten physiological ageing. The impact of mechanical damage in commercial Kenyan potato supply chains is great: traditional harvesting with hand hoes results in extensive tuber scarring, and poor packaging and handling practices cause considerable bruising during transit to markets. Mechanical damage mitigation is well recognized as a high-value, low-cost measure to reduce rotting losses.

Curing is a post-harvest treatment where newly harvested potatoes are held at warm temperatures (15-20°C) and high relative humidity (85-95% RH) for 7-14 days to induce the formation of a wound-periderm (a corky layer of suberized cells) on cut surfaces, bruises and other mechanical injuries from harvesting and handling (Gálvez & Palmero, 2022; Xue et al., 2023). This response prevents pathogens from entering the potato and minimizes water loss from damaged tissues and greatly reduces rotting losses in subsequent storage. Curing is not practiced by Kenyan smallholders due to a lack of storage facilities and the need to quickly get the product to market (Pineda-Vélez et al., 2024). This contributes to high rotting losses in the Kenyan potato value chain. The benefits of curing are likely to be greatest at high levels of mechanical damage, suggesting the need to include a curing duration and mechanical damage interaction effect in the response surface model (Jahani et al., 2022).

2.9 Varietal differences in post-harvest loss

Significant varietal differences in post-harvest loss susceptibility have been reported for the major commercial potato varieties grown in Kenya. Potato variety

characteristics that control post-harvest behavior include skin texture and thickness (which affects water loss and resistance to pathogen penetration), dry matter content, dormancy period, and disease resistance (Petrescu et al., 2020; Zhu et al., 2023).

The major commercial variety in Kenya, Shangi, which accounts for 60-70% of the total marketed production, has intermediate skin thickness, dry matter content and duration of dormancy, and is therefore prone to moderate losses along all three pathways (Gikundi et al., 2023; Komen et al., 2017). Tigoni, a breeding line released by the KALRO, is known for its relatively good storage performance in national trials, due to its thick skin and long dormancy. Kenya Mpya is a newer variety which has consistently shown good post-harvest performance in KALRO storage trials, with low losses across all three pathways. Asante, with its short dormancy period and relatively thin skin, is highly susceptible to early sprouting and pathogen-induced rotting and tends to have the highest post-harvest losses of all major varieties grown in Kenya. Dutch Robjin, Markis, Unica and Sherekea are intermediate performers.

2.10 Composite Indices and Multivariate Loss Measurements

The weight loss, sprouting, and rotting are the three main indicators of post-harvest loss, and are products of different biological processes, but are empirically correlated, and tend to increase under unfavorable storage conditions, and tend to decrease under favorable conditions. Univariate analyses that model the three indicators separately neglect this shared information and generate disjointed, even conflicting, recommendations for management when predictors of one indicator contradict those of another. Combining the indicators into a composite index that defines a single continuous response variable is a way to overcome this problem and allow integrated analysis.

The most common approach to constructing a composite index from multiple correlated continuous variables in the social sciences, economics and environmental science is Principal Component Analysis (PCA) (Fakher et al., 2021; Wang & Chen, 2022). PCA finds orthogonal linear combinations of variables (principal components) that account for the largest possible variance in the data, in a sequential order. The first principal component (PC_1) is the co-variation between the original variables that has the largest variance and is generally interpreted as the overall level of the latent

construct in this case, the overall severity of post-harvest loss. Weighting *PCs* retains information about the underlying structure. PCA-based approach to composite index formulation has been used in various agricultural studies for food quality assessment and soil health index for potato post-harvest loss integration (Omrani et al., 2019).

2.11 Response Surface Methodology

Response Surface Methodology is a commonly used statistical tool for optimizing and analyzing the interactions of numerous factors in various research areas, including engineering, agriculture, and manufacturing (Zhang & Wu, 2021). However, the usage of RSM is often limited by missing values, which can originate from different factors, such as limitations in the conduct of experiments, errors in measurements, or dropout in the context of longitudinal research. The management of such missing or incomplete data is an important asset in relation to the dependability and credibility of the subsequent enhancement of RSM models. Thus, there is the need to deal with missing data situation in RSM to mitigate the impacts that may arise from the experimentations and enhance the precision of the predictions. There are several approaches recommended for addressing incomplete or missing data in case of RSM. The simplest and most often applied are imputation methods such as mean imputation, regression imputation, and multiple imputations (Jerez et al., 2010).

Response Surface Methodology (RSM) consists of statistical methods of exploring the correlation among multiple explanatory variables and one or more response variables. RSM is a method to optimize a response by approximating the true relationship between the input variables and response with a regression model. The technique is especially applicable in experimental cases where the aim is to determine the most significant factors affecting the dependent variable and to optimize such factors to have an optimal outcome. RSM has found extensive application in agricultural research to optimize a wide range of processes including application of fertilizers, irrigation scheduling and post-harvest manipulations. An example is the use of RSM to model the optimal temperature, the humidity, and of storage time of potatoes so as to reduce losses due to spoilage of potatoes. By allowing systematic variations of these factors, RSM would determine the most economical conditions to maintain the quality of potatoes (Dave et al., 2017).

The major advantage of RSM is that it can be used to analyze complex systems where several input variables are involved. Conventional optimization algorithms tend to get stuck in multivariate domains, but RSM has the capability to explore the interactions between the variables and evaluate their combined impact on the response. It also renders RSM a perfect method to streamline agricultural processes, as several variances come into play that can be interactively related to each other such as soil quality, the rate of irrigation, and the use of fertilizers. RSM consists of statistical methods of exploring the correlation among multiple explanatory variables and one or more response variables. RSM is a method to optimize a response by approximating the true relationship between the input variables and response with a regression model. The technique is especially applicable in experimental cases where the aim is to determine the most significant factors affecting the dependent variable and to optimize such factors to have an optimal outcome (Romsomsa et al., 2010).

RSM has found extensive application in agricultural research to optimize a wide range of processes including application of fertilizers, irrigation scheduling and post-harvest manipulations. By allowing systematic variations of these factors, RSM would determine the most economical conditions to maintain the quality of potatoes. The major advantage of RSM is that it can be used to analyze complex systems where several input variables are involved. Conventional optimization algorithms tend to get stuck in multivariate domains, but RSM has the capability to explore the interactions between the variables and evaluate their combined impact on the response. It also renders RSM a perfect method to streamline agricultural processes, as several variances come into play that can be interactively related to each other such as soil quality, the rate of irrigation, and the use of fertilizers (Jadhav et al., 2019).

Despite the relative ease of implementation and efficacy, it has systematically overestimated the variance and generates unbiased estimates of missing not at random data (Horton & Lipsitz, 2001). However, multiple imputations which include creating at least two reasonable complete datasets from the observed data with results combined have been known to yield better estimate in case of missing value. The existence of machine learning improves the ability of RSM in solving problems when the missing data is a function of other variables of the experiment (Jadhav et al., 2019; Li et al., 2022). Furthermore, the incorporation of Bayesian methods is very effective

for the issues of missing data for RSM (Li et al., 2022). These methods use information from other studies and models based on probabilities to address the issues producing from lack of data particular the missing not at random type. Thus, when there is a limited amount of information available, the use of the Bayesian RSM approach in solving engineering optimization problems improves the accuracy of the predictions. Such methods include the sequential estimation of the model parameters and the imputation of the missing data that can provide unbiased and efficient estimates if the missing data occur in some systematic fashion. These techniques also simultaneously conduct model selection and missing value imputation and offer unbiased and efficient estimation even if data are systematically missing (Jadhav et al., 2019).

In a study to extend the shelf life of Pears fruit and enhancing the quality using RSM by Dave et al., (2017) the response variables which were considered were total soluble solid, pH, weight loss and titratable acidity. Considering the results of the Soy protein isolate, Potassium Sorbate and Olive oil concentrations in the edible coating of pear fruit, the experiment concluded that RSM was an ideal method for optimization. They concluded that for optimal outcome, much importance must be accorded to the edible coating as it delays and improves the shelf life of the pear by reducing evaporation through respiration and the softening of the fruit. In the study, they also affirmed that the titratable acidity decreased with increase in storage time because organic acid was being transformed into sugar.

As Visse-Mansiaux et al., (2021) pointed out, the tubers develop dormancy during storage, and the sprouting process starts. Some factors that can hinder the sprouting of the tubers include the potato cultivars, the temperature of storage, and the use of suppressants. In Europe and other first-world countries, sprout inhibitors are utilized to maintain quality and address the cons that supply throughout the year is always available. However, since the applications of these chemicals also pose health hazards to people, they opted to ban them. To assess dormancy of the various varieties, the researchers used a pre-harvest chemical, known as maleic hydrazide (MH), and other post-harvest chemicals: 3-Decen-2-one and 1,4-dimethyl naphthalene and Chlorpropham (CIPC). The work also revealed that while Chlorpropham (CIPC) was effective for different potato varieties after six months, 1,4-dimethyl naphthalene and

maleic hydrazide had no helpful impact on the potential of minimizing sprouting rate, in the process, these chemicals accumulate dangerous toxins into potatoes.

Romsomsa et al., (2010) employed Plackett-Burman design and response surface methodology to optimize silk degumming protease from *Bacillus Subtilis* C4 to establish the most significant factors affecting enzyme production. This study employed the Plackett-Burman design to identify factors that significantly affect the system the most. This design did not help determine the right places on the graph for each predictor variable. To enable the definition of the optimum values, the mathematical technique known as RSM was used to define the optimum values of several factors, including the nutritional and physiological parameters of the selected probiotics and their interactions. RSM was noticed to perform more effectively than the Plackett-Burman design since the former may be used to construct models in addition to checking the interaction effects between the various independent variables.

2.12 Gradient Boosting

Gradient Boosting has emerged as one of the most valuable algorithms in the field of machine learning and big data analysis, particularly with structured data (Ali et al., 2024). Originally, this algorithm was targeted at minimizing the total prediction error as a statistical tool and has grown larger in its application and in the dealing with issues in theoretical development as well as in actual application. It stems from an accumulative mechanism that involves constructing a strong learning model using several weak models like a decision tree. Models developed early in gradient boosting, such as those proposed by Friedman in 2001 (Friedman, 2001), formalized how successive models reduce errors made by decision trees in converging towards the optimal solution by minimizing the loss function, such as the mean squared error.

The gradient boosting's adaptability and robustness have helped address the overfitting problem through imputation techniques. The advantage of gradient boosting is its ability to adapt to various loss functions, making it suitable for regression and classification tasks. Its application extends to agriculture, medicine, finance and even engineering. It also has the advantage of handling missing, complex and nonlinear relationships existing in the data and can also perform well with high dimensional data. Due to the inherent risk of overfitting the data, GBM utilizes various

regularization techniques, thus maintaining the model's generalizability (Khamis et al., 2024).

Gradient Boosting is the ensemble-based learning algorithm that constructs the models iteratively, with a new model trying to rectify the mistakes made by its predecessor. The algorithm functions by iterating through decision trees in which each tree targets the given residuals of earlier trees to reach accuracy in prediction. This approach has been extensively used in many other areas including agriculture as it has shown good performance in making predictions and has the capacity to manage multiple, nonlinear relationships may exist in a set of data (Gök & Olgun, 2021).

The robustness of the algorithm lies in the fact that through predicting the instances in the data which are most difficult to predict, one reduces the error in prediction, which leads to higher levels of accuracy of a model in general. A major strength of gradient boosting in comparison with non-tree-based machine learning models is that it allows modeling complex interactions among variables. Applications to agricultural research This is especially pertinent in farm research where the effects of the interactions between the environmental conditions, crop management, and soil conditions are often not straightforward (Bifarin, 2023).

In agriculture, gradient boosting has been applied in predicting crop yield, detection of pests, forecasting crop diseases and optimization of agricultural processes. As an example, gradient boosting models have been used to estimate crop yield given the climate conditions, types of soil, and the farming activities. These models have proven to be very accurate and predictive hence they can be used in forecasting post-harvest losses among potatoes (Mahesh & Soundrapandiyam, 2024). The prediction of moisture in soil is another most vital application since it is involved directly in the management of irrigation. Gradient boosting has been used to develop predictive models that provide farmers with input data on factors that require attention to optimize the schedule of irrigation, such as weather forecasts and soil sensors. The resulting application of gradient boosting has a high possibility of being especially useful in areas with limited water resources, allowing a more efficient use of water (Domingues et al., 2022).

Despite the benefits, the use of gradient boosting also has challenges. A major disadvantage is that it can overfit when not tuned (Domingues et al., 2022). Overfitting happens when the model learns the random fluctuations in the data as patterns, thus the predictions made by the model when applied to unseen data become inaccurate. It is thus very important to do hyperparameter tuning and cross-validation so that the model can generalize to real world data. In gradient boosting, regularization can be different, such as using a learning rate to scale each new tree's contribution and penalize the trees' complexity. The shrinkage technique is applied to shrink the prediction errors of each decision tree with a set learning rate, which helps to prevent overfitting while learning from the residuals of the previous iteration (Huber et al., 2022). Gradient boosting has been used to handle complex data such as time series and spatial data. The literature shows that time series data have strong autocorrelation, which tends to affect the effectiveness of GBM (Effrosynidis et al., 2023). As such, adjusted gradient boosting model, such as the temporal gradient boosting, directly incorporate temporal information into the feature set, making the model adaptable and capable of learning from the previous historical patterns. They have leveraged their strengths to build a powerful, accurate and robust ensemble model when integrated with other models like support vector machines and neural networks. In an ensemble model, gradient boosting handles the structured data, while deep learning models like neural networks handle unstructured data like texts or images.

Karatas et al., (2020) stated that gradient boosting is used when high accuracy is needed and when the data set is either missing or unbalanced. In medicine for instance, where some data may not be complete, gradient boosting is used to determine trends that are required in identifying the correct dimension. It can also be applied in disease diagnostics, treatment and prediction of patient re-admission into the health care facility. In the complex models, gradient boosting models can detect and predict the risk level of diseases like heart diseases and diabetes with better accuracy than the logistic base models. Gradient boosting has proved to be highly versatile in disease surveillance issues that form the basis of several emerging developments in manifold sectors and research. The study by Hou et al., (2020) further affirms that a XGBoost can be used to predict sepsis in a particular hospital since the algorithm achieved an accuracy rate of over 90 percent, thus supporting the proposed algorithm that operates simultaneously with multivariate scatterplots. In addition, gradient boosting is applied

in genomics for diagnosing cancer risks based on genes and outperforms conventional approaches including logistic regression (Malakouti et al., 2024). Gradient boosting has been helpful in risk assessments, algorithmic trading, or fraud detection as the model can deal with nonlinear and complex relationship functions. Gradient boosting is used in credit scoring models, fraud models and risk assessment models. It can be used to detect fraudulent transactions since it can cope with imbalanced data which is one of the predictors of success for most banking systems in today's world. Furthermore, by employing gradient boosting, it also allows for the determination of the trend of the change in the price of a particular stock regarding several characteristics of the financial market and their data (Botchey et al., 2020).

Malakouti et al., (2024) argue that gradient boosting produces more accurate results than statistical methods such as linear and logistic regression models when tested with high-dimensional data. Gradient boosting in healthcare has been useful for the diagnosis of patterns in diseases, disease development, genomic analysis, and potential refinement of the treatment plan. While developing the chronic kidney disease progression, Islam et al., (2023) used Light GBM and got comparatively higher AUC-ROC than other traditional models like random forests. Gradient boosting was applied in cancer research to estimate the interactions of genes while predicting the survival probability of patients (Ma et al., 2020). GBM has been applied in predicting crop yield, precision agriculture, and controlling and regulating pests (Domingues et al., 2022). This model was found to be much more accurate and quicker compared to the statistical-based models for predicting the yield or the productivity of the crops under different climates and weather. They pointed out that proper determination of the yields is very important in the effective management of commodities, feeding the nation, and in the formulation of the national economy. Gradient boosting has been applied in the monitoring of pests and diseases, in resource allocation, and in the prediction of weather using time series data (Domingues et al., 2022).

According to Xu et al., (2024), Gradient boosting has been adopted in exploring nutrient uptake by crops, making it easier for farmers to apply farm inputs efficiently and effectively to prevent wastage and increase productivity. Under the irrigation management system where water conservation forms part of management, GBM has been used to forecast water demands regarding the type of soil, weather conditions,

crops' physiological characteristics, and soil moisture contents, thus minimizing wastage, ultimately resulting in increased crop productivity. Gradient boosting and RSM are general-purpose techniques used in the machine learning and statistical modelling domains. Gradient boosting focuses on building models from weak learners to build a strong predictive model. RSM is a statistician tool used to model and optimize processes and find optimal solutions in the data. Leveraging the strengths of these two models will help in handling complex data (Karatas et al., 2020).

Gradient boosting can thus perform better than other machine learning methods in most cases due to its strategy and versatility (Sibindi et al., 2023). Compared to Random Forests, a sequential learning method in the gradient boosting model produces a higher accuracy for the price of high computational complexity. In contrast, Random Forests are faster and more computationally efficient in tasks that involve parallel construction of multiple trees and aggregation of results by simple voting or averaging. This trade-off makes gradient boosting the algorithm suitable for any task in which precision is paramount (Hoof & Vanschoren, 2021).

Compared with Neural Networks, gradient boosting is better at handling structured data. Neural networks, particularly deep learning models, perform well when taking tasks with unstructured data, including images and speech, as they take so much time, resources, and large sets. Gradient boosting is advantageous as it has an understandable structure and can perform well given the limited sample size often experienced in real-world problems (Manoharan et al., 2022). SVMs are also effective when classifying smaller data sets in which the classes are well defined and when the features' number is restricted. On the other hand, gradient boosting is more efficient when classifying data that is densely packed in higher dimensions, thus overcoming the noise. The contrast shows how the algorithm can solve various problems within the machine learning domain, establishing its position as one of the essential instruments in predictive modelling (Upadhyay et al., 2021).

Gradient boosting models have emerged as the state-of-the-art model for fraud detection in finance. For example, PayPal uses gradient boosting to predict the nature of transactions and detect fraud, and it achieves this with little to no false positives, thus improving customer confidence (William et al., 2024). Also, the fast and accurate credit risk assessment tools driven by LightGBM have helped banks cut down the loan

approval time by sorting out the applicants from high-risk to low-risk applicants, as explained by Chi (2022).

2.13 Integrated Hybrid Approach

Hybrid modelling approaches integrate the statistical framework of conventional experimental designs like CCD with machine learning. These hybrid models have been demonstrated to be more effective than either RSM or machine learning alone in determining the most important predictor variables and estimating non-linear response surfaces in complex factor spaces in post-harvest science (Sibindi et al., 2023). Gradient boosting (GB) is a type of machine learning algorithm that fits an additive model of weak learners (usually decision trees) via stage-wise minimization of a differentiable loss function (Friedman, 2001). The RSM-GB hybrid framework uses CCD to design the experiment and fit second-order surface models, with GB providing agnostic rankings of predictor importance that are free from assumptions of normality of the error terms or the functional form of the response. Predictor rankings obtained from the RSM and GB components are cross validated to reinforce the selection of the top predictors.

Hybrid models that integrate RSM and machine learning (ML) approaches have become increasingly popular in agricultural and food science research to harness the power of both the traditional statistical models and the novel machine learning models for improved model interpretation and inference (Li et al., 2022). The parallel hybrid architecture fits a RSM model and an ML model to the same response as the RSM and aggregates the RSM and ML model predictions using weighted averaging or stacking, in order to harness the predictive diversity of the models to reduce predictive error (Zhang & Jánošík, 2024; Zhang & Haghani, 2015). This approach is easy to implement but does not divide the modelling task and may not be an efficient use of the ML component.

The serial hybrid architecture first fits the parametric model (RSM), then fits the ML model to the residual error of the parametric model (Gad et al., 2025; Hou et al., 2020; Li et al., 2022). The residual correction approach is theoretically motivated by the decomposition of the response variable into a parametric component and a non-parametric residual: the ML component is only directed at that part of the response not explained by the quadratic surface, rather than being directed at the same information

twice. The hybrid forecast is the sum of RSM and the ML residual correction, maintaining transparency in the RSM surface, and supplementing it with residual correction. Such architecture is not new to econometrics modelling but is a novelty in modelling agricultural post-harvest losses (Gad et al., 2025).

Botchey et al., 2020 examined integrated RSM-ML approaches in food process optimization and found that hybrid approaches outperform standalone RSM or standalone ML approaches for a variety of food quality and process responses, with hybrid models showing even greater performance benefits as dataset complexity and response non-linearity increase. The authors noted the residual-correction sequential architecture is particularly suited to situations where RSM interpretability is desired, but the polynomial model is insufficient to capture the response.

The literature reviewed shows that post-harvest losses continue to be a significant problem in agricultural production, food security and livelihood of the farmers, especially in developing countries like Kenya. Several factors have been identified in the past which affect potato post-harvest losses such as temperature, relative humidity, storage duration, transportation mode and handling, market access and varietal factors. There are also previous studies demonstrating that appropriate storage conditions and good post-harvest management can greatly limit deterioration and enhance the shelf life. Moreover, in agricultural modelling and optimization, different statistical and machine learning models including linear regression, Poisson regression, negative binomial regression, beta regression, RSM, and GBM have been applied.

Nevertheless, there are some methodological gaps in literature. The data characteristics of post-harvest loss data make the use of conventional statistical methods questionable. Post-harvest loss data is characterized by bounded proportional outcomes, hierarchical data structure and nonlinear interactions and these make it questionable that can be handled by conventional statistical methods. RSM is structured optimization and interpretability but does not work very well if relationships between variables become very nonlinear. On the other hand, machine learning methods like GBM have predictive power but may be opaque and may not provide a structured optimization of responses. Furthermore, numerous studies have considered each of the post-harvest loss indicators individually and not combined in a composite index that can be used to reflect overall loss. There is little research that

combines the two approaches of machine learning and statistical optimization in modelling losses in potato post-harvest, especially in the Kenyan context. Moreover, post-harvest loss data has been limitedly studied using multilevel beta regression models to take into consideration proportional and hierarchical nature of the data.

2.14 Chapter Summary

Chapter two reviewed literature on potato post-harvest losses, their causes, and the statistical and machine-learning methods applicable to their analysis. This chapter explored the variables that affect losses along the potato value chain such as storage conditions, handling, transportation, market access, and varietal traits. It also examined the application of weight loss, sprouting and rotting as post-harvest degradation indicators and the use of Principal Component Analysis to create composite measures.

The chapter also examined RSM, gradient boosting (GB) and hybrid modelling. RSM was discussed in the context of modelling linear, quadratic and interaction effects and optimizing responses, and Gradient Boosting was discussed about for its ability to model complex nonlinear associations. The review confirmed that both methods have their own beneficial and complementary features but there is limited available research that combines the two methods for modelling potato post-harvest losses, specifically in Kenya. The chapter thus outlined the methodological gap that is the driving force behind the proposed development of the GB-RSM framework and emphasized the consideration of proportional and hierarchical nature of post-harvest loss data.

To fill these gaps, this study developed an integrated hybrid GB-RSM framework that combines the predictive ability of gradient boosting with the interpretability and optimization ability of Response Surface Methodology. Multilevel beta regression is also used to model proportional losses while considering clustering at the county level and value chain stages. A CLI is also developed using Principal Component Analysis, to model multiple dimensions of post-harvest deterioration together. The proposed framework consequently provides methodological and practical guidance in enhancing post-harvest management and mitigating losses of potatoes in Kenya.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Introduction

The research utilized a mixed research design incorporating quantitative and qualitative methodologies. Surveys and interviews were conducted to evaluate the storage methods and difficulties encountered by potato farmers in Nakuru and Nyandarua counties in Kenya. To evaluate the impact of temperature, humidity, and duration of storage on post-harvest losses and potato quality, data was simulated in R to mimic the actual storage conditions. A questionnaire survey was carried out to identify distinct factors associated with post-harvest losses.

This study used both simulated experimental data and real-world survey data, and each data set was used for a different research purpose. The framework of integrated gradient boosting –Response Surface Methodology was developed and tested under controlled nonlinear conditions using simulated Central Composite Design (CCD) experimental data in objectives one and two. For objective three, real-world cross-sectional survey data was collected from potato value chain actors in Nakuru and Nyandarua counties and modeled using multilevel beta regression to estimate the proportional losses during post-harvest. Objective four used the hybrid GBM-RSM approach to examine the CLI derived from several simulated postharvest loss indicators. The use of simulated data enabled the evaluation of the methodological performance of the hybrid framework, and the field survey data gave practical insight into the factors affecting the post-harvest losses in the Kenyan potato value chain.

Simulated experimental data were generated using a CCD and employed to develop and assess the performance of hybrid GB-RSM framework along with field survey data. This simulated experimental design allowed the effects and interactions of storage temperature, relative humidity, storage duration, variety, and mechanical damage on post-harvest deterioration to be evaluated. The simulated data was then specifically used for testing, validating, and comparing the predictive ability of the hybrid GB-RSM model with the standalone RSM and GBM models.

The study therefore comprised two analytical parts that were complementary to each other. The methodological development and evaluation of the hybrid GB-RSM model was the first component, which involved the generation of experimental data through

a central composite design. The second part centered on the use of multilevel beta regression on empirical field survey data of potato value chain actors in Kenya. Both approaches were combined, so that the study could focus on methodological development and application for modeling losses after harvesting potato.

The methodological contribution of this study is the development of an integrated hybrid modelling framework which combines the strengths of the RSM, in terms of optimization and interpretability of the model, and the strength of the GBM in nonlinear prediction. This hybrid model uses gradient boosting to predict the remaining nonlinear structures that the second order-polynomial surface does not explain, while the conventional RSM models are limited to capturing linear, quadratic and low order interactions. Further, the study employs multilevel beta regression to model post-harvest loss data that are inherent proportions with hierarchical clustering structure, and CLI is used to represent post-harvest loss data in a multidimensional way by combining loss indicators into a single response variable. Combining these methodologies offers a more flexible, interpretable and complete analytical approach to model complex agricultural post-harvest systems.

3.2 Sampling Design and Data Collection

The collection of data was done in Nakuru and Nyandarua counties, which are among the most significant potato growing areas in Kenya. Both counties are in the highlands with cool weather, fertile volcanic soil, and consistent rainfall that are very suitable for the growth of potatoes. Nakuru county is a place that is highly productive in agriculture and is also inhabited by high density of wholesale and retail markets along the Nakuru-Nairobi highway. Nyandarua in the Aberdare Ranges is among the highest potato-producing counties in the country and contributes a significant percentage of potatoes being consumed in Nairobi and the other urban areas (Olakiumide, 2021).

The two counties were chosen purposely with respect to the large volumes of potato production, the variety of value chain players that operate in these two counties, and the presence of both the traditional and the improved storage facilities. This gives adequate geographic and operational diversity to produce results that are widely applicable to the potato value chain in the highland production areas of Kenya. Among the 628 respondents sampled 388 (61.8%) were sampled in Nakuru County and 240 (38.2%) in Nyandarua County.

A questionnaire survey was carried out to identify distinct factors associated with post-harvest losses, planting time, quality of seeds, harvesting time, insects' infestation, and climatic conditions pattern of the selected or sampled region. The data collected provides the needed variables for modeling using multilevel beta regression. A stratified sampling technique was applied to select farmers from different regions that produce potatoes. A stratified sampling technique was used since it divides a population into small groups called strata based on geographical regions where potatoes are produced. A sample from each stratum was selected, and the sample size was determined using Cochran's formula for population proportions since the population size is large and the total number of growers cannot be determined (Karimi & Niknami, 2020).

The formula is given by:

$$n = \frac{z^2 p(1 - p)}{e^2} \quad 3.1$$

Where n is the sample size Z is the Z-score which is based on the confidence level p is the estimated population and e the error margin.

The questionnaire was tested for reliability and consistency to assess consistency and accuracy. While reliability focuses on the consistency of the results, validity, on the other hand, focuses on the accuracy of the results. The most common reliability test, internal consistency, is evaluated by Cronbach's Alpha formula (Hamed & Lumpur, 2016). This formula was used to examine the correlation between various items in the questionnaire and is considered significant if it is greater or equal to 0.70. Validity was used to assess the questionnaire's accuracy and ensure that it covers all relevant aspects. Validity was evaluated using metrics such as the content validity index (CVI). R statistical software was used to analyze the data.

3.3 Model Evaluation

The significance test for the second-order model was performed using the analysis of variance (ANOVA) procedure. The adequacy of the model was determined by the probability value (p-value) from the ANOVA together with the coefficient of determination R^2 , mean squared error (MSE), an average between the observed and the predicted values, and the standard deviation of the predicted errors also called the root mean squared error (RMSE). R^2 was used to evaluate how well the model fits the

data. It informs how much of the variance in the outcome variable can be explained by the independent variables. The smaller the value of the RMSE, the better the model. A model was considered desirable and with no overfitting if the RMSE for the test and training are closer to one another. Also, the mean absolute error (MAE) was used to evaluate the model's prediction accuracy. MAE is the average absolute difference between the actual and the predicted values. The p-value was considered significant if it is less or equal to 0.05 significance level.

3.4 Ethical Consideration

This study was conducted with strict adherence to ethical considerations to protect the participants' rights and confidentiality. All the farmers and retailers were involved in the study, other stakeholders were given detailed information about the research objectives, roles and responsibilities, and any risks involved in participation in the study. While conducting the study, all participants' personal information remained anonymous. Besides, the study ensured to take into consideration cultural values and norms in the community. A permit to conduct the research was acquired from National Commission for Science, Technology and Innovation (NACOSTI) that allowed the gathering of data from farmers and retailers in Nakuru and Nyandarua counties.

3.5 Chapter Summary

This chapter presented the materials and methods used to achieve the objectives of the study. A mixed design of research was used, with a mix of simulated experimental data and cross-sectional survey data. The proposed GB-RSM framework was built and tested on simulated CCD data under controlled nonlinear conditions, and survey data were gathered from potato value-chain actors in Nakuru and Nyandarua counties to examine factors impacting potato post-harvest losses.

The sampling processes, data collection, study variables, experimental design and data preprocessing procedures described in the chapter. It also summarized the statistical methods applied such as RSM, gradient boosting, multilevel beta regression, and Principal Component Analysis for building Composite Loss Index. The procedures of model estimation, model validation, model diagnostics and performance evaluation were also introduced. The chapter gave the methodological framework for developing the hybrid GB-RSM model and analyzed the proportional post-harvest losses to assess the causes of overall potato post-harvest losses.

CHAPTER FOUR

DEVELOPMENT AND EVALUATION OF HYBRID RSM-GBM MODEL

4.1 Introduction

This chapter describes the methodological framework adopted to achieve the study objectives and develop the proposed gradient boosting -based response surface methodology (GB-RSM) for modelling complex nonlinear systems. It provides research philosophy, research design, study area, target population, sampling process, sample size determination, data collection methods and variables which were taken into consideration in the study. The experimental design using the four-factor rotatable Central Composite Design (CCD), data processing techniques and statistics to develop the model are also described. Special focus is on the development of the hybrid GB-RSM framework that combines the advantages of optimization and interpretability of the Response Surface Methodology and the nonlinear learning power of GB. Further, the chapter specifies the estimation, validation and performance evaluation procedures of the model with suitable statistical measures. Lastly, ethical issues and the underlying hypotheses of the analytical techniques are discussed to establish the validity, reliability and replicability of the proposed modelling framework.

4.2 Study Design

The research employs a computational modeling design which integrates RSM with GBM. An experimental design of a simulated experiment was taken to determine the performance of the GB-RSM hybrid framework. Four major stages were followed in the modeling process. Preprocessing was done using normalization and K-Nearest Neighbor (KNN) imputation of the missing data and predictors were standardized to mean zero and unit variance. Secondly, second-order regression model was estimated by RSM to include linear, quadratic and interaction effects. Third, the RSM model generated the residuals, which were modeled by the gradient boosting, where regression trees were used to identify nonlinear patterns. The grid search and cross-validation were used to reduce the root-mean-square error (RMSE) by optimizing the learning rate and maximum depth. Lastly, the findings of the RSM and GBM components were incorporated in the GB-RSM hybrid model, with RSM providing a form of prediction that could be interpreted and GBM shrinking predictions by adjusting the residual.

4.3 Response Surface Methodology

The first approach was to apply response surface methodology to model the relationship between the response variable and predictor variables. A second order polynomial approximation for the response variable y was fitted to capture the linear, quadratic and interaction terms as shown below:

$$y = f_{RSM}(X) + \varepsilon \quad 4.1$$

where $f_{RSM}(X)$ is a second order polynomial.

$$f_{RSM}(X) = \beta_0 + \sum_{i=1}^n \beta_i x_i + \sum_{i=1}^n \beta_{ii} x_i^2 + \sum_{i < j}^n \beta_{ij} x_i x_j + \varepsilon \quad 4.2$$

where β_0 is the constant coefficient, β_i the linear coefficient, β_{ii} is the regression coefficient for the quadratic term, and β_{ij} is the coefficient for the interaction term, $x_{i,j}$ are the independent variables and ε is the error term. These coefficients were estimated using maximum likelihood estimation. Then, the residuals were computed as the difference between the observed and the predicted values as shown below:

$$r_i = y_i - f_{RSM}(X) \quad 4.3$$

where r_i is the residual from the i^{th} data point, y_i is the observed value of the response while $f_{RSM}(X)$ is the predicted value from the RSM model.

4.4 Gradient Boosting

Gradient boosting refines the prediction by modeling the residuals from the RSM model. It sequentially builds the models from weak learners to predict the residuals. Systematic Hyperparameter tuning was used to optimize the gradient boosting component. The learning rate, number of trees, interaction depth, minimum observations per terminal node, and subsampling fraction were tested to find the combination that would minimize prediction error without overfitting. The cross-validation performance of the parameters was used to select optimal values, which enabled the final hybrid model to represent a balance between the predictive accuracy and generalizability of the model.

For a GB model, the initial step was to fit the first decision tree $f_1(x)$ to the errors (residuals) from the RSM model with the aim to minimize the errors. The following steps were as follows:

First, the initial decision tree was as follows:

$$F_1(x) = F_0(x) + \lambda f_1(x) \quad 4.4$$

where $F_0(x)$ is the initial prediction from the RSM model, $f_1(x)$ is the initial decision tree fitted to the RSM errors and λ is the learning rate determines the contribution of each decision tree in the model. The residuals at t –iterations are given by:

$$r_i^{(t)} = y_i - F_{t-1}(x_i) \quad 4.4.1$$

where y_i is the observed value of the i^{th} observation, $r_i^{(t)}$ is the residual of the i^{th} –observation at t iteration and $F_{t-1}(x_i)$ is the prediction of the i^{th} observation at $t - 1$ iterations. The model was sequentially updated by fitting weak learners (decision trees) on the errors of the previous models. The resulting model is given below:

$$F_t(x) = F_{t-1}(x) + \lambda f_t(x) \quad 4.4.2$$

where λ is the learning rate, $f_t(x)$ is the new decision tree at t iteration trained on residuals, and $F_{t-1}(x)$ is the prediction from $t - 1$ iteration and $F_t(x)$ is the prediction at iteration t . After T iterations, the final prediction is the sum of the prediction from the RSM model and the subsequent adjustments made by the weak learners.

This study adopted a leaf-based approach over the node-based approach because in a node-based gradient boosting, the correction is applied across the entire decision tree after it has been trained on the residuals to make a general correction to all observations in the dataset. The leaf-based gradient boosting on the hand applies to the correction term at the leaf level of the tree and not the entire tree to update the model for all observations. Correction is made on each individual leaf. Each leaf is assigned a correction term based on the average residual of the training data points in

that leaf. For each learner $f_t(x)$, data is split into K_m leaves and the update for each K leaf is computed based on the residuals for the data points in the leaf such that:

$$C_{t,k} = \underset{x \in R_{t,k}}{\operatorname{argmin}} \sum L(y_i, F_{t-1}(x) + c) \quad 4.4.3$$

$C_{t,k}$ is the correction term. It represents the adjustments made to the prediction for the k^{th} leaf at t iteration. It is determined based on minimizing the loss function that falls within a particular region. $L(y_i, F_{t-1}(x) + c)$ is the loss function which measures the difference between the observed and the predicted values and c is the correction applied to the leaf at t –iteration. $R_{t,k}$ is the region or leaf of the decision tree where k represent each node.

After T iterations the final model is given by:

$$F_T(x) = F_{t-1}(x_i) + \sum_{t=1}^T \sum_{k=1}^{K_m} C_{t,k} I(x \in R_{t,k}) \quad 4.4.4$$

$F_T(x)$ is the final prediction at T iteration, $C_{t,k}$ is the correction term for the k^{th} leaf at T iteration, $R_{T,k}$ is the subset of the data points for the k^{th} leaf at T iteration and $I(x \in R_{t,k})$ is the indicator function that applies adjustments in the regions $R_{t,k}$.

$$I(x \in R_{t,k}) = \begin{cases} 1, & \text{for } x \in R_{t,k} \\ 0, & \text{otherwise} \end{cases}$$

The integrated gradient boosting with the RSM model is given by:

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^k \hat{\beta}_i x_i + \sum_{i=1}^k \hat{\beta}_{ii} x_i^2 + \sum_{i < j}^k \hat{\beta}_{ij} x_i x_j + F_T(x) \quad 4.5$$

$$F_T(x_i) = \hat{\beta}_0 + \sum_{i=1}^k \hat{\beta}_i x_i + \sum_{i=1}^k \hat{\beta}_{ii} x_i^2 + \sum_{i < j}^k \hat{\beta}_{ij} x_i x_j + F_{t-1}(x) \quad 4.6$$

$$+ \sum_{t=1}^T \sum_{k=1}^{K_m} C_{t,k} I(x \in R_{t,k})$$

$$\hat{y} = \hat{\beta}_0 + \sum_{i=1}^k \hat{\beta}_i x_i + \sum_{i=1}^k \hat{\beta}_{ii} x_i^2 + \sum_{i < j}^k \hat{\beta}_{ij} x_i x_j + F_{t-1}(x) \quad 4.7$$

$$+ \sum_{t=1}^T \sum_{k=1}^{K_m} C_{t,k} I(x \in R_{t,k})$$

The decision to adopt an additive model within gradient boosting-based response surface methodology (GB-RSM) framework was informed by several factors that would fit the objectives of the study. Additive models such as the Generalized Additive Models (GAMs) are flexible as they model nonlinear effects with smooth functions of the predictors. This attribute renders them optimal when it comes to capturing the nonlinearity relationships that may exist in agricultural data like the interaction between humidity and temperature and the rate of crop spoilage. Moreover, additive models are computationally effective and simpler to apply than more complex models, which makes them a popular choice when a large amount of data is involved, as is common with agricultural research.

Conversely, convolution models, and particularly Convolutional Neural Networks (CNNs) typically handle tasks that require processing of images or sequential data, where spatial or temporal dependencies are dominant. CNNs are good at finding local patterns in data arranged in grids, including images where pixels are spatially connected or time-series data. But these models are not very suitable to work with tabular data which lacks spatial or sequential patterns. In agricultural studies, the predictors are commonly independent variables with no natural spatial relationships (e.g., temperature and humidity are independent measurements), implying that a neural network, such as CNN, would be inefficient. Also, although convolution models are useful in extracting complex features of high-dimensional data, they are not always easy to interpret and the problem with them is that they require large

datasets to be trained so they are less convenient when the dataset is not very large or when the relationship between variables is not inherently hierarchical or spatial.

4.5 Model Evaluation

The model was evaluated using the mean squared error (MSE), root mean squared error (RMSE) and the R-squared. The MSE measures how well the model fits the data and a lower MSE value indicates a better fit. R-squared also called the coefficient of determination shows the proportion of variance in the response explained by the model. The coefficient of determination value close to 1 or 100% indicates a good fit. To evaluate the statistical properties, robustness and predictive power of the hybrid model, metrics such as biasness, variance and model's assumptions were evaluated. Bias is the difference between the actual value and the average prediction of the model. High bias means that the model will underfit the data (making strong assumptions about the data which not actually be true) while low bias means that model is more flexible and can fit the data more closely. Variance on the other hand refers to how sensitive the model is to fluctuations and variation in the training dataset. A low variance shows that the model is stable and generalizes the data well while high variance leads to overfitting (capturing noise hence reducing the generalizability power of the model). For comparison, low bias and low variance are appropriate as they indicate the model is flexible and generalizes the data well.

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad 4.8.1$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad 4.8.2$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad 4.8.3$$

4.6 Experimental Design and Data Simulation

This study used simulated experimental data to identify the strengths of the integrated GB-RSM model using a rotatable four factor central composite design. Four factors x_1, x_2, x_3, x_4 were studied at 5 levels $(-\alpha, -1, 0, +1, +\alpha)$ where α is the axial points which facilitate rotatability. The axial distance was $\alpha = 2$ and the number of

experimental runs was 30 consisting of center points, axial and factorial points. The outcome variable y was generated on a nonlinear combination of factors and their interactions. The K-Nearest Neighbor (KNN) technique was used to impute the missing values (5% of the data) and all variables were normalized prior to modelling. The RSM was estimated as a second-order polynomial with ordinary least squares (OLS) estimating linear, quadratic, and two-factor interaction terms, whereas the GBM model was trained on five-fold cross-validation, learning rate of 0.05, tree depth of three to four and 1500 trees in R version 4.4.0. The hybrid GB-RSM model was developed in two steps, where the RSM was first fitted to yield residuals which were later fitted to the GBM to describe the local nonlinear trends that the surface of polynomials was unable to capture. The resulting hybrid prediction was determined by adding the RSM and GBM components. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination were used as the evaluation of model performance on an 80% – 20% train-test split. To evaluate the bias, variance and stability, residual diagnostics, Q-Q plots and predicted-versus-observed plots were performed. All visualizations and simulations such as 2D contour and 3D surface plots were performed in R with caret, rsm, gbm, dplyr, plotly and viridis packages.

4.6.1 Regression Model Results

The regression analysis performed with the help of a second-order Response Surface Model (RSM) offers a thorough insight into the nonlinear and interactive relationships between the four predictor variables (x_1, x_2, x_3, x_4). This model has both linear, quadratic, and interaction terms, thus they can depict both curvature and multi-variable complex interactions that cannot be represented in a pure second order (quadratic) model. Higher-order interaction terms, especially two-way combinations increase the power of the model to explain complex dynamics of systems, including behavior between multiple variables which are synergistic.

$$\begin{aligned}
\hat{y} = & 0.4564 + 0.1055x_1 + 1.2321x_2 - 0.0783x_3 + 8.394x_4 \\
& + 2.2207x_1^2 + 2.7594x_2^2 - 0.4470x_3^2 \\
& - 1.6523x_4^2 + 3.900x_1x_2 - 1.217x_1x_3 \\
& + 0.5177x_1x_4 + 4.1523x_2x_3 - 0.0135x_2x_4 \\
& + 5.679x_3x_4
\end{aligned}
\tag{4.9}$$

Table 4.1: RSM Output

Parameter Estimates with Standard Errors, t-values, and p-values				
Term	Estimate	Std. Error	t-value	p-value
(Intercept)	0.4564	0.0825	5.5331	0.0000
x1	0.1055	0.1726	0.6110	0.5414
x2	1.2321	0.1774	6.9447	0.0000
x3	-0.0783	0.1743	-0.4491	0.6535
x4	8.3940	0.1729	48.5547	0.0000
x1:x2	3.9004	0.4470	8.7253	0.0000
x1:x3	-1.2172	0.4449	-2.7361	0.0064
x1:x4	0.5177	0.4363	1.1865	0.2358
x2:x3	4.1523	0.4533	9.1592	0.0000
x2:x4	-0.0135	0.4451	-0.0303	0.9758
x3:x4	5.6795	0.4433	12.8116	0.0000
x1^2	2.2207	0.2963	7.4952	0.0000
x2^2	2.7594	0.3123	8.8349	0.0000
x3^2	-0.4470	0.2978	-1.5010	0.1338
x4^2	-1.6523	0.2966	-5.5698	0.0000
Residual Std. Error = 1.6405 on NA DF R ² = 0.7914, Adjusted R ² = 0.7876 F-statistic = 208.38 on 14 and NA DF (p < <2e-16)				

Table 4.1 above shows that the model yielded a R^2 of 0.7914 and adjusted R^2 of 0.788, which means that about 78.8% of the total variance in the response variable is accounted by the predictors in the model. Adjusted R^2 is used to explain the complexity of the model and ensures that most predictors also have significant explanatory power and do not overfit the data. The large F-statistic 208.3, $p < 2.2e - 16$) also confirms the statistical significance of the overall regression model. These findings indicate that the second-order model gives a good fit and sufficiently represents the underlying structure of the system, but there is still residual variance that cannot be explained.

The model coefficients as interpreted in the second-order RSM reflect an insightful and detailed comprehension of how each predictor affects a system response separately and in combination with other predictors. The analysis was conducted on the basis of RSM model, estimated by the Maximum Likelihood Estimation (MLE).

The method is common in process optimization and experimental design to investigate the relationships between several predictor variables and responses. This research investigates four predictors x_1, x_2, x_3 , and x_4 and the model incorporates both the interaction effects (two-way interactions) and quadratic terms (second-order polynomials) to explain both linear and nonlinear effects. The model focused on the main effects and interactions to reveal the significant factors that affect the response and determine whether there is a relationship between the predictors or not. The aim of the analysis was to determine the importance of these factors and the combined effect of these factors on the response. Interaction and quadratic terms enable a more precise approach to understanding the behavior of the system, especially in its relation to variable influence upon one another, and the variation in response to various conditions.

These findings show that some predictor variables such as x_2 and x_4 have a strong relationship with the response variable while x_1 and x_3 exhibited a weak association. The variables x_2 had a coefficient of 1.2321 and a p-value of 0.00, indicating a strong and positive relationship between x_2 and the response. Also, x_4 had a coefficient of 8.390 and a p-value of 0.00 demonstrating a highly significant and positive relationship with the outcome variable. On the other hand, x_1 and x_3 did not show a significant relationship with the response variable. The x_1 variable with a coefficient of 0.1055 and a p-value of 0.5414 that it does not have a significant effect on the response. Also, x_3 has coefficient of -0.0783 and a p-value of 0.6535 which indicates that the variable is statistically insignificant. These two variables do not have a significant impact on the outcome variable when analyzed individually.

Nevertheless, the terms of interaction provide a more complex situation. Even though x_1 and x_3 do not significantly impact the response, their interaction ($x_1: x_3$) significantly impacts it, with an estimate of -1.2172 and a p-value of 0.0064. This implies there is a negative relationship between these two predictors; that is, the x_1 impact on the response is reduced when x_3 is high. Conversely, the relationship between x_2 and x_3 ($x_2: x_3$) significantly impacts the response, and the coefficient of this relationship is 4.1523, accompanied by a p-value of 0.00, which means that the two variables together increase the response. The interaction effects such as $x_1: x_2$

and $x_3 : x_4$ also reveal how complex the way these predictors interact to determine the result.

The quadratic terms are also extremely important in modeling. All the coefficients of x_1^2, x_2^2 and x_4^2 are significant and the p-value of each of these is 0.00. These quadratic terms note that the association between the predictors and the response is not entirely linear but is curvilinear in nature. Particularly, the coefficients of x_1^2 and x_2^2 are positive, which implies that their influence on the response is greater as the levels of these variables rise, signifying a convex association. Conversely, the negative coefficient of x_4^2 indicates that at increased levels of x_4 , the relationship between the x_4 and the response can change to a concave form.

In the case of two-way interactions, the results of the analysis demonstrate a number of significant and important relationships between pairs of predictors. The interaction between x_1 and x_2 is positive, its coefficient is 3.9004 and p-value is 0.00, which means that there is a synergistic effect of the interaction between these two variables. This implies that an increase of both x_1 and x_2 will yield higher response relative to the individual contributions indicating that the two predictors synergize to improve the system performance. In the same way, the response is positively and significantly affected by the interaction of x_2 and x_3 , with a coefficient of 4.1523 and p-value of 0.00 indicating that both x_2 and x_3 jointly have an additional impact. This result indicates that interaction terms are also significant to consider in modeling because they can demonstrate combined effects that would be omitted in an analysis of individual predictors at any given moment.

Conversely, the interaction of x_1 and x_3 exhibits a negative impact with a coefficient of -1.2172 and p-value of 0.0064. It implies that x_1 and x_3 have an inverse relationship, the larger they are, the less the effect of the two. It means that there is a negative correlation between x_1 and x_3 , as the positive impact of the increase in one variable is weakened by the positive impact of the increase in the other variable. The relationships between x_1 and x_4 and between x_2 and x_4 were, however, revealed as statistically insignificant (*p values* > 0.05), and so the two groups of predictors do not exert a significant impact on the response when they interact. This implies that x_1 ,

x_2 , and x_4 alone are significant; however, their combinations do not result in any meaningful synergies.

In a nutshell, the RSM (MLE) model has been able to portray the linear and nonlinear relations between the variables under study. The strong quadratic and interaction effects indicate that the response surface is nonlinear and highly interdependent; this implies that the effect of one factor can change the effect of other factors on the response, which forms a complex, but predictable response pattern throughout the design space. These results, with good goodness-of-fit measures and a very significant overall model, indicate a dependable and interpretable model of the system behavior, and thus the RSM (MLE) method can be used not only to make predictions but also to optimize the system.

Table 4.1: MLE Model Performance

RSM (MLE) Model Performance Metrics	
Overall model fit statistics	
Metric	Value
Log-Likelihood	-1,889.3080
AIC	3,810.6170
BIC	3,888.8170
R ²	0.7798
Adjusted R ²	0.7764
F-test p-value	0.0000

Table 4.1 provides some of the main performance measurements of a model using Maximum Likelihood Estimation (MLE). The log-likelihood of -1889.3080 gives some idea of how well the model fits the data, with larger values desired to give a better fit. Both the AIC of 3810.6170 and the BIC of 3888.817 are model comparison measurements, with a lower value indicating a superior model, but they are to be compared to other models. The p-value F is 0.00, which validates the overall statistical significance of the model, i.e., the regression model is a good predictor. All these metrics indicate good performance of the model, and the balance between the explanatory power and model complexity is good. The AIC and BIC criteria are not

suitable for gradient boosting as they are designed for parametric models where the number of parameters is clearly defined. However, K-fold cross validation is preferred for ensemble models.

4.6.2 Model Performance Summary

A rotatable Central Composite Design (CCD) was employed in the study to achieve the equal prediction variance of all directions of the experimental region, thereby facilitating effective estimation of the linear, quadratic, and interaction effects. This rotatable central composite structure of design contributed to the stability and accuracy of the estimated response surface to give an optimal source of modeling nonlinear responses.

The K-Nearest Neighbors (KNN) method was used in data preprocessing stage to impute several missing values. This method was favored because it maintains multivariate relationships and reduces the bias over mean or median imputation, which enhances the quality of data and the overall performance of the model. Three different models were then fit using the imputed data: a second-order Response Surface Model (RSM), a gradient boosting machine (GBM) model and a hybrid GB-RSM model combining both methods.

Table 4.2: Model Summary

Model Performance Table			
Model	RMSE	MAE	R^2
RSM	0.475	0.277	0.821
GBM	0.176	0.118	0.975
Hybrid GB-RSM	0.090	0.049	0.993

Table 4.2 shows that RSM model yielded an RMSE of 0.475, MAE of 0.277, and an R^2 of 0.821, which is a good fit to the data. RSM was found to be very effective in capturing main effects factors and interaction of the factors, with a clear interpretation ability in regression coefficients and surface plots. To evaluate the adequacy and diagnostic validity of the second-order RSM, residual analysis was performed; after

fitting the quadratic model, the residuals were computed as differences between the observed and predicted response values as a measure of the unexplained variation, and several diagnostic plots, such as those of the residuals versus fitted values, normal Q-Q plot, and residuals versus predictor plot, were plotted to assess the assumptions and performance of the model. The analysis of the residual values revealed slight curvature patterns which meant that the quadratic model could not.

The GBM model was much better, with $RMSE = 0.176$, $MAE = 0.118$, and $R^2 = 0.975$. Nonlinear dependencies and higher-order interactions among the factors were well captured by the ensemble learning mechanism of GBM which minimized residual errors sequentially using boosted decision trees. In addition, GBM was found to be very robust to the imputed data, adapting to missingness patterns and noise in the predictors. Although the model had better predictive power, it was limited in transparency over the effects of factors compared to RSM due to its non-parametric, black-box nature.

The hybrid GB-RSM model that combines the parametric format of RSM with the nonlinear learning ability of gradient boosting machines (GBM) showed the highest overall predictive performance ($RMSE = 0.089$, $MAE = 0.049$, $R^2 = 0.993$). In this context, the response surface in the RSM component was able to capture the global and continuous curvature of the response surface due to its use of quadratic terms and was also able to rotate to estimate main effects as well as second-order interactions. But since RSM is limited by its fixed quadratic functional form, the residual analysis indicated the existence of other nonlinear patterns on the data. To overcome such shortcomings, GBM was trained on the RSM residuals and enables it to model local deviations and higher order non-linear dependencies. This stepwise integration played a major role in improving predictive power, interpretative power, and strength. The hybrid model preserved the explanatory power of RSM giving it explicit regression coefficients and response surfaces that can be visualized and the flexibility of GBM to deal with nonlinear and high-dimensional data. This means that the GB-RSM hybrid framework will provide a sensible and interpretable modeling approach, which is ideal in optimization studies of agricultural, environmental, and engineering systems that are described by nonlinear, complex, and multivariate interactions.

4.6.3 Model Interpretation

The better predictive capability of the hybrid GB-RSM model could be explained by the synergistic combination of RSM and GBM models. RSM model provides a parametric base by estimating a quadratic regression function to capture the linear, quadratic and two-way interaction effects of the predictor variables. This parametric methodology is interpretable and can be used to visualize the world response surface. But Conversely, GBM model adopts a non-parametric, and ensemble-based learning algorithm which minimizes residual errors by sequential tree boosting. It enables it to model nonlinear relationships and higher-order interactions that could not have been modeled by traditional regression approaches such as RSM.

The hybrid GB-RSM model combines the strengths of the two approaches. The RSM model estimates the structure of the response and the residuals, or the unexplained variance is then modeled using GBM. GBM removes the systematic error of RSM by minimizing the residual component and characterizes nonlinear trends that cannot be explained by the linear terms. This hierarchical modeling approach is successful in minimizing bias and variance and leads to the best predictive power and more stable models. This means that the hybrid framework preserves interpretability of RSM but adds to it the flexibility and adaptive learning ability of GBM. This integration, empirically, results into a statistically sound model which minimizes the prediction error and enhances the generalization of the model under different data conditions.

4.6.4 Residual and Diagnostic Analysis

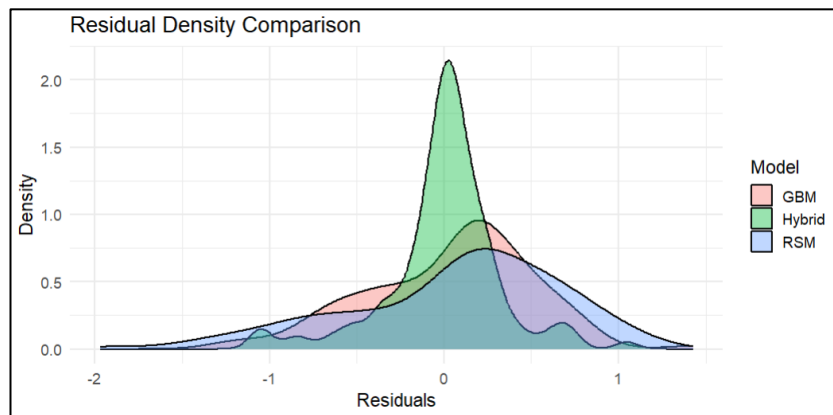


Figure 4.1: Residual Density Analysis

The comparison plot of the residual density in Figure 4.1 shows the nature of the distribution of the residuals of the three models RSM, GBM and hybrid GB-RSM framework. The residuals of the hybrid model have a narrow and sharply peaked distribution around zero indicating that most of the prediction errors are small and evenly distributed. This trend is typical of a high predictive precision model with low bias. Conversely, a broader distribution of the residuals is observed in the RSM model which shows greater variability and systematic errors of the actual response surface. Residual values of the GBM model are not too far in either direction compared to RSM, although they are more concentrated than the hybrid model. These findings suggest that, although GBM is superior to the pure RSM in capturing nonlinear effects, the hybrid offers a better balance because the residual dispersion is minimized and the predictions are more likely to cluster around the actual mean.

The RSM model has additional limitations explained in the diagnostic plots. The scatterplot of the residuals versus fitted indicates no strong curvature, which is desirable; however, the distribution of the residuals increases slightly with the fitted values, thus heteroscedasticity. The fact that the residuals cluster in some fitted parts also points to the possibility of model misspecification, possibly because the quadratic RSM form cannot capture complex interactions or other nonlinearities in the data. The Q-Q plot indicates that although the data are distributed around the normal line, the tails are far to the right and to the left, which indicates the existence of heavy-tailed behavior and outliers. Collectively, these findings suggest that, despite its reasonable first order approximation, RSM breaks the normality and homoscedasticity assumptions of regression and hence may not be as reliable in predicting more complex phenomenon.

The predicted-versus-observed plots also give more insight into the relative performance of the three models. In the case of the RSM model, the points predicted trace the line of perfect fit with moderate accuracy but apparently scattering around the line suggesting weak predictive capabilities. On the contrary, the GBM model is less aligned, and the dispersion and bias appears to be significant, especially at the edges of the observed range. This trend indicates that though GBM is adaptable in terms of nonlinear trends, it can have a problem of overfitting or underfitting in certain areas because it uses iterative boosting trees in its analysis without a clearly defined

parameter structure. Conversely, the hybrid model shows a great, almost perfect linear relationship between observed and predicted values. The close concentration of the points in the 45° line shows that the hybrid method effectively combines the interpretability and smoothness of RSM with the nonlinear adaptive ability of GBM. It therefore reduces bias and variance between the prediction range.

In general, the comparative diagnostics prove that the hybrid GB-RSM model is the most correct and consistent in terms of performance compared to the other two methods. Its residual values are normally distributed with the least variance, and its forecasts relate well with the observed results. These findings indicate that the hybridization approach is useful in resolving the shortcomings of the traditional RSM strategy and the pure machine-learning strategy, offering a more well-founded and predictive general model on response surface estimation and predictive analysis.

4.6.5 Residual Diagnostics of the RSM Model

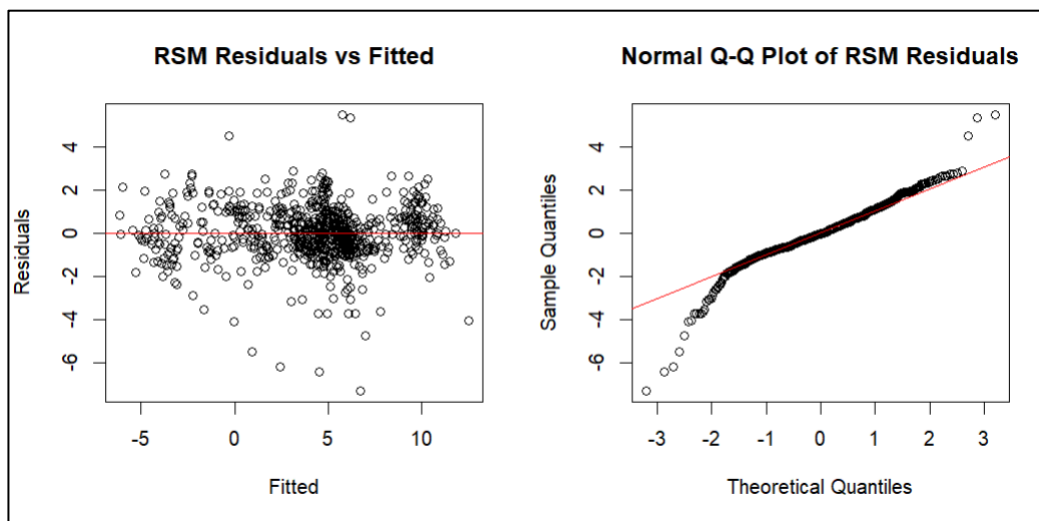


Figure 4.2: RSM Residuals

The diagnostic plots of the RSM model in Figure 4.2 offer understanding of the suitability of the fitted quadratic surface, as well as whether the employed regression is being met. To determine whether the residuals have systematic patterns which may indicate model misspecification, nonlinearity or the residuals versus fitted values plot (left panel) is utilized. The residuals, in this case, seem to be randomly distributed around the horizontal zero line, and it is a sign that the RSM model is not strongly biased in its structure. The fitted values, however, increase the variability of residuals

to a slight degree, which is exhibited by a mild funnel-shaped dispersion. This non-homogeneous variance indicates that there is a heteroscedasticity, which means that the model is more likely to generate high prediction error when the fitted response is large. The plot also shows some individual clusters and outliers, which can be because of some of the experimental runs with extreme factor combinations or interactions that are not well represented by the quadratic model.

The right panel of the normal Q-Q plot also tests the normality assumption of the residuals, which is a major requirement that valid statistical inference in RSM requires. Ideally, when there is a normal distribution in the residuals the points would be close to the 45-degree reference line. The middle part of the data in this plot compares well with the theoretical curve implying that most of the residuals are normally distributed. There are, however, observable deviations at both tails, with the residuals moving downwards (on the left) and upwards (on the right). Such departures imply heavy tails and the existence of a handful of extreme residuals which do not follow a normal behavior. The common occurrence of such deviations in the RSM models is when the true response surface possesses a greater nonlinearity or interaction than can be represented by a second order polynomial.

All these diagnostics suggest that the RSM model can be used to explain a significant portion of the data, but not with all assumptions of homoscedasticity and normality. The fact that the heteroscedasticity is not insignificant and there are tail deviations suggests that the quadratic structure is too severe in relation to the data used in the study. As a result, although RSM represents the overall trend of the response, it does not represent some of the intricate trends especially in the areas where the response surface changes more sharply. This weakness justifies the incentive toward the more flexible GBM or the Hybrid GB-RSM model to better adapt to nonlinearities and lead to a smaller residual variance.

4.6.6 3D Surface with Contour Plots

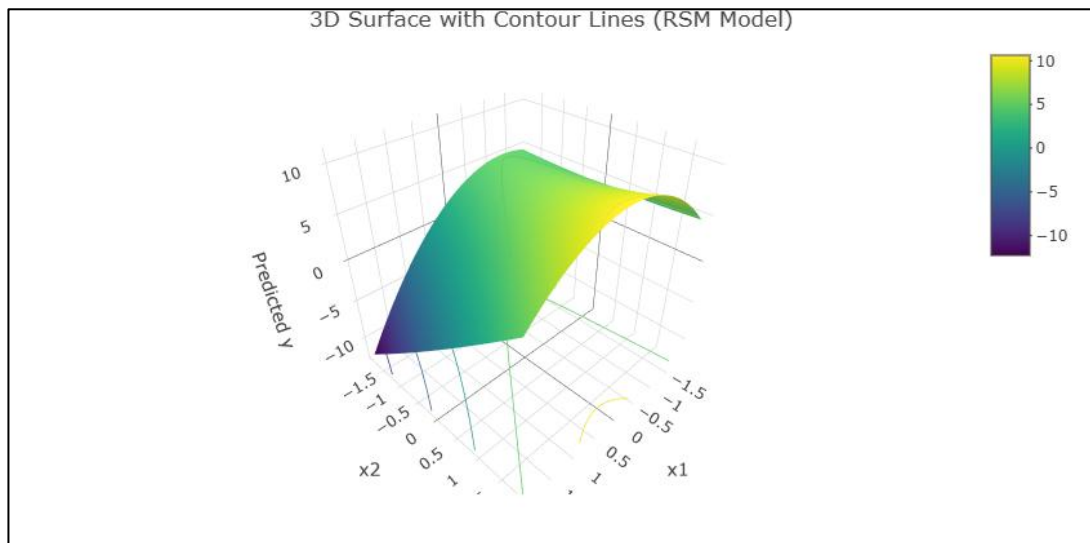


Figure 4.3: Response Surface

Figure 4.3 is a 3D surface plot with contour lines that show the response surface produced by the fitted RSM model of factors x_1 and x_2 . The plot reflects a smooth, curved surface, which depicts that the model has successfully captured both straight and quadratic lines of relationships between the independent variables and the predicted response Y . The gradient of colors between purple (low response values) and yellow (high response values) shows areas of the lowest and highest predicted responses, demonstrating the joint effect of changes in the factors on the outcome. The curvature and the form of the contours indicate that the interaction between x_1 and x_2 is nonlinear. The darker lines of contour in some areas suggest a steep response surface, where minimal changes in the level of factors cause significant changes in the response. Generally, the surface suggests that the area with the highest response increase or decrease is around one side of the design space, affirming that the RSM model can effectively elicit effects of critical factors and possibly the optimum operating conditions.

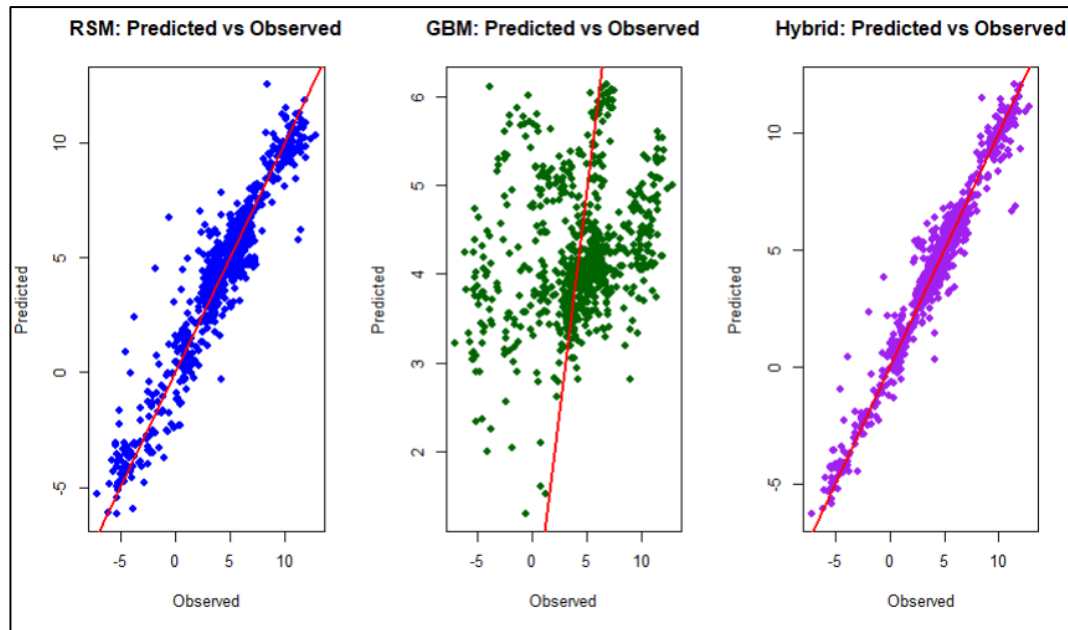


Figure 4.4: Predicted vs Observed Values for RSM, GBM and Hybrid GB-RSM Models

Figure 4.4 is a comparative analysis of three models; RSM, GBM, and a Hybrid model, using the observed vs. predicted scatter graphs. The primary goal of this kind of plot is to determine the fit between the model predictions and the real-life data. In a perfect model, all scatter points would fall on the line, and all predictions would be 100% accurate and reliable but this is not the case, and this means that the red diagonal line in each scatter plot indicates that the predictions of the model are not 100% accurate and reliable. When a point is not on the line, it is an indication of an error in prediction by the model.

The points in the RSM plot (left) are clustering well and closely around the red line, which indicates a strong linear relationship between predicted and observed values. This shows that the underlying trend of the data is well represented by the RSM model with comparatively low errors of prediction. The RSM plot shows that most of the points are close to the red diagonal line which means that the model has good predictive ability at the major (mid) range of the data. When we go into the extremes being observed by the values, one can see that the points begin to be dispersed away from the red line. The GBM plot (middle) however demonstrates a scattered distribution of the points around the red line, which means less predictive consistency. The GBM model appears to fit the overall trend but with a higher amount of noise and

variance meaning that it may be overfitting or that it is not proportional in addressing different data ranges. This may be because of the data-driven nature of GBM that allows minimizing error by sequential boosting but does not provide interpretability or stability without appropriate tuning.

The Hybrid one (right) combines the advantages of both RSM and GBM, and it is also evident in its scatter pattern. The data points are well clustered on the diagonal which implies high predictive accuracy and strength. The hybrid model is probably a combination of the parametric structure of RSM (with interpretability and smoothness) and the nonlinear learning capability of GBM (with the ability to capture complex patterns in the data). This makes it reduce the error of prediction and more generalizes across the data space. In general, this value proves that although RSM provides good baseline performance and GBM brings flexibility, Hybrid model beats both as it provides the most perfect balance between accuracy, stability, and generalization ability. It makes the most of statistical and machine learning schemes, making statistically and machine-learned predictions as reliable as possible.

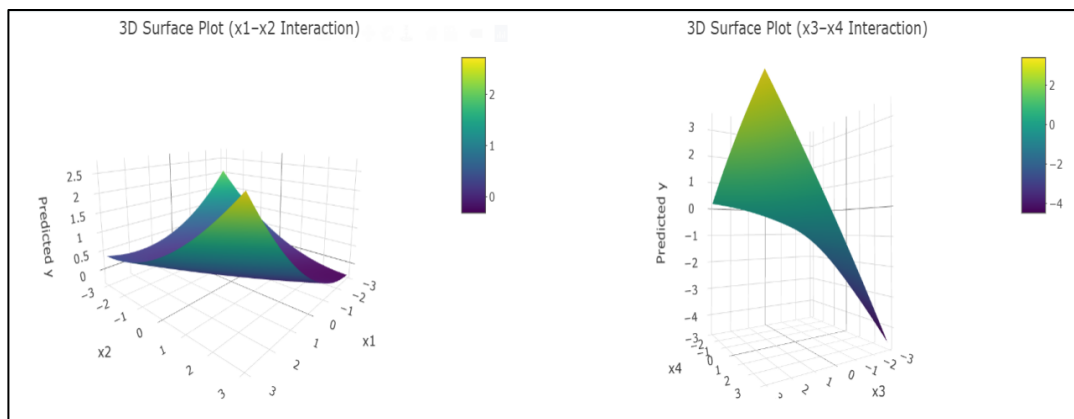


Figure 4.5: RSM Contour Plots

The two 3D response surfaces in Figure 4.5 above show different patterns of interaction in the fitted response surface model. The left surface (the relationship between x_1 and x_2) shows a strong curvature which is a typical characteristic of a strong second-order effect. This curvature shows that the effect of these factors is not additive but synergistic, i.e., the overall effect of x_1 and x_2 on the response variable is

even larger than the sum of the effects of the two factors. The surface increases evenly on the extremes of the two factors, indicating a nonlinear dependence on which the predicted response is growing with the deviation of the two variables out of the central levels. This is statistically significant quadratic and interaction terms which lead to model curvature and a stationary region due to an optimal or critical point in the experimental domain.

Conversely, the surface by x_3 and x_4 , has a regular gradient structure, which indicates that the behavior of the response has first-order behavioral influence with a minimal nonlinear deviation. The slope on the surface shows opposite direction effects in that an increase in x_3 leads to increased predicted response levels and a corresponding increase in x_4 translates to weakening of the response, indicating opposite main effects and a negative interaction in the system being modeled. Statistically, the association between x_3 and x_4 is almost null and that linear coefficients are the leading factor in explaining the response pattern. These contrasting surfaces explain the statistical heterogeneity of factor interactions, meaning that the response function is statistically heterogeneous with sensitivities that vary spatially because of interaction between nonlinear curvature and directional linearity. The $x_1:x_2$ pair interaction shows a strong non-linear variability whereas the $x_3:x_4$ interaction shows a comparatively nonlinear effect exhibiting a dependent structure hence suggesting that response surface is hierarchically driven by the $x_1:x_2$ pair. The fact that x_1 and x_2 define a region with localized curvature and a global linearity implies that the model should be optimized on the region where x_1 and x_2 are most sensitive to changes and predictive uncertainty is high.

The models were robustly tested across diverse conditions, such as the addition of noise, missing data, and outliers to the dataset. GB-RSM model was found to be stronger than RSM and could withstand data imperfections. The former is attributed to the fact that gradient boosting can deal with nonlinear relationships as well as its resistance to overfitting. In the case of addition of outliers to the data, the GB-RSM model did not exhibit much deterioration in performance, as compared to RSM and GBM models that experienced more serious losses in performance. This observation stresses the strength of GB-RSM, especially in situations when the data in the real-world is noisy or contains missing values. Imputation and regularization techniques of

gradient boosting enable the model to achieve high accuracy in making predictions despite missing values in some data points. The RSM model on the other hand is more sensitive to missing or incomplete data, as it makes a heavy assumption of linearity between predictors and response, which is more difficult to estimate when some data are missing.

The findings validate that the GB-RSM hybrid model has greater predictive power compared to both the traditional RSM model and other machine learning methods. The principal benefit of the GB-RSM model is that it combines the strengths of RSM and gradient boosting. The RSM model forms an initial model of response variable by capturing linear, interaction, and quadratic effects of the predictors, with the gradient boosting component refining the initial model by estimating nonlinearities and complex interactions among the variables.

4.7 Chapter Summary

This chapter developed and evaluated the proposed Gradient Boosting-Based Response Surface Methodology (GB-RSM) framework using simulated experimental data. Nonlinear experimental data were generated using a rotatable Central Composite Design and missing data were filled with K-Nearest Neighbor (KNN) imputation with normalization of the predictors. A second order Response Surface Model was initially fitted to describe the linear, quadratic and interaction effects of the predictors. The residuals of the RSM model were then modelled with Gradient Boosting to represent non-linear structures and local patterns that were not well captured by the quadratic response surface. The Gradient Boosting component was improved in terms of predictive performance and generalizability by using hyperparameter tuning and cross-validation. The predictive performance measures and the diagnostic procedures were used to evaluate the performance of standalone RSM, standalone GBM and hybrid GB-RSM model. The results showed that the RSM gave useful interpretability of factor effects and response surfaces and GBM gave good nonlinear predictive ability. The hybrid model has integrated these complementary strengths and has yielded the highest prediction accuracy of the three models. The residual analysis also showed that the Gradient Boosting component could capture the nonlinear patterns after RSM fit. Consequently, the chapter gave empirical evidence of the proposed residual-correction architecture where the global polynomial structure is represented

by the RSM, and the remaining nonlinear structure is represented by the Gradient Boosting models.

CHAPTER FIVE

MULTILEVEL BETA REGRESSION ANALYSIS

5.1 Introduction

The chapter introduces the use of multilevel beta regression modelling for the analysis of determinant and variability of potato post-harvest losses in the farm-to-market value chain in Kenya. This chapter presents statistical modelling and analysis of the survey data gathered in the potato value chain in Nakuru and Nyandarua Counties, including the general research methodology and data collection procedures undertaken. The chapter outlines the modelling, estimation, diagnostic testing and interpretation of the results of proportional post-harvest losses for various value chain stages. The chapter also compares various methods of count and proportional modelling such as Poisson regression, negative binomial regression and multilevel beta regression to identify the best suited model for modelling potato post-harvest losses. Fixed effects, random effects, model diagnostics and county-level variability have been discussed in detail to gain the full picture of factors affecting post-harvest losses for potato value chain actors.

5.2 Study Design

The study follows analytical design that focuses on actors at the various stages of the potato value chain in Kenya. Cross-sectional designs suit the gathering of information about a specific population at one time and are commonly applied in agricultural economics and food systems research to investigate relationships among exposures and outcomes in many units of observation. The population of interest is divided into potato farmers, transporters and retail traders. All these categories depict different stages of the field to consumer movement of potatoes, and each category is linked to a particular kind of handling practice, storage conditions and the nature of market interface which determines the degree of the post-harvest losses. It relies on primary data, which was collected through a structured questionnaire, which was distributed to 628 value chain actors in the two counties (Nakuru and Nyandarua). Post harvest loss experiences, storage, transportation experiences, market access, receipt of extension services, and socioeconomic characteristics were also captured in the questionnaire. Post harvest loss was modeled as percentages of overall produce processed and produced continuous bounded results in (0, 1) at counties and at value chain stages.

5.3 Definition of variables

The dataset is hierarchically structured in three levels due to the nesting of observations. At the first level (Level 1), individual observations are associated with farmers or traders, respectively, each of whom reports on the volume of potatoes processed and the volume of potatoes lost at a particular point of the value chain. On the second level (Level 2), the grouping of observations is based on the stage of value chain, indicating that the losses incurred at the same stage have common operating determinants, including storage infrastructure, packaging materials, and handling procedures. The third level (Level 3) includes such groupings as county whereby observation includes similarities in terms of infrastructure, agroecological situations and market situations based on which farmers and traders are located.

Let Y_{ijk} represent the number of lost units of potatoes by individual i ($i = 1, 2, \dots, n$) in county j , ($j = 1, 2$) at the stage of the value chain number k ($k = 1, 2, \dots, k$). The overall number of respondents was 628 and are distributed between Nakuru ($j = 1$; $n = 388$) and Nyandarua ($j = 2$; $n = 240$). The observations within each county are further subdivided into each value chain stage which is where a respondent is majorly operating. This multilevel nesting contravenes the assumption of independence of classical regression models and drives the multilevel approach to modelling.

The dependent variable Y_{ijk} is defined as the number of potato units (in kilograms) lost by the respondent i in the county j at the stage of the value chain k as a percentage of the total amount produced, given as a loss rate through physical damage, disease, sprouting, rotting, or loss of weight in the reference period. This proportion-based operationalization complies with beta regression specifications and was also consistent with post-harvest loss literature procedures. Rotting and decay (181 respondents; 28.8%), sprouting (129 respondents; 20.5%), and weight loss (68 respondents; 10.8%), were the most reported types of losses in the survey. The rotting and decay rates are also high due to the high usage of traditional storage methods among the respondents that lack proper temperature and humidity control. The explanatory variables are grouped into three categories namely, socioeconomic, operational, and market access variables which are hypothesized to affect the level of post-harvest losses by different causal mechanisms as shown in Table 5.1 below.

Table 5.1: Types of Losses

Type of Loss	Frequency	Percentage (%)
Rotting and decay	181	28.8
Sprouting	129	20.5
Weight loss	68	10.8
Greening	2	0.3
Other	1	0.2

The socioeconomic variables depict human capital and institutional attributes of the respondents. The measure of access to agricultural extension training is considered a binary measure as shown in Table . Among the surveyed respondents, 321 (51.1%) of the respondents indicated that they had been trained by agricultural extension officers, and 307 (48.9%) had never been trained by agricultural extension officers. The use of chemicals (339) and proper storage techniques (331) and pest and disease management (299) were the most covered topics among those trained.

Operational variables represent the practices and technologies employed by respondents on how they manage, store and transport potatoes. The most crucial variable of operation is storage method, which could be traditional storage (pits and open spaces), dark room storage, and cold storage. The overwhelming majority of the respondents 355 (56.5%) used traditional storage, 213 (33.9%) dark room storage and only 58 (9.2%) had access to cold storage. Conventional storage is known to have ineffective control of temperature and humidity, which hastens spoilage, sprouting and disease formation. Storage time is also recorded; 223 (35.5%) stored less than two weeks, 203 (32.3%) less than three weeks, 118 (18.8%) less than four weeks, and 78 (12.4%) more than one month. The use of chemicals- the application of sprout suppressants and other post-harvest chemicals is operationalized as a binary response. Only 193 out of 628 respondents (30.7%) said that they used chemicals to manage the post-harvest, and 435 (69.3%) said they did not. The fact that the adoption of chemical inputs remains low despite its effectiveness in prolonging storage and minimizing losses due to sprouting is indicative of the knowledge gap and cost among the smallholder farmers and traders (Mugo et al., 2021).

Table 5.2: Storage Methods of Individual Respondents

Variable/ category	Frequency	Percentage (%)
Storage Method		
Traditional (pits/open spaces)	355	56.5
Dark room	213	33.9
Cold storage	58	9.2
Other	2	0.3
Storage Duration		
Less than 2 weeks	223	35.5
2–3 weeks	203	32.3
3–4 weeks	118	18.8
Over a month	78	12.4
Chemical Use		
Yes	193	30.7
No	435	69.3

Market access variables represent the spatial and logistical aspects of the value chain that influence the rate and circumstances of market access of potatoes. The distance to the closest market is a key factor that defines transportation-related losses; the further the distance, the more exposed it is to heat, physical damage, and delays that speed up the process of spoilage. Among the 628 respondents, 316 (50.3%) said they were within the market ranges 6-10 km, 166 (26.4%) said outside that range and 146 (23.2%) within the 5 km market distance. The transportation method of moving potatoes is also captured with 359 (57.2%) respondents confirming the use of motorcycles and 250 (39.8%) using trucks. Motorcycle transport, though quicker, is linked to increased incidences of mechanical damage owing to lack of cargo protection and poor road conditions.

Transportation issues were measured on a three-point scale on their effect on the quality of potatoes. A large proportion of the respondents, 497 (79.1%) respondents

reported that transportation issues had a big impact on the quality of potatoes, with 128 (20.4%) respondents reporting that it had a small impact and 3 (0.5%) reporting no impact whatsoever. The major transportation issues cited included high transportation costs (498 respondents; 79.3%), time delays in getting to the market (414 respondents; 65.9%), and mechanical damage on potatoes (269 respondents; 42.8%) as shown in Table 5. below.

Table 5.3: Market Access for respondents

Variable/ category	Frequency	Percentage (%)
Distance to Nearest Market		
Less than 5 km	146	23.2
6–10 km	316	50.3
Over 10 km	166	26.4
Mode of Transportation		
Motorcycle	359	57.2
Truck	250	39.8
Other	19	3.0
Transportation Impact on Quality		
Significantly	497	79.1
Slightly	128	20.4
Not at all	3	0.5
Transportation Challenges (Multi-response)		
High transport costs	498	79.3
Delays in reaching market	414	65.9
Mechanical damage	269	42.8

5.4 Statistical modeling framework

Classical linear regression is essentially not appropriate because the dependent variable Y_{ijk} is a proportional loss rate which is the fraction of total produce lost at each stage and the proportional loss rate is a strictly positive and open interval (0, 1). Linear regression is not applicable due to the bounded nature of the response variable that would violate the assumption of normality and homoscedasticity (Ferrari and Cribari-Neto, 2004).

5.4.1 Multilevel Negative Binomial Model

Post harvest data are often expressed as count data. Poisson regression is a standard method for such data but has a limitation of assuming that the mean and variance are equal hence there is need to explore other methods that take care of overdispersion. Negative binomial model relaxes the equi-dispersion assumption by introducing a dispersion parameter θ . This allows for greater flexibility in modeling over dispersed counts. Multilevel models also provide a good alternative to loss data since it is hierarchical. They include individual nested within counties and the potato value chain stages. This model allows for quantification of variability at each level, adjustment of standard errors for clustering and random intercepts for counties (μ_k) and stages (v_k). Combining negative binomial regression with multilevel structure accounts for hierarchical independence and overdispersion we have: $Y_{ijk} \sim NB(\mu_{ijk}, \theta)$ where $\mu_j \sim N(0, \sigma_u)$ is the random effects for county $j = 1, 2$ and $v_k \sim N(0, \sigma_v)$ is the random effects for stage k . The model allows researchers to quantify variability across counties and the potato value chain while accounting for overdispersion in the data.

Let Y_{ijk} represent the number of lost units of potatoes by individual i ($i = 1, 2, \dots, n$) in county j , ($j = 1, 2$) at the stage of the value chain number k ($k = 1, 2, \dots, k$). Then, $Y_{ijk} \sim Poisson(\mu_{ijk})$ hence the probability mass function is given by:

$$p(Y_{ijk} = y) = \frac{e^{-\mu_{ijk}} (\mu_{ijk})^y}{y!} \quad \text{for } y = 0, 1, 2, \dots$$

The Poisson distribution assumes that the mean and variance are the same $E(Y_{ijk}) = Var(Y_{ijk})$. The assumption implies equidispersion where the variance is equal to the mean. However, agricultural data exhibit overdispersion where the variance is larger than the mean where $Var(Y_{ijk}) > E(Y_{ijk})$. When overdispersion is present, Poisson regression underestimates standard errors leading to biased statistical inference. As such, negative binomial distribution is used as it introduces an additional dispersion parameter. This distribution is derived from a Poisson-Gamma mixture so that $Y_{ijk} | \lambda_{ijk} \sim Poisson(\lambda_{ijk})$ where $\lambda_{ijk} \sim Gamma(\alpha, \beta)$. Integrating the Gamma distributed random variable leads to negative binomial distribution. This implies that

$$Y_{ijk} \sim NB(\mu_{ijk}, \theta), \text{ where } E(Y_{ijk}) = \mu_{ijk} \text{ and } var(Y) = \mu_{ijk} + \frac{\mu_{ijk}^2}{\theta}.$$

θ controls degree of dispersion and as θ approach ∞ , the model approaches the Poisson distribution. The conditional mean is linked to independent variables using the log link function

$$\mu_{ijk} = E(Y_{ijk}|X)$$

$$\log(\mu_{ijk}) = \eta_{ijk} \text{ where } \eta_{ijk} = \beta_0 + \sum_{r=1}^p \beta_r X_{ijk}$$

To account for hierarchical clustering random effects are introduced so that the multilevel model becomes

$$\log(\mu_{ijk}) = \beta_0 + \sum_{r=1}^p \beta_r X_{ijk} + \mu_j + v_k \quad 5.2$$

where β_0 is the intercept, β_r are regression coefficients, μ_j are random effects for counties and v_k are value chain stage random effects. These random effects capture unobserved heterogeneity across counties at various stages.

5.4.2 Likelihood Function

To estimate the unknown parameters such as the dispersion parameter, random effects and regression coefficients, likelihood function was used.

Let $f(Y_{ijk} | \mu_{ijk}, \theta)$ be the negative binomial probability function. The conditional likelihood is $L_c = \prod_{j=1}^2 \prod_{k=1}^k \prod_{i=1}^n f(Y_{ijk} | \mu_{ijk}, \theta)$. The marginal likelihood integrates over random effects so that:

$$L(\theta) = \iint L_c \phi(\mu_j) \phi(v_k) d\mu_j dv_k = \theta = (\beta, \sigma_\mu^2, \sigma_v^2, \theta) \quad 5.3$$

Therefore,

$$L = \prod_{k=1}^K \int \prod_{j=1}^2 \prod_{i=1}^n f(Y_{ijk} | \mu_k, v_k) g(\mu_k) g(v_{jk}) dv_{jk} d\mu_k \quad 5.4$$

where $f(Y_{ijk} | \mu_k, v_k)$ is conditional negative binomial density and $g(\mu_k), g(v_{jk})$ represent normal densities of a random effect where $\mu_k \sim N(0, \sigma_\mu^2)$ and

$v_{jk} \sim N(0, \sigma_{v_{jk}}^2)$. The log likelihood is $l(\beta, \sigma_\mu^2, \sigma_v^2, \theta) = \text{Log}l$. Since the likelihood involves integrals over the random effects, numerical approximation techniques are therefore employed and hence this study employs Laplace transformation.

5.4.3 Laplace Transformation.

This approximates likelihood integral by expanding the log-likelihood around the maximum using a second-order Taylor expansion. For integrals of the form $L = \int e^{h(u)} du$, the Laplace approximation gives:

$$L = \int e^{h(u)} \sqrt{\frac{2\pi}{|h''(\hat{\mu})|}} \quad 5.6$$

$\hat{\mu}$ is the value that maximizes $h(u)$ where $h(u)$ is the joint log-density of the data and the random effects evaluated as a function of u . This method is computationally efficient, stable for large datasets and works well for generalized linear models. After estimating parameters, statistical tests are conducted to determine the significance of independent variables. Fixed and random effects significance are tested to determine whether a particular explanatory variable affects the postharvest losses significantly or not. The hypothesis to be tested is such that $H_0: \beta_r = 0$ vs $H_1: \beta_r \neq 0$ where β_r are the estimated coefficients. The significance of this parameter estimate is assessed using the Wald Test statistic $Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$, $SE(\hat{\beta})$ is the standard error of the estimate. Under the null hypothesis, the Wald test statistic approximately follows a standard normal distribution $Z \sim N(0,1)$. A large Z value indicates strong evidence against the null hypothesis.

5.4.4 Testing Random Effects

The significance of random effects is evaluated using the likelihood ratio test (LRT). This test compares two nested models (the null model without the random effect and the full model including the random effects). The likelihood ratio test is defined by $2(l_0 - l_1)$ where l_0 is the log-likelihood of the reduced model and l_1 is the log-likelihood of the full model. Under the regularity conditions, the test statistic

approximately follows a chi-square distribution $\Lambda \sim \chi^2_{df}$ where df is the difference in the number of parameters between the models.

5.4.5 Interclass Correlation Coefficients (ICC)

This measures the proportion of the total variation in the response that is attributable to clustering effects. In multilevel models, ICC provides insight into the degree of similarity among observations within the same cluster. An ICC value greater than 0.1 is considered enough to warrant a multilevel approach. The county-level ICC provides the proportion of variance by difference between counties. It is given by:

$ICC_{County} = \frac{\sigma_{\mu}^2}{\sigma_{\mu}^2 + \sigma_v^2 + \sigma_e^2}$ where σ_{μ}^2 is the county-level variance, σ_v^2 is stage-level variance and σ_e^2 is the residual variance. A higher value shows stronger clustering within counties.

The stage-level ICC also provides the proportion of variance by the difference between various stages in the value chain

$ICC_{stage} = \frac{\sigma_v^2}{\sigma_{\mu}^2 + \sigma_v^2 + \sigma_e^2}$ quantifies the proportion of variability attributable to differences between stages of the potato-value chain.

5.4.6 Model Diagnostics

After estimating the model, diagnostic procedures are performed to evaluate its adequacy and validity. This study adopted information criterion procedures which balance the model fit and complexity. These criteria are the Akaike and Bayesian information criteria. The Akaike Information Criterion (AIC) is given by $-2l + 2k$ where l is the log-likelihood and k is the number of parameters. Bayesian information criterion is given by $-2l + 2k \log(n)$ where n is the sample size. Lower AIC and BIC indicate a better model.

5.5 Multilevel Beta Regression Model

The multilevel negative binomial model is useful in explaining the overdispersion in absolute count results, but post-harvest loss data is better explained in terms of proportions i.e. the ratio of quantity lost to total quantity harvested. The proportional outcomes are constrained on the open interval (0, 1) but are generally skewed and

heteroscedastic, which contravenes the fundamental assumptions of classical linear regression. The multilevel beta regression model is thus chosen as a supplementary and theoretically suitable model of the proportional loss variable. This methodology has the flexibility of the beta distribution and the hierarchies of multilevel modelling to simultaneously estimate the individual-level determinants of losses, and the group-level variance elements at the county and value chain stages. The data of post-harvest loss are often measured in proportional rates within the open interval (0, 1) that is the proportion of produce lost at each value chain. Such data cannot be handled using linear regression because they have bounded interval, skewness, and heteroscedasticity. To solve these problems, Beta regression is used to model the outcome based on the beta distribution, whose mean is logit-linked, and the model has a single parameter ϕ that represents the dispersion of the variables. The beta regression model does not assume constant variance but instead adds a precision parameter ϕ . It enables more flexibility in the modeling of skewed and heteroscedastic proportional results. The hierarchical structure of loss data makes multilevel models an efficient way of describing loss data since there are individuals within the counties and value chain stages. Under this model, variability at each level can be quantified, standard errors can be adjusted to clustering, and random intercepts may be used to account for counties and value chain stages. Beta regression combined with multilevel structure takes into consideration hierarchical dependence and limited proportionality of the outcome.

5.5.1 Beta Distribution

Let P_{ijk} denote the proportional loss incurred by individual i in county j at the value chain stage k where $P_{ijk} \in (0,1)$. The response variable is a beta distributed variable under the beta regression parameterization suggested by (Ferrari & Cribari-Neto, 2004) such that $P_{ijk} \sim \text{Beta}(u_{ijk}, \phi)$ where u_{ijk} is the conditional mean also called the proportional loss and $\phi > 0$ is the precision parameter. The probability density function is given by:

$$\frac{\Gamma \phi}{\Gamma(U \phi) \Gamma((1 - U) \phi)} P^{U\phi-1} (1 - P)^{(1-U)\phi-1}, \quad 0 \leq P \leq 1 \quad 5.7$$

where P is the dependent variable bounded between 0 and 1, U is the mean parameter, $\Gamma(.)$ Is the Gamma function and ϕ is the precision parameter which controls the

concentration of observations around the mean where large values indicate lower variability. The mean and variance of this distribution under parametrization are $E(P_{ijk}) = \mu_{ijk}$ and $var(P_{ijk}) = \frac{\mu_{ijk}(1-\mu_{ijk})}{1+\phi}$. A large value of the precision parameter ϕ implies a lower variance and a greater contribution of observations around the mean reflecting a more homogeneous loss distribution. A smaller value implies high dispersion which is usually a common feature of agricultural data. The conditional mean μ_{ijk} is linked to the independent variables through a logit link function which maps the bounded (0,1) response to the real line and assures that the predicted proportions remain within the admissible bounds. The logit function is given by:

$$logit(\mu_{ijk}) = \log\left(\frac{\mu_{ijk}}{1 - \mu_{ijk}}\right) = \eta_{ijk}$$

To account for the hierarchical nesting of the data, the linear predictor η_{ijk} is extended to include county level and value chain stage-level random effects. The full multilevel beta regression model is specified as follows:

$$logit(\mu_{ijk}) = \beta_0 + \sum_{r=1}^p \beta_r X_{ijk} + \mu_j + w_k \quad 5.8$$

where β_0 is the intercept, β_r is the fixed effects regression coefficients for X_{ijk} and w_k are the stage-level regression coefficients. $\mu_j \sim N(0, \sigma_u^2)$, $w_k \sim N(0, \sigma_w^2)$ and the random effects μ_j are independent of w_k . The standard beta regression model assumes a constant precision parameter ϕ across all observations. To allow for heteroscedasticity which is a common feature of post-harvest loss data where variance differs across value chain stages the precision parameter is modelled as observation-specific using a log link is $\log(\phi_{ijk}) = \gamma_0 + \sum_{s=1}^q \gamma^s Z_{ijk}^s$ where γ_0 is the baseline log-precision, γ^s are dispersion coefficients and Z_{ijk}^s are covariates hypothesized to influence the variability of proportional losses. The storage methods and the value chain stages are the primary covariates included in the sub-model.

5.5.2 Marginal Likelihood and Estimation

Maximum likelihood estimation was used to estimate the parameters of the multilevel beta regression model. The marginal likelihood was computed by integrating over the distributions of the random effects, which are included at the count level and at each

stage of the value chain. The Laplace approximation method was used to numerically estimate this integration because there is no closed form solution for it. The marginal likelihood of the multilevel beta regression model is obtained by integrating out the random effects. Given the random effects μ_j and w_k , the marginal likelihood is:

$$L(\beta, \gamma, \sigma_\mu^2, \sigma_w^2) = \iint L^c\left(\beta, \frac{\gamma}{\mu_j}, w_k\right) \phi(\mu_j, \sigma_\mu^2) \phi(w_k, \sigma_w^2) d\mu_j dw_k \quad 5.9$$

where L^c is the conditional likelihood given the random effects and $\phi(\cdot)$ denotes the standard normal density. Since the integral has no closed form, Laplace approximation is used to compute maximum likelihood estimates of the parameters. The parameter vector to be estimated is $\phi(\beta, \gamma, \sigma_\mu^2, \sigma_w^2, \phi)$. Once these parameters are estimated, interclass correlation coefficients for Beta regression were estimated to measure the extent of total variance that can be attributed to each level (Mondal et al. 2024). As beta regression is based on the logit scale, components of variance are viewed on the transformed scale. In the case of logit-normal, the level 1 residual

variance of the latent is $\sigma_\varepsilon^2 = \frac{\pi^2}{3}$. The county and stage level ICC is given by:

$$ICC_u = \frac{\sigma_u^2}{\sigma_\mu^2 + \sigma_w^2 + \frac{\pi^2}{3}} \quad 5.9.1$$

and

$$ICC_w = \frac{\sigma_w^2}{\sigma_\mu^2 + \sigma_w^2 + \frac{\pi^2}{3}} \quad 5.9.2$$

The ICC of more than 0.05 on the logit scale is enough evidence to support the multilevel structure of proportional outcomes (Asmare et al. 2025). These ICC estimates tell us about the relative significance of county-level and stage-level variables in the description of variability in proportional post-harvest losses.

5.5.3 Interpretation of random effects

The estimated fixed-effect coefficient β_r is the change in the log-odds of the proportion loss that occurs with a one unit increase in covariate X_r fixed while other variables and random effects are kept constant. The delta method was used to get the marginal effect on the proportional scale such that:

$$\frac{\partial u}{\partial X_r} = \beta_r u(1 - u) \quad 5.10$$

The marginal effect is computed at an estimated mean μ and gives a viable estimate of the overall proportional loss change that is attributed to each covariate. Both the multilevel negative binomial and multilevel beta regression models were used complementarily to achieve an appropriate level of methodological robustness. The multilevel negative binomial model was used to analyze the absolute count of post-harvest losses, while the multilevel beta regression model was employed to examine the proportional magnitude of losses. The two modelling approaches were complementary and gave a complete statistical description of the potato post-harvest loss patterns across Kenya's potato value chain.

5.6 Results and Interpretation of Multilevel Beta Regression Model

This chapter reports the statistical findings of the multilevel modelling of post-harvest losses across the potato farm-market value chain in Kenya. A sample size of 628 respondents was surveyed in Nakuru ($n=388$) and Nyandarua ($n=240$) counties, which comprised farmers and retailers within the farm/transport stage ($n=426$) and the retail stage ($n = 196$) of the value chain. Following the use of listwise deletion of missing values in key modelling variables, the analytic sample consisted of 622 respondents. There were 5 models that were estimated sequentially; a Poisson baseline, a single-level Negative Binomial (NB), a multilevel NB, a single-level Beta regression, and the primary model - the multilevel Beta regression. Findings are reported on model selection and fit, intraclass correlation coefficients, Poisson baseline estimates, the complete multilevel Beta regression fixed effects with marginal effects, Wald significant tests, county-level random (intercept) results and model diagnostics.

5.6.1 Model Selection and Fit statistics

The Poisson baseline model generated a deviance dispersion ratio of 0.244 and a *Pearson's $\chi^2 / degree of freedom$ ratio* of 0.243 ($p = 1.00$) indicating that no overdispersion was found. The NB dispersion parameter deviated to infinity ($\theta = 2,393,074$) showing the NB family is not better at the count outcome than Poisson, and the likelihood ratio test was non-significant ($\chi^2 = -0.006, p = 1.000$). In the case where the loss is proportional, the single-level Beta regression was at pseudo- Pearson $R^2 = 0.7612$. The multilevel Beta model was a better fit (*log – likelihood was 1375.65 vs 1362.12*) shown in **Table 5.4** below.

Table 5.4: Model Comparison Statistics

Model	Log-Likelihood	AIC	BIC	Parameters
Poisson (SL)	-1718.41	3470.82	3546.18	17
NB (SL)	-1718.41	3472.83	3552.62	17
NB (ML) *	NA	NA	NA	20
Beta Regression (SL)	1362.12	-2688.23	-2608.44	21
Multilevel Beta Regression ✓	1375.65	-2711.30	-2622.64	20

The negative binomial multilevel model information criteria are not provided since the Hessian is not positive definite, as the dispersion parameter tends to infinity, the highlighted row identifies the model chosen, Poisson and negative binomial models and Beta models work on different scales of outcomes and cannot thus be directly compared since Poisson and NB models count outcome while Beta regression models proportional outcome. SL and ML denote single level and multilevel models respectively.

5.6.2 Intraclass Correlation Coefficients (ICC)

The null multilevel beta model on the logit scale was computed to generate intraclass correlation coefficients (ICC) using the recommended latent level 1 residual variance of $\sigma_\varepsilon^2 = \frac{\pi^2}{3} = 3.290$ as recommended by (Devine et al., 2024). The variance at the county level was $\sigma_u^2 = 0.000817$ ($ICC = 0.02\%$) and the variance at the stage level was $\sigma_v^2 = 0.001342$ ($ICC = 0.04\%$). Though the estimated values are lower than the standard significance level of 0.05, this is to be taken with care. This has been largely due to the demonstrations of few units grouping up in the upper levels of the model, especially where two groups only could be observed (Nugent & Kleinman,

2021). Small units at higher levels in multilevel modeling lower the reliability of the estimates of variance components and can result into unstable or erroneous statistical inferences. Consequently, the observed statistical significance might not be necessarily a manifestation of strong underlying effects but a phenomenon of the nature of the data structure. Thus, the focus should be made on the extent and usefulness of the estimates instead of focusing on statistical significance. The correct criterion of justification is the likelihood ratio test (LRT), which was very significant ($\chi^2 = 27.07, df = 2, p = 0.001$), and proves the multilevel structure to be a significant factor that enhances the model fit. ICC estimates are shown in Table 5. below.

Table 5.5: Intraclass Correlation Coefficient

Grouping Level	Variance (σ^2)	ICC	ICC (%)	Justified?
County (j = 1, 2)	0.000817	0.0002	0.02%	NO
Value Chain Stage (k = 1, 2)	0.001342	0.0004	0.04%	NO
Level 1 residual variance $(\sigma_\varepsilon^2 = \frac{\pi^2}{3})$	3.2899	—	—	—
LRT: Multilevel vs Single-level Beta	$\chi^2 = 27.07$	df = 2	p < 0.001	YES ✓

5.7 Poisson Regression Results

The count-based loss outcome was estimated using the Poisson regression as a baseline. The model explained significant variation (*null deviance* = 591.91, *residual deviance* = 147.80, *df* = 605). All predictor effects showed consistent direction and significance values with those of the multilevel beta regression, which cross-model validated the fixed effect structure. Figure 5.1 shows the incidence rate ratio (IRR) plot of the multilevel NB model, which graphically summarizes the predictors that significantly increase or decrease the incidence rate of count-based losses. IRR is a measure used in Poisson and Negative binomial regression (models for count data, like number of potato units lost) that tells you how a predictor changes the expected count (or rate) of the outcome, relative to a reference

category. The most significant loss reducing and risk-increasing effects are cold storage ($IRR = 0.651$), dark room storage ($IRR = 0.804$), storage over one month ($IRR = 1.174$) and market distances greater than 10 km ($IRR = 1.112$) respectively. Full estimates are in Table 5.6 below.

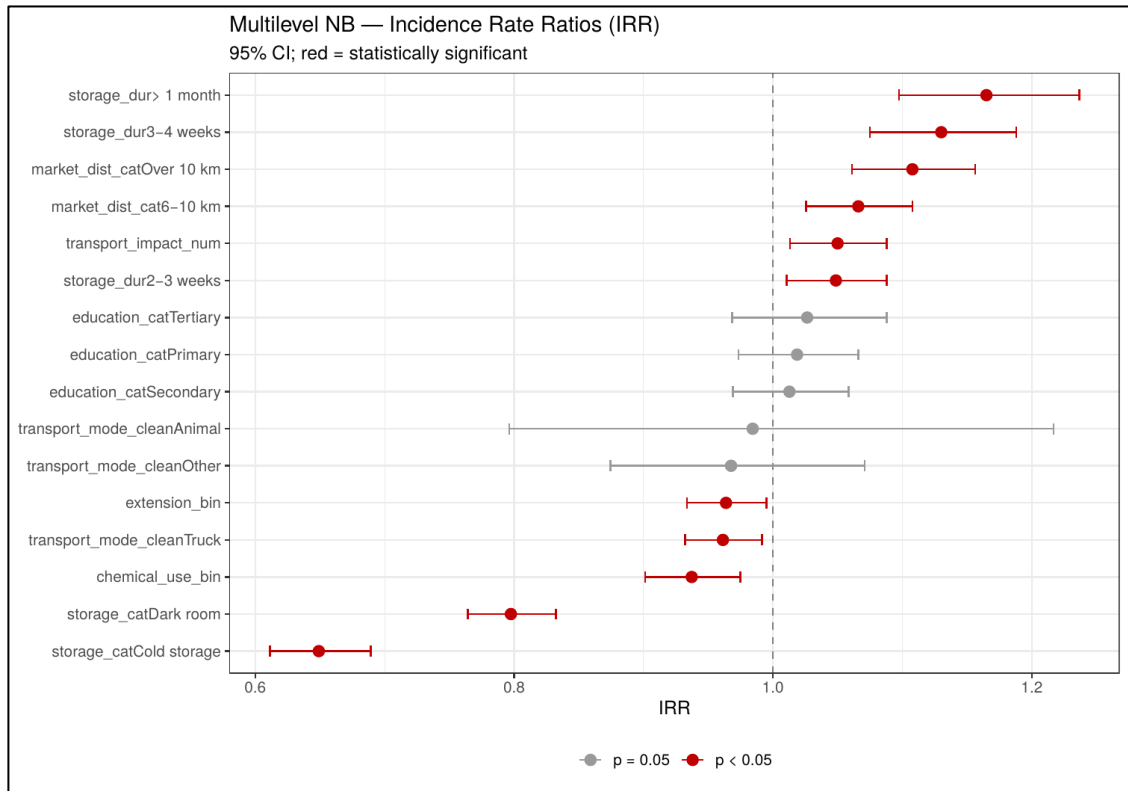


Figure 5.1: Multilevel Negative Binomial model

Figure 5.1 gives Multilevel Negative Binomial model that is; Incidence Rate Ratios (IRR) with 95% confidence level. Markers with red color denote significant predictors (p value less than 0.05). Reference categories: Traditional storage, less than 2 weeks storage, less than 5 km market distance, no education, Motorcycle transport.

Table 5.6: Multilevel Negative Binomial Fixed Random Effects

Variable	β	SE	Z	P-value	Sig.	IRR
(Intercept)	3.344	0.058	57.38	< 0.001	***	28.31
Storage: Dark room	-0.218	0.020	-10.72	< 0.001	***	0.804
Storage: Cold storage	-0.429	0.030	-14.12	< 0.001	***	0.651
Duration: 2–3 weeks	0.052	0.018	2.90	0.004	**	1.053
Duration: 3–4 weeks	0.130	0.024	5.36	< 0.001	***	1.139
Duration: > 1 month	0.161	0.029	5.51	< 0.001	***	1.174
Chemical use (Yes)	-0.060	0.019	-3.11	0.002	**	0.942
Extension training (Yes)	-0.033	0.015	-2.16	0.031	*	0.967
Market distance: 6–10 km	0.065	0.020	3.30	0.001	***	1.067
Market distance: > 10 km	0.106	0.022	4.86	< 0.001	***	1.112
Transport impact (severity)	0.048	0.018	2.69	0.007	**	1.049
Education: Primary	0.015	0.023	0.67	0.506	—	1.016
Education: Secondary	0.012	0.022	0.55	0.586	—	1.012
Education: Tertiary	0.026	0.030	0.89	0.376	—	1.027
Transport mode: Truck	-0.041	0.015	-2.64	0.008	**	0.960
Transport mode: Animal	-0.030	0.108	-0.28	0.781	—	0.971
Transport mode: Other	-0.037	0.052	-0.71	0.476	—	0.964

Table 5.6: *The Incidence ratio rate $IRR = \exp(\beta)$, The reference categories are; the traditional storage method, storage duration less than 2 weeks, market distance less than 5km, no education, mode of transport is motorcycle.*

The multilevel negative binomial (NB) model was used to model the absolute number of potato units lost as a function of the value chain. This is because the loss count data are over dispersed (the variance is higher than the mean), meaning that they do not follow the Poisson distribution assumption and thus its standard errors and significance level would be underestimated. The multilevel structure provides for clustering of respondents in counties (Nyandarua and Nakuru) with variance being partitioned into within-cluster and between-cluster variance.

The most important output of the model is the Incidence Rate Ratio (IRR) which is a multiplicative measure that quantifies the change in the expected number of lost units associated with a one unit increase in a predictor variable (or between the categories of a binary predictor). An IRR of less than 1 means the losses are reduced proportionately and IRR more than 1 means the losses are increased proportionately. The model intercept ($\beta = 3.344, IRR = 28.31, p < 0.001$) indicates how many potato units are lost by a farmer in the reference category (traditional storage, storage duration less than two weeks, no chemical use, no extension training, market distance less than 5 km, motorcycle or foot transport, no formal education). Under these conditions, the model simulations yield an estimated loss of potato of about 28 kg, giving an empirical anchor to assess the relative contribution of each predictor.

The absolute post-harvest losses in the NB model are best predicted by storage method. The two methods of alternative storage (dark room and cold storage) had highly significant reductions in the expected number of losses when compared to the traditional storage system. The absolute losses of potato actors stored in dark room storage was about 20% less than those stored in traditional storage ($Z = -10.72, \beta = -0.218$). The Z statistic value is very high, indicating a strong and well-measured effect. Dark rooms minimize light exposure, which slows the greening process and chlorophyll production, and create a more consistent and cooler microclimate that prevents sprouting and microbial activity compared to open air or pit storage.

The effect of cold storage is highest in the negative binomial (NB) model, with a reduction of about 34.9% of expected absolute losses ($Z = -14.12, \beta = -0.429$). This is the most significant and statistically sound result in the model. The process is simple, as the lower temperature reduces the rate of enzymatic activity, microbial growth and physiological processes in tubers. This leads to significantly reduced key deterioration pathways, sprout, rot, and dehydration, accounting for the observed effect.

Nevertheless, the use of cold storage is still very low, as only 9.2% of the respondents use it. This leads to the greater implementation gap that the most effective intervention is the least used intervention. In terms of policy, this is the greatest leverage point for

minimizing absolute post-harvest losses, provided there is a way to deal with infrastructure and access constraints.

Storage duration shows a clear, statistically significant and monotonic relationship with increased losses when compared to the reference category of less than two weeks. This is a pattern representing the overall biological deterioration of the tubers with the passage of time. 2–3 weeks: $IRR = 1.053$ (+5.3%, $p = 0.004$). This is a slight but statistically significant rise in losses, which is typical of early-stage sprouting and moderate moisture losses for 3–4 weeks whose $IRR = 1.139$ (+13.9%, $p < 0.001$). Losses rise steeply, reflecting heightened sprouting activity and faster dehydration and initial rotting processes. For more than 1 month the $IRR = 1.174$ (+17.4%, $p < 0.001$). This category shows the highest loss, indicating that there is a greater risk of deterioration for longer storage periods, and this effect is significant regardless of the storage method applied.

Of particular interest, 12.4% of respondents kept potatoes for over a month, frequently because of poor market prices or delayed sales. This suggests that a portion of the loss burden is due to market inefficiency, which is more structural than just post-harvest handling. Some interventions like better market information systems, forward contracting, and better linkages between buyers and farmers may make it easier to reduce the storage duration, thus reducing the duration-related losses.

The use of post-harvest chemicals (including sprout suppressants and fungicides) is significantly associated with lower expected absolute losses of approximately 5.8% ($\beta = -0.060$, $IRR = 0.942$, $Z = -3.11$, $p = 0.002$). The estimate is accurate but small compared with the effects of different methods of storing the potatoes. Adoption rates are relatively low, at 30.7%, which may be due to financial considerations and differential access in rural communities. If losses could be reduced at relatively low per unit cost, it could be done through expanded access, which might be accomplished via targeted subsidies or enhanced distribution systems.

Losses were significantly reduced by extension training ($IRR = 0.967$, $\beta = -0.033$, $Z = -2.16$, $p = 0.031$), with a 3.3% decrease in the number of losses among trained actors. The effect is significant at the 5% level, but is the smallest of all the NB predictors, indicating a relatively small impact of training on absolute volumes

of losses. This is because it is more pronounced in the Beta Regression model, which suggests that training has more impact on the consequences of loss management (what is lost as a proportion of what is managed than on the actual loss of counts).

There is a clear and statistically significant response relationship between market distance and absolute losses in the NB model. (6–10 km) $IRR = 1.067$ (+6.7%, $p = 0.001$). There is a significant increase in losses with moderate distances above the reference, as it is at longer transit times and higher physical stresses on tubers. More than 10 km, $IRR = 1.112$ (+11.2%, $p < 0.001$). The increased distance also amplifies the effect, with the extra 4.5 percentage points compared to the 6–10 km range showing a clear non-linear escalation. The more direct policy responses are better rural roads and strategic placement of collection centres.

The transport impact severity index, which reflects vibration, rough roads and handling during transport, is independently related to absolute losses ($\beta = 0.048$, $IRR = 1.049$, $Z = 2.69$, $p = 0.007$). This 4.9% rise in expected losses per unit of severity is independent of both distance and the quality of transit handling as measured by the transport mode, suggesting that the quality of the handling itself is a key driver of expected losses beyond proximity to market. This pathway would be addressed by packaging standards and improvements in the road quality.

Truck transport is statistically significant in the NB model ($IRR = 0.960$, $\beta = -0.041$, $Z = -2.64$, $p = 0.008$), meaning that there is a 4% decrease in the absolute losses expected for this mode of transport compared to the reference (motorcycle/foot). When potatoes are properly secured, they are better protected, loaded and transported quicker, and less vibration damage to the trucks. Other transport ($IRR = 0.971$, $p = 0.781$) and animal transport ($IRR = 0.964$, $p = 0.476$) have negative, though not significant, associations, possibly because of small subgroup sizes rather than a lack of effect. These estimates are inaccurate ($SE = 0.108$ for animal transport), which restricts the inferences that can be drawn.

The level of education did not significantly predict post-harvest losses in the NB model. The coefficients of primary ($p = 0.506$), secondary ($p = 0.586$), and tertiary education ($p = 0.376$) are all non-significant and have very small effect sizes ($IRRs$ of 1.012 – 1.027). This finding suggests that formal education does not

lead to a substantial reduction in the amount of produce lost. Specific and action-oriented interventions (such as better storage facilities, access to post-harvest chemical inputs and reinforced extension services) seem to be more effective in reducing losses as compared to education level.

The NB model statistically identifies the factors that explain the absolute potato post-harvest losses. The dominant structural determinants are storage method (especially cold storage), length of storage, market distance and transport mode. The protection effects of chemical use and/or extension training are smaller but more certain. There is no significant difference between the various education levels. The results of the NB model (incidence rate ratio) are particularly beneficial for policy implementation: The loss rate is significantly lower in cold storage compared to conventional storage, while the extension of potato storage time to over one month leads to an increase in losses by about 17%.

5.8 Multilevel Beta Regression Results

Multilevel Beta regression model was estimated using the value chain stage and county as the random intercept grouping factors. The model converged ($AIC = -2711.30$, $BIC = -2622.64$, $\loglikelihood = 1375.65$, $df = 2$, precision parameter $\phi = 308$). The high precision parameter implies that the proportional loss distribution is narrowly clustered around the mean, which is consistent with the narrow range of losses observed (15% to 45%). The Smithson and Verkuilen compression $P^* = \frac{P(n-1)+0.5}{n}$ was used to transform proportional losses to ensure values strictly lie in $(0, 1)$ where n is sample size, P is original proportion and P^* is the transformed proportion. All fixed effects are reported with odds ratios ($OR = \exp(B)$) and marginal effects (in percentage points) computed at the sample mean (0.319, 31.9%) which is the average proportion that was observed in the sample. Figure 5.2 shows the odds ratio tree plot of the multilevel beta model, which gives a visual representation of the direction and significance of all the predictors on the proportional loss outcome. Full estimates are in Table 5.7 below.

Multilevel Beta Regression model was used to analyze the proportional rate of post-harvest losses which is the ratio of the total produce lost. This is because the outcome variable is continuous, and bounded between 0 and 1, so linear regression is not

suitable because it can lead to predicted values that lie outside of the range 0 to 1. Similarly, Poisson and Negative Binomial models are not appropriate because they are for count (not proportional) outcomes. The most statistically appropriate framework is therefore the beta regression model where the mean of a beta distributed response is directly modeled via a logit link function.

The model generates Odds Ratios which are based on the logit scale and indicate a multiplicative effect of a change in the odds of losing a unit of the outcome variable associated with a one-unit change in a predictor variable. Marginal Effects (ME), in percentage points (pp), also measure the average absolute change in the probability of loss. For the reasons of direct and policy-relevant interpretation of practical impact, the marginal effects are the basis of discussion and interpretation in this research.

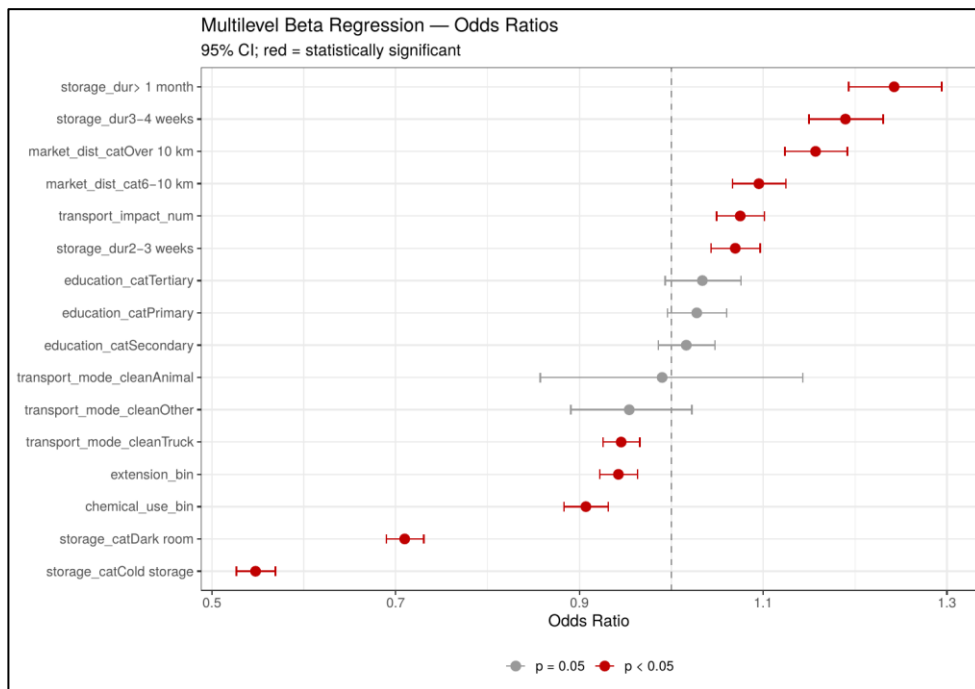


Figure 5.2: Multilevel Beta Regression

Figure 5.2 gives the Multilevel Beta Regression summaries: Odds ratios with 95% confidence interval. Red markers represent statistically significant predictors ($p < 0.05$). Odds Ratio less than one means a corresponding proportional decrease in the loss; $OR > 1$ means corresponding proportional increase in the loss.

Table 5.7: Multilevel Beta Regression Coefficient Random Effects

Variable	β	SE	z	p-value	Sig.	OR	ME (pp)
(Intercept)	-0.908	0.052	-17.53	< 0.001	***	0.403	—
Storage: Dark room	-0.343	0.015	-23.41	< 0.001	***	0.710	-7.44
Cold storage	-0.602	0.020	-30.31	< 0.001	***	0.548	-13.07
Duration 2–3 weeks	0.067	0.013	5.26	< 0.001	***	1.070	+1.46
Duration: 3–4 weeks	0.174	0.017	10.03	< 0.001	***	1.190	+3.77
Duration: > 1 month	0.217	0.021	10.45	< 0.001	***	1.240	+4.71
Chemical use (Yes)	-0.098	0.014	-7.21	< 0.001	***	0.907	-2.12
Extension training (Yes)	-0.059	0.011	-5.36	< 0.001	***	0.942	-1.29
market distance:6–10 km	0.091	0.013	6.78	< 0.001	***	1.100	+1.98
Market dist: > 10km	0.146	0.015	9.73	< 0.001	***	1.160	+3.17
Transport impact (severity)	0.072	0.012	5.84	< 0.001	***	1.080	+1.57
Education: Primary	0.027	0.016	1.71	0.087	.	1.030	+0.59
Education: Secondary	0.016	0.016	1.04	0.298	—	1.020	+0.35
Education: Tertiary	0.033	0.020	1.63	0.102	—	1.030	+0.72
Transport mode: Truck	-0.056	0.011	-5.20	< 0.001	***	0.945	-1.22
Transport mode: Animal	-0.010	0.073	-0.14	0.891	—	0.990	-0.22
Transport mode: Other	-0.047	0.035	-1.33	0.183	—	0.954	-1.02
<i>The Odds ratio OR = $\exp(\beta)$, ME (pp) are the marginal effects in percentage points at $\mu = 0.319$ (31.9%). The reference categories: Storage duration less than 2 weeks, distance less than 5km, mode of transport is motorcycle.</i>							

The intercept term ($\beta = -0.908, OR = 0.403, p < 0.001$) represents the log-odds of proportional post-harvest losses for factors in the reference category. The estimate, on the probability scale, indicates that the loss probability is less than 50%, meaning that retained produce is still under the baseline conditions included in the model. This gives a realistic and statistically plausible reference point for interpreting the subsequent marginal effects and predictor impacts.

The two most significant marginal effects in the entire Beta Regression model come from storage method, underlining its importance in determining the proportional post-harvest losses outcomes. Both estimates have very high Z-statistics, a high degree of statistical precision and reaffirm the robustness and reliability of these effects.

The largest marginal effect observed across the entire study was for cold storage, with proportional loss probability decreasing by 13.07 percentage points ($\beta = -0.602, Z = -30.31, OR = 0.548$). The Z-statistic of -30.31 is the highest recorded in the model, meaning there is very strong statistical proof for the robustness of the effect. The estimated odds ratio also suggests that the chances of proportional losses being incurred are about 45.2% less on the cold storage users than the traditional storage users. The magnitude of the effect, in view of the low adoption rate (9.2%), makes cold storage expansion the top intervention to minimize proportional losses post-harvest. In practical terms, a decrease in the loss fraction of 13 percentage points means that a trader would have about 130 kg more potatoes in every 1000kg available for sale in the market after storage and handling compared to when no measures were taken to reduce the loss fraction, emphasizing the potential impact the wider use of cold storage could have.

As expected from the Negative Binomial model, all increases in storage duration up to more than two weeks are statistically significant in terms of proportional losses when compared to the reference category of less than two weeks. Z-statistics are even higher for these variables, and the Beta Regression model has increased statistical power to detect effects associated with duration on proportional outcomes. Even relatively small increases to short-term storage durations significantly increase the proportion of produce lost, with a 2–3week extension to storage associated with a 1.46 percentage point increase in the proportional loss probability ($OR = 1.070, Z = 5.26, p < 0.001$). The effect becomes significant for 3–4 weeks of storage, increasing the probability of proportional loss by 3.77 percentage points ($OR = 1.190, Z = 10.03, p < 0.001$), because the sprouting, dehydration and resultant deterioration occur faster in these few weeks.

The longest duration effect is observed in those respondents who store potatoes for more than one month. This category has 4.71 percentage points higher loss probability than storage periods of less than two weeks ($OR = 1.240, Z = 10.45, p < 0.001$). The results suggest that the percentage of loss during extended storage is significantly higher for both types of storage. From a practical perspective, an increase of approximately five percentage points in the loss fraction translates to about 47 kg of additional losses per 1,000 kg of potatoes. This is a reminder of the need for

interventions aimed at improving value chain logistics, coordination of markets, linkages between buyers and better market access to reduce storage duration.

The results also demonstrate a progressive decrease in expected loss with the use of dark-room and cold-storage storage, ranging from an average of 30% for traditional storage conditions to around 25% for dark-room storage and even lower to around 20% for cold-storage conditions, with all other predictors held constant.

Chemical use is associated with a 2.12 percentage point reduction in proportional loss probability ($\beta = -0.098, OR = 0.907, Z = -7.21, p < 0.001$). The Beta Regression model shows a relationship with much greater statistical precision than the Negative binomial model, which has a corresponding Z-statistic of -3.11 . This indicates that the major role of the post-harvest chemicals is to minimize the percentage of deterioration in the produce during handling and storage, and not necessarily to reduce the absolute loss. The estimated odds ratio suggests that the odds of proportional post-harvest losses are about 9.3% lower for chemical users than for non-users. Even with this protection, many farmers still do not use post-harvest chemicals, as 69.3% said they did not use any. The adoption gap is quite large, reflecting significant untapped potential for reduction in losses by increasing access to chemical inputs. Input subsidies, better rural transportation networks and extension strategies to encourage the use of inputs in a safe and effective manner have the potential to significantly affect input use rates and minimize proportional losses through the value chain.

The proportional loss probability is 1.29 percentage points lower with extension training ($OR = 0.942, Z = -5.36, p < 0.001$). This effect is stronger in the Beta regression model than the NB model ($Z = -5.36$ vs -2.16 respectively) and points to a greater effect of training being on the proportional efficiency of loss management (i.e. actors losing a smaller proportion of what they handle), and not so much on the absolute losses themselves. The potential to expand extension reach to a larger target population (51.1% of trained) may result in another proportional loss reduction along the value chain.

The market distance shows a clear response relationship with proportional post-harvest losses in the Beta regression model, where both the distance categories have

highly significant effects. The proportional loss probability is 1.98 percentage point higher for farmers that live 6–10 km away from the market ($OR = 1.100, Z = 6.78, p < 0.001$), which means that the odds of them having a proportional loss increase of 10% compared to those of who live within 5 km to the market. This is probably due to the increased distance, multiple handling and physical stress of tubers during transport. This effect is significantly greater when the respondents are more than 10 km away from the market, and their proportional loss probability rises by 3.17 percentage points ($OR = 1.160, Z = 9.73, p < 0.001$). In real terms, this comes to around 32kg loss for every 1,000kg as distance exceeds 10km.

The Z statistics generated by the beta regression model are more significant for the effects of transport distance than those from the Negative binomial model, indicating that transport distance has a greater effect on the proportion of produce successfully making it to market than on the volume of losses. Transportation is likely to increase the percentage of unmarketable tubers due to greater exposure of vulnerable tubers to selective deterioration, bruising, dehydration, and spoilage during transit.

There is a statistically significant and independently estimated effect of transport severity on proportional losses ($\beta = 0.072, OR = 1.080, ME = +1.57 pp, Z = 5.84, p < 0.001$). For every increase in the severity index, there is a 1.57% increase in the proportional loss probability, and an 8% increase in the odds of a proportional loss outcome. The precision of this estimate in the Beta model ($Z = 5.84$ vs. 2.69 in the NB model) suggests that transport severity acts primarily through the process of proportional loss, affecting tubers by poor handling, poor road surfaces and inadequate cushioning in such a way as to increase the proportion of produce that is no good for eating even if total volume is unchanged. The most direct interventions are package material improvement, handling training of the transporters and road maintenance.

There are only 2 transport modes that are statistically significant in the Beta regression model; truck transport ($\beta = -0.056, OR = 0.945, ME = -1.22$ percentage points, $Z = -5.20, p < 0.001$), and all other transport ($\beta = -0.077, OR = 0.924, ME = -1.51$ percentage points, $Z = -1.03, p = 0.31$). This suggests that the proportional loss probability is reduced by 1.22 percentage points by those using trucks instead of motorcycles or foot transport, and that the probability of a proportional loss outcome

is reduced by 5.5% for those using trucks compared to those using motorcycles or foot transport. Other transport modes ($ME = -1.02$ percentage points, $p = 0.183$) and animal transport ($ME = -0.22$ percentage points, $p = 0.891$) are not statistically significant, with large standard errors in the animal transport mode since there were very few respondents in this category. The truck effect is estimated more accurately in the Beta model than in the NB model ($Z = -5.20$ versus -2.64), supporting the interpretation that the reduced proportion of produce loss due to better protection and reduced damage in transit are mainly responsible for the reduction in the truck effect, rather than a simple decrease in total loss volumes.

In the Beta Regression model, there is no meaningful association between education level and proportional loss outcomes. The secondary education qualification is not statistically significant ($p = 0.298$) and has a small positive marginal effect (+0.59 per cent), whereas primary education is marginally statistically significant ($p = 0.087$) and has a small positive marginal effect (+0.56 per cent), and tertiary education is not statistically significant ($p = 0.102$). But this weak positive association could be spuriously driven by unobserved confounding and is not robust in the specifications, nor consistent. This general trend is the same for both models and suggests that formal education does not lead to better outcomes of proportional loss management. The policy implication of this finding is that post-harvest losses can be best reduced if the interventions are specifically targeted and practice oriented, such as investments in storage facilities, distribution of post-harvest chemicals, enhancement of market linkages and extension services, then by general increase of formal education. For loss reduction programs, therefore, it is important to consider behavioral and infrastructure-based criteria over educational attainment.

5.9 Independent variables

5.9.1 Storage method and storage duration

The single most significant predictor was storage method ($Wald \chi^2 = 1182.71, df = 2, p = 0.001$). Dark room storage substantially lowered the risk of proportional potato losses ($\beta = -0.343, OR = 0.710, ME = -7.44pp, p = 0.001$), and cold storage further lowered the risk ($\beta = -0.602, OR = 0.548, ME = -13.07pp, p = 0.001$). Dark room and cold storage reduced the risk of losses by approximately 29.0% and 45.2% respectively compared to traditional storage, as

suggested by the odds ratios. The use of the dark room storage was predicted to have a 7.44 percentage-point lower probability of having proportional losses compared to traditional storage $Z = -23.41$, while cold storage was predicted to have a 13.07 percentage-point lower probability of a proportional loss. These findings indicate that losses by potatoes after harvest can be significantly reduced by adopting better storage technology within the range of the simulated conditions in the study. The very high Z statistic value shows that the effect is very large, and the precision of the estimation is also high due to the size of the sample. The odds ratio ($OR = 0.710$) suggests that the risk for proportional post-harvest losses is about 29% less for those who use dark room storage. Figure 5.3 shows that storage time also had a monotonically increasing effect on losses ($Wald \chi^2 = 137.39, df = 3, p < 0.001$): +1.46pp at 23 weeks, and +4.71pp after more than one month.

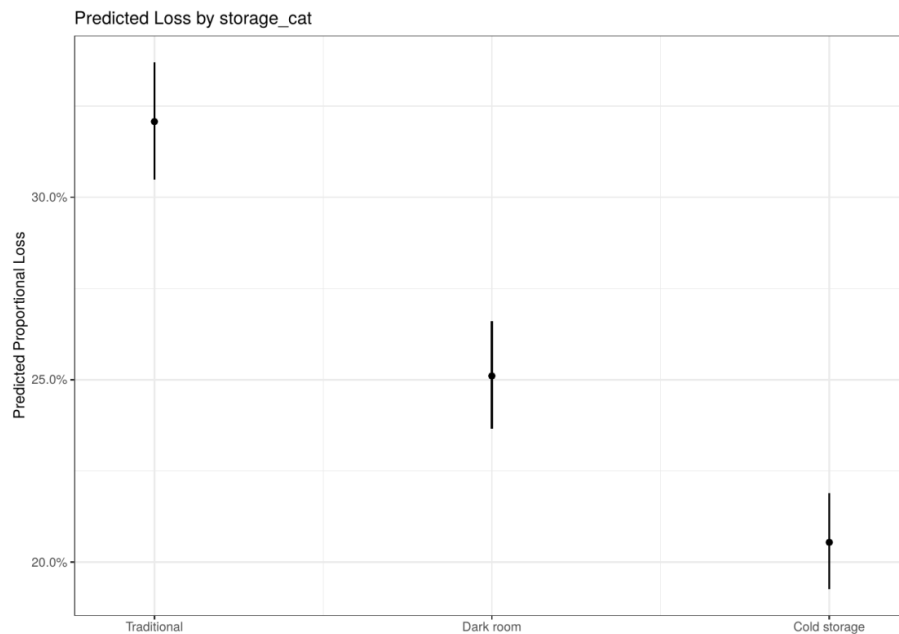


Figure 5.3: Proportional loss by method of storage

Figure 5.3: Predicted proportional loss to storage method adjusted on all other covariates. Error bars denote confidence intervals of 95% intervals.

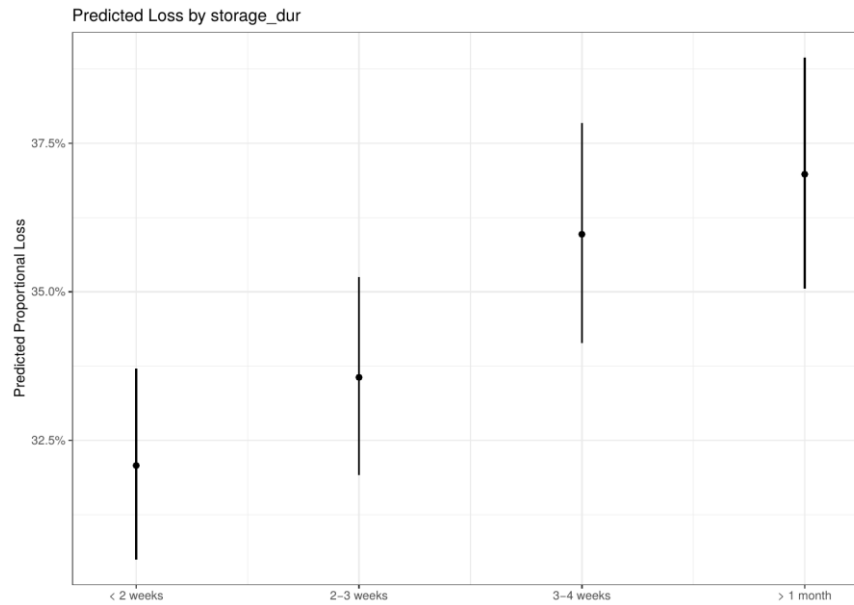


Figure 5.4: Proportional loss by Duration of storage

Figure 5.4: *Predicted proportional loss by length of storage predicted by other covariates, other covariates held constant. The gradual, unvarying increase is an indication of deterioration*

5.9.2 Extension Services and Education level

Proportional losses were also reduced significantly by extension training ($\beta = -0.059, OR = 0.942, ME = -1.29pp, Wald \chi^2 = 28.68, p = 0.001$), and trained respondents were estimated to lose about 31% versus 35% in the case of an untrained respondent Figure 5.5 in contrast, the level of formal education was not a statistically significant predictor ($Wald \chi^2 = 4.06, 3, p = 0.255$). The predicted education-based losses as denoted in Figure 5.6 indicate a relatively linear trend in education variables where the predicted proportional losses vary within a tight range of 31% to 34% once the variables of operation are held constant across the education groups. This contrasts with the unadjusted means where the relationship was non-linear, and it is therefore confirmed that the apparent education effect in the raw data is explained by the confounding effects of operational variables like storage method and chemical use rather than by a direct effect of education on loss behavior.

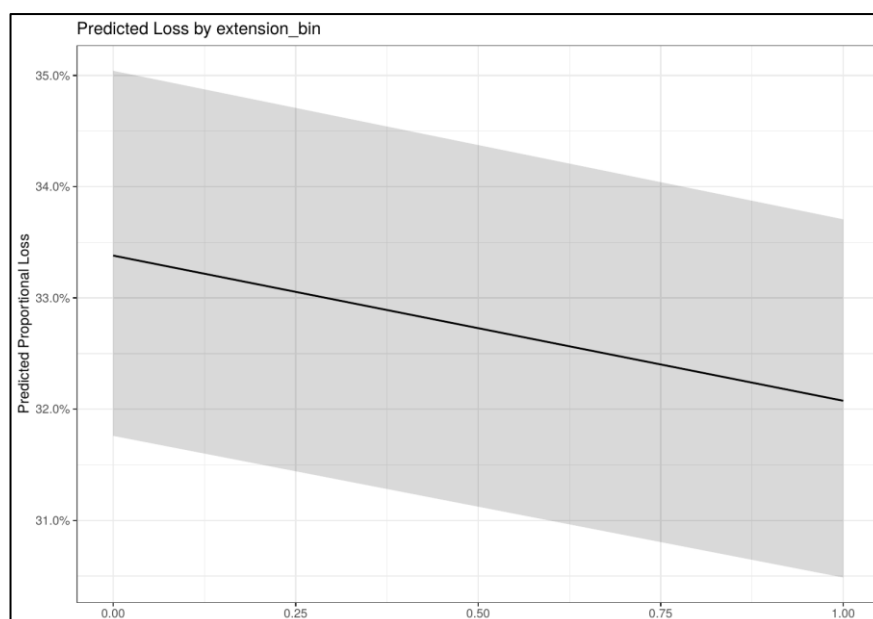


Figure 5.5: Model-predicted proportional loss by extension training

Figure 5.5 shows *Model-predicted proportional loss by extension training status* (0 = no training, 1 = training received), adjusted for all other covariates.

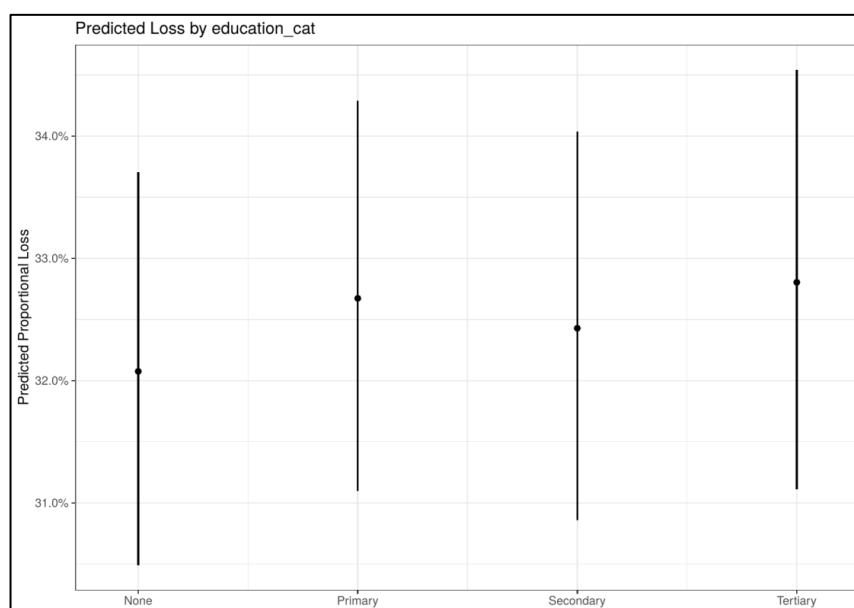


Figure 5.6: Proportional loss predicted by education Level

Proportional loss predicted by education level through the model, with all other covariates controlled. The flat-shaped profile shows that education is not significant in the model.

5.9.3 Market Distance and Mode of Transport

Market distance was an important positive predictor ($Wald \chi^2 = 95.97, df = 2, p = 0.001$). Compared to those within 5 km of market, those within 6-10 km experienced +1.98pp ($\beta = 0.091, OR = 1.10, p < 0.001$), and those outside the market beyond 10 km experienced +3.17pp ($\beta = 0.146, OR = 1.160, p < 0.001$). Figure shows model-predicted losses by market distance, and it is clear that there is a step increase in the percentage of approximately 32% to 36% across the three distance categories. Perceived severity of transport also significantly increased losses by 1.57pp per unit ($Wald \chi^2 = 34.10, p < 0.001$). Overall transport mode was also important ($Wald \chi^2 = 27.46, df = 3, p = 0.001$) as the mode of transport generated a 1.22pp reduction in comparison to motorcycle Figure 5.8. There was no significance in animals and other modes with their very small samples ($n = 3$ and $n = 14$ respectively).

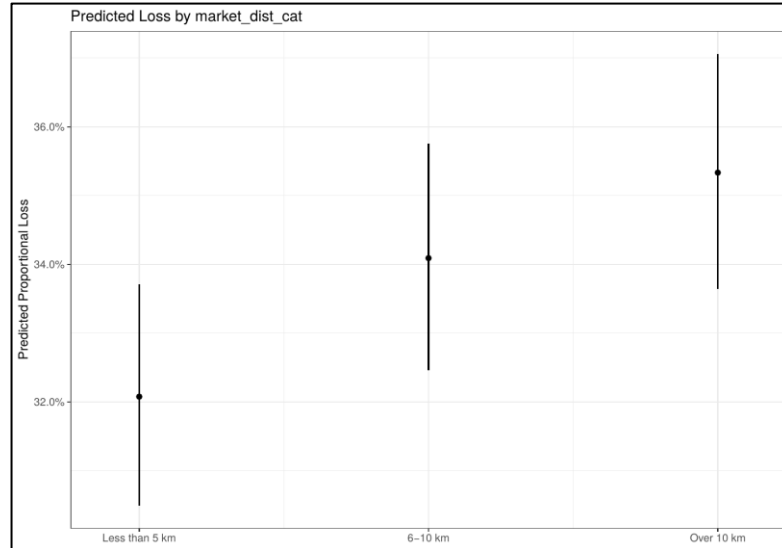


Figure 5.7: Proportional loss modeled through market distance category

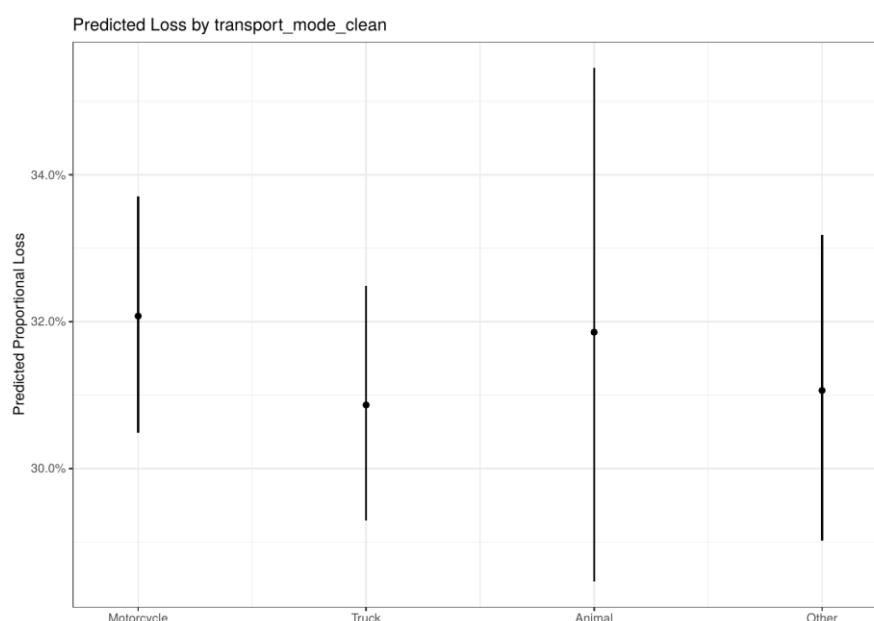


Figure 5.8: Model-predicted proportional loss by mode of transportation

The predicted model proportional storage and chemical use losses are shown in Figure 5.9 and Figure 5.10 respectively. Figure 5.9, the storage duration plot, demonstrates the non-linear response relationship that is monotonically increasing, with the predicted losses increasing as much as 32.5% with a storage period of less than two weeks and 37.5% with a storage period of more than one month. The loss percentage predicted for chemical users (about 29%) is significantly less than that of non-users (about 33%) and is statistically and practically significant, given the low price of sprout suppressants in comparison with the value of losses avoided, as illustrated by the chemical use plot in Figure 5.10.

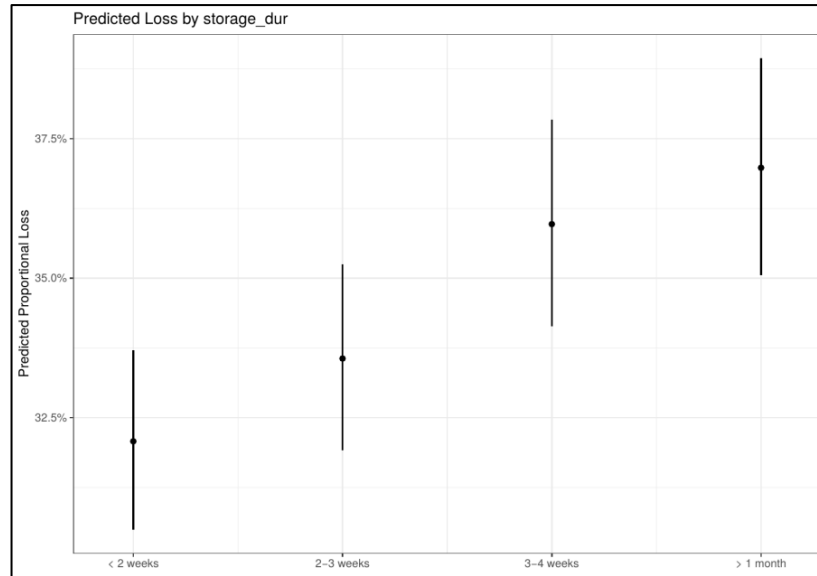


Figure 5.9: Proportional loss by duration of storage

Predicted proportional loss by duration of storage category and adjusted, all other covariates held constant. The cumulative physiological deterioration is reflected through the monotonic increase.

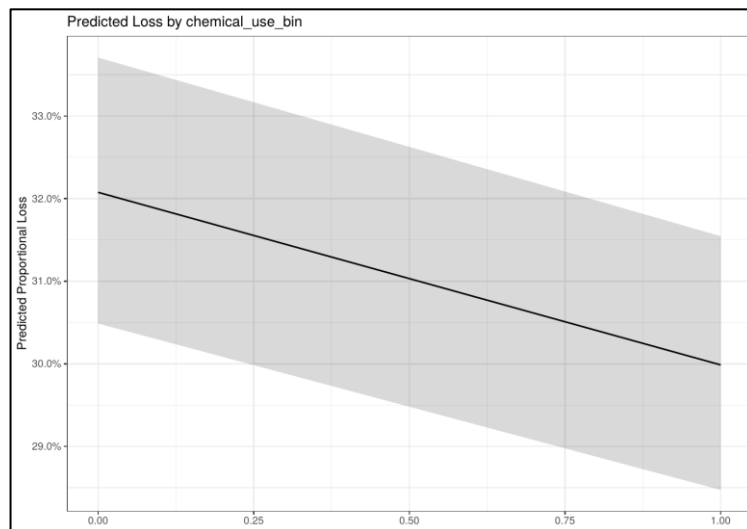


Figure 5.10: Proportional loss predicted based on chemical use

This figure gives the proportional loss predicted using the model based on status of using chemicals (0 = no chemicals, 1 = chemicals used) as adjusted by all other covariates.

5.9.4 Wald Tests

Type III Wald Chi-square tests were used to evaluate the overall significance of each predictor in the model. This method is especially suitable for categorical predictors for which multiple dummy variables are used, as it tests the overall null hypothesis that

all the coefficients for a given predictor are zero. This one of the advantages of Type III tests is that they measure the contribution of each predictor to the model, but not the effects in each category relative to a reference category.

Type III Wald Chi-square tests compared the overall significance of each predictor at all levels of predictor described by one or more dummies. Operational and market access variables were very significant. Storage method had the largest statistic ($\chi^2 = 1182.71, df = 2, p = 0.001$), then storage duration ($\chi^2 = 137.39$), market distance ($\chi^2 = 95.97$), chemical use ($\chi^2 = 51.93$), transport impact ($\chi^2 = 34.10$), extension training ($\chi^2 = 28.68$), and transport mode ($\chi^2 = 27.46$). Education level was not a significant predictor ($\chi^2 = 4.06, 3 df = 0.255$). Full results are in Table below.

Table 5.8: Type III Chi-Square Tests. Multilevel Beta Regression

Predictor	Wald χ^2	df	p-value	Significance
Storage method	1182.71	2	< 0.001	*** Highly significant
Storage duration	137.39	3	< 0.001	*** Highly significant
Chemical use	51.93	1	< 0.001	*** Highly significant
Extension training	28.68	1	< 0.001	*** Highly significant
Market distance	95.97	2	< 0.001	*** Highly significant
Transport impact	34.10	1	< 0.001	*** Highly significant
Transport mode	27.46	3	< 0.001	*** Highly significant
Education level	4.06	3	0.255	Not significant

5.9.5 County Level Random Effects

Figure 5. displays the county-level random intercepts estimated using the fitted multilevel Beta regression model. The random intercept of Nyandarua county was negative while that of Nakuru was positive; the two random intercepts were within the range of 0.02 on the logit scale. The near-zero magnitudes are also in line with the low values of the ICC, and they mean that beyond individual-level predictors, the county-level residual heterogeneity in proportional losses is not significant. The raw unadjusted disparity between the counties 31.5% and 32.4% in Nakuru and Nyandarua

is thus caused by differences in the distribution of storage methods, access to the markets and transport behavior of the two counties rather than structural factors at county level. The likelihood ratio test, however, established the fact that the random intercepts significantly contribute to the fit of the model ($\chi^2 = 27.07, p < 0.001$), which is why they should be left in the model. Figure 5.11 gives multilevel random intercepts at the county level of the Beta regression model (logit scale). A positive value denotes above-average proportional loss as compared to the fixed-effects prediction whereas a negative value denotes below-average.

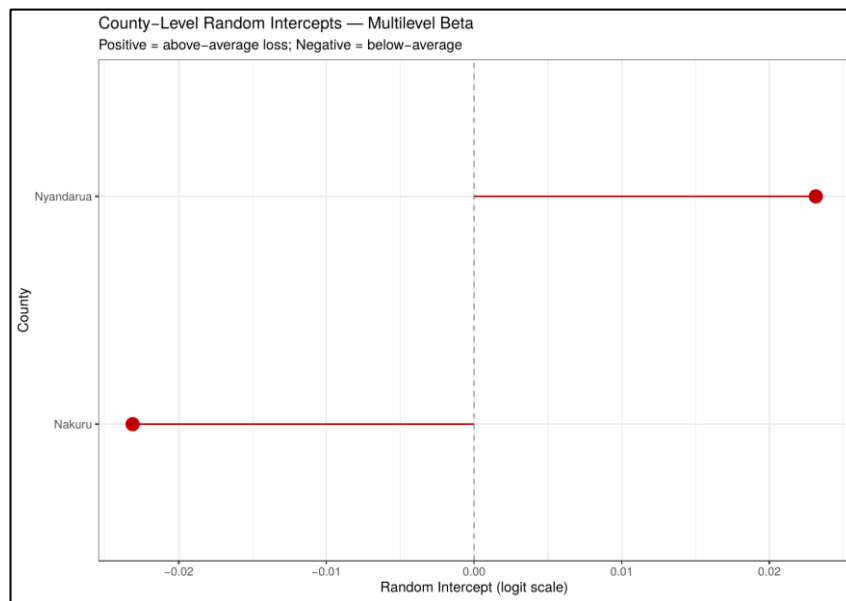


Figure 5.11: Multilevel random intercepts at the county level of the BRM

5.9.6 Model Diagnostics

The multilevel negative binomial and multilevel beta regression models were subjected to *DHARMA* simulation-based residual diagnostics to determine the appropriateness of their distributional assumptions. Figure 5.12 and Figure 5.13 shows the *DHARMA* diagnostic plots of each model respectively. In the multilevel NB model (Figure 5.12), the Q-Q plot of residuals shows a clear S-shaped deviation to the diagonal reference line; the Kolmogorov-Smirnov test ($p = 0.00$) and the dispersion test ($p = 0.00$) are both statistically significant. The plot of residual versus fitted indicates systematic curvature, and the overall adjusted quantile test is also significant. These findings indicate that the multilevel NB model cannot be used on this dataset because overdispersion is not present in the count-based outcome (*Pearson's* χ^2 /

$df = 0.243$) and the NB dispersion parameter θ approaches infinity. This gives a solid empirical validation that the NB family does not fit the count-based loss outcome in this research, and the Poisson and multilevel Beta regression specifications are the right modelling options.

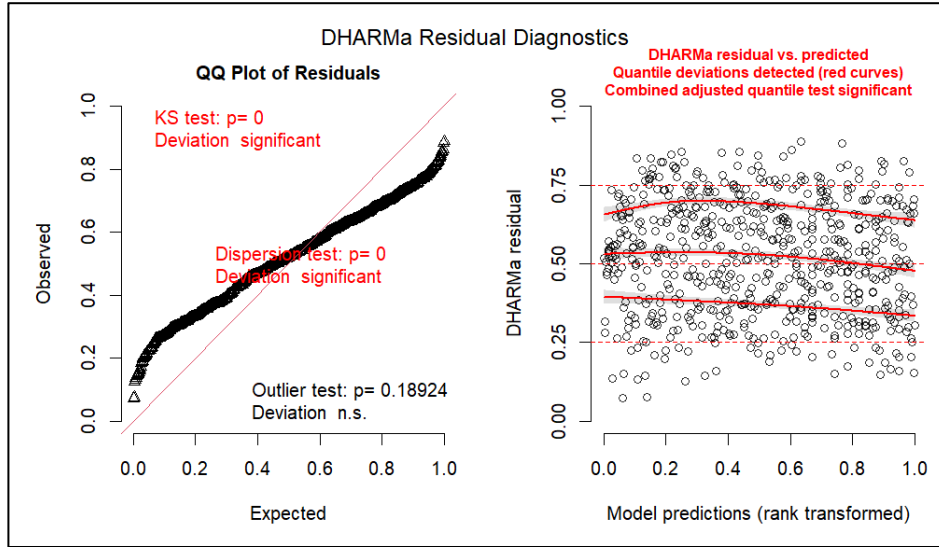


Figure 5.12: Dharma Residual Diagnostics Multilevel Negative Binomial Model

Figure 5.12 shows *DHARMA* residual diagnostics multilevel negative binomial models. Left panel: Q-Q plot of simulated residuals with a large deviation in the expected uniform distribution (KS test $p = 0.006$, dispersion test $p = 0.612$ not significant, outlier test = 0.189 not significant). Right panel shows the results of the residual versus model prediction (rank transformed) with systematic curvature and significant sum test of quantile. These two panels confirm model misspecification, which is also in line with the lack of overdispersion in the count outcome and the departure of the NB dispersion parameter toward infinity ($\theta \rightarrow 2,393,074$).

The high deviations in the multilevel NB diagnostics plots are not surprising considering the nature of the data, the count-based loss outcome is not overdispersed ($Pearson's \chi^2 / df = 0.243$) and the NB dispersion parameter was brought to non-zero in the estimation process, effectively reducing the NB to a Poisson. The *DHARMA* residual diagnostics of the multilevel Negative Binomial model shows obvious model misspecification. The Q-Q plot indicates a systematic departure of the distribution and the Kolmogorov-Smirnov (KS) test, and the dispersion test are both significant. Further, the residual-versus-fitted plot indicates that there are organized

trends, implying that the model assumptions are violated. These findings verify that the Negative Binomial model cannot fit the data as it does not have overdispersion and the dispersion parameter tends to infinity. These findings also justify the need to use proportional beta regression framework as a main analytical framework as seen in Figure 5.13.

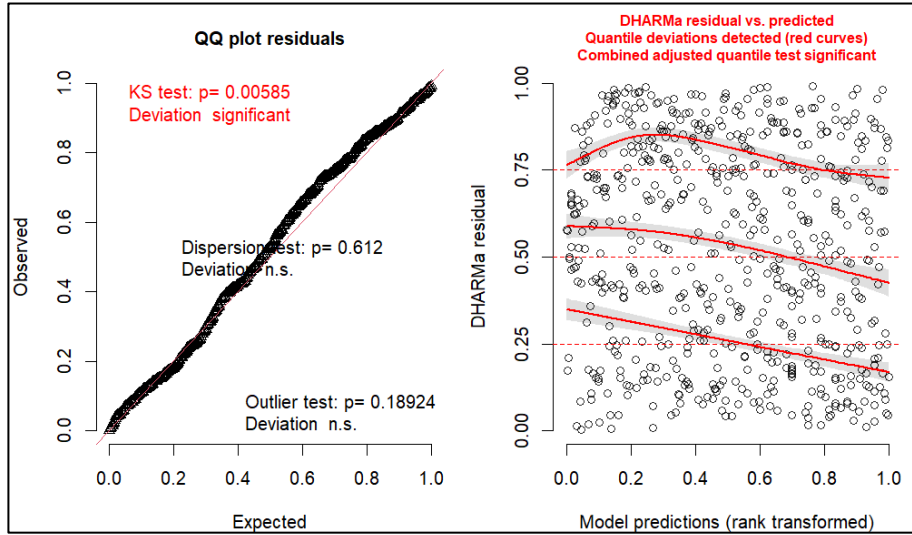


Figure 5.6: Multilevel Beta Regression Model, DHARMA Measure of Residual Diagnostics

Figure 5.13 shows Multilevel Beta Regression model, *DHARMA* measure of residual diagnostics. Left panel: Q-Q plot, there is a strong adherence to the diagonal reference line with slight upper-tail deviations (KS test $p = 0.006$; dispersion test $p = 0.612$, not significant.; outlier test $p = 0.189$, not significant.). Right panel shows Residuals versus model predictions with acceptable scattered results with no systematic curvature as shown in the NB model. These diagnostics establish that the proportional loss outcome has been specified correctly in terms of its beta distribution.

Unlike the NB model, the multilevel Beta regression model gives a much better fit to data. Since the non-significant dispersion test ($p = 0.612$) and outlier test ($p = 0.189$) indicate that the beta distribution is specified correctly in the proportional loss outcome, there is no systematic over- or underdispersion. Although the KS test attains marginal significance ($p = 0.006$), this is typical with large samples ($n = 622$) where even insignificant changes in uniformity are statistically significant and need not be construed as an indication of substantial misspecification. The plot of

the residual's vs fitted values indicates that there is a satisfactory values distribution without the steep turn as noticed in the NB model. This combination of diagnostic results largely confirms the hypothesis that the multilevel Beta regression is the appropriate and well-fitted model to use with this data, and all the inferences made based on parameter estimates are valid statistically. The single level *betareg* model provided a $pseudo R^2 = 0.7612$, which implies that the fixed effects explain the proportional losses variation of the 622 respondents in the analytic sample.

5.10 Chapter Summary

This chapter introduced the analysis of proportional potato post-harvest losses in a multilevel beta regression model. Two key properties of the response variable were important to the analysis: the bounded proportionality of post-harvest losses, and the hierarchical structure of the data by county and value chain setting. Alternative modelling techniques such as Poisson, negative binomial and beta regression models were compared, and the appropriateness of the multilevel beta regression framework assessed. The beta distribution was appropriate for representing the outcome of proportional loss, and the multilevel structure was useful to allow for the clustering and heterogeneity of observations. Model estimation, likelihood-based procedures, random effects, intraclass correlation coefficients, model selection and diagnostic assessment were discussed. The findings showed that storage methods are an important determinant of proportional post-harvest losses with improved storage methods yielding lower losses as compared to the traditional storage method. Other factors that were found to have a significant effect on post-harvest losses were storage duration, chemical use, extension training, market distance, and transportation conditions while formal education was not significant. The analysis, in addition, illustrated the importance of considering hierarchical structure of the data. In general, it was found that multilevel beta regression was appropriate for analyzing proportional potato post-harvest losses, and that evidence was presented on the factors that contribute to losses along the potato farm-market value chain in Kenya.

CHAPTER SIX

APPLICATION OF HYBRID RSM-GBM MODEL USING CLI

6.1 Introduction

This study aimed to present a methodological framework that simulates the integration of the Principal Component Analysis (PCA) and Response Surface Methodology (RSM) together with gradient boosting machines (GBM) in a controlled computational environment for demonstration purposes. The study did not involve any field observations or laboratory measurements as is typically done in empirical agricultural investigations, but instead involved synthetic experimental data based on informed parameter assumptions from potato post-harvest literature. It is important to understand that the framework is not considered a fully validated tool for agricultural decision making and should be taken as a first step computational model to test the methods. This simulation-based approach was chosen to reduce the costs of conducting large-scale potato storage experiments, the necessary equipment and space for controlled storage, and the length of the experimental periods needed where multiple environmental variables, storage durations, and interactions between different varieties were evaluated. The use of simulation-based modelling was thus a suitable and flexible approach to testing the performance of the hybrid analytical framework before its future empirical use. Thus, any results produced in the context of this should be understood as preliminary analytical results from simulated conditions and not as final conclusions for the agricultural sector.

Random environmental effects were included in the simulation by adding stochastic noise terms from a Gaussian probability distribution. These random factors were added to represent random environmental variations and random noise that occur in an actual storage system. Model included interaction effects between predictor variables to account for complex interdependencies of deterioration dynamics. The simulation framework was thus designed to mimic the kind of storage behavior found in the real world, but with enough variability to assess the analytical capabilities of the hybrid RSM-GBM modelling architecture. The observed predictive performance, however, may be partly due to the characteristics of the process used to generate the simulations rather than to real agricultural variability.

6.2 Experimental Design

This study used a Central Composite Design (CCD) as an experimental design. The CCD is a response surface design of second order, allowing estimation of the linear, quadratic and interaction effects of the factors without a full factorial design. Literature review was conducted to identify the predictor variables, which were: Temperature (°C), relative humidity (%), storage duration (days), light exposure (lux), mechanical damage (%), and curing duration (days). Five levels (-2, -1, 0, +1, +2) of each factor were coded for the axial, low factorial, center, high factorial and axial high points, respectively. The design space was established based on plausible storage conditions.

A rotatable Central Composite Design (CCD) comprising six predictor variables was developed to investigate the experimental conditions systematically and efficiently. For six factors, the fundamental CCD structure consisted of 64 factorial points (2^6), 12 axial points (2×6), and replication center points to allow for estimation of experimental error and assessment of model adequacy. The addition of axial and center points increased the capacity of the design to account for curvature effects and ensured rotatability of the design, thus allowing for equal prediction variance. Stochastic variation of biological storage systems was introduced by additional Monte Carlo replications. The CCD is an effective tool for estimating linear, interaction and quadratic effects when the storage conditions are carefully controlled, however, real storage conditions involve random variation due to biological heterogeneity, changes in the environment, and uncertainties in measurement. The Monte Carlo replications did not affect the CCD structure and statistical properties. In contrast, repeated observations were collected around each CCD design point to ensure the orthogonality and rotatability of the original design and achieve sufficient variability to assess the predictive ability of the hybrid approach. The real levels of the factors used are shown in Table 6.1.

Table 6.1: CCD Factor Levels for the Experiment

variable	–2 (α) (extremely low)	–1 (low)	0 (center)	1 (high)	2 (α) (extremely high)
Temp (°C)	4	8	12	14	18
RH	55%	65%	75%	90%	95%
Storage duration	Short period (15) days	30	60	90	>90 days
Light exposure (lux)	Complete darkness	50 lux	250 lux	500 lux	Excess exposure >500
Mechanical damage %	Minimal (0%)	5%	10%	20%	>20%
Curing duration	No curing (0 days)	5 days	10 days	14 days	>14 days

The complete CCD design matrix included 512 factorial points, 96 axial points and 80 center point replicates, for a total of 688 design points after 8 replications. For each design point, Monte Carlo simulation was done in R, with one simulated value per variety per design point. Eight potato varieties known to be grown in Kenya were simulated: Shangi, Dutch_Robjin, Markis, Tigoni, Asante, Sherekea, Unica and Kenya_Mpya, each with variety-specific mean loss function parameterization for their varied post-harvest traits.

6.3 Variables and Response Measurements

Three outcome variables were generated in each design unit that reflected different, but complementary aspects of post-harvest deterioration. Weight loss (%) was the percentage drop in fresh tuber weight after the storage period compared to the fresh tuber weight at harvest. This is mostly due to physiological losses caused by transpiration and respiration processes that cause progressive dehydration and loss of metabolic reserves over time. The second outcome, the Sprouting Index, was designed as a composite, weighted index of the development of sprouts during storage. It combines information about the number of sprouts, the mean length of sprouts, and the observed vigor, and then is converted onto a common scale for comparison of the experimental units. It is a conceptual operational surrogate for dormancy break and

physiological shift of tubers from dormancy to active growth, a major factor contributing to marketability loss. Rot loss (%), a third response variable, reflects the percentage of potato tubers that had some storage rots (fungal or bacterial). Rot loss differs from weight loss and sprouting, as it is irreversible and is a true measure of biological cause of post-harvest losses. The variable represents the combined effects of storage hygiene and microbial contamination, storage environmental conditions and disease pressure in the storage system. Rot loss is therefore an important tool to assess the effectiveness of storage conditions and post-harvest management to reduce the biological deterioration of potato tubers.

The three response variables were simulated at the experimental points of the Central Composite Design (CCD) in R. The simulation model was based on published empirical relationships showing how storage conditions affect potato post-harvest deterioration, and parameters from expert opinion appropriate to Kenyan potato varieties. Gaussian random error terms were added to each response variable to induce the natural variability found in storage and field experiments. The simulation procedures are carried out with the packages *rsm* and *MASS* in R.

6.4 Composite Loss Index

The Composite Loss Index (CLI) was developed to give a single quantitative value that represents multidimensional potato loss after harvest. Post harvest losses are a combination of multiple interrelated processes such as physiological weight loss, sprouting development and rotting progress therefore a single model response may be inadequate to describe the overall process of loss. To overcome this limitation, the Principal Component Analysis (PCA) was used as a dimensionality reduction method to combine the standardized loss indicators as one analytical response variable.

The post-harvest losses: Weight loss (%), sprouting index (%), and rotting loss (%) were first standardized to a mean zero and unit variance (Z-score) using the R function *scale()*. All factors were standardized to remove the scale dependency and to make all factors equally weighted in the covariance structure estimation before the implementation of PCA. Principal Component Analysis (PCA) was performed on the standardized data matrix using the *prcomp()* function. The first two principal components were used to create a composite response variable for the analysis, which

we call the Composite Loss Index (CLI) and is defined as: $CLI = PC_1 + 0.4PC_2$ where PC_1 and PC_2 are the scores of the observations on the first and second principal components respectively.

The first principal component (PC_1) represents the major direction of variation among the three post-harvest loss indicators and is thus considered as the main indicator of the overall impact of post-harvest losses. The addition of the second principal component (PC_2), with a coefficient of 0.4, ensures variability that is not captured by PC_1 is preserved in the composite index, and so, the representational accuracy of the composite index is enhanced without introducing undue weight to secondary structure. The CLI is a continuous index expressed in the scale of the PCA score space and higher values represent higher post-harvest losses. This index was the only response variable used in subsequent modelling. The second principal component was given a weighting coefficient of 0.4 to keep the secondary variation meaningful and to make sure that the principal deterioration structure (PC_1) was the main component of the composite index. This weighting scheme aimed to optimize information retention and interpretability while adding some extra variation that might not result in instability in the modelling of a response surface, thereby avoiding the effect of secondary orthogonal patterns. The value of 0.4 was also tested using sensitivity analysis to check the robustness of the CLI formulation. Several alternative weighting options using $PC_2 = 0.2, 0.6$, and 1.0 were tested and compared in terms of the stability of the predictors, the behaviour of the response surface, and overall hybrid model performance. Storage duration, relative humidity, and temperature remained the most important predictor variables in all weighting scenarios, with the rankings very little changed and the performance measure metrics showing relatively small changes. The results indicate that the hybrid modelling approach is robust with respect to small shifts in the weighting structure for the secondary principal components.

Observed stability in the results of weighting schemes suggests that the proposed Composite Loss Index has a satisfactory level of robustness within the simulated context as a composite loss measure. Moreover, incorporation of PC_2 enabled a more comprehensive multidimensional information to be captured in the index, improving representation of complex deterioration behaviour after harvest, without

compromising the analytical tractability for the subsequent RSM and gradient boosting analyses.

6.5 Integrated hybrid framework

The hybrid model used in this study was built in a residual-correction approach, using RSM and gradient boosting. Prior to model fitting, the three post-harvest loss indicators (Weight Loss, Sprouting Index, and Rotting Loss) were condensed into a single CLI via PCA. The three response variables were initially standardized (mean of 0) and scaled (SD of 1) (Z-score) and PCA was applied to the transformed matrix. The CLI was then defined as: $CLI = PC_1 \times 0.4PC_2$ where PC_1 and PC_2 are the first and second principal component scores, respectively. This index represents the primary axis of shared variation among the three loss indicators and is used as the response variable in the modelling. The coefficient of 0.4 on PC_2 retains some second order variation not captured by the first component. The data were divided into training (80%) and a test (20%) set using the `createDataPartition()` function from the *caret* package (*seed* = 42), ensuring the CLI distributions were representative in each set. The hybrid integrated model included three components:

The first component is the Response Surface Methodology (RSM): A second-order polynomial response surface methodology (RSM) model was applied to the CLI on the training set using the `rsm()` function with the `SO()` operator from the *rsm* package, which automatically generates all linear, quadratic and two-way interaction terms for the six coded predictor variables (`temp_coded`, `rh_coded`, `storage_coded`, `light_coded`, `damage_coded`, `curing_coded`).

Secondly, a GBM model was trained on the CLI using the *gbm* package on the training set with the actual (uncoded) predictor values and variety. The model settings were: *n.trees* = 4,000, *interaction.depth* = 5, *shrinkage* = 0.01, *bag.fraction* = 0.8 and *cv.folds* = 5. The number of trees was chosen by 5-fold cross-validation (*gbm.perf()*). This GBM is used to assess the predictive power of a machine learning model, as compared to the RSM and hybrid models. The third component is hybrid RSM-GBM Model (Residual Correction) using a sequential (residual correction) approach. The RSM model (Component 1) was first used to make predictions on the training set, and the RSM residuals. The RSM residuals represent

the systematic structure in the CLI that was not captured by the second-order polynomial RSM model, due to higher-order non-linearities, interaction effects and variety-specific effects. A second GBM model was then trained on these RSM residuals using the true values of the predictors, and variety with the number of trees again determined by cross-validation. The hybrid model prediction on the test set was then made by adding the residuals. This design strategy is deliberate: RSM captures the smooth, interpretable surface within the CCD design space, while the GBM residual model captures the remaining non-parametric complexity. The performance of the three models (RSM, pure GBM, and Hybrid) was assessed on the test set by Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R^2).

6.6 Statistical Analysis

All data simulation and statistical analyses were performed in R (version 4.3.3) using the packages: *rsm* for CCD design and RSM model fitting; *gbm* for gradient boosting; *caret* for train/test split; *car* for Type II ANOVA; *tidyverse* for data manipulation and visualization; and *gt* for table output. The *rsm* package was used to simulate the CCD matrix with Monte Carlo response. To determine predictor significance, a separate Type II ANOVA (*car::Anova()*) was fitted using *lm()* on the full response/uncoded predictors data set plus variety, reporting F-statistics and p-values for the effect of each predictor in explaining variance in CLI. The *persp()* function in the *rsm* package was used to visualize the response surface, with CLI as a response to Temperature and Relative humidity (coded predictors) while other factors are set at the center point levels (*coded* = 0). Typical RSM diagnostic plots (residuals vs fitted, Q-Q, scale-location, leverage) were created via *plot(rsm_{model})*. The significance level for all statistical tests was $\alpha = 0.05$.

6.7 Results and Discussions

The chapter provides the empirical findings of the hybrid RSM-GBM modelling framework used on the CLI of post-harvest losses. The results are presented in the following order: RSM model fit statistics and residual diagnostics, standalone GBM and residual GBM variable importance, model performance comparison between RSM, standalone GBM and hybrid RSM-GBM on held-out test set, predictor significance via Type II ANOVA with effect sizes, and the response surface analysis of the CLI across the most all numerical data are directly obtained through validated R output; numbers are displayed with their labels and directly preceded by discussion.

The analytical pipeline had four stages which were interrelated. The first stage involved deriving a CLI by the PCA to convert the three correlated outcome variables into a single scalar response that represents a combination of all three factors of deterioration in a statistically principled and variance-maximizing way. The entire data was then analyzed with a Type II ANOVA to determine which variables predicting storage were significant, and to assess the size of the predictive effect using standardized measures of effect size. The third step involved the fitting of a second-order Response Surface Model (RSM) to the coded predictor variables to describe the shape of the polynomial response surface, approximate the magnitudes and directions of linear, quadratic and interaction effects, and determine areas in the design space that correspond to minimum CLI. The final step involved training a GBM as an independent nonparametric model and then combining its residual predictions with RSM structural predictions in a hybrid RSM-GBM model.

6.7.1 Principal Component Analysis

6.7.1.1 PCA loadings

The PCA loadings reveal the underlying structure of relationships among the three post-harvest loss indicators. The first principal component (PC_1) has positive loadings of all variables, weight loss (0.4919), sprouting index (0.6372) and rotting loss (0.5933). This homogenous loading pattern suggests that PC_1 reflects the common variation which is shared among all three indicators and can be interpreted as a general measure of overall post-harvest deterioration. In practice, any increase in any of the individual components of the losses will have a positive impact on PC_1 thus confirming its appropriateness as the core dimension of the CLI.

In comparison, the second principal component (PC_2) shows a mixed loading pattern with a strong positive contribution of weight loss (0.8436) and negative contributions by sprouting index (-0.1803) and rotting loss (-0.5058). This means that PC_2 does not reflect the magnitude of the losses but distinguishes between the types of losses. It encapsulates an opposition between physical loss (weight loss) and biological driven processes of deterioration (sprouting and rotting). This difference is critical in the sense that various mechanisms can predominate under varying circumstances. The third major component (PC_3) explains the residual variation and it has opposing loadings across the variables and its interpretation is not as straightforward and its contribution to the overall variance is relatively lower. This component therefore does not make a significant contribution to the CLI. The overall loading structure provides support for use of (PC_1) as a primary measure of total post-harvest deterioration and addition of (PC_2) enables meaningful secondary variations that are not captured by the first principal component as shown in Table below.

Table 6.2: PCA Loadings

Response Variable	PC_1	PC_2	PC_3
Weight Loss	0.4919	0.8436	0.2153
Sprouting Index	0.6372	-0.1803	-0.7494
Rotting Loss	0.5933	-0.5058	0.6262

6.7.1.2 PCA Variance Explained

The amount of variance explained by the principal components provides a quantitative basis to assess how many components are to be retained. The total variance of the three indicators of loss can be summarized along a single dimension (PC_1) which represents 57.84% of the total variance. This confirms that there is a strong common structure among weight loss, sprouting and rotting and this supports the interpretation of PC_1 as an overall loss index.

The second principal component (PC_2) explains an extra 26.13% of the variance, and the total variance explained by the first two components is 83.97%. This cumulative percentage is large, well beyond the generally accepted dimensionality cutoff levels

(e.g. 70-80%), enough so that it can be regarded as a strong reason to keep both PC_1 and PC_2 in the analysis. Inclusion of PC_2 helps to maintain important variability not captured by PC_1 alone, especially variation based on differences in loss mechanisms. The third major component (PC_3) only explains 16.03% of the variance and does not add much more explanatory power than the first two components. The addition of PC_3 would be an unnecessary complexity with no significant improvement in interpretability or reflection of the data structure. Thus, the optimal balance between dimensionality reduction and information retention is obtained by retaining only the first two major components. The combination of the variance structure would justify the construction of the Composite Loss Index (CLI) based on PC_1 and PC_2 since they capture most of the variation in post-harvest losses and maintain both the overall severity and the secondary structural dissimilarity as shown in Table below.

Table 6.3: PCA Variance

Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC_1	1.3173	0.5784	0.5784
PC_2	0.8854	0.2613	0.8397
PC_3	0.6935	0.1603	1.0000

6.7.2 RSM Model Estimation and Diagnostics

6.7.2.1 RSM Model Fit

A second-order RSM model was estimated on the training sample. Table provides a summary of the important fit statistics of the RSM model on the full data. The standalone RSM model accounted for 47.54% of the variance in CLI ($R^2 = 0.4754$) and the adjusted $R^2 = 0.4484$ estimated 27 parameters. The general model was statistically significant (F test $p < 0.001$) which verifies that the second-order model is representative of real systematic variation in CLI. The remaining standard error of 1.0068 with 524 degrees of freedom is the size of unknown variability in the RSM framework. AIC (1603.23) and BIC (1728.33) are the values of models that are used to compare the models.

Although the RSM R^2 of 0.4754 is a statistically significant parametric fit to a 27-parameter model of complex agricultural data, it implies that about 52% of the CLI variance is not accounted for by the second-order polynomial approximation. This unaccounted variation, which is in the higher-order effects, the variety of complex interaction of various environments, and other nonlinearities is what the GBM residual component of the hybrid model is meant to explain. The hybrid modelling extension is thus justified by the moderate RSM fit.

Table 6.4: RSM Model Fit Statistics

RSM Model Fit Statistics	
Second-Order Response Surface Model (CLI) [†]	
Metric	Value
Log-Likelihood	-772.6161
AIC	1603.2321
BIC	1728.325
R^2	0.4754
Adjusted R^2	0.4484
F-test p-value	5.10×10^{-57}
Residual Std Error	1.0068
Residual df	524
[†] R^2 and Adjusted R^2 are key fit indices; F-test p-value reported in scientific notation.	

The model performance comparison of RSM, GBM, and hybrid models are shown in Table . The hybrid RSM-GBM model has the highest predictive performance on all the three metrics of evaluation: RMSE = 0.4031, MAE = 0.3231, and R^2 = 0.9169. The hybrid model reduced the standalone RSM by 57.75% in RMSE (0.9541 to 0.4031) and 37.66% in R^2 (0.5403 to 0.9169) which validates the significant value of the GBM residual correction in restoring the variance not accounted by the parametric structure of the RSM. The hybrid model was further reducing the RMSE (0.4453 to 0.4031) and the R^2 (0.8990 to 0.9169) by a further 9.47% and 1.79% points, respectively, as compared to the standalone GBM, indicating that the RSM structural component is an additional predictive feature that cannot be adequately explained

These findings align with hybrid modeling literature which has repeatedly shown that the use of parametric structured models and machine learning residual correction outperform both models on their own (Bock et al., 2021). The R^2 of 0.9169 obtained by the hybrid model shows that this model accounts for more than 91% of CLI variation in the test set; a good predictive result of such complicated agricultural data as potato variety and environment interaction and non-linear response dynamics. The nonlinear, variety-specific patterns that are not approximated by the second-order RSM are measured by the difference between the RSM alone ($R^2 = 0.5403$) and the hybrid model ($R^2 = 0.9169$). This 37.66 percentage-point increase in the amount of explained variance is the informational contribution of the GBM residual term, that reflects the rich genetic and environment interaction structure that is systematically missed by the parametric RSM framework.

Table 6.5: Model Performance on CLI

Model Performance (CLI Prediction)			
RSM vs GBM vs Hybrid			
Model	RMSE	MAE	R2
RSM	0.9541	0.7376	0.5403
GBM	0.4453	0.3378	0.8990
Hybrid	0.4031	0.3231	0.9169

6.7.2.2 Predictor Significance – Type II ANOVA

The predictor significance was determined using a Type II ANOVA modeled using `car::Anova()` on the complete data with CLI as the response and all six storage predictors and potato variety. ANOVA Type II tests the marginal contribution of each predictor, holding all other factors constant, to give order-independent rankings of importance. The results from the ANOVA must be interpreted with care since they were not obtained from the actual observations but from simulated observations. Thus, reported F-statistics, p-values and effect size measure the relative performance of variables in the simulated environment and are not empirical estimates of real-world potato storage systems. The results are presented mainly to show how the proposed modelling framework was able to identify and quantify the relationship between the

predictors and the response variable in a controlled way during simulation. The results of ANOVA are given in Table .

Table 6.6: Predictor Effects on CLI

Predictor Effects on CLI						
Significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$						
Predictor	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Significance
temperature_c	1	71.939	71.939	198.624	0.000	***
relative_humidity_pct	1	87.017	87.017	240.257	0.000	***
storage_duration_days	1	310.598	310.598	857.567	0.000	***
light_exposure_lux	1	49.689	49.689	137.192	0.000	***
mechanical_damage_pct	1	7.158	7.158	19.765	0.000	***
curing_duration_days	1	0.007	0.007	0.019	0.8902	
variety	7	507.926	72.561	200.342	0.000	***
Residuals	674	244.113	0.362	NA	NA	

Storage duration was the most important predictor of CLI in the simulated experimental context. This result is consistent with existing postharvest literature but is to be interpreted as a property of the simulated data-generating process, not as empirical evidence. The main contribution of this result is the ability to quantify the importance of the predictors in a controlled simulation environment. This variable had the largest F-statistic, $F = 857.567$ ($p < 0.001$) and a large Sum of Squares of 310.598. This ascertains that storage time is the prevailing determinant of total cumulative post-harvest losses in a variety of ways: sustained transpiration and respiration lead to progressive loss of weight, dormancy break leads to sprouting over time, and long-term exposure results in increased cumulative probability of pathogen-induced rotting. Its large F-statistic, which is significantly bigger than that of all other predictors, indicates its central role in the determination of CLI variation. Relative humidity is also highly significant, with $F = 257$, ($p < 0.001$) and a Sum of Squares of 87.017. It has a complex effect, acting through two opposite pathways: low humidity conditions enhance vapour pressure deficit, accelerating moisture loss, and high humidity promotes condensation and microbial growth, increasing rotting losses. This twofold impact on various elements of CLI adds to its great overall impact.

Temperature ranks among the most influential predictors, with $F = 198.624$ ($p < 0.001$) and a Sum of Squares of 71.939. It causes loss by controlling the rate of transpiration, sprouting and pathogen growth dynamics. Its lower rank in the ANOVA compared to storage period in the experimental range, and not a reduced biological significance. The most important effect is observed in Variety, $F = 200.342$, $p < 0.001$) and the largest Sum of Squares (507.926) of all predictors. This suggests that intrinsic varietal features have a significant contribution to variation in post-harvest loss, regardless of storage conditions. This effect is probably due to variability in physiological characteristics like dormancy duration and pathogen resistance.

Light exposure is statistically significant, with $F = 137.192$, ($p < 0.001$) and a Sum of Squares of 49.689. Its action is mainly related to the stimulation of sprouting by light-triggered physiological conditions, which lead to further deterioration during storage. Mechanical Damage is a relatively minor contributor to the total CLI variability, as it has a relatively small Sum of Squares (7.158). Its effect is mainly confined to rotting losses since physical injury offers avenues through which pathogens can enter.

Curing duration does not significantly explain the variation in CLI within the examined range, and the F value of 0.019 ($p = 0.8902$) is not statistically significant. This implies its impact can be small or context specific compared to the overriding predictors. Besides the difference in the importance of each predictor, the distribution of Sum of Squares across variables can be used to learn an important structural lesson about the determinants of post-harvest losses. Although the contribution of storage duration is the highest of all the continuous predictors, the contribution of variety is exceptionally large when considered as a whole. This implies that an optimal post-harvest management strategy should be one which combines both selection and control measures instead of one which relies on environmental regulation.

The findings also indicate that there are nonlinear and interaction-driven processes that are not necessarily well represented by the additive ANOVA model. Indicatively, the interaction behavior of relative humidity and the role of curing duration which is, in any case, implies that their impacts may be different at different levels of other predictors. This gives a high methodological justification of the subsequent

application of the hybrid RSM-GBM model where RSM reflects the structured polynomial relations and GBM reflects the remaining nonlinearities and higher order interactions.

Practically, the significance of predictor hierarchy offers clear information on how to prioritize intervention. In minimizing losses, reduction of storage period proves to be the best approach to minimization. However, there are some limitations in the storage facilities, and it is important to keep humidity and temperature as optimal as possible levels to minimize the rate of deterioration, as a secondary control measure. The strong impact of light exposure also supports the importance of simple and low-cost interventions, like creating dark storage conditions, whereas the role of variety can explain why it is important to choose cultivars with better storage resilience.

Lastly, the fact that the residual variance is relatively low as compared to the combined contribution of major predictors indicates that the model captures a significant portion of the underlying structure in the data. However, the unexplained variation that remains after the explanation justifies the addition of more flexible modelling procedures, like gradient boosting, to capture or explain complex patterns, which are not explainable by the classical parametric models.

6.7.2.3 ANOVA Predictor Effects on CLI with Effect Sizes η^2 and ω^2

A linear model was used to test Type II ANOVA of the relationship between CLI and all six continuous predictors, and variety on the entire dataset ($n = 688$). The ANOVA results are provided in Table , including partial eta-squared and omega-squared effect sizes. The statistical significance of the ANOVA findings represents conclusive statistical evidence on the relative significance of each predictor in determining CLI. Discussion of the findings is presented below:

Table 6.7: Predictor Effects on CLI with Effect Sizes

Predictor Effects on CLI with Effect Sizes							
Partial η^2 and ω^2 alongside F-statistics							
Predictor	Df	Sum Sq	F value	Pr(>F)	Partial_Eta_Sq	Omega_Sq	Significance
temperature_c	1	71.938702482	198.62431822	0.000	0.2276	0.0560	***
relative_humidity_pct	1	87.017483704	240.25716030	0.000	0.2628	0.0678	***
storage_duration_days	1	310.597618116	857.56676184	0.000	0.5599	0.2426	***
light_exposure_lux	1	49.688907640	137.19215197	0.000	0.1691	0.0386	***
mechanical_damage_pct	1	7.158436398	19.76459819	0.000	0.0285	0.0053	***
curing_duration_days	1	0.006904967	0.01906476	0.8902	0.0000	-0.0003	
variety	7	507.926393770	200.34225605	0.000	0.6754	0.3952	***
Residuals	674	244.112533187	NA	NA	0.5000	0.0001	
Partial η^2 : proportion of variance explained controlling for other predictors. ω^2 : bias-corrected effect size.							

Besides the individual effect sizes, the relative magnitude of partial η^2 and ω^2 give valuable information about the structure of variance in CLI. Partial η^2 indicates the percentage of variance in CLI accounted for by the predictor after controlling for the variance of other predictors, and ω^2 is a more conservative, bias-corrected estimate of the actual effect of the predictor in the general population. ω^2 corrects for the bias and gives better estimates of what the effect would be if the study is repeated with many different samples. The relatively small ω^2 for all predictors (particularly for temperature and relative humidity) suggest that all variables are important in explaining the variation in CLI, but that the actual population effect may be less than is suggested by the partial η^2 and may be more a function of the specific sample from which the data were collected. The relatively large values of ω^2 for the storage type and storage duration, however, suggest that both factors can account for a considerable amount of variance even after bias correction, pointing to greater stability, robustness, and generalizability of the effects within and across populations and settings.

Storage duration proved to be the one most statistically significant and practically important environmental predictor of CLI ($F = 857.57, p < 0.001, \text{partial } \eta^2 = 0.5599, \omega^2 = 0.2426$). The partial $\eta^2 = 0.5599$ indicates that all other predictors are already accounted and storage duration is the only predictor that explains 55.99% of the variance in CLI. Each of the three main loss mechanisms, weight loss from continuous transpiration and respiration, sprouting due to dormancy breakage and

rotting due to an accumulation of pathogens, tend to increase in severity as storage time progresses. These processes are all contributing to the exponential increase of the composite loss index with storage time. Practically, it indicates that one of the most effective and economically viable strategies to reduce the post-harvest losses is to reduce the storage time. This can be done by better market coordination, increased processing capacity and efficient demand-management systems which speed up the flow of products through the value chain. The results also highlight the need to increase the efforts for market development and the need for contract farming system as a strategy to reduce the time lag between storage and sales. These interventions are especially important for Kenyan smallholder potato farmers who often store their produce for extended periods of time in the face of poor and irregular market access.

The large overall partial eta-squared of all predictors was found to be potato variety ($F = 200.34, p < 0.001, \text{partial } \eta^2 = 0.6754, \omega^2 = 0.3952$) which led to the conclusion that genetic variation between varieties explains 67.54% of CLI variance after controlling all storage conditions. The 7 degrees of freedom associated with variety correspond to the 8 potato varieties used in the study and each variety had a different baseline Composite Loss Index (CLI) and a different response pattern to the storage conditions. This discovery offers quantitative proof for the well-established agronomic observations that the quality of storage and post-harvest management is largely due to varietal properties.

Potato varieties have shown significant variation in skin-set quality, dormancy period, dry matter composition, and fungal and bacterial resistance, which all have an impact on storage performance. A variety with good skin-set will tend to be more resistant to mechanical injury and moisture loss and will have a longer dormancy period. Likewise, dry matter content will have an impact on respiration rates and water loss, while disease resistance will be the deciding factor in susceptibility to rot. All these physiological and biochemical differences contribute to the large differences in post-harvest deterioration among varieties.

Relative humidity was the second most significant environmental predictor ($F = 240.26, p < 0.001, \text{partial } \eta^2 = 0.2628, \omega^2 = 0.0678$). The partial eta-squared of 0.2628 is a huge effect, meaning that the humidity control can explain

26.28% of the residual CLI variance. Relative humidity controls the gradient of vapor pressure between the tuber tissue and the atmosphere around: low humidity facilitates quick transpiration weight loss, whereas too much humidity facilitates condensation, formation of surface moisture, and growth of bacterial and fungal pathogens that cause rotting. Moderate humidity (between 85 and 95% RH) is the best humidity to store potatoes because too much humidity lowers the weight of the potato and too little prevents pathogenic rotting (which needs enough ventilation). The high level of significance and the large effect size of relative humidity in the present study confirm the fact that the humidity management systems, including the simple hygroscopic materials used in traditional shops as well as electronic humidity controllers used in modern ones are of high priority as post-harvest investment.

Temperature was the third most significant environmental predictor ($F = 198.62, p < 0.001, \text{partial } \eta^2 = 0.2276, \omega^2 = 0.0560$), with a partial η^2 of 0.2276 representing a large effect. Virtually all metabolic activities of stored potato tubers, such as respiration rate, transpiration rate, dormancy break, and pathogen activity, are controlled by temperature. High temperatures enhance respiration and dry matter losses, early sprouting and pathogen growth, all of which exacerbate CLI. On the other hand, refrigerated storage at 4 – 8°C is more effective in inhibiting all three loss mechanisms but needs energy infrastructure and capital investment that smallholder producers can seldom afford. The significant effect size of temperature confirms the underlying significance of temperature control in potato storage, and justifies investment in low-cost temperature control technologies, such as underground storage, shaded structures, evaporative cooling systems, and night ventilation strategies, especially in the Highland areas in Kenya where diurnal temperature change can be used to passively cool.

Light exposure demonstrated significant effects on CLI ($F = 137.19, p < 0.001, \text{partial } \eta^2 = 0.1691, \omega^2 = 0.0386$), with a moderate-to-large partial η^2 . Exposure to light stimulates the production of glycoalkaloids in potato tubers, leading to greening, bitter, and toxicity issues, and associated with rapid sprouting. The effect of light on CLI through the sprouting component is well-established in the post-harvest literature. The high, moderate-large impact of light suggests that eliminating

light during storage, which is a low-cost and feasible intervention, has a significant impact on reducing composite post-harvest losses.

Mechanical damage (percent bruising) was significant ($F = 19.77, p < 0.001, \text{partial } \eta^2 = 0.0285, \omega^2 = 0.0053$), though with a small practical effect size (partial $\eta^2 = 0.0285$ by Cohen's standards). The relatively low η^2 relative to the length of storage, the variety of products stored, the humidity and temperature of storage, and the fact that the development of pathogens and the loss of weight at the site of damage are both strongly driven by the humidity and temperature of the storage environment, independent of physical damage, explains this. Although the effect size is smaller, mechanical damage is still a significant target of post-harvest loss reduction, especially since it can be addressed by relatively low implementation costs compared to the losses avoided.

The only predictor that did not achieve statistical significance ($F = 0.019, p = 0.8902, \text{partial } \eta^2 = 0.00$) was curing duration. This observation implies that the range of curing time (0-14 days) does not significantly change the CLI when the other storage factors are held constant. A number of explanations are plausible; the fact that curing may be the primary determinant of the quality of skin sets, which is not directly captured by the three CLI components, the fact that the other storage factors may be the critical determinants, not based on the duration per se, and the fact that the optimal curing conditions (temperature and humidity during curing) may be the critical determinants, not based on the duration per se. This finding does not imply that the duration of curing was not important factor in the development of composite loss under the specific experimental conditions of this study and the CLI response variable of interest delineated.

The magnitude of the effects caused by storage duration and variety points to two sources of variation that can affect CLI: the effect of storage conditions over time and the biological differences between varieties. The large effect size and dominant F-statistic for storage duration may be due to the cumulative nature of this factor, as the impact of different deterioration processes builds up over the storage period. Variety, on the other hand, is a natural type of physiological variability which can affect the rate of occurrence of such deterioration processes. The significant overall effect size

for variety indicates that the deterioration process is highly genotype dependent and its expression is influenced by storage time.

Large effect sizes and a non-negligible residual variance also reveal that there is some amount of complexity that is not modelled within the linear ANOVA framework. Specifically, relative humidity has opposing effects, and the context-dependent role of curing duration indicates that relative humidity may be nonlinear and have interaction effects. These limitations have led to the implementation of the hybrid RSM-GBM modelling framework, with RSM, capturing the structured and nonlinear relationships between predictors, and GBM, which considers higher-order, non-linear, and interaction-driven residual patterns. In this regard, the results of the ANOVA would not only identify the key predictors but also provide a diagnostic basis of the subsequent modelling strategy.

Systematically, the findings support the idea that biological, environmental and temporal factors interact to contribute to post-harvest loss. Therefore, interventions that focus solely on one of the dimensions (e.g., temperature control alone) are unlikely to lead to optimal reductions unless they are combined with strategies that address storage duration and varietal suitability. This justifies a systems approach to the management of post-harvest, wherein interventions are synchronized within the production, storage, and access of market stages. Lastly, the high levels of effect size of the controllable environmental factors show the possibility of cost-effective technological interventions within the context of smallholder environments. Although the advanced cold storage systems might not be universally available, the dark storage, shading, moisture control, and improved ventilation can all be used to produce meaningful reductions in CLI. Together with enhanced market integration and adoption of storage-resilient varieties, these interventions provide a viable intervention pathway to reduce post-harvest losses on a scale.

6.7.3 Response Surface Analysis

6.7.3.1 3-D Response Surface (Temperature×Relative Humidity)

The RSM model was fitted to produce a 3D response surface plot of CLI as a function of temperature and relative humidity, with all other coded predictors held at their center point values (coded = 0). This surface is shown in Figure 6.1.

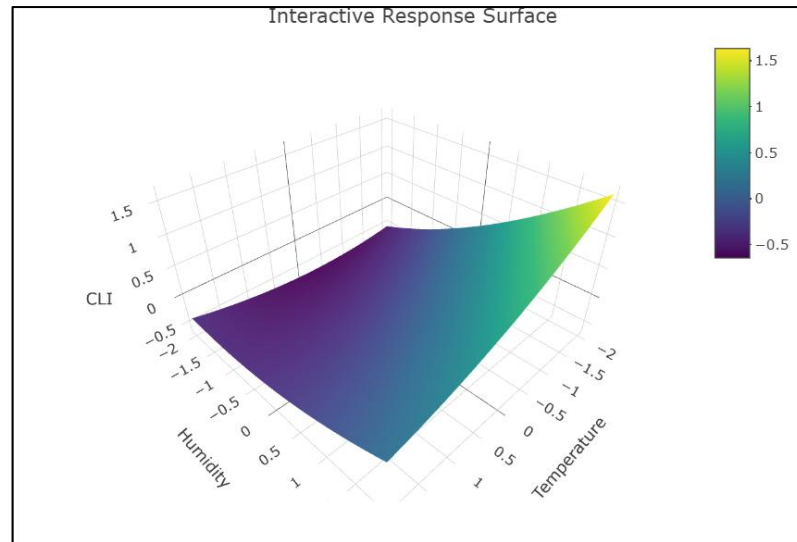


Figure 6.1: 3D Response Surface

Figure 6.1: 3D Response Surface: CLI as a function of Temperature (coded) and Relative Humidity (coded), with Storage Duration, Light Exposure, Mechanical Damage and Curing Duration held at center point values (coded -0). Created through interactive visualization of the fitted RSM model.

The response surface exhibits a curved shape, which captures joint effect of temperature and humidity on CLI. The surface is not a flat plane but has curvature in both temperature and humidity dimensions and this confirms the importance of the quadratic terms incorporated in the second order polynomial model. In the temperature dimension, the surface rises positively with increasing temperature across all levels of humidity, consistent with the multi-pathway positive effect of temperature on weight loss, sprouting and rotting. The temperature dependence of the rate of CLI increase with temperature is non-linear: the surface steepens at higher temperature values, indicating the non-linear acceleration of the deterioration processes with which temperature dependence of the rate of change is described. On the humidity dimension, the surface shows the typical saddle-shaped form that encodes the dual-risk structure of the effects of relative humidity. At the low end of humidity range (corresponding to dry conditions less than 65%), CLI is increased via the weight loss pathway as the large vapor pressure deficit stimulates rapid transpiration loss of moisture as tuber tissue. When humidity rises (higher than 90 percent), CLI rises due to enhanced rotting process as moisture condensation on tuber surfaces encourages the

germination of fungi and the growth of bacteria. The intermediate humidity range of 75–90% seems to be best for controlling weight loss due to dehydration, while still preventing an environment for pathogen growth. This middle optimum is consistent with the long-established empirical recommendation of 85-90% relative humidity to store potatoes which here is quantified precisely through the fitted second-order surface.

The surface geometry shows that the CLI response rises as both temperature and relative humidity increase, highlighting a strong relationship between the two. This implies that high humidity has an interactive effect on the rot and is more pronounced in warm conditions of storage. The observed interaction between temperature and humidity has practical implications, indicating that the adverse effects of high humidity on rotting are moderated by storage temperature. High humidity during low-temperature storage (4–12 °C) is correlated with lower CLI as compared to high humidity at ambient temperature (18°C or above), indicating that low-temperature storage inhibits pathogen growth and the course of deterioration despite the favorable moisture conditions. The optimal storage zone defined as a region where CLI surface is minimized is clearly identified at low temperature (*coded* -1 to -2 , $4 - 8^{\circ}\text{C}$) at moderate-to-high humidity (*coded* 0 to $+1$, $65 - 90\%$).

6.7.3.2 Contour Plot (Temperature and Storage Duration)

Figure 6.2 provides the interactive contour plot of CLI as a function of the temperature and storage duration coded variables and visualizes the top-down view of the response surface of CLI on the biologically significant predictors.

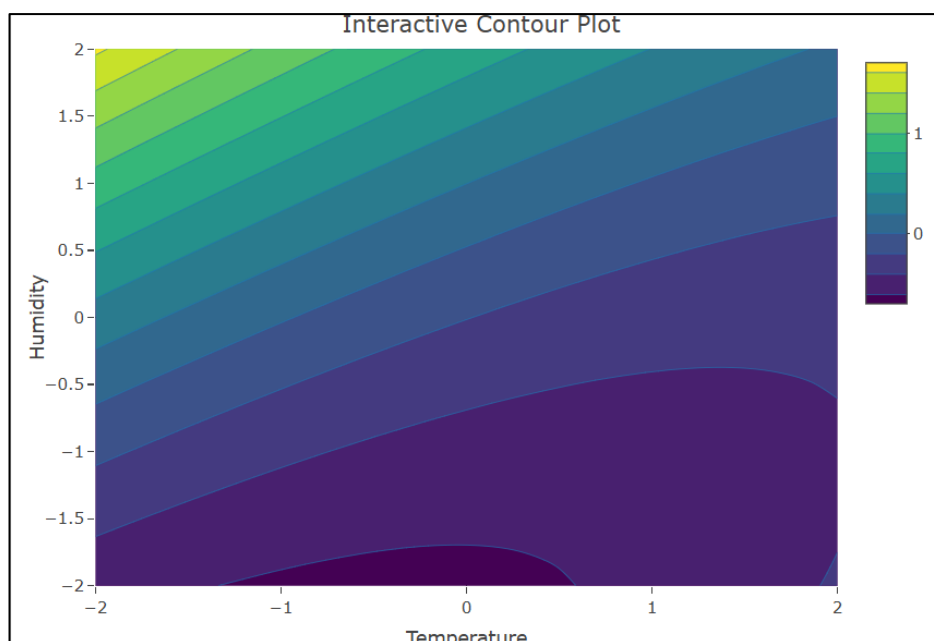


Figure 6.2: Interactive Contour Plot

Figure 6.2: *Interactive contour plot: CLI versus temperature (coded) and storage duration (coded), with relative humidity, light exposure, mechanical damage and curing duration held at center point values. contour lines indicate the same level of CLI; warmer colors (yellowish or greenish) represent a higher level of CLI.*

The contour plot indicates that there exists a strong positive interaction between temperature and storage duration in determining CLI. CLI contour lines are not parallel to either axis but instead converge to the upper-right corner of the plot space. The contour lines in the lower-left quadrant of the plot (cool temperatures, short storage times) are widely spaced, and CLI values are low, indicating the safe storage zone in which the losses are minimized. Mobility along the temperature axis at any given storage period, CLI increases monotonically, which is consistent with the positive temperature effect on all three loss factors. There is also an increase in CLI moving along the storage duration axis at any given temperature, and this is cumulative post-harvest deterioration. The slope of the increase in storage duration of CLI is, however, significantly steeper at high temperatures than at low temperatures confirming the super additive interplay between these two predictors identified in the ANOVA. At the lowest level of temperature (*coded as -2 , approximately 4°C*) the contour lines are widely spaced along the axis of storage time, indicating that CLI changes only slowly with time where cool storage effectively suppresses all the three

loss pathways and extends the safe storage window dramatically. The contour lines at the highest temperature level (*coded* + 2) are closely spaced along the storage duration axis, indicating that CLI rises significantly with time at ambient temperatures. Figure 6.1 and Figure 6.2 establish that the two most actionable management strategies in reducing CLI are temperature reduction and minimization of storage duration and that the two interventions interact synergistically.

6.7.4 Residual GBM Residual Importance

After the RSM fitting, RSM training-set residuals were calculated and used as the response variable in a second GBM model; the residual GBM component of the hybrid architecture. The following model (*n.trees* = 3,000, *interaction.depth* = 4, *shrinkage* = 0.01, *bag.fraction* = 0.8, *cv.folds* = 5) was fitted to retrieve systematic non-parametric CLI structure that cannot be represented by the second-order polynomial surface. The relative importance of variables in the residual GBM was measured using the relative influence statistic; the average decrease in squared error that can be attributed to splits on each predictor across all trees. The residual GBM variable importance is presented in Table 6.8.

Table 6.8: Residual GBM Variable Importance

GBM Variable Importance	
Residual GBM (Hybrid component)	
Variable	Relative Influence (%)
variety	92.7456
relative_humidity_pct	3.2391
storage_duration_days	2.4047
temperature_c	0.6593
curing_duration_days	0.3715
mechanical_damage_pct	0.3240
light_exposure_lux	0.2558

Table 6.8: *Residual GBM Variable Importance: Relative Influence (%) of each predictor in the GBM fitted to RSM residuals (hybrid component; n.trees = 3,000, interaction.depth = 4, shrinkage = 0.01, cv.folds = 5).*

The most methodologically informative output in this study because it directly quantifies which predictors have systematic CLI structure which is not captured by the second-order RSM *polyplot* of variable importance. The residual GBM importance is dominated by variety-specific non-linear response patterns with a relative influence of 92.75% confirming that the primary source of RSM residual structure is variety-specific non-linear response patterns. This very high dominance of variety in residual importance has a clear mechanistic explanation: the RSM model fits a single surface of the form of a polynomial, to data on eight different varieties all with fundamentally different dormancy periods, skin permeability, dry matter contents, and pathogen resistance profiles. The necessary condition of the linear and variety-independent functions of the predictors of the storage condition, is that the surface represented by the polynomials must be an average response of the varieties, and that the systematic differences of the responses of the individual varieties with the average response, which are themselves non-linear and variety-specific functions of the predictors of the storage condition, represent the dominant structure in the residuals.

6.7.5 Hybrid Model Actual vs Predicted

Figure 6.3 shows the scatter plot of the actual and predicted values of CLI in the hybrid RSM-GBM model on the test set, and it provides a visual evaluation of the quality of prediction throughout the entire CLI range.

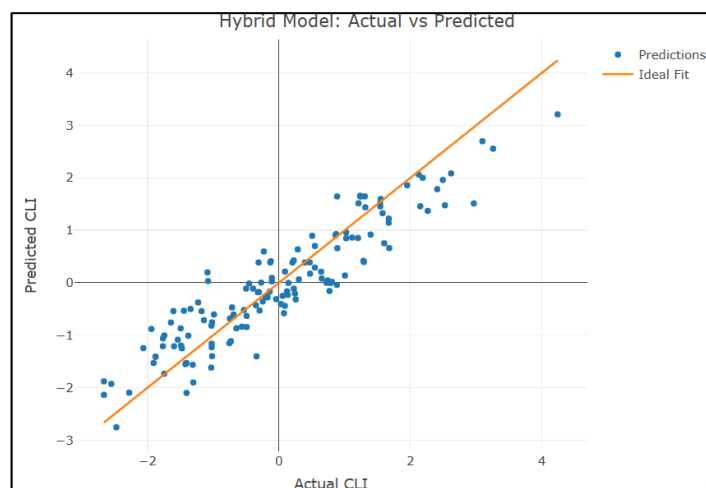


Figure 6.3: Hybrid Model: Predicted vs Actual CLI on the test set

The actual-verses-predicted scatter plot shows that the hybrid model is highly predictive. The points are tightly concentrated around the ideal-fit diagonal line over the entire range of CLI values between about -2 and +4, with no systematic curvature, offset, or deviation of the diagonal anywhere in the CLI scale. The high clustering is consistent with the R^2 of 0.9169 reported in the performance table: 91.69% of the variance in actual CLI values on the test set is explained by the hybrid predictions. The residual scatter around the diagonal is, approximately, symmetric, evenly distributed, and with no discernible pattern of heteroscedastic dissemination (the width of the scatter does not systematically increase or decrease across the CLI range). Only a few points deviate more substantially than the diagonal; these are visible at the high-CLI end (real CLI is approximately 3-4) and reflect observations that have been made in Asante under the most adverse storage conditions (high temperature, long storage period, high light exposure), where the random variability prescribed by the simulation is at its maximum and the exact combination of variety-specific and environmental extremes is difficult to predict precisely. The fact that the residual GBM component has managed to compensate for the systematic positive bias that the standalone RSM is exhibiting at the extreme values of CLI is confirmed.

6.7.6 Hybrid Model Residuals Diagnostics

6.7.6.1 Residuals vs Predicted

The plot of the residuals-versus-predicted-CLI of the hybrid model on the test set is shown in Figure 6.4.

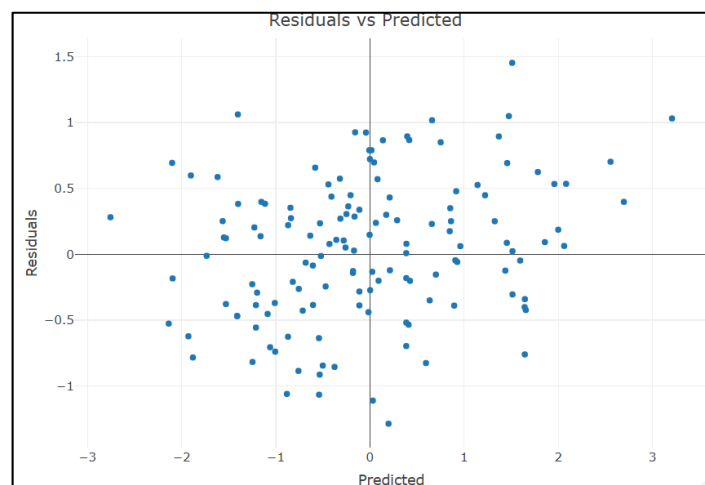


Figure 6.4: Hybrid Model: Residuals vs Predicted CLI on the test set

The residual versus predicted plot has verified that the hybrid model does not have any systematic predictive bias in the entire range of values that the model is trying to predict. The residual cloud is centered very close to the horizontal zero reference line throughout the predicted CLI range between about -1.5 to +3, and there is no discernible upward trend or downward trend. This confirms that the hybrid model does not systematically over or under predict at any point along the CLI scale, which confirms the effectiveness of the RSM residual-correction in removing the systematic over or under prediction at any point in the CLI scale. The distribution of the residuals is approximately constant over the predicted range, except that somewhat larger scatter values are obtained at the high end of the predicted-CLI range. The largest residual magnitude is about 1.5 CLI units, which occurs at a small number of the high-CLI observations that are the most extreme. These outlying residuals do not characterize a systematic pattern and are like stochastic aspect of the simulation as opposed to misspecification of the model. The approximately constant variance across the predicted range of the test-set predictions of the hybrid model confirms both unbiasedness and homoscedasticity.

6.7.6.2 Normality of Hybrid Residuals

The two supplementary tests of the residuals of hybrid models normality (an interactive Q-Q plot and a standard Normal Q-Q plot respectively) are presented in Figure 6.5 and Figure 6.6 respectively.

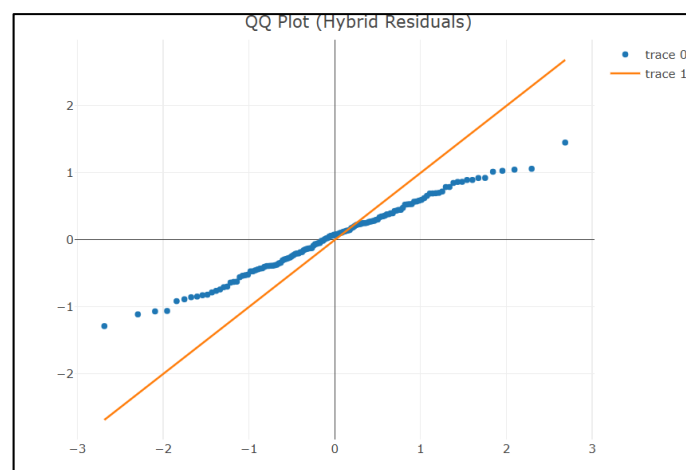


Figure 6.5: Interact Q-Q Plot Hybrid Model Residuals

Figure 6.5: *Interact Q-Q Plot Hybrid Model Residuals: real quantiles (points) versus ideal normal quantiles (line). The hybrids (test-set residuals of a hybrid model) were generated.*

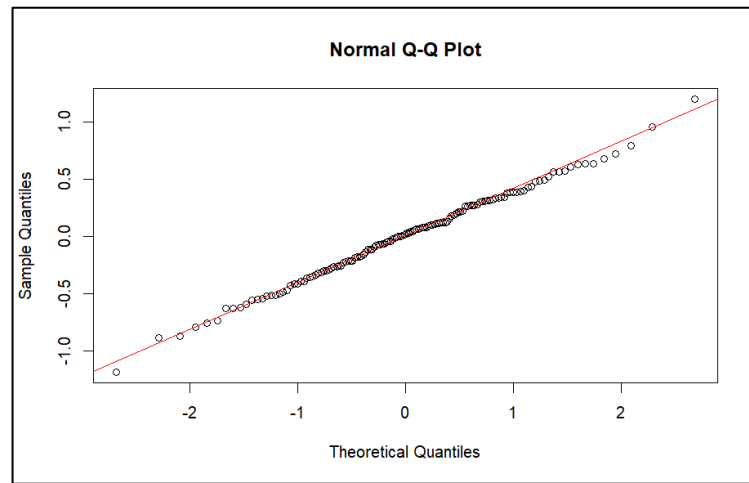


Figure 6.6: Normal Q-Q Plot of Hybrid Model Residuals

Figure 6.6: Normal Q-Q Plot of Hybrid Model Residuals. The points are standardized residuals that are plotted against the theoretical normal quantiles; the reference line is an indication of perfect normality.

Both Q-Q plots reveal that the residuals of the hybrid model are almost normally distributed. Figure 6.6 (interactive Q–Q) shows the observed residual quantile points (trace 0) closely follow the theoretical normal quantile line (trace 1) through the entire distribution between about -2.5 and +2.5 standard deviations. This alignment is especially fine in the center of the body of the distribution (between -1.5 to +1.5 standard deviations), where most of the observations lie, to confirm that the core residual distribution is well-modeled by a normal distribution. The small deviations at the ends of the distribution slightly heavier tails than the theoretical normal can be seen at both ends of the distribution, indicating that the residuals have slightly increased kurtosis as compared to a perfect Gaussian. It is a typical characteristic of residual distributions of mixed-effects and ensemble models fitted to heterogeneous data and does not have a material impact on the validity of inference. This interpretation is supported by Figure 6.6 (standard Q-Q plot); the standardized residuals are close to the diagonal in the central range with slight S-shaped deviations at the extremes that are consistent with light tail heaviness. Formal evidence that the

residual distribution does not significantly violate the assumption of normality at the 0.05 level is the Shapiro-Wilk test on the hybrid residuals, which is applied to a random subsample of up to 5,000 observations.

6.7.6.3 RSM Four Panel Residual Diagnostics

Figure 6.7 gives the RSM model four-panel diagnostic plot generated by the standard plot (rsm model) command in R. These four plots together test the major assumptions of the linear model framework on which the RSM is based: zero conditional mean of residuals, homoscedasticity, normality of residuals, and no influential observations.

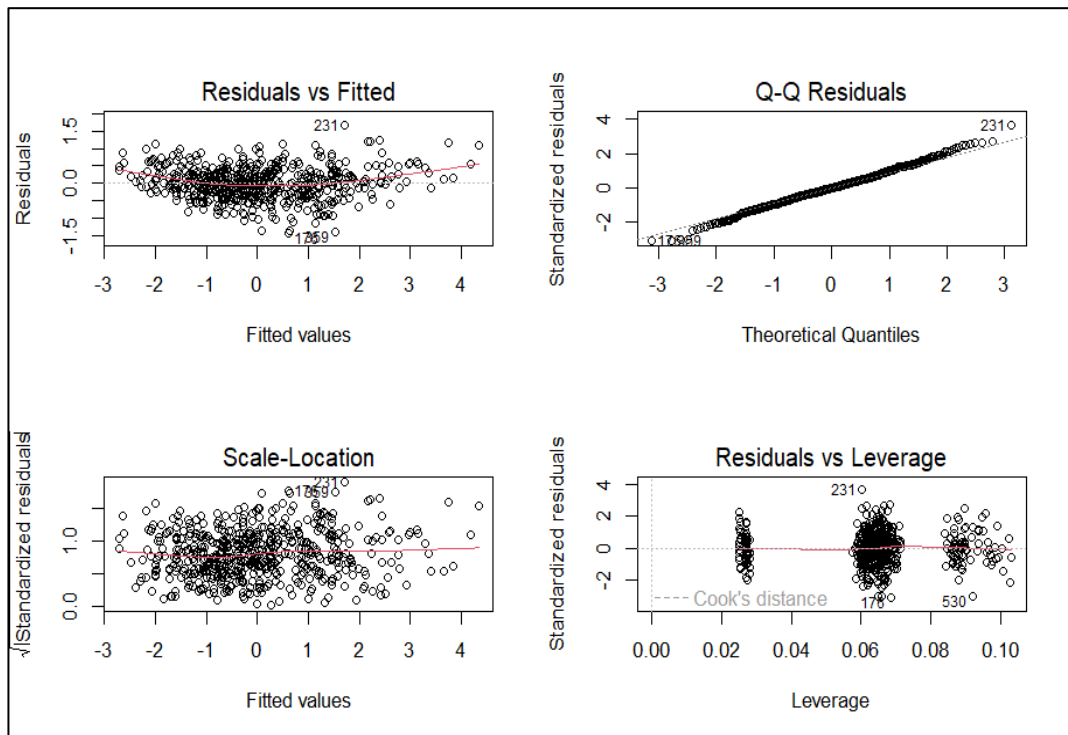


Figure 6.7: RSM Four-Panel Diagnostic Plots

Figure 6.7: RSM four-panel diagnostic plots of (top-left) Residuals vs Fitted, (top-right) Normal Q-Q plot of residuals, (bottom-left) Scale-Location plot and (bottom-right) Residuals vs Leverage. The red curves are trend lines of Locally Estimated Scatterplot Smoothing (LOESS), which are local regressions to subsets of the data to demonstrate underlying patterns without assuming a certain functional form. There are labelled observations (231, 1769, 1099, 17, 530) that represent points with relatively large residuals and/or large leverage.

To evaluate the adequacy of the fitted Response Surface Methodology (RSM) model, a series of standard diagnostic plots were examined, including Residuals vs Fitted, Normal Q-Q, Scale-Location, and Residuals vs Leverage plots. These diagnostics

were applied to test the major regression assumptions of linearity, normality, homoscedasticity and the existence of influential observations.

The Residuals vs Fitted plot shows that the residuals are usually concentrated around zero, which implies that the model accounts for a significant part of the systematic variation in the data. Nevertheless, the fact that a slight curvature on the smoothing line is present is an indication of mild nonlinearity and therefore, the form of the RSM as a second-order polyform does not fully explain all underlying relationships. This implies that there might be more nonlinear structure in the data. The normal Q-Q plot indicates that the residuals are closely clustering around the normal theoretical distribution with some small deviations observed at the tails. This shows that the assumption of normality is reasonably met, making it possible to infer statistically using the model. The minor deviations in the tail could be due to few extreme observations but not to such an extent that they can undermine the reliability of the models.

The Scale Location plot shows that the dispersion of the standardized residuals is constant across the fitted values, indicating approximate homoscedasticity. However, there is a slight upward trend in the smoothing line indicating that there is slight heteroscedasticity whereby the variance in the residuals has a slight upward trend with increasing fitted values. This deviation is small but shows that the homogeneity of variance assumption is not satisfied.

The Residuals vs Leverage plot shows that most of the observations lie within the acceptable Cooks distance limits, revealing that there are no highly influential data points. Even though some of the observations have moderate leverage, their residuals are within reasonable limits, implying that they do not have undue influence on model estimates. This validates the overall stability of the model fitted.

These diagnostics show that RSM gives a satisfactory but not optimal representation of the data. The normality and independence assumptions are fairly met, with minor violations of the assumptions of linearity and homoscedasticity indicating existence of residual structure that is not represented by the model. These results can be attributed to the fact that the processes of postharvest deterioration are complex, nonlinear and interaction-driven mechanisms influence them. Notably, these

diagnostic outcomes give a good methodological justification to adopting the hybrid RSM-GBM model. Although the RSM component is effective in capturing the interpretable global structure of the data in the experimental design space, the observed residual patterns suggest that there are still additional nonlinear relationships. The gradient boosting machine (GBM) component is thus used to model these residual patterns so that both structured parametric effects and unstructured nonlinear variation are captured in the hybrid model. The hybrid approach is especially well-suited to modelling complex agricultural systems as it provides both predictive accuracy and interpretability.

6.8 Chapter Summary

This chapter applied the proposed hybrid RSM-GBM framework to an integrated measure of potato post-harvest losses represented by the Composite Loss Index (CLI). The correlated indicators of weight loss, sprouting and rotting were first combined using Principal Component Analysis to produce a composite response measure for overall post-harvest deterioration. This chapter then used a second order Response Surface Model to estimate the effect of the experimental factors and to determine the response surface shape, linear, quadratic and interaction. Model diagnostics were performed to check the suitability of the RSM specification. The remaining nonlinear structure of the residuals after the RSM fit was then modelled using Gradient Boosting. These models were then contrasted with each other with the help of held-out test data and predictive performance measures: the standalone RSM, the standalone GBM and the hybrid RSM-GBM. The relationships between the important predictors and the CLI were visualized using the response surface and contour analyses, and the variable-importance analysis was used to select the predictors that had the highest contribution to the residual nonlinear structure. The results demonstrated that the hybrid model outperformed the other models in terms of predictive performance, leveraging the structural interpretability of RSM and the nonlinear learning ability of GBM. The chapter therefore showed the application of the proposed framework to an integrated measure of potato post-harvest losses and the significant factors associated with deterioration during storage.

CHAPTER SEVEN

CONCLUSION AND RECOMMENDATIONS

7.1 Conclusion

Gradient boosting-based response surface methodology (GB-RSM) model yielded better results which was a significant enhancement over the traditional RSM and gradient boosting model. The GB-RSM model was robust in dealing with high-dimensional, nonlinear and data with missing values by integrating the interpretability of RSM, which models linear, quadratic, and interaction effects, and the adaptive learning ability of gradient boosting. Higher-order interactions and nonlinearities that could not be explained by RSM alone could be captured by the model, which made it a more accurate and robust prediction. The diagnostics of the residual showed that the GB-RSM model was stable with low bias and variance, and thus highly predictive. This is a hybrid method that takes advantage of the strengths of both methods and is especially applicable in complex uses where high accuracy, flexibility, and interpretability are needed.

The GB-RSM hybrid model is recommended to be used in agriculture, engineering, environmental sciences, and healthcare, where the nonlinear relationships, high-dimensional data, and missing values are inherent due to its better performance in these areas. The model is particularly useful in optimization problems where it is important to know how variables interact, and where both model interpretability and predictive power are of interest. Further studies should be conducted on the optimization of hyperparameters, including the learning rate, count of trees, and the depth of the tree, to enhance the effectiveness and performance of the model. Lastly, the incorporation of other machine learning methods, including neural networks or deep learning, might be considered to work with even more complex data and enhance the performance of a model in a particular application.

This study establishes that storage type, storage duration, market access conditions, transportation infrastructure, chemical use, and extension training are critical determinants of post-harvest loss along Kenya's potato farm-market value chain. Using a multilevel beta regression (MLBR) framework which is appropriate for bounded, nested, proportional loss data, we quantify these factors and derive clear policy implications. Three integrated policy pillars emerge. First, public and donor

investment should prioritize accessible and affordable storage including dark rooms promoted through farmer groups, and solar-powered cold storage at aggregation points along major transportation corridors. Second, post-harvest extension services require restructuring through an integrated co-delivery model that bundles training on approved sprout suppressants with simultaneous access via licensed agricultural dealers. This would close the current gap between extension coverage (51.3%) and actual adoption (30.7%). Third, upgrading transportation infrastructure and introducing low-cost protective packaging for motorcycle transport is essential to reduce physical damage and distance-related losses.

Due to the cross-sectional design and regional sampling frame, causal inference and national generalizability are constrained. Future research should employ longitudinal or experimental designs, nationally representative geographic coverage, and objective post-harvest loss measurements. Without coordinated action on these three pillars i.e. storage, extension reform, and transport, technological interventions alone are unlikely to reduce post-harvest losses at scale in Kenya's smallholder potato sector.

The findings from this study are interpreted in the framework of a simulation-based analytical approach under which all the responses in the experiment were produced. The relationships between the predictor variables and the response outcomes observed were not derived from empirical potato storage experiments but were biologically informed, based on known post-harvest literature that forms a mathematical model. It is therefore important to note that the reported results are computational simulations of methodological behavior and do not provide conclusive empirical evidence of agricultural processes.

In the simulated storage environment, storage duration proved to be the most significant factor affecting the Composite Loss Index, as it significantly increased the overall loss from moisture loss, sprouting development, and rotting progression in the range of parameters tested. Moreover, the simulated response surfaces of temperature and relative humidity showed strong nonlinear effects, which indicated that these two parameters have interactive effect on post-harvest deterioration dynamics. The results are generally in line with the known physiology of post-harvest, indicating that the simulated deterioration patterns are biologically plausible and that the hybrid

modelling approach was able to capture these deterioration patterns in the simulation environment.

This research showed that temporal, environmental, and biological factors are the cause of post-harvest losses occurring in potatoes with clear differences in their relative importance. storage duration was found to be the most significant environmental predictor of CLI which is the cumulative nature of deterioration processes over time. The overall size of the effects indicated that variety had the strongest overall effect with respect to defining storage performance. Relative humidity, temperature and light exposure were also found to have significant and practically meaningful effects and mechanical damage contributed to losses on a lesser scale. The duration of curing, in the range of the study, did not have a significant effect on CLI.

The results also affirmed that no single modelling approach can be used to sufficiently capture the complexity of post-harvest systems. The hybrid RSM-GBM framework was considered a strong solution, as it was based on the interpretability of parametric models and the flexibility of machine learning. RSM was able to capture the structured relationships in the experimental design space, whereas GBM was able to capture nonlinear and interaction-driven residual patterns, resulting in improved predictive performance. In general, the research indicates that to achieve effective post-harvest loss reduction, a systems-based approach is necessary, which involves storage time management, environment control, and varietal selection. The findings present both conceptual and practical contributions, providing a more detailed framework of the analysis and management of post-harvest losses in agricultural systems.

The results of this study indicate that an integrated and coordinated approach is required to reduce losses incurred after harvest of the potato crop by addressing both the operational practices and structural constraints in the potato value chain. Minimizing the storage duration should be given priority as it is the most powerful determinant of loss accumulation. This can be done by increasing market accessibility, enhanced coordination of supply chains and increased processing capacity to minimize time lag between harvest and sale. In this regard, the emergence of structured markets, aggregation systems, and the contract farming arrangements is of

particular importance to the small-scale farmers, as it can substantially reduce the storage-to-market interval.

7.2 Recommendations

7.2.1 Recommendations for Policy

The use of advanced hybrid statistical modelling techniques for the analysis of complex nonlinear systems, where accurate prediction and optimization are crucial, should be encouraged by policymakers and research funding agencies. Research efforts that involve both classical statistical and machine learning methods and techniques should be encouraged for improved evidence-based decision-making efforts, with a particular focus on interdisciplinary research. Moreover, national research institutes should help to develop statistical modelling frameworks that are more predictive but also remain interpretable and thus aid policy making in agriculture, health, environmental management, manufacturing and other data-intensive sectors.

Hybrid statistical-machine learning based models like the proposed framework of GB-RSM are recommended for modelling complex nonlinear processes in industries and organizations where predictive analytics and process optimization are crucial. By combining Response Surface Methodology with gradient boosting, there is a manageable trade-off between the accuracy of predictions, optimization, and interpretability, supporting organizations in making sound operational decisions. It can be used in the agricultural sector to improve storage conditions, minimize post-harvest losses, in the manufacturing sector to optimize manufacturing processes, allocate resources, and minimize risk, and in the quality control sector, to minimize risk and optimize manufacturing processes.

County governments of Nakuru and Nyandarua, in liaison with the national Ministry of Agriculture, Livestock and Fisheries, should focus on the investment of accessible potato storage infrastructure, especially encouraging the development of dark room storage infrastructure under community storage center initiatives, where access to electricity allows, cold storage facilities powered by solar energy in the aggregation points along major marketing routes since storage method was found to be the one most effective determinant of proportionate losses. At the same time, Kenya Agricultural and Livestock Research Organization (KALRO) and county extension

departments would need to redesign post-harvest extension programs so that they integrate knowledge transfer with input supply into single co-delivery model, so that in a single model, training on approved sprout suppressant application is accompanied by facilitated access to chemicals through licensed agro-dealer networks as the current adoption gap between extension training coverage (51.3%) and chemical use (30.7%) indicates that knowledge alone cannot be used to cause changes in behavior. Individual farmers and traders are to be advised to switch to dark room storage, to adopt immediate post-harvest curing, and to utilize better packaging material to be carried on motorcycles to minimize mechanical damage during transit, and to combine consignments with other actors where roads allow, to divide the truck transport expenses. Lastly, longitudinal and experimental designs, geographic expansion into a nationally representative sample of Kenyan counties, objective measures of loss, and retesting the multilevel Beta regression model in other perishable crop value chains in sub-Saharan Africa should be utilized in future studies to enhance the evidence base on post-harvest loss reduction policy.

The overall findings lead to a policy agenda of reducing the losses in the Kenyan potato value chain after harvest in a coherent and evidence-based manner. Storage method is the leading cause of losses, and thus, the most important intervention is investment in affordable and reachable storage infrastructure. In the short-term, extension programs with specific subsidies on construction materials would be promoted to build dark rooms for most farmers and traders that cannot afford cold chain infrastructure. Public investment in large-scale community storage facilities using solar powered refrigeration in large potato growing wards in Nakuru and Nyandarua in the medium term would enable cold storage on a large scale without the individual household bearing the entire cost of capital. There is an infrastructure investment implication in the transport and market access results, namely that they indicate that infrastructure investment is needed to support actor-level interventions: the amount of mechanical damage caused by rough roads and the distance-related indirect impact of long storage would be reduced. Short-term solutions might include switching to more robust wrapping materials to protect motorcycle-based transportation and managing smallholder shipments so that they can be pooled to allow the use of trucks economically.

7.2.2 Recommendations for Future Work

To establish the generalizability and robustness of the proposed framework of GB-RSM, further research is needed to test it in other application areas and with a broader set of nonlinear modelling problems. Further insight into the relative strengths of hybrid modelling approaches would be gained through comparative studies with other advanced machine learning algorithms, such as Extreme gradient boosting (XGBoost), LightGBM, CatBoost, Random Forests, Gaussian Process Regression, and deep learning models. Further research might explore alternative approaches to combining statistical response surface models with machine learning algorithms, derive theoretical properties of the hybrid estimator such as consistency and asymptotic behavior, and generalize the approach to high-dimensional data, mixed response variables, longitudinal data, and adaptive design. Such studies would help to develop hybrid statistical modelling techniques for complex nonlinear systems.

Future research should extend the proposed GB-RSM framework by incorporating other machine-learning algorithms, such as Random Forest, XGBoost, neural networks, and deep learning, and by optimizing key Gradient Boosting hyperparameters to improve predictive performance and computational efficiency. The models should also be expanded to include objective post-harvest losses, seasonal and climatic factors, and other supply-chain variables in the future to enhance the accuracy and comprehensiveness of the models. Finally, the framework should be tested with high dimensional data, different adaptive experimental designs, and other applications than potato post-harvest losses to ensure its wider applicability to complex nonlinear systems.

REFERENCES

- Abonazel, M. R., Algamal, Z. Y., Awwad, F. A., & Taha, I. M. (2022). A New Two-Parameter Estimator for Beta Regression Model: Method, Simulation, and Application. *Frontiers in Applied Mathematics and Statistics*, 7. <https://doi.org/10.3389/fams.2021.780322>
- Agarwal, M., Gupta, S. K., & Biswas, K. K. (2020). Development of Efficient CNN model for Tomato crop disease identification. *Sustainable Computing: Informatics and Systems*, 28, 100407. <https://doi.org/10.1016/J.SUSCOM.2020.100407>
- Ahmad, A., Yadav, A. K., Singh, A., & Singh, D. K. (2024a). A comprehensive machine learning-coupled response surface methodology approach for predictive modeling and optimization of biogas potential in anaerobic Co-digestion of organic waste. *Biomass and Bioenergy*, 180. <https://doi.org/10.1016/j.biombioe.2023.106995>
- Ahmad, A., Yadav, A. K., Singh, A., & Singh, D. K. (2024b). A comprehensive machine learning-coupled response surface methodology approach for predictive modeling and optimization of biogas potential in anaerobic Co-digestion of organic waste. *Biomass and Bioenergy*, 180, 106995. <https://doi.org/10.1016/J.BIOMBIOE.2023.106995>
- Akhtar Gul, Muhammad Ghulam Shabeer, Rija Ahmad Abbasi, & Abdul Wahab Khan. (2022). Africa's Poverty and Famines: Developmental Projects of China on Africa. *PERENNIAL JOURNAL OF HISTORY*, 3(1), 165–194. <https://doi.org/10.52700/pjh.v3i1.109>
- Alamar, M. C., Tosetti, R., Landahl, S., Bermejo, A., & Terry, L. A. (2017a). Assuring potato tuber quality during storage: A future perspective. In *Frontiers in Plant Science* (Vol. 8). Frontiers Media S.A. <https://doi.org/10.3389/fpls.2017.02034>
- Alamar, M. C., Tosetti, R., Landahl, S., Bermejo, A., & Terry, L. A. (2017b). Assuring potato tuber quality during storage: A future perspective. In *Frontiers in Plant Science* (Vol. 8). Frontiers Media S.A. <https://doi.org/10.3389/fpls.2017.02034>
- Alhassan, A.-K., Danso-Abbeam, G., Boateng, V. F., Setsoafia, E. D., & Mabe, F. N. (2025). Sustainable Agricultural Practices in the Savannah Region of Ghana: Factors shaping Smallholder Adoption. *Ghana Journal of Science, Technology and Development*, 11(1), 40–45. <https://doi.org/10.47881/464.967x>
- Ali, A., Jayaraman, R., Azar, E., & Maalouf, M. (2024). A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework. *Building and Environment*, 252, 111268. <https://doi.org/10.1016/J.BUILDENV.2024.111268>
- Ali, A., Xia, C., Ouattara, N. B., Mahmood, I., & Faisal, M. (2021). Economic and environmental consequences' of postharvest loss across food supply Chain in the developing countries. *Journal of Cleaner Production*, 323, 129146. <https://doi.org/10.1016/J.JCLEPRO.2021.129146>
- Almasi, A., Eshraghian, M. R., Moghimbeigi, A., Rahimi, A., Mohammad, K., & Fallahigilan, S. (2016). Multilevel zero-inflated Generalized Poisson regression

- modeling for dispersed correlated count data. *Statistical Methodology*, 30, 1–14. <https://doi.org/10.1016/j.stamet.2015.11.001>
- Alzu'bi, A., Amira, A., & Ramzan, N. (2015). Semantic content-based image retrieval: A comprehensive study. *Journal of Visual Communication and Image Representation*, 32, 20–54. <https://doi.org/10.1016/J.JVCIR.2015.07.012>
- Andati, P., Majiwa, E., Ngigi, M., Mbeche, R., & Ateka, J. (2023). Effect of climate smart agriculture technologies on crop yields: Evidence from potato production in Kenya. *Climate Risk Management*, 41. <https://doi.org/10.1016/j.crm.2023.100539>
- Angeli, V., Silva, P. M., Massuela, D. C., Khan, M. W., Hamar, A., Khajehei, F., Graeff-Hönniger, S., & Piatti, C. (2020). Quinoa (*Chenopodium quinoa* Willd.): An overview of the potentials of the “golden grain” and socio-economic and environmental aspects of its cultivation and marketization. *Foods*, 9(2). <https://doi.org/10.3390/foods9020216>
- Anteneh, A., & Endalew, B. (2023). Determinants of teff commercialization among smallholder farmers: Beta regression approach. *Cogent Social Sciences*, 9(1). <https://doi.org/10.1080/23311886.2023.2209987>
- Ariong, R. M., Okello, D. M., Otim, M. H., & Paparu, P. (2023a). The cost of inadequate postharvest management of pulse grain: Farmer losses due to handling and storage practices in Uganda. *Agriculture and Food Security*, 12(1). <https://doi.org/10.1186/s40066-023-00423-7>
- Ariong, R. M., Okello, D. M., Otim, M. H., & Paparu, P. (2023b). The cost of inadequate postharvest management of pulse grain: Farmer losses due to handling and storage practices in Uganda. *Agriculture and Food Security*, 12(1). <https://doi.org/10.1186/s40066-023-00423-7>
- Asmare, A. A., Tegegne, A. S., Belay, D. B., & Agmas, Y. A. (2025). Coexisting predictors for undernutrition indices among under-five children in West Africa: application of a multilevel multivariate ordinal logistic regression model. *BMC Nutrition*, 11(1). <https://doi.org/10.1186/s40795-025-01099-x>
- Ayieko, M., Odame, H., & Olwande, J. (n.d.). *WPS 71/2021 Strengthening Seed Systems and Market Development in Kenya: Perspectives on Political Economy and Policy Processes*. Retrieved <http://www.tegemeo.org>
- Bacanin, N., Zivkovic, M., Stoean, C., Antonijevic, M., Janicijevic, S., Sarac, M., & Strumberger, I. (2022). Application of Natural Language Processing and Machine Learning Boosted with Swarm Intelligence for Spam Email Filtering. *Mathematics*, 10(22). <https://doi.org/10.3390/math10224173>
- Bahrami, S., Rouhi Rad, M., & Nayga, R. M. (2025). Dollars for Drops: Abatement Cost of Water for Irrigation in the Colorado River Basin. *Applied Economic Perspectives and Policy*. <https://doi.org/10.1002/aepp.70026>
- Bassey, N. E. (2018). *Post-harvest losses in Sub-Saharan Africa measurement concepts and economic assessment of reduction strategies*.
- Bayuo, J., Abukari, M. A., & Pelig-Ba, K. B. (2020). Optimization using central composite design (CCD) of response surface methodology (RSM) for biosorption of hexavalent chromium from aqueous media. *Applied Water Science*, 10(6). <https://doi.org/10.1007/s13201-020-01213-3>

- Bedini, G., Haff, R. P., Benelli, A., Bandiera, A., Taormina, E., Massantini, R., & Moschetti, R. (2024). Potatoes (*Solanum tuberosum* L.) grown at “Patata dell’alto Viterbese” PGI have different quality characteristics and storage responses. *Postharvest Biology and Technology*, 214. <https://doi.org/10.1016/j.postharvbio.2024.112991>
- Bendinelli, W. E., Su, C. T., Péra, T. G., & Caixeta Filho, J. V. (2020). What are the main factors that determine post-harvest losses of grains? *Sustainable Production and Consumption*, 21, 228–238. <https://doi.org/10.1016/J.SPC.2019.09.002>
- Bezerra, M. A., Santelli, R. E., Oliveira, E. P., Villar, L. S., & Escalera, L. A. (2008). Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta*, 76(5), 965–977. <https://doi.org/10.1016/J.TALANTA.2008.05.019>
- Bifarin, O. O. (2023). Interpretable machine learning with treebased shapley additive explanations: Application to metabolomics datasets for binary classification. *PLoS ONE*, 18(5 May). <https://doi.org/10.1371/journal.pone.0284315>
- Bisht, A., & Singh, S. P. (2024). Postharvest Losses and Management of Horticultural Produce: A Review. *Journal of Scientific Research and Reports*, 30(3), 305–320. <https://doi.org/10.9734/jsrr/2024/v30i31881>
- Bock, F. E., Keller, S., Huber, N., & Klusemann, B. (2021). Hybrid modelling by machine learning corrections of analytical model predictions towards high-fidelity simulation solutions. *Materials*, 14(8). <https://doi.org/10.3390/ma14081883>
- Botchey, F. E., Qin, Z., & Hughes-Lartey, K. (2020). Mobile money fraud prediction-A cross-case analysis on the efficiency of support vector machines, gradient boosted decision trees, and Naïve Bayes algorithms. *Information (Switzerland)*, 11(8). <https://doi.org/10.3390/INFO11080383>
- Busungu, C. (2025). Invasive Insect Species Devastating Crop Production in Tanzania: Implications on Management and Agriculture Resilience. In *Journal of Agricultural Sciences - Sri Lanka* (Vol. 20, Number 3, pp. 501–525). Faculty of Agricultural Sciences, Sabaragamuwa University of Sri Lanka. <https://doi.org/10.4038/jas.v20i3.11077>
- Chen, V. Y. J., & Yang, Y. C. (2024). Geographically weighted regression analysis for nonnegative continuous outcomes: An application to Taiwan dengue data. *PLoS ONE*, 19(12 December). <https://doi.org/10.1371/journal.pone.0315327>
- Cherono, K. (2016). Development and Optimization of an Integrated Postharvest Management System For Storage And Handling of Fresh Market Tomatoes in South African Supply Chains.
- Chi, Y. (2022). *Managing Credit Risk and the Cost of Equity with Machine Learning Techniques*.
- Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, 34(2), 1–24. <https://doi.org/10.18637/jss.v034.i02>
- Dave, R. K., Ramana Rao, T. V., & Nandane, A. S. (2017a). Improvement of post-harvest quality of pear fruit with optimized composite edible coating formulations. *Journal of Food Science and Technology*, 54(12), 3917–3927. <https://doi.org/10.1007/s13197-017-2850-y>

- Dave, R. K., Ramana Rao, T. V., & Nandane, A. S. (2017b). Improvement of post-harvest quality of pear fruit with optimized composite edible coating formulations. *Journal of Food Science and Technology*, 54(12), 3917–3927. <https://doi.org/10.1007/s13197-017-2850-y>
- Degebasa, A. C. (2020). Prospects and Challenges of Postharvest Losses of Potato (*Solanum Tuberosum* L.) in Ethiopia. *Global Journal of Nutrition & Food Science*, 2(5). <https://doi.org/10.33552/gjnfs.2020.02.000550>
- Devaux, A., Goffart, J. P., Kromann, P., Andrade-Piedra, J., Polar, V., & Hareau, G. (2021). The Potato of the Future: Opportunities and Challenges in Sustainable Agri-food Systems. In *Potato Research* (Vol. 64, Number 4, pp. 681–720). Springer Science and Business Media B.V. <https://doi.org/10.1007/s11540-021-09501-4>
- Devine, S., Uanhoro, J. O., Otto, A. R., & Flake, J. K. (2024). Approaches for Quantifying the ICC in Multilevel Logistic Models: A Didactic Demonstration. *Collabra: Psychology*, 10(1), 3175–3199. <https://doi.org/10.1525/collabra.94263>
- Domingues, T., Brandão, T., & Ferreira, J. C. (2022a). Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. In *Agriculture (Switzerland)* (Vol. 12, Number 9). MDPI. <https://doi.org/10.3390/agriculture12091350>
- Domingues, T., Brandão, T., & Ferreira, J. C. (2022b). Machine Learning for Detection and Prediction of Crop Diseases and Pests: A Comprehensive Survey. In *Agriculture (Switzerland)* (Vol. 12, Number 9). MDPI. <https://doi.org/10.3390/agriculture12091350>
- Dong, N., Kelcey, B., Spybrook, J., Nickodem, K., & Sui, N. (2026). Statistical Power for Moderation in Three-Level Multisite Individual Randomized Trials and Consequences of Ignoring a Level of Nesting. *American Journal of Evaluation*. <https://doi.org/10.1177/10982140251394304>
- Effrosynidis, D., Spiliotis, E., Sylaios, G., & Arampatzis, A. (2023). Time series and regression methods for univariate environmental forecasting: An empirical evaluation. *Science of The Total Environment*, 875, 162580. <https://doi.org/10.1016/J.SCITOTENV.2023.162580>
- Endalew, B., Aynalem, M., Anteneh, A., & Mossie, H. (2023). Sources of wheat production technical inefficiency among smallholder farmers in Northwestern Ethiopia: Beta regression approach. *Cogent Economics and Finance*, 11(1). <https://doi.org/10.1080/23322039.2023.2208895>
- Fakher, H. A., Panahi, M., Emami, K., Peykarjou, K., & Zeraatkish, S. Y. (2021). New Insight into Development of Environmental-Economic Model Based on a Composite Environmental Quality Index: An Application of Principal Components Analysis. *Tasmingiri va Tahqiq Dar Amaliyyat*, 6(2), 183–209. <https://doi.org/10.22105/dmor.2021.276820.1334>
- Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31(7), 799–815. <https://doi.org/10.1080/0266476042000214501>
- Figueroa-Zuñiga, J. I., Arellano-Valle, R. B., & Ferrari, S. L. P. (2012). *Mixed Beta Regression: A Bayesian Perspective*. <http://arxiv.org/abs/1201.2375>

- Firdous, N. (2020). Post-harvest losses in different fresh produces and vegetables in Pakistan with particular focus on tomatoes. *JOURNAL OF HORTICULTURE AND POSTHARVEST RESEARCH*, 2021(1), 71–86. <https://doi.org/10.22077/jhpr.2020.3168.1125>
- Fokkema, M., & Zeileis, A. (2024). Subgroup detection in linear growth curve models with generalized linear mixed model (GLMM) trees. *Behavior Research Methods*, 56(7), 6759–6780. <https://doi.org/10.3758/s13428-024-02389-1>
- Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. In *Source: The Annals of Statistics* (Vol. 29, Number 5).
- Gad, M. S., Soliman, M. S., & Helal, E. B. (2025). Hybrid XGBoost-RF-MLP model and PSO optimization for performance and emissions of CI engine using waste cooking biodiesel blends. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-29269-8>
- Gálvez, L., & Palmero, D. (2022). Fusarium Dry Rot of Garlic Bulbs Caused by *Fusarium proliferatum*: A Review. In *Horticulturae* (Vol. 8, Number 7). MDPI. <https://doi.org/10.3390/horticulturae8070628>
- Geissinger, E. A., Khoo, C. L. L., Richmond, I. C., Faulkner, S. J. M., & Schneider, D. C. (2022a). A case for beta regression in the natural sciences. *Ecosphere*, 13(2). <https://doi.org/10.1002/ecs2.3940>
- Geissinger, E. A., Khoo, C. L. L., Richmond, I. C., Faulkner, S. J. M., & Schneider, D. C. (2022b). A case for beta regression in the natural sciences. *Ecosphere*, 13(2). <https://doi.org/10.1002/ecs2.3940>
- Gikundi, E. N., Buzera, A. K., Orina, I. N., & Sila, D. N. (2023). Storability of Irish Potato (*Solanum tuberosum* L.) Varieties Grown in Kenya, Under Different Storage Conditions. *Potato Research*, 66(1), 137–158. <https://doi.org/10.1007/s11540-022-09575-8>
- Githumbi, R. O. S. E., Mburu, J. O. H. N., Ogutu, A., & Ambuko, J. A. N. E. (2024). Assessment of factors influencing adoption of postharvest loss reduction practices along the mango value chain in Embu, Machakos, and Nairobi counties, Kenya. *Heliyon*, 10(4). <https://doi.org/10.1016/j.heliyon.2024.e25146>
- Gök, E. C., & Olgun, M. O. (2021). SMOTE-NC and gradient boosting imputation based random forest classifier for predicting severity level of covid-19 patients with blood samples. *Neural Computing and Applications*, 33(22), 15693–15707. <https://doi.org/10.1007/s00521-021-06189-y>
- Grudzińska, M., Boguszevska-Mańkowska, D., & Zarzyńska, K. (2022). Drought stress during the growing season: Changes in reducing sugars, starch content and respiration rate during storage of two potato cultivars differing in drought sensitivity. *Journal of Agronomy and Crop Science*, 208(5), 609–620. <https://doi.org/10.1111/jac.12498>
- Gunathilake, D. M. C. C., Senanayaka, D. P., Adiletta, G., & Senadeera, W. (n.d.). 4 *Drying of Agricultural Crops*.
- Hamed Taherdoost, A., & Lumpur, K. (2016). Validity and Reliability of the Research Instrument; How to Test the Validation of a Questionnaire/Survey in a Research. In *International Journal of Academic Research in Management (IJARM)* (Vol. 5, Number 3). <https://hal.science/hal-02546799v1>

- Hameed, A., Zaidi, S. S. e. A., Shakir, S., & Mansoor, S. (2018). Applications of new breeding technologies for potato improvement. In *Frontiers in Plant Science* (Vol. 9). Frontiers Media S.A. <https://doi.org/10.3389/fpls.2018.00925>
- Harbecke, J., Grunau, J., & Samanek, P. (2024). Are the Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) Applicable in Determining the Optimal Fit and Simplicity of Mechanistic Models? *International Studies in the Philosophy of Science*, 37(1–2), 17–36. <https://doi.org/10.1080/02698595.2024.2304487>
- He, Y., Jiang, T., Wen, J., & Xu, G. (2021). Likelihood Ratio Test in Multivariate Linear Regression. *31*(3), 1215–1238. <https://doi.org/10.2307/27034815>
- Hoffman, L., & Walters, R. W. (2026). *Annual Review of Psychology Catching Up on Multilevel Modeling Cross-level interaction: a model-estimated constant slope for the multiplicative combination of two predictor variables measured for different dimensions of sampling*. 34, 29. <https://doi.org/10.1146/annurev-psych-020821>
- Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *American Statistician*, 55(3), 244–254. <https://doi.org/10.1198/000313001317098266>
- Hou, N., Li, M., He, L., Xie, B., Wang, L., Zhang, R., Yu, Y., Sun, X., Pan, Z., & Wang, K. (2020). Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *Journal of Translational Medicine*, 18(1). <https://doi.org/10.1186/s12967-020-02620-5>
- Huber, F., Yushchenko, A., Stratmann, B., & Steinhage, V. (2022). Extreme Gradient Boosting for yield estimation compared with Deep Learning approaches. *Computers and Electronics in Agriculture*, 202, 107346. <https://doi.org/10.1016/J.COMPAG.2022.107346>
- Islam, M. A., Majumder, M. Z. H., & Hussein, M. A. (2023). Chronic kidney disease prediction based on machine learning algorithms. *Journal of Pathology Informatics*, 14, 100189. <https://doi.org/10.1016/J.JPI.2023.100189>
- Jadhav, A., Pramod, D., & Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intelligence*, 33(10), 913–933. <https://doi.org/10.1080/08839514.2019.1637138>
- Jahani, Y., Baena, M., Barris, C., Perera, R., & Torres, L. (2022). Influence of curing, post-curing and testing temperatures on mechanical properties of a structural adhesive. *Construction and Building Materials*, 324, 126698. <https://doi.org/10.1016/J.CONBUILDMAT.2022.126698>
- Jakubowski, T., & Królczyk, J. B. (2020). Method for the reduction of natural losses of potato tubers during their long-term storage. *Sustainability (Switzerland)*, 12(3). <https://doi.org/10.3390/su12031048>
- Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2), 105–115. <https://doi.org/10.1016/J.ARTMED.2010.05.002>
- Ju, K., Lin, L., Chu, H., Cheng, L. L., & Xu, C. (2020). Laplace approximation, penalized quasi-likelihood, and adaptive Gauss-Hermite quadrature for generalized linear

- mixed models: Towards meta-analysis of binary outcome with sparse data. *BMC Medical Research Methodology*, 20(1). <https://doi.org/10.1186/s12874-020-01035-6>
- Kaguongo, W. (n.d.). *Post-harvest losses in potato value chains in Kenya: Analysis and recommendations for reduction strategies*. *Post-harvest losses in potato value chains in Kenya Analysis and recommendations for reduction strategies*. <https://doi.org/10.13140/2.1.3761.3764>
- Karatas, G., Demir, O., & Sahingoz, O. K. (2020). Increasing the Performance of Machine Learning-Based IDSs on an Imbalanced and Up-to-Date Dataset. *IEEE Access*, 8, 32150–32162. <https://doi.org/10.1109/ACCESS.2020.2973219>
- Karimi, E., & Niknami, M. (2020). Analyzing Impacts of Farmer Field School on the Economic, Social, Production, and Knowledge Status of Greenhouse Owners: Evidence from Tehran Province and its Surrounding Counties. In *J. Agr. Sci. Tech* (Vol. 22, Number 1).
- Kasali, J., & Adeyemi, A. A. (2022). Model-Data Fit using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and The Sample-Size-Adjusted BIC. *Square : Journal of Mathematics and Mathematics Education*, 4(1), 43–51. <https://doi.org/10.21580/square.2022.4.1.11297>
- Khamis, G. S. M., Mohammed, Z. M. S., Alanazi, S. M., Mahmoud, A. F. A., Abdalla, F. A., & Bkheet, S. A. (2024). Prediction of Myocardial Infarction Complications using Gradient Boosting. *Engineering, Technology and Applied Science Research*, 14(6), 18550–18556. <https://doi.org/10.48084/etasr.9076>
- Khan, Md. K., & Hasan, Md. T. (2022). A Poisson Regression Approach to Modeling Traffic Accident Frequency In Urban Areas. *American Journal of Interdisciplinary Studies*, 03(04), 117–156. <https://doi.org/10.63125/wqh7pd07>
- Khanal, S., Karimi, K., Majumdar, S., Kumar, V., Verma, R., Bhatia, S. K., Kuca, K., Esteban, J., & Kumar, D. (2024). Sustainable utilization and valorization of potato waste: state of the art, challenges, and perspectives. In *Biomass Conversion and Biorefinery* (Vol. 14, Number 19, pp. 23335–23360). Springer Science and Business Media Deutschland GmbH. <https://doi.org/10.1007/s13399-023-04521-1>
- Khorramifar, A., Rasekh, M., Karami, H., Lozano, J., Gancarz, M., Łazuka, E., & Łagód, G. (2023a). Determining the shelf life and quality changes of potatoes (*Solanum tuberosum*) during storage using electronic nose and machine learning. *PLoS ONE*, 18(4 April). <https://doi.org/10.1371/journal.pone.0284612>
- Khorramifar, A., Rasekh, M., Karami, H., Lozano, J., Gancarz, M., Łazuka, E., & Łagód, G. (2023b). Determining the shelf life and quality changes of potatoes (*Solanum tuberosum*) during storage using electronic nose and machine learning. *PLoS ONE*, 18(4 April). <https://doi.org/10.1371/journal.pone.0284612>
- Komen, S. K., Ngeny, J. M., & Osen, E. (2017). Bridging the potato variety gap: A review of the Kenya - The Netherlands potato project (2012-2015): Its success, challenges and opportunities. *Open Agriculture*, 2(1), 267–273. <https://doi.org/10.1515/opag-2017-0030>
- Levaj, B., Pelaić, Z., Galić, K., Kurek, M., Ščetar, M., Poljak, M., Dite Hunjek, D., Pedisić, S., Balbino, S., Čošić, Z., Dujmić, F., & Repajić, M. (2023). Maintaining

- the Quality and Safety of Fresh-Cut Potatoes (*Solanum tuberosum*): Overview of Recent Findings and Approaches. In *Agronomy* (Vol. 13, Number 8). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/agronomy13082002>
- Li, J., Wu, Z., & Chen, J. (2022). An advanced Bayesian parameter estimation methodology for concrete dams combining an improved extraction technique of hydrostatic component and hybrid response surface method. *Engineering Structures*, 267, 114687. <https://doi.org/10.1016/J.ENGSTRUCT.2022.114687>
- Lu, J., Tan, L., & Jiang, H. (2021). Review on convolutional neural network (CNN) applied to plant leaf disease classification. In *Agriculture (Switzerland)* (Vol. 11, Number 8). MDPI AG. <https://doi.org/10.3390/agriculture11080707>
- Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., & Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, 121, 103761. <https://doi.org/10.1016/J.COMPBIOMED.2020.103761>
- Ma, Y., Sanchez, V., & Guha, T. (2025a). ZIP: Scalable Crowd Counting via Zero-Inflated Poisson Modeling. <http://arxiv.org/abs/2506.19955>
- Ma, Y., Sanchez, V., & Guha, T. (2025b). ZIP: Scalable Crowd Counting via Zero-Inflated Poisson Modeling. <http://arxiv.org/abs/2506.19955>
- Macharia, I., & Mutungi, C. (2015). *Postharvest losses in Africa-Analytical review and synthesis: the case of Kenya*. <https://www.researchgate.net/publication/285594327>
- Mahesh, P., & Soundrapandiyan, R. (2024). Yield prediction for crops by gradient-based algorithms. *PLoS ONE*, 19(8). <https://doi.org/10.1371/journal.pone.0291928>
- Malakouti, S. M., Esfandiari, A., Menhaj, M. B., Suratgar, A. A., & Doustmohammadi, A. (2024). WITHDRAWN: Enhancing Cancer Detection Through DNA Sequences and Ensemble Method. *Intelligent Systems with Applications*, 200371. <https://doi.org/10.1016/J.ISWA.2024.200371>
- Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: an investigation using PISA sampling structures. *Large-Scale Assessments in Education*, 9(1). <https://doi.org/10.1186/s40536-021-00099-0>
- Manoharan, A., Begam, K. M., Aparow, V. R., & Sooriamoorthy, D. (2022). Artificial Neural Networks, Gradient Boosting and Support Vector Machines for electric vehicle battery state estimation: A review. *Journal of Energy Storage*, 55, 105384. <https://doi.org/10.1016/J.EST.2022.105384>
- Marina Pare, B. A. (2025). Farmers' groups and the intensity of cereal-based agroforestry practice in Burkina Faso. *Forests Trees and Livelihoods*, 34(4), 301–314. <https://doi.org/10.1080/14728028.2025.2526344>
- Martín, I., Gálvez, L., Guasch, L., & Palmero, D. (2022). Fungal Pathogens and Seed Storage in the Dry State. In *Plants* (Vol. 11, Number 22). MDPI. <https://doi.org/10.3390/plants11223167>
- Mburu, J. M., Oyugi, M. A., & Makenzi, P. M. (2023a). Influence of Access to Extension Services on Smallholder Farmers' Use of Certified Seed Potatoes (CSPs) in

- Kipipiri Sub-County, Kenya. *Asian Journal of Agricultural Extension, Economics & Sociology*, 41(4), 169–177. <https://doi.org/10.9734/ajaees/2023/v41i41891>
- Mburu, J. M., Oyugi, M. A., & Makenzi, P. M. (2023b). Influence of Access to Extension Services on Smallholder Farmers' Use of Certified Seed Potatoes (CSPs) in Kipipiri Sub-County, Kenya. *Asian Journal of Agricultural Extension, Economics & Sociology*, 41(4), 169–177. <https://doi.org/10.9734/ajaees/2023/v41i41891>
- McDonald, J. (2009). Using least squares and tobit in second stage DEA efficiency analyses. *European Journal of Operational Research*, 197(2), 792–798. <https://doi.org/10.1016/j.ejor.2008.07.039>
- Misango, V. G., Nzuma, J. M., Irungu, P., & Kassie, M. (2022). Intensity of adoption of integrated pest management practices in Rwanda: A fractional logit approach. *Heliyon*, 8(1). <https://doi.org/10.1016/j.heliyon.2022.e08735>
- Misra, N. N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., & Martynenko, A. (2022). IoT, Big Data, and Artificial Intelligence in Agriculture and Food Industry. *IEEE Internet of Things Journal*, 9(9), 6305–6324. <https://doi.org/10.1109/JIOT.2020.2998584>
- Mohd Ali, M., Hashim, N., Abd Aziz, S., & Lasekan, O. (2022). Characterisation of Pineapple Cultivars under Different Storage Conditions Using Infrared Thermal Imaging Coupled with Machine Learning Algorithms. *Agriculture (Switzerland)*, 12(7). <https://doi.org/10.3390/agriculture12071013>
- Mondal, D., Vanbelle, S., Cassese, A., & Candel, M. J. J. M. (2024). Review of sample size determination methods for the intraclass correlation coefficient in the one-way analysis of variance model. In *Statistical Methods in Medical Research* (Vol. 33, Number 3, pp. 532–553). SAGE Publications Ltd. <https://doi.org/10.1177/09622802231224657>
- Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Multivariate Statistical Machine Learning Methods for Genomic Prediction. In *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer International Publishing. <https://doi.org/10.1007/978-3-030-89010-0>
- Morris, D. S., & Sellers, K. F. (2022). A Flexible Mixed Model for Clustered Count Data. *Stats*, 5(1), 52–69. <https://doi.org/10.3390/stats5010004>
- Mugo, J. N., Karanja, N. N., Gachene, C. K., Dittert, K., Gitari, H. I., & Schulte-Geldermann, E. (2021). Response of potato crop to selected nutrients in central and eastern highlands of Kenya. *Cogent Food and Agriculture*, 7(1). <https://doi.org/10.1080/23311932.2021.1898762>
- Musita, C. N., Okoth, M. W., & Abong, G. O. (2019). Postharvest handling practices and perception of potato safety among potato traders in Nairobi, Kenya. *International Journal of Food Science*, 2019. <https://doi.org/10.1155/2019/2342619>
- Mwakidoshi, E. R. (2021). Economic Importance, Ecological Requirements and Production Constraints of Potato (*Solanum tuberosum* L.) in Kenya. *International Journal of Bioresource Science*, 8(2). <https://doi.org/10.30954/2347-9655.02.2021.1>

- Ndaka, D., Macharia, I., Mutungi, C., & Affognon, H. (2012). *Postharvest losses in Africa-Analytical review and synthesis: the case of Kenya*.
- Nemati, M., Ansary, J., & Nemati, N. (2020). Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data. *Patterns*, 1(5). <https://doi.org/10.1016/j.patter.2020.100074>
- Njogu, H. W. (2022). Effects of droughts on the delivery of infrastructure services in Kenya. *Natural Resources Forum*, 46(2), 221–244. <https://doi.org/10.1111/1477-8947.12251>
- Nohamba, S., Musara, J. P., Bahta, Y. T., & Ogundeji, A. A. (2022). Drivers of Postharvest Loss among Citrus Farmers in Eastern Cape Province of South Africa: A Zero-Inflated Poisson (ZIP) Regression Model Analysis. *Agriculture (Switzerland)*, 12(10). <https://doi.org/10.3390/agriculture12101651>
- Nugent, J. R., & Kleinman, K. P. (2021). Type I error control for cluster randomized trials under varying small sample structures. *BMC Medical Research Methodology*, 21(1). <https://doi.org/10.1186/s12874-021-01236-7>
- Ogheneruemu, O.-E., & Ifoga, J. O. (2023). Nigerian Agricultural Journal Effect of Post-Harvest Losses on Food Security among Yam Farmers in Nigeria. In *Nigerian Agricultural Journal* (Vol. 54, Number 2). Retrieved <http://www.ajol.info/index.php/najhttps://www.naj.asn.org.ng>
- Olakiumide, O. (2021). *Post-Harvest Loss Reduction: Enhancing Food Security and Economic Sustainability*. <https://hdpublication.com/index.php/jss>
- Omran, H., Valipour, M., & Jafari Mamakani, S. (2019). Construct a composite indicator based on integrating Common Weight Data Envelopment Analysis and principal component analysis models: An application for finding development degree of provinces in Iran. *Socio-Economic Planning Sciences*, 68, 100618. <https://doi.org/10.1016/J.SEPS.2018.02.005>
- Onditi, J., Nyongesa, M., & van der Vlugt, R. (2022). Prevalence, distribution and control of potato virus Y (PVY) strains in Kenyan potato cultivars. *Tropical Plant Pathology*, 47(5), 659–671. <https://doi.org/10.1007/s40858-022-00520-1>
- Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, 208, 109244. <https://doi.org/10.1016/J.PETROL.2021.109244>
- Petrescu, D. C., Vermeir, I., & Petrescu-Mag, R. M. (2020). Consumer understanding of food quality, healthiness, and environmental impact: A cross-national perspective. *International Journal of Environmental Research and Public Health*, 17(1). <https://doi.org/10.3390/ijerph17010169>
- Pineda-Vélez, E., Yadalam, P. K., & Ardila, C. M. (2024). Efficacy of the finite element analysis in assessing the effects of light curing on the mechanical properties of direct restorative composites: A systematic review. In *Journal of Clinical and Experimental Dentistry* (Vol. 16, Number 11, pp. e1411–e1421). Medicina Oral S.L. <https://doi.org/10.4317/jced.62021>
- Post-Harvest-Losses-of-Cereals-in-Developing-Countries-A-Review*. (n.d.).

- Qange, S., Mdoda, L., & Mditshwa, A. (2024). Modeling the critical causal factors of postharvest losses in the vegetable supply chain in eThekweni metropolitan municipality: The log-linear regression model. *Heliyon*, 10(20). <https://doi.org/10.1016/j.heliyon.2024.e39565>
- Rabiah, P., Adawiah, A., Azrianingsih, R., & Mastuti, R. (2019). Effect of Place and Time Storage on the Quality of Tubers *Amorphophallus muelleri* Blume. *Life Sci*, 9(1). <http://blynk.ub.ac.id/microclimate.php>
- Romsomsa, N., Chim-anagae, P., & Jangchud, A. (2010). Optimization of silk degumming protease production from *Bacillus subtilis* C4 using Plackett-Burman design and response surface methodology. *ScienceAsia*, 36(2), 118–124. <https://doi.org/10.2306/scienceasia1513-1874.2010.36.118>
- Sáenz-Baños, M., Latorre-Biel, J. I., Martínez-Cámara, E., Jiménez-Macías, E., Longo, F., & Blanco-Fernández, J. (2022). Methodology for energy demand reduction of potato cold storage process. *Journal of Food Process Engineering*, 45(10). <https://doi.org/10.1111/jfpe.14127>
- Saha, A., Gandrakota, A., & Morozov, A. V. (2026). *Poisson Log-Normal Process for Count Data Prediction*. <http://arxiv.org/abs/2602.05926>
- Santa-Ramírez, H. A., Otálvaro-Castro, G. J., Joost, S., Melgar-Quinonez, H., Bilal, U., & Stringhini, S. (2023). Small area vulnerability, household food insecurity and child malnutrition in Medellín, Colombia: results from a repeated cross-sectional study. *The Lancet Regional Health - Americas*, 23. <https://doi.org/10.1016/j.lana.2023.100521>
- Schrenk, D., Bignami, M., Bodin, L., Chipman, J. K., del Mazo, J., Hogstrand, C., Hoogenboom, L., Leblanc, J. C., Nebbia, C. S., Nielsen, E., Ntzani, E., Petersen, A., Sand, S., Schwerdtle, T., Vleminckx, C., Wallace, H., Brimer, L., Cottrill, B., Dusemund, B., ... Grasl-Kraupp, B. (2020). Risk assessment of glycoalkaloids in feed and food, in particular in potatoes and potato-derived products. *EFSA Journal*, 18(8). <https://doi.org/10.2903/j.efsa.2020.6222>
- Sedi, B. G., Boukaka, A., & Azzarri, C. (n.d.). *Assessment of potato tuber quality in Kenya Baseline quality assessment survey report Strengthening Food Systems to Promote Increased Value Chain Employment Opportunities for Youth*.
- Shahbazi, Z., Hazra, D., Park, S., & Byun, Y. C. (2020). Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches. *Symmetry*, 12(9). <https://doi.org/10.3390/SYM12091566>
- Sharma, R., Kamble, S. S., Gunasekaran, A., Kumar, V., & Kumar, A. (2020). A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Computers & Operations Research*, 119, 104926. <https://doi.org/10.1016/J.COR.2020.104926>
- Sibindi, R., Mwangi, R. W., & Waititu, A. G. (2023). A boosting ensemble learning based hybrid light gradient boosting machine and extreme gradient boosting model for predicting house prices. *Engineering Reports*, 5(4). <https://doi.org/10.1002/eng2.12599>
- Sirba, H. Y., Yesuf Sirba, H., & Chimdessa, T. B. (2021). Review of Impact of Climate Change on Food Security in Africa. *International Journal of Research and*

- Spencer, P., Kraay, A., Wang, D., & Andree, B. P. J. (2020). Predicting Food Crises. In *Predicting Food Crises*. World Bank, Washington, DC. <https://doi.org/10.1596/1813-9450-9412>
- Stürner, L., Ross, T., Querengässer, J., & Traub, H. J. (2024). Institutional influence on length of stay in German forensic hospitals: a multilevel analysis of patients with schizophrenia spectrum disorders. *Frontiers in Psychiatry*, 15. <https://doi.org/10.3389/fpsyt.2024.1456363>
- Sunday, C. O., Sowunmi, F. A., Obayelu, O. A., Awoyemi, A. E., Omotayo, A. O., & Ogunniyi, A. I. (2022). Disentangling Drivers of Food Waste in Households: Evidence from Nigeria. *Foods*, 11(8). <https://doi.org/10.3390/foods11081103>
- Swartz, J., Etuhole, M., & Mwahi, M. (2022). *Spatial Modelling of Malaria Incidence and its Risk Factors in Namibia*.
- Trabelsi, A., Elouedi, Z., & Lefevre, E. (2023). An ensemble classifier through rough set reducts for handling data with evidential attributes. *Information Sciences*, 635, 414–429. <https://doi.org/10.1016/J.INS.2023.01.091>
- Trizoglou, P., Liu, X., & Lin, Z. (2021). Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. *Renewable Energy*, 179, 945–962. <https://doi.org/10.1016/J.RENENE.2021.07.085>
- Upadhyay, D., Manero, J., Zaman, M., & Sampalli, S. (2021). Gradient Boosting Feature Selection with Machine Learning Classifiers for Intrusion Detection on Power Grids. *IEEE Transactions on Network and Service Management*, 18(1), 1104–1116. <https://doi.org/10.1109/TNSM.2020.3032618>
- van Hoof, J., & Vanschoren, J. (2021). *Hyperboost: Hyperparameter Optimization by Gradient Boosting surrogate models*. <http://arxiv.org/abs/2101.02289>
- Vieira, J. T., Pereira, R. B. D., Lauro, C. H., Brandão, L. C., & Ferreira, J. R. (2024). Multi-objective evolutionary optimization of extreme gradient boosting regression models of the internal turning of PEEK tubes. *Expert Systems with Applications*, 238, 122372. <https://doi.org/10.1016/J.ESWA.2023.122372>
- Visse-Mansiaux, M., Tallant, M., Brostaux, Y., Delaplace, P., Vanderschuren, H., & Dupuis, B. (2021). Assessment of pre- and post-harvest anti-sprouting treatments to replace CIPC for potato storage. *Postharvest Biology and Technology*, 178. <https://doi.org/10.1016/j.postharvbio.2021.111540>
- Wang, B., & Chen, T. (2022). Social Progress beyond GDP: A Principal Component Analysis (PCA) of GDP and Twelve Alternative Indicators. *Sustainability (Switzerland)*, 14(11). <https://doi.org/10.3390/su14116430>
- Watson, S. I. (2024). *Generalised Linear Mixed Model Specification, Analysis, Fitting, and Optimal Design in R with the glmmr Packages*. <http://arxiv.org/abs/2303.12657>
- William Clement Aaron, Obehi Irekponor, Ngozi Tracy Aleke, Linda Yeboah, & Jennifer E Joseph. (2024). Machine learning techniques for enhancing security in financial

- technology systems. *International Journal of Science and Research Archive*, 13(1), 2805–2822. <https://doi.org/10.30574/ijrsra.2024.13.1.1965>
- Williams, N. J., Cardamone, N. C., Beidas, R. S., & Marcus, S. C. (2024). Calculating power for multilevel implementation trials in mental health: Meaningful effect sizes, intraclass correlation coefficients, and proportions of variance explained by covariates. *Implementation Research and Practice*, 5. <https://doi.org/10.1177/26334895241279153>
- Wu, J., Zhang, J., Ni, W., Xu, X., George, M. S., & Lu, G. (2023). Effect of Heat Treatment on the Quality and Soft Rot Resistance of Sweet Potato during Long-Term Storage. *Foods*, 12(23). <https://doi.org/10.3390/foods12234352>
- Wubetie, H. T. (2024). *Comprehensive Modeling of Household Food Insecurity Levels in Ethiopia*.
- Xu, X., Xiao, C., Dong, Y., Zhan, L., Bi, R., Song, M., Pan, J., & Xiong, Z. (2024). Machine learning algorithms realized soil stoichiometry prediction and its driver identification in intensive agroecosystems across a north-south transect of eastern China. *Science of The Total Environment*, 906, 167488. <https://doi.org/10.1016/J.SCITOTENV.2023.167488>
- Xue, H., Liu, Q., & Yang, Z. (2023). Pathogenicity, Mycotoxin Production, and Control of Potato Dry Rot Caused by *Fusarium* spp.: A Review. In *Journal of Fungi* (Vol. 9, Number 8). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/jof9080843>
- Yaremych, H. E., & Preacher, K. J. (2024). Understanding the Consequences of Collinearity for Multilevel Models: The Importance of Disaggregation Across Levels. *Multivariate Behavioral Research*, 59(4), 693–715. <https://doi.org/10.1080/00273171.2024.2315549>
- Yonas, M., Ashenafi, F., Kidane, D., & Welegerima, K. (2021). Stored-grain losses and management practices among smallholder cereal farmers in Northern Ethiopia. *International Journal of Pest Management*, 68(1), 59–68. <https://doi.org/10.1080/09670874.2020.1795301>
- Zhang, L. (2025). Estimating school and teacher effects on students' academic performance using multilevel models with three or more levels: a literature review. In *Discover Education* (Vol. 4, Number 1). Discover. <https://doi.org/10.1007/s44217-024-00392-4>
- Zhang, L., & Jánošík, D. (2024). Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. *Expert Systems with Applications*, 241, 122686. <https://doi.org/10.1016/J.ESWA.2023.122686>
- Zhang, W., Xu, M., Feng, Y., Mao, Z., & Yan, Z. (2024). The Effect of Procrastination on Physical Exercise among College Students—The Chain Effect of Exercise Commitment and Action Control. *International Journal of Mental Health Promotion*, 26(8), 611–622. <https://doi.org/10.32604/ijmhp.2024.052730>
- Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, 58, 308–324. <https://doi.org/10.1016/J.TRC.2015.02.019>
- Zhang, Y., & Lai, M. H. C. (2024). Evaluating two small-sample corrections for fixed-effects standard errors and inferences in multilevel models with heteroscedastic,

- unbalanced, clustered data. *Behavior Research Methods*, 56(6), 5930–5946. <https://doi.org/10.3758/s13428-023-02325-9>
- Zhang, Y., & Wu, Y. (2021). Introducing Machine Learning Models to Response Surface Methodologies. In *Response Surface Methodology in Engineering Science*. IntechOpen. <https://doi.org/10.5772/intechopen.98191>
- Zhou, X., & Gilmour, S. G. (2024). *\$Q_{B}\$ Optimal Two-Level Designs for Baseline Parameterization*. <http://arxiv.org/abs/2409.01926>
- Zhu, T., Pei, H., Li, Z., Zhang, M., Chen, C., & Li, S. (2023). The Postharvest Application of Carvone, Absciscic Acid, Gibberellin, and Variable Temperature for Regulating the Dormancy Release and Sprouting Commencement of Mini-Tuber Potato Seeds Produced under Aeroponics. *Plants*, 12(23). <https://doi.org/10.3390/plants12233952>
- Zhu, T., Pei, H., Li, Z., Zhang, M., Chen, C., & Li, S. (2023). The Postharvest Application of Carvone, Absciscic Acid, Gibberellin, and Variable Temperature for Regulating the Dormancy Release and Sprouting Commencement of Mini-Tuber Potato Seeds Produced under Aeroponics. *Plants*, 12(23). <https://doi.org/10.3390/plants12233952>
- Zhu, X., Hui, S., Huang, H., Liu, R., Wang, S., & Huang, C. (2024). Antimicrobial mechanism of chlorine dioxide and its impacts on postharvest management in horticultural produce: A review. *Postharvest Biology and Technology*, 213, 112921. <https://doi.org/10.1016/J.POSTHARVBIO.2024.112921>
- Zsögön, A., Peres, L. E. P., Xiao, Y., Yan, J., & Fernie, A. R. (2022). Enhancing crop diversity for food security in the face of climate uncertainty. *Plant Journal*, 109(2), 402–414. <https://doi.org/10.1111/tpj.15626>

APPENDICES

Appendix 1: Questionnaire for Farmers and Retailers

For Farmers

Section 1: Farming and Storage Practices

1. What type of potato varieties do you grow? (Open-ended)
2. How do you store potatoes after harvest?
 - ☐ Traditional storage (e.g., pits, open spaces)
 - ☐ Cold storage
 - ☐ Other (Specify)
3. How long do you typically store potatoes after harvest before selling or using them?
 - ☐ Less than two weeks
 - ☐ 2–3 weeks
 - ☐ 3–4 weeks
 - ☐ Over a month
4. What challenges do you face in maintaining potato quality during storage? (Open-ended)
5. Which factors do you think contribute to potato losses after harvest?

Section 2: Use of Chemicals and Techniques

6. Do you use chemicals (e.g., sprout suppressants) to prolong potato storage?
 - ☐ Yes
 - ☐ No
- If yes, which chemicals do you use? (Open-ended)
7. Are you aware of any health or environmental concerns related to the use of these chemicals?
 - ☐ Yes
 - ☐ No

Section 3: Role of Extension Services

8. Have you received training or advice from agricultural extension officers on potato storage?
 - ☐ Yes
 - ☐ No
9. If yes, what topics were covered in the training?
 - ☐ Proper storage techniques
 - ☐ Use of chemicals
 - ☐ Pest and disease management
 - ☐ Other (Specify)
10. Could more extension services help reduce post-harvest losses?
 - ☐ Yes
 - ☐ No

Section 4: Transportation and Market Access

11. How far is the nearest market from your farm?
 - ☐ Less than 5 km
 - ☐ 6–10 km
 - ☐ Over 10 km
12. What challenges do you face during transportation of potatoes?
 - ☐ Mechanical damage to potatoes
 - ☐ High transport costs

- ☐ Delays in reaching the market
 - ☐ Other (Specify)
13. How do transportation challenges affect the quality of your potatoes?
- ☐ Significantly
 - ☐ Slightly
 - ☐ Not at all

Section 5: Education and Awareness

14. What is your highest level of education?
- ☐ Primary
 - ☐ Secondary
 - ☐ No formal education
 - ☐ Post-secondary
15. Would more education or training help improve your storage practices?
- ☐ Yes
 - ☐ No
16. What topics would you like to learn more about?
- ☐ Post-harvest handling techniques
 - ☐ Chemical use and safety
 - ☐ Advanced storage technologies
 - ☐ Other (Specify)
17. How long do you typically store potatoes before they are sold?
- ☐ Less than a week
 - ☐ 1–2 weeks
 - ☐ 2–4 weeks
 - ☐ Over a month
18. What are the most common causes of potato spoilage during storage?
- ☐ Sprouting
 - ☐ Greening
 - ☐ Rotting
 - ☐ Weight loss

Section 2: Transportation

19. Where do you source your potatoes?
- ☐ Directly from farmers
 - ☐ Wholesalers
 - ☐ Other (Specify)
20. How are the potatoes transported to your location?
- ☐ By truck
 - ☐ By motorcycle
 - ☐ Other (Specify)
21. What issues do you encounter during transportation of potatoes?
- ☐ Mechanical damage
 - ☐ Delays
 - ☐ High costs
 - ☐ Bad weather
 - ☐ Other (Specify)

Section 3: Use of Chemicals and Techniques

22. Do you use or require your suppliers to use chemicals (e.g., sprout suppressants) to extend potato shelf-life?

- ☐ Yes
- ☐ No

If yes, which chemicals are used? (Open-ended)

23. Are customers concerned about chemical use on potatoes?

- ☐ Yes
- ☐ No

Section 4: Role of Education and Extension Services

24. Have you received training on potato handling or storage from extension officers or other organizations?

- ☐ Yes
- ☐ No

25. Do you think training would help reduce your losses?

- ☐ Yes
- ☐ No

26. What topics would you like training on?

- ☐ Proper storage techniques
- ☐ Transportation best practices
- ☐ Chemical use and safety
- ☐ Other (Specify)

27. What percentage of potatoes goes to waste after harvest?

Appendix 2: Directorate of Postgraduate Approval Letter



MURANG'A UNIVERSITY OF TECHNOLOGY

Office of the Director, Postgraduate Studies

E-mail: pgs@mut.ac.ke
Website: www.mut.ac.ke

P.O. Box 75-10200
Murang'a, Kenya

Ref: MUT/ARP/PGS/19/2020 /VOL.1

Date: 12th February, 2025

Mr. Erick Cheruiyot Kirui- AS500/5772/2022

School of Pure, Applied and Health Sciences,
Murang'a University of Technology,
P.O. Box 75-10200
Murang'a.

Dear Mr. Kirui,

RE: APPROVAL OF RESEARCH PROPOSAL AND SUPERVISORS.

I am pleased to inform you that the Directorate of Postgraduate Studies on 10th February 2025 considered and approved your Doctoral research proposal entitled *"Gradient Boosting -Based Response Surface Methodology with Application in Modelling Potato Post-Harvest Losses."* and appointed the following as supervisors:

1. Dr. Ayubu Anapapa- Murang'a University of Technology
2. Dr. John Mutuguta- Murang'a University of Technology

You may now proceed with your data collection subject to obtaining research permit from NACOSTI, if required. You should also begin consulting your supervisors and submit through them quarterly progress reports to the Director Postgraduate Studies through your CoD and School Dean. Progress Reports forms can be accessed in the Student Portal.

It is the policy and regulations of the University that you observe deadlines. The guidelines on Postgraduate supervision can be accessed in the post graduate Handbook.

Your responsibilities as a student will include, among others:

- i. Maintain regular consultation with your supervisor(s), at least once a month
- ii. Submit quarterly reports on time, through your supervisors, CoD, Dean and to the Director of Postgraduate Studies;

- iii. Ensure quality work all through;
- iv. Present your research findings at 2-3 seminars/conferences prior to thesis examination.
- v. Publish two articles from your research findings in a refereed journal prior to thesis examination.

For any further clarification please contact the undersigned.

Yours Sincerely,

Prof. Geoffrey Muchiri, PhD
Director, Postgraduate Studies

Copy to:

1. Registrar – Academics and Students Affairs
2. Dean, School of Pure, Applied and Health Sciences



Appendix 3: NACOSTI Research Permit

 REPUBLIC OF KENYA	 NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION
Ref No: 557857	Date of Issue: 09/June/2025
RESEARCH LICENSE	
	
This is to Certify that Mr. ERICK Cheruiyot KIRUI of Murang'a University of Technology, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in Kiambu on the topic: Gradient Boosting-Based Response Surface Methodology with Application in Modelling Potato Post-Harvest Losses for the period ending : 09/June/2026.	
License No: NACOSTI/P/25/4174725	
Applicant Identification Number: 557857	Deputy Director Innovation NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION
	Verification QR Code 
NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.	
See overleaf for conditions	

Appendix 4: Publications

This thesis has produced the following publications:

Cheruiyot Kirui, E., Anapapa, A., & Mutuguta, J. (2026). Hybrid gradient boosting-based response surface methodology for modeling nonlinear data. *Journal of Statistical Sciences and Computational Intelligence*, 2(2), 475–487.

<https://doi.org/10.64497/jssci.168>

Kirui, E., Anapapa, A., & Mutuguta, J. (2026). Multilevel beta regression model for potato post-harvest losses along the farm-market value chain in Kenya. *Asian Journal of Probability and Statistics*, 28(5), 46–60.

<https://doi.org/10.9734/ajpas/2026/v28i5894>