

GENERALIZED LINEAR MIXED MODELS (GLMM), GENERALIZED MIXED EFFECT TREE (GMET), AND GENERALIZED MIXED EFFECT RANDOM FOREST (GMERF) WHEN THE RESPONSE VARIABLE FOLLOWS THE FOUR PARAMETER BETA DISTRIBUTION

DIAN KUSUMANINGRUM



**STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2025**



@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



STATEMENT REGARDING THE DISSERTATION AND INFORMATION SOURCES AND COPYRIGHTS

I hereby declare that the dissertation entitled " Generalized Linear Mixed Models (GLMM), Generalized Mixed Effect Tree (GMET), and Generalized Mixed Effect Random Forest (GMERF) when the Response Variable Follows the Four parameter Beta Distribution " is my work under supervision from the supervisory committee and has not been submitted in any form to any tertiary institution. Sources of information derived or quoted from published and unpublished works by other authors have been mentioned in the text and included in the Bibliography at the end of this dissertation.

I hereby assign the copyright of my writing to the IPB University, Indonesia.

Bogor, April 2025
Dian Kusumaningrum
G1601201001



@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.



RINGKASAN

DIAN KUSUMANINGRUM. *Generalized Linear Mixed Models (GLMM)*, *Generalized Mixed Effect Tree (GMET)*, dan *Generalized Mixed Effect Random Forest (GMERF)* Ketika Peubah Respon Memiliki Sebaran Beta Empat Parameter. Dibimbing oleh HARI WIJAYANTO, ANANG KURNIA, and KHAIRIL ANWAR NOTODIPUTRO.

Pengembangan model prediksi berdasarkan karakteristik data sangatlah penting karena berpengaruh terhadap akurasi prediksi di berbagai bidang. Pengembangan ini mencakup model yang dirancang untuk data yang memiliki batas maksimum dan minimum tertentu, seperti halnya sebaran beta empat parameter, yang dikenal fleksibel dalam mengakomodasi berbagai bentuk sebaran, termasuk kemenjuluran dan ekor berat (*heavy tails*).

Selanjutnya, model prediksi juga perlu mempertimbangkan keragaman data yang umum terjadi akibat perbedaan geografis, fluktuasi temporal, keragaman lingkungan, serta disparitas sosial-ekonomi. Selain itu, pengembangan model prediksi harus mempertimbangkan kompleksitas hubungan yang muncul ketika mengaplikasikan berbagai jenis dataset. Khususnya dalam bidang pertanian, pengembangan model prediksi dapat memperoleh manfaat dari integrasi antara data survei dan data satelit. Memahami serta memodelkan hubungan kompleks ini menjadi kunci dalam meningkatkan akurasi prediksi.

Oleh karena itu, kami telah mengembangkan tiga kajian yang membahas metode dan model yang sesuai untuk memprediksi peubah respon yang memiliki distribusi beta empat parameter, keragaman data yang tinggi, serta hubungan kompleks yang terjadi pada peubah respon dan peubah bebas. Studi ini secara khusus menyoroti pentingnya pengembangan tersebut dalam memprediksi produktivitas padi, dengan fokus pada penerapannya dalam pengembangan asuransi tanaman padi *Area Yield Index (AYI)*. Prediksi produktivitas padi yang akurat sangat penting dalam menentukan premi asuransi dan menilai risiko.

Kajian pertama dalam Bab 3 mengembangkan model GLMM beta empat parameter dengan menerapkan proses transformasi. Transformasi ini memetakan peubah respon aktual y , yang memiliki rentang interval (a, b) ke y^* yang memiliki rentang interval $(0, 1)$. Proses ini memungkinkan pemodelan dan prediksi data menggunakan model beta GLM atau GLMM. Model GLMM digunakan jika peubah respon diukur dalam kelompok atau area tertentu ($i=1, 2, \dots, q$) dan $j = 1, 2, \dots, n_i$, sedangkan model GLM diterapkan jika struktur data lebih sederhana. Hasil penelitian menunjukkan bahwa model GLMM lebih baik dibandingkan pendekatan GLM, yang mengindikasikan bahwa efek acak dan tetap diperlukan dalam memprediksi produktivitas padi. Meskipun model GLMM beta empat parameter tersebut menunjukkan hasil yang cukup baik, namun proses transformasi dapat menyebabkan bias dalam estimasi parameter serta menimbulkan kesulitan dalam interpretasi nilai koefisien.

Dalam kajian berikutnya yang dipresentasikan dalam Bab 4, kami mengembangkan lebih lanjut model regresi mean beta empat parameter yang diperkenalkan oleh Zhou dan Huang (2022) dengan menambahkan efek acak ke dalam model tersebut. Model ini dikembangkan menggunakan pendekatan Bayesian melalui paket *Stan* di perangkat lunak *R*. Studi simulasi menunjukkan bahwa estimasi parameter model ini relatif tidak bias, kecuali untuk parameter presisi ($\hat{\phi}$). Selain itu, studi empiris menunjukkan bahwa prediksi GLMM beta empat parameter yang dikembangkan lebih akurat dibandingkan model acuan Zhou dan Huang.



Kajian ketiga akan dibahas dalam Bab 5, akurasi prediksi lebih lanjut ditingkatkan dengan menangani hubungan data yang lebih kompleks. Sebagai contoh, Ketika kita mengintegrasikan data survei petani dan data satelit ke dalam model, muncul hubungan linier maupun non-linier. Oleh karena itu, model GLMM beta empat parameter dikembangkan lebih lanjut menjadi *Generalized Mixed Effect Tree* (GMET) dan *Generalized Mixed Effect Random Forest* (GMERF). Dari studi kasus empiris, model-model ini terbukti lebih sesuai untuk memprediksi produktivitas padi dibandingkan model GLMM beta empat parameter, terutama dalam penggunaan data satelit dan survei petani.

Pada Bab 5, kami juga mengevaluasi akurasi prediksi dari model yang dikembangkan serta memilih model terbaik. Dengan mengkalibrasi model terbaik terhadap data empiris, kami akan memperoleh hasil prediksi produktivitas padi yang akurat dan melakukan studi *Bootstrap* secara ekstensif untuk mengestimasi premi murni dan *Value at Risk* (VaR) dari AYI. Hasil penelitian menunjukkan bahwa desain AYI di tingkat kabupaten lebih tepat ketika terdapat keragaman produktivitas antar wilayah. Selain itu, perlu pertimbangan dalam mendefinisikan produktivitas acuan, terutama ketika distribusi produktivitas padi mengikuti distribusi beta empat parameter. Penggunaan data satelit dalam model terbukti bermanfaat karena menyediakan informasi berskala luas, konsisten secara temporal, dan lebih efisien dibandingkan survei lapangan yang berskala besar. Namun, data satelit tetap perlu dikombinasikan dengan data survei untuk menggambarkan faktor spesifik lokal yang mungkin tidak sepenuhnya dapat dijelaskan oleh data satelit saja.

Secara umum, premi AYI umumnya dihitung menggunakan data produktivitas historis rata-rata, yang kurang fleksibel dan tidak mempertimbangkan faktor dinamis seperti perubahan iklim atau serangan hama. Sebaliknya, model prediktif seperti GLMM beta empat parameter, GMET, dan GMERF menawarkan pendekatan yang lebih baik karena prediksi diperoleh berdasarkan berbagai data yang terkait. Hal ini akan meningkatkan adaptabilitas, akurasi, serta responsivitas terhadap perubahan dalam sektor pertanian. Dengan demikian, risiko dapat dinilai dengan lebih baik, sehingga pada akhirnya menghasilkan produk asuransi yang lebih efektif. Hal ini meningkatkan kemungkinan bahwa petani akan mendapatkan kompensasi yang adil dan memadai dalam kasus gagal panen, sementara pihak asuransi tetap dapat menjaga stabilitas finansial produknya.

Kata Kunci: Sebaran Beta Empat Parameter, *Generalized Linear Mixed Model* (GLMM), *Generalized Mixed Effect Tree* (GMET), *Generalized Mixed Effect Random Forest* (GMERF), Produktivitas Padi, Asuransi Berbasis Indeks Hasil (*Area Yield Insurance*).



SUMMARY

DIAN KUSUMANINGRUM. Generalized Linear Mixed Models (GLMM), Generalized Mixed Effect Tree (GMET), and Generalized Mixed Effect Random Forest (GMERF) when the Response Variable Follows the Four Parameter Beta Distribution. Supervised by HARI WIJAYANTO, ANANG KURNIA, and KHAIRIL ANWAR NOTODIPUTRO.

Advancements in developing prediction models based on the characteristics of data are critical as they influence the accuracy of predictions across various fields. These advancements include models designed for data constrained by specific maximum and minimum values, such as the Four Parameter Beta distribution, which is recognized for its flexibility in accommodating diverse shapes, skewness, and heavy tails. Next, prediction models should also consider the variability of data that is commonly found in conditions where data is effected by geographical differences, temporal fluctuations, environmental variability, and socio-economic disparities. Furthermore, the development of prediction models should also account for the complexity of relationships that may occur when applying different datasets. Particularly in agriculture, developing prediction models can benefit from the integration of survey data and satellite data. Understanding and modelling these complex relationships is key to improving prediction accuracy.

Therefore, we have developed three studies that discuss the methods and models suitable for predicting response data characterized by a Four Parameter Beta distribution, high variability, and inherit complex relationships within datasets. This study specifically highlights the significance of these advancements in predicting paddy productivity, with a particular focus on their application to the development of Area Yield Index (AYI) crop insurance for paddy. Accurate predictions of paddy productivity are essential for determining insurance premiums and assessing risks.

The first study in Chapter 3 develops the Four Parameter Beta GLMM model by implementing a transformation process. The transformation maps the actual response variable y that has an interval (a, b) to y^* with the interval $(0,1)$. This process enables us to model and predict data by applying beta GLM or GLMM models. We use a GLMM model when the response variable is measured groups/areas ($i=1, 2, \dots, q$) and $j = 1, 2, \dots, n_i$ or apply a GLM model if the data structure is more straightforward. Results show that the GLMM model is better than the GLM approach, indicating that random effects and fixed effects are needed for predicting paddy productivity. Eventhough, the results of the Four Parameter Beta GLMM model is promising, the transformation process can cause bias in parameter estimates and complications in the interpretation of coefficient values.

In the next study presented in Chapter 4, we have further developed Zhou and Huang's (2022) which means Four Parameter Beta regression model by introducing a random effect within the model. This model was developed based on a Bayesian approach through a Stan package in R software. Simulation studies showed that the parameter estimates of the model are considered relatively unbiased, except for precision parameter $(\hat{\phi})$. Furthermore, empirical study shows that the proposed Four Parameter Beta GLMM predictions are more accurate than Zhou and Huang's benchmark model.

For the third study in Chapter 5, we can further improve the prediction accuracy by addressing more complex data. As an example, when integrating farmer survey and satellite data in our model, both linear and non linear relationships emerge. Thus, the four-parameter beta GLMM has been further developed into a Generalized Mixed Effect Tree (GMET) and a Generalized Mixed Effect Random Forest (GMERF). Empirical case study wise, these models proved to be more suitable for predicting paddy productivity compared to the Four Parameter Beta GLMM when using satellite data and farmer surveys.

At the end of Chapter 5, We have also evaluated the developed model's prediction accuracy and selected the best model. By calibrating the best model to empirical data, extensive Bootstrap studies were performed to estimate the pure premium and VaR of AYI. It was shown that designing AYI at district level is more appropriate when productivity among areas vary. Consideration must also be given in defining the benchmark productivity when there is proof that the distribution of paddy productivity follows a Four Parameter Beta distribution. The use of satellite data in the model has proven a beneficiary as it provides valuable, large-scale, and temporally consistent information. It is also more efficient compared to conducting massive field surveys. However, satellite data still may need to be combined with survey data to capture localized, context-specific factors that satellites alone might not fully address

Typically, AYI premiums are calculated using average historical productivity data, which lacks flexibility and does not account for dynamic factors like climate or pest outbreaks. In contrast, predictive models such as the Four Parameter Beta GLMM, GMET, and GMERF offer a more refined, data-driven approach to estimating paddy productivity, improving adaptability, accuracy, and responsiveness to agricultural changes. Hence, enhancing risk assessment and leading to more effective insurance products. Consequently, farmers are ensured fair and adequate compensation in cases of crop failure, while insurers maintain financial stability.

Keywords: Four Parameter Beta Distribution, Generalized Linear Mixed Models (GLMM), Generalized Mixed Effect Tree (GMET), Generalized Mixed Effect Random Forest (GMERF). Paddy Productivity, Area Yield Insurance



@Hak cipta milik IPB University

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

© Copyrights Belong to IPB University, 2025

Copyright protected by law

It is forbidden to cite part or all this manuscript without acknowledging or mentioning the source. Citations are only for the purposes of education, research, writing scientific papers, reports, criticism or reviewing a problem and the citation is not detrimental to IPB University.

It is prohibited to publish and reproduce part or all this manuscript in any form without permission from IPB University.



@Hak cipta milik IPB University

IPB University



IPB University
— Bogor Indonesia —

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

GENERALIZED LINEAR MIXED MODELS (GLMM), GENERALIZED MIXED EFFECT TREEE (GMET), AND GENERALIZED MIXED EFFECT RANDOM FOREST (GMERF) WHEN THE RESPONSE VARIABLE FOLLOWS THE FOUR PARAMETER BETA DISTRIBUTION

DIAN KUSUMANINGRUM

DISERTATION

as one of the requirements for obtaining a
Doctoral degree on
Statistics and Data Science Study Program

**STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2025**





@Hak cipta milik IPB University


Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

External Examiners on Closed Examination:

1. Dr. Bagus Sartono, S.Si, M.Si 
2. Dr. Yenni Angraini, S.Si, M.Si 

External Promoters on Doctoral Promotion:

1. Dr. Bagus Sartono, S.Si, M.Si 
2. Dr. Stevanus Wisnu Wijaya



Dissertation Title : Generalized Linear Mixed Models (GLMM), Generalized Mixed Effect Tree (GMET), and Generalized Mixed Effect Random Forest (GMERF) when the Response Variable Follows the Four Parameter Beta Distribution

Name : Dian Kusumaningrum

NIM : G1601201001

Approved by

Supervisor :
Prof. Dr. Hari Wijayanto, M.Si.



Co-Supervisor :
Prof. Dr. Anang Kurnia, S.Si., M.Si.



Co-Supervisor :
Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S.

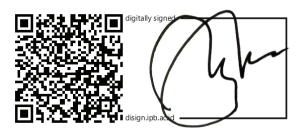


Acknowledged by

Head of Doctoral Program Study:
Dr. Kusman Sadik, M.Si
NIP 196909121997021001



Dean of School of Data Science, Mathematics,
and Informatics :
Prof. Dr. Ir Agus Bueno, M.Si, M.Kom
NIP 196607021993021001



Date of Closed Examination : January 30th, 2025

Graduate Date :



FOREWARD

Alhamdulillah Robbil 'Alaamiin. The author would like to praise and thank Allah subhanaahu wa ta'ala for all His gifts so that this dissertation was successfully completed. This dissertation is entitled "A Study Of Generalized Linear Mixed Models (GLMM) , Generalized Mixed Effect Tree, and Generalized Mixed Effect Random Forest (GMEFR) When The Response Variable Follows The Four parameter Beta Distribution". The author prepared this dissertation as one of the requirements for obtaining a Doctor Degree in Statistics and Data Science at the Program Study of Statistics and Data Science, School Of Data Science, Mathematics, And Informatics, IPB University.

This dissertation was completed thanks to prayers, guidance, input, suggestions and assistance from many parties. Therefore, on this occasion the author would like to expresses his appreciation and thanks to:

1. Supervisory committee who has provided a lot of guidance, direction, input and motivation to the author with great patience: Mr. Prof. Dr. Hari Wijayanto, M.Sc., Mr. Dr. Anang Kurnia, S.Si., M.Sc., and Mr. Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S
2. Examiners outside the commission, Mr. Prof. Dr. Nur Aidi, Mr. Dr. Bagus Sartono, and Mr. Prof. Dr. Asep Saefuddin who have provided valuable suggestions and input to improve the research proposal. Also, Mr. Dr. Bagus Sartono, Mrs. Dr. Yenni Angraini, Mrs. Dr Indahwati, Mr Dr Stevanus Wisnu Wijaya, Mr Dr Kusman Sadik, and Prof Dr Agus Buwono who has provided valuable suggestions and input to improve research dissertation.
3. Prasetya Mulya University who have given the author the opportunity and support to continue my doctoral education at the IPB University.
4. The Southeast Asian Regional Center for Graduate Study and Research in Agriculture (SEARCA) for the doctoral full scholarship opportunities, training facilities and other supports, as well as the motivational encouragement that has been provided
5. Japan Student Services Organization (JASSO) for SUIJI PhD Joint Research Program and The United Graduate School of Agricultural Sciences (UGAS) Ehime University Japan for the opportunity to conduct joint research in Ehime University Japan
6. The Late Prof Ken Seng Tan, Asoc. Prof Islam MD Parvez, Prof Dr. Anang Kurnia, Prof Dr. Asep Saefuddin, Prof Dr. Nur Aidi, The Late Prof Andi Hakim Nasution my inspiring mentors and supervisors from bachelor's to master's Degree Program.
7. IPB Postgraduate School and its staff who have helped facilitate the author's studies.
8. Head of the Study Program and lecturers at the Department of Statistics and Data Science who have provided a lot of knowledge, enlightenment, direction, and assistance during the lecture period.
9. The Secretariat of Department of Statistics and Data Science Study Program, IPB University, who always patiently assists with all administrative and information processes.
10. The author's family, who always supports, motivates and pray for the success of the author's studies with full sincerity and patience
11. Doctoral students' class of 2020, and other Doctoral and Master' s Degree students from the Department of Statistics and Data Science Study Program, IPB University for their sincere prayers, assistance, encouragement and cooperation with the author.

This dissertation still has limitations and shortcomings. However, the author hopes that this research can be useful for all parties.

Bogor, April 2025



LIST OF CONTENTS

RINGKASAN	i
SUMMARY	iii
LIST OF CONTENTS	iv
LIST OF TABELS	vi
LIST OF FIGURESS	vii
LIST OF APPENDICES	viii
I. INTRODUCTION	1
1.1. Background	1
1.2. Research Questions	3
1.3. Objective	3
1.4. Benefit	3
1.5. Research Scope	4
1.6. Novelty	6
1.7. Dissertation Systematic Plan	7
II. LITERATURE REVIEW	9
2.1. Beta Distribution and Four Parameter Beta Distribution	9
2.2. Beta Generalized Linear Model (GLM)	12
2.3. Beta Generalized Linear Mixed Model (GLMM)	12
2.4. Four Parameter Beta Regression Model	15
2.5. Generalized Mixed Effect Tree (GMET)	15
2.6. Generalized Mixed Effect Random Forest (GMERF)	16
2.7. Mean Square Error (MSE) of Parameter Estimates	17
2.8. Paddy Productivity Prediction for Area Yield Index (AYI) Crop Insurance	18
2.9. Crop Cutting Experiment (CCE)	19
2.10. Farmer Survey Data	20
2.11. Sentinel 2A Satellite Data	21
III. Modelling the Four Parameter Beta Distribution Response VariableThrough a Transformation Process	24
3.1. Introduction	24
3.2. Method	26
3.2.1. Transformation Process for a Four Beta Parameter Distribution	26
3.2.2. Predicted Values	26
3.3. Empirical Case Study	27
3.3.1. Empirical Case Study Dataset	27
3.3.2. Data Collecting, Processing, and Modelling for Paddy Productivity	28
3.3.3. Variable Selection	32
3.3.4. Paddy Productivity Distribution	32
3.3.5. Model Evaluations based on Key Performance Indicators	34
3.3.6. Analyzing the Best Prediction Models	36
3.4. Effect of the Transformation Process	39
3.5. Model Limitations and Further Development	48
3.6. Conclusion and Policy Recommendations	51

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

IV. The Four Parameter Beta GLMM Based on Bayesian Approach	53
4.1. Introduction	53
4.2. Proposed Method	55
4.2.1. Model Specifications	55
4.2.2. Prior Distributions	56
4.2.3. Bayesian Parameter Estimation	57
4.2.4. Predictions of the Response Variable	58
4.3. Simulation Study	60
4.3.1. Simulation Design and Process	60
4.3.2. Simulation Results and Discussion	64
4.4. Empirical Case Study	66
4.4.1. Research Process	66
4.4.2. Paddy Productivity in Central Kalimantan	68
4.4.3. Applying the Four Parameter Beta GLMM and Evaluations	69
4.5. Model Limitations and Further Development	73
4.6. Conclusion and Recommendations	73
V. Four Parameter Beta Generalized Mixed Effect Tree and Random Forest	75
5.1. Introduction	75
5.2. The Proposed Method	76
5.3. Methodology	78
5.4. Results and Discussion	79
5.4.1. Model Evaluation	79
5.4.2. Significant Variables for Predicting Paddy Productivity	80
5.4.3. Estimating Pure Premium and Risks of the Area Yield Crop Insurance	86
5.5. Conclusion	87
VI. General Discussion	88
VII. Conclusion and Recommendation	92
1.1. Conclusion	92
1.2. Recommendations	93
References	94



LIST OF TABELS

Table 2. 1 Formula Mean, Variance, dan Skewness of the Four parameter beta Distribution	11
Table 2. 2. Variables in the Farmer Survey Data	20
Table 2. 3 Spectral Wavelength Characteristics and Spatial Resolution of Sentinel 2 Bands	22
Table 3. 1 Independent Variables Used in the Prediction Model	32
Table 3. 2 Distribution Fit Test and Parameter Estimates	33
Table 3. 3 GLMM Model's Prediction Accuracy Based on RRMSE	35
Table 4. 1. Data Generating Scenarios in the Simulation Process	61
Table 4. 2 Simulation Results for Four Parameter Beta Mean GLMM with Stan (4P Beta GLMM (1)) and for Four parameter beta Mean GLMM with brms (4P Beta GLMM (2))	65
Table 4. 3 Variables Used in the Proposed Four Parameter Beta Mean GLMM	67
Table 4. 4 Model Evaluation for the Four Parameter Beta Mean GLMM	70
Table 4. 5 The Mean, Standard Deviation (Stdev), and the 95% Credible Interval Estimate for the Best Fit Four Parameter Beta GLMM	72
Table 5. 1 Goodness of Fit and Prediction Evaluation for Developed Models	80
Table 5.2 AYI Premium and <i>VaR</i> based on Four Parameter Beta GMERF Estimates (in Rupiah)	87
Table 6.1 Root Means Square Error of the Predictions (RMSEP) of the Purposed Methods	90



LIST OF FIGURESS

Figure 1. 1 Research Stages	5
Figure 1. 2 Research Framework	6
Figure 2.1 Comparison of a Beta Distribution and a Beta Four Parameter Beta Distribution with $\omega_1=2$ and $\omega_2=5$	9
Figure 2. 2 Comparison of a Beta Distribution and a Beta Four Parameter Beta Distribution with $\omega_1=2$ and 5 and $\omega_2=2$ and 5	10
Figure 2. 3 Research Road Map of Paddy Productivity Prediction	19
Figure 2. 4 Sentinel-2 Satellite Orbital Configuration (Drusch <i>et al.</i> 2012)	22
Figure 3. 1 Distribution of Survey Plots and Data Used in Central Kalimantan and Karawang (2020)	27
Figure 3. 2 Research Process Flowchart of the Empirical Study	31
Figure 3. 3 Paddy Productivity Distribution	33
Figure 3. 4 Four Parameter Beta GLM and GLMM Goodness of Fit Model	35
Figure 3. 5 Band 4, Band 8, and NDVI Index Trend	36
Figure 3. 6 Pest Attack Severity in Central Kalimantan and Karawang	37
Figure 3. 7 Climate Change Impact in Central Kalimantan and Karawang	38
Figure 3. 8 Clustered Boxplots of Farmer Survey Data and Paddy Productivity (a)	38
Figure 3. 9 Clustered Boxplots of Farmer Survey Data and Paddy Productivity (b)	39
Figure 3.10 Correlation Plot of lagged NDVI and Paddy Productivity	48
Figure 3.11 Scatterplot Between the Paddy Productivity, Current NDVI, And Lag One to Lag Four NDVI Values	49
Figure 3.12. Scatterplot Between the Paddy Productivity, Number Of Seeds Used, and the Amount ff Fertilizer Components Used	50
Figure 4. 1 Simulation Process Flowchart	60
Figure 4. 2 Empirical Study Process Flowchart	66
Figure 4. 3 Paddy Productivity Season 1and Per Sub District Paddy Productivity in Central Kalimantan	69
Figure 4. 4 Evaluations on Bayesian Approach on a Four Parameter Beta Mean GLMM	71
Figure 5.1. Example of a Four-Parameter Beta GMET Model in Central Kalimantan	81
Figure 5.2 . Band 4, Band 8, and NDVI Scatterplots with Productivity Groups	82
Figure 5.3. Estimated Random Effects of the Best Four Parameter Beta GMET Model	85
Figure 6. 1 Predicted and Actual Scatter Plots and Correlation of GLM and GLMM Models	89



LIST OF APPENDICES

Appendix 1. GEE Script to Download Band 4, Band 8 and NDVI Data	99
Appendix 2. Best Beta GLMM Model Output for Central Kalimantan and Karawang	103
Appendix 3. Scatterplot between Paddy Productivity of Sentinel 2A's NDVI, Band 4m and Band 8 data	105

I. INTRODUCTION

1.1. Background

The Generalized Linear Mixed Model (GLMM) has been developed based on the Linear Mixed Model (LMM) when the response variable does not follow a normal distribution but comes from the exponential distribution family. In contrast to the Generalized Linear Model (GLM), the GLMM model doesn't only have fixed effects but also has random effects (McCulloch and Searle 2001). The GLMM model is usually used when researchers have data with more than one source of variability. For example, in an experiment the results can be measured more than once on the same object (repeated measurements are taken overtime). In such experiments, researchers need to consider both within-object and between-object variability. If the researcher only uses the residual variance, then this value cannot explain both. Apart from repeated observations, the GLMM model is also often used in cases of multilevel data, longitudinal data, group/clustered data, or cross section data. Factors that need to be considered when preparing an appropriate GLMM model are (1) the distribution of the response variable (Y_{ij}), (2) selecting the appropriate link function, (3) the selection of the independent variables that will be used in the model, as well as the distributional assumption of the random effect.

Currently, the GLMM model is often used when the response variable follows a Bernoulli, Binomial, Poisson, and Beta Binomial distribution. The GLMM model has also been used to model response variables following a Beta distribution (Bonat et al. 2015). The Beta distribution is one of the continuous variable distributions within an interval of (0, 1) and has two parameters, ω_1 and ω_2 . Next, there exists a notable research gap in the modeling of response variables that are naturally bound within a specific interval, typically between a minimum value (a) and a maximum value (b). Variables with these characteristics are known to have a Four Parameter Beta distribution. Current modeling approaches, such as Beta GLMM, are often insufficient in capturing the full characteristics of the Four Parameter Beta data. The Four Parameter Beta distribution has higher flexibility and accuracy in capturing the distributional properties of bounded response variables, especially when those bounds vary across observations or subgroups. An example of data following a Four Parameter Beta distribution is the productivity of paddy which tends to skew to the left and has a specific minimum and maximum value (Hennessy 2009). The minimum and maximum values are determined by agronomic and environmental factors, including soil quality, water availability, and access to agricultural inputs, which limit both the minimum and maximum achievable yields.

Despite its flexibility, the development of both theoretical frameworks and applied modeling techniques for the Four Parameter Beta distribution remains insufficiently explored. In particular the development of models based on the Four Parameter Beta distribution that account for random effects or hierarchical data structures has never been addressed in previous studies. This research gap is due to the mathematical complexity of the Four Parameter Beta likelihood function, which presents challenges in parameter estimation, especially in the presence of random effects or hierarchical data structures. To address this gap, there is a need to develop advanced modelling framework based on the Bayesian approach. The Bayesian approach is suitable because it has advantages in accommodating hierarchical or multilevel structures, it allows for the inclusion of prior knowledge or expert beliefs through prior distributions, it provides posterior distributions for parameter estimates, and it is more robust in estimating parameters for complex models by using MCMC (Markov Chain Monte Carlo) (Zou et al. 2022).



As a first step, this study applies a transformation procedure to enable the use of the Beta GLMM framework proposed by Bonat et al. (2015) for response variables that follow a Four Parameter Beta distribution. However, such transformations may introduce bias in parameter estimates and complicate the interpretation of model coefficients. To overcome these limitations, this study further extends the mean regression model for the Four Parameter Beta distribution developed by Zhou and Huang (2022) by incorporating random effects, thereby proposing a novel Four Parameter Beta GLMM that can accommodate hierarchical data structures without relying on potentially bias-inducing transformations.

The development of the GLMM model for response variables having a Four Parameter Beta distribution represents the main novelty contribution in this study. The process begins with specifying the model, determining parameter estimation methods, formulating model inferences, and evaluating the developed model's performance through comprehensive simulation studies and empirical studies. Simulation studies were designed based on the characteristics of a real population of interest. Meanwhile, application wise, the proposed model will be applied to predict paddy productivity at the individual farmer level. The independent variables used will be from farmer surveys and satellite sentinel 2A index data. Various studies have successfully used satellite data to build predictive models. Said *et al.* (2015) have successfully used the value of plant greenness (Vegetation Index) obtained through the analysis of satellite Sentinel 2A index data to develop a model for estimating paddy productivity.

The GLMM model requires a linear relationship between independent and response variables (Agresti 2018). Fontana *et al.* (2021) have developed a GLMM model for multilevel data modelling with the assumptions of a nonlinear relationship between response variables and independent variables and it is called the Generalized Mixed Effects Tree (GMET). The GMET model is built based on the CART and GLMM algorithms. Next, Generalized Mixed Effect Random Forest (GMERF) was also introduced by Pellagatti *et al.* (2021), where random forest was applied along with GLMM. The random forest was opted because it is generally known to be more robust and accurate. These developed models were applied to predict student dropout cases at Politecnico colleges in Italy. Therefore, the model has a response variable that has a Bernoulli distribution.

In this research, these models have been applied to cases where the response variable follows a Four Parameter Beta distribution. Therefore, the GMET and GMERF will also be applied to predict paddy productivity at the farm level when there is an indication of a non-linear relationship between the response variables and the response variables. Accurate predictions obtained from the developed model can also be used by insurance companies to develop alternative productivity-based insurance policies in Indonesia. In addition, local governments can use the accurate prediction results of paddy productivity to identify local food supplies for the population.

This research is crucial because the Four Parameter Beta distribution is reasonably flexible and allows various shapes and skewness at a limited interval. If the transformation process is carried out into a standard beta distribution, it will cause various problems, namely a complicated back-transform process, difficulties in interpreting the results obtained, and will also cause bias in the estimated parameters. Hence, we can highlight that this research will provide a solution for modelling a response variable that has a Four Parameter Beta distribution, high variability, and complex linear and nonlinear relationships. This research has not only provided benefits for the development of theoretical statistics but also developed accurate paddy productivity prediction methods at the individual farmer level, which can be used for the development of the potential alternative Area Yield Index (AYI) crop insurance policy.

1.2. Research Questions

The main research questions are as follows:

1. How can we develop a Four Parameter Beta GLMM model that can be applied for predictions purposes?
2. How can we integrate the Four Parameter Beta GLMM model with tree regression to develop a Four Parameter Beta GMET and integrate the Four Parameter Beta GLMM model with random forest to develop a Four Parameter Beta GMERF model?
3. How can we evaluate the prediction results of the Four Parameter Beta GLMM, Four Parameter Beta GMET, and the Four Parameter Beta GMERF model?
4. How can the Four Parameter Beta GLMM, GMET, and GMERF model be applied to predict paddy productivity? Which is the best model that can be used for calculating premiums and risks in a Crop Insurance policy?

1.3. Objective

Therefore, the research objectives are as follows:

1. Develop the Four Parameter Beta GLMM model: selecting the parameter estimation method, choosing the link functions, defining the inference process, calculate predictions, and model evaluation.
2. Integrate the Four Parameter Beta GLMM model with regression tree to develop a Four Parameter Beta GMET and integrate the Four Parameter Beta GLMM model with random forest to develop a Four Parameter Beta GMERF model.
3. Applying the Four Parameter Beta GLMM, GMET, and GMERF model to predict paddy productivity at the individual farmer level.
4. Compare the predictions and model fit of the Four Parameter Beta GLMM, GMET, and GMERF model to find the best model that can be used to predict paddy productivity at the individual farmer level. Further on the best model can be used for calculating premiums and risks in a Crop Insurance policy

1.4. Benefit

This research is important to be carried out because it is beneficial for scientific development in the fields of statistics and data science as well as developing models for the application of paddy productivity predictions at the individual farmer level which can be utilized by the Indonesian Ministry of Agriculture/Statistics Indonesia/Agricultural Insurance Companies.

The benefits for the field of statistics and data science are: (1) providing information about modeling methods that can be used to model data when the response variable follows a Four Parameter Beta distribution and the relationship between the response variable and the independent variable is linear. Thus, the most appropriate model is the Four Parameter Beta GLMM model. This model can estimate fixed effects and estimate random effects. (2) Along with regression tree or random forest, the Four Parameter Beta GLMM model is also the base of developing a Four Parameter Beta distribution GMET and GMERF non-linear models. The models can also estimate fixed effects and random effects based on GLMM modeling integrated with regression trees and random forests applied.

Furthermore, there are also apparent benefits for the Indonesian Ministry of Agriculture/Statistics Indonesia/Agricultural Insurance Company, which are: (1) obtaining more accurate estimates of paddy productivity levels at the area or individual farmer level (2) the results of productivity estimates at the area or individual farmer level can be used by local governments to identify local food supplies for the population, and (3) the resulting paddy

productivity estimates can be used to calculate premiums and risks in developing Area Yield Index (AYI) Insurance policies.

1.5. Research Scope

The scope of this research consists of the development of three models: a Four Parameter Beta GLMM, a Four Parameter Beta GMET, and a Four Parameter Beta GMERF. Following the development, these models will be applied to predict paddy productivity in Central Kalimantan or Karawang. Subsequently, the most suitable model will be selected, and its predictions will be utilized to calculate the risk and premium for the AYI crop insurance policy.

The process of developing the Four Parameter Beta GLMM model includes defining model specifications, applying the selected link function, selecting the most suitable parameter estimation methods, evaluating parameter estimates, developing inferences for the Four Parameter Beta GLMM model, and evaluating the developed model through a simulation study. To ensure that the research is focused and manageable, the following research limitations were determined : (1) the model is restricted to Four Parameter Beta GLMM with a random intercept, (2) a logit link function was employed in the development approach, (3) the model was based on a mean central tendency, and (4) it was assumed that the observations of the response variable are independent.

Two approaches have been adopted to develop the Four Parameter Beta GLMM model. In the first approach, for simplicity we have conducted a transformation process of the Four Parameter Beta distributed response variable into a beta distributed response variable. Afterwards we will apply the beta GLMM proposed by Bonat *et al.* (2015) where the parameter estimates were based on Maximum Likelihood (ML) estimators. Next, we will analytically analyze the potential bias that may occur due to the transformation process. Transformation can cause bias in the estimated parameters, complicates the back-transformation procedure, and make it challenging to interpret the results. Thus, the second approach is applied, where we will further develop a Four Parameter Beta GLMM model based on the Four Parameter Beta regression model introduced by Zou *et al.* (2022). A random effect will be incorporated into the current model to handle data that tends to have high variability across group/area. Parameter estimates in this model will be derived based on Bayesian approach.

Next, to address the challenges posed by complex linear and nonlinear relationships between the response and independent variables, we will outline the process of developing the Four Parameter Beta GMET model, which was first presented by Fontana *et al.* (2021). Similarly, the process of developing the Four Parameter Beta GMERF model will also be explained, following the methodology introduced by Pellagatti *et al.* (2021). Initially, the researchers need to estimate the response variable (μ_i) based on the Four Parameter Beta GLM model. Next apply a regression tree or random forest. In this process, we will have obtained terminal nodes (R_ℓ), where $\ell = 1, 2, \dots, L$. Each observation will belong to one of the terminal nodes that will be described by a certain set of independent variables. Thus, we can define a set of indicators in the form of a dummy variable for each node. Moving forward, we apply the Four Parameter Beta GLMM model to estimate random effects and fixed effects by incorporating the indicators as independent variables that have been estimated by the regression tree in the previous stage. Moving forward in developing the Four Parameter Beta GMERF, we will apply the Random Forest algorithm introduced by Breiman (2001), which is a variation of bagging that uses regression tree as its base learners.

The Four Parameter Beta GLMM, GMET, and GMERF are then applied to predict paddy productivity at the individual farmer level in Central Kalimantan or Karawang. The response variables are farmers' paddy productivity data for specific paddy field plots during a certain planting season period in 2020. This data is obtained from the CCE (SI). The independent

variables are collected from (1) farmers surveys conducted during the CCE (SI), (2) sentinel 2A satellite index data, encompassing single band values and vegetation index, and (3) one year lagged sentinel 2A satellite index data.

The stages of this research can be seen in Figure 1.1. First, we start by developing a Four Parameter Beta GLMM model. Then, we develop the Four Parameter Beta GMET and GMERF by using the previous Four Parameter Beta GLMM model. In the development process, we have conducted a simulation study and we have also apply the models to predict paddy productivity at a certain area level.

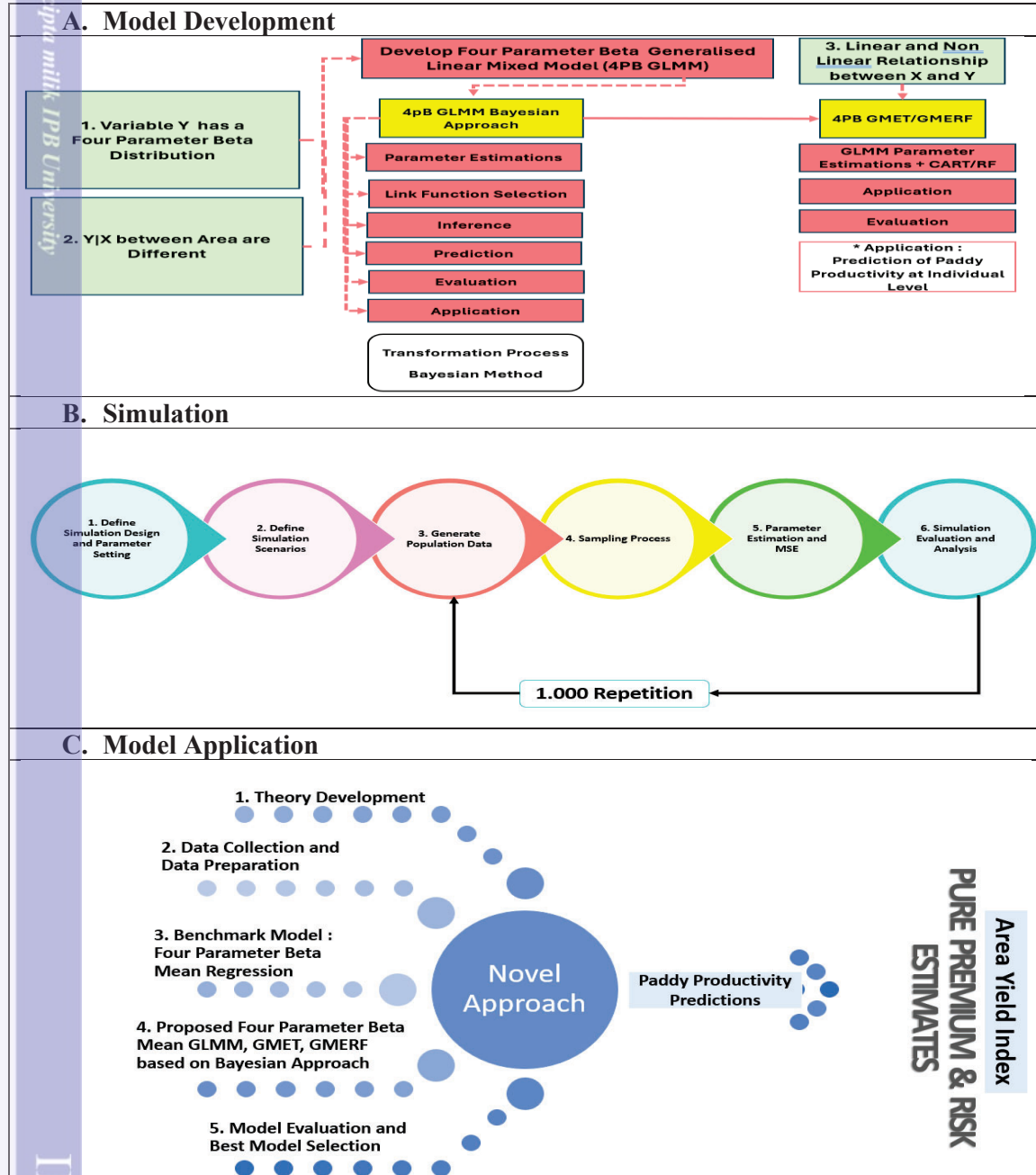


Figure 1. 1 Research Stages

1.6. Novelty

There are four novelties in this research, namely

1. The first novelty contributes to addressing the research gap by providing a Four Parameter Beta GLMM model specification and parameter estimates based on the Bayesian approach. This model is designed for response variables that have a Four Parameter Beta distribution, and it incorporates both fixed effects and random effects. Detailed discussions on this model's development are provided in Chapter 4.
2. Next, we have developed inference and prediction methods in the Four Parameter Beta GLMM model. The inference process of the Four Parameter Beta GLMM model includes the inference process for hypothesis testing and the inference of the interval estimator for prediction results. The model's inference and prediction methods are provided in Chapter 4.
3. We have also conducted further development of the Four Parameter Beta GMET and GMERF by combining the Four Parameter Beta GLMM with regression tree/ random forest. This model can be applied when the relationship between the independent variables and the response variable is complex. Chapter 5 provides a detailed explanation of the development process of the Four Parameter Beta GMET and GMERF.
4. Lastly, the application of the Four Parameter Beta GLMM, GMET, and GMERF model to predict paddy productivity at a certain area level. Furthermore, we can estimate the premium and risks of AYI can be predicted based on the best model with the highest prediction accuracy

In general, an overview of the stages of research to achieve the proposed novelty is shown in Figure 1. 1. It shows how the model was developed, how in general the simulation process was carried out, and how we have applied the model to predict paddy productivity up to estimating the premium and risks of the AYI crop insurance. Meanwhile, the research framework and theoretical foundation used to develop the GLMM Four Parameter Beta GLMM, GMET, and GMERF models can be seen in Figure 1. 2 below.

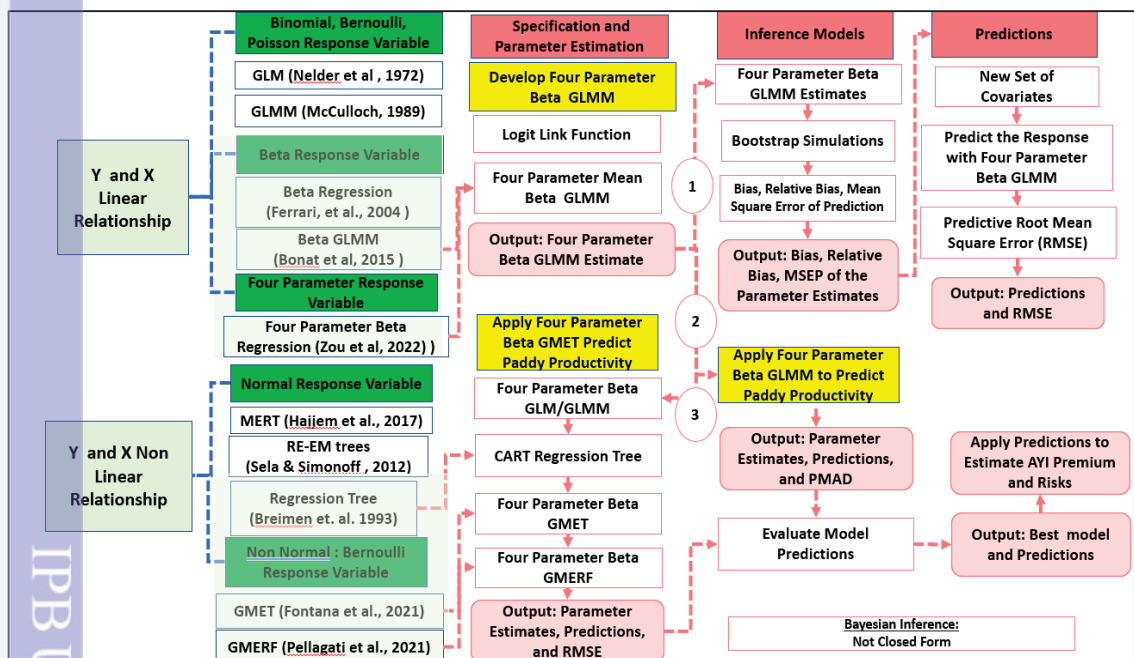


Figure 1. 2 Research Framework

1.7. The Systematic of the Dissertation

This dissertation consists of 7 chapters. Chapter 1 is the Introduction Chapter, which includes: background, problem formulation, research objectives and benefits, research scope, research framework, and novelty of the research. The writing framework for the following chapters is as follows:

1. Chapter 2 is the Literature Review Chapter. The material presented in Chapter 2 includes a brief explanation of the Four Parameter Beta Distribution, Generalized Linear Model (GLM) with Response Variables that Have Beta Distribution, Generalized Linear Mixed Model (GLMM) with Response Variables that Have Beta Distribution, Link Function, Generalized Mixed Effect Tree (GMET), Classification and Regression Trees (CART) Algorithm, Generalized Mixed Effect Random Forest (GMERF), Estimation of Mean Square Error (MSE), Mean Square Error Prediction (MSEP), Prediction using GLMM, GMET, and GMERF, Area Yield Insurance (AYI), Crop Cutting Experiment, Farmer Survey, and Sentinel 2A Satellite Data.
2. Chapter 3 is the study of conducting a transformation process for developing the Four-parameter beta GLMM. In this chapter we proposed a prediction model based on the Four Parameter Beta distribution. Initially, we developed a Four Parameter Beta GLMM by transforming the Four Parameter Beta distribution into a beta distribution and applied the Beta GLMM model introduced by Bonat *et al.* (2015). This method is then applied to predict paddy productivity with three types of independent variable dataset conditions, including (a) just using the farmer survey data set, (b) just using the Sentinel 2A satellite data set, and (c) using a combination of farmer survey and Sentinel 2A satellite data set. Next, we will also select the most appropriate random effect level for the empirical data set. Then, we select the best model. Last, in this chapter we have also pointed out the advantages and disadvantages of applying a transformation process.
3. Chapter 4 is the study of developing a Four Parameter Beta GLMM model through a Bayesian Approach. This model extends the four-parameter beta regression model (Zou *et al.* 2022) by incorporating both fixed and random effects. These effects are considered crucial when modeling the response variable, which follows a four-parameter beta distribution. The main highlight of Chapter 4 includes defining the specification of the developed model, parameter estimations based on Bayesian approach, inference process, and predictions. The inference process of the Four Parameter Beta GLMM model includes inference processes for hypothesis testing and interval estimation for parameter estimate results. A simulation study has been done to measure the goodness of fit of the developed model, and evaluation of the predictions in the developed Four Parameter Beta GLMM model. Further evaluation will also be carried out by comparing the Four Parameter Beta GLMM model with the Four Parameter Beta mean regression model in an empirical case study.
4. Chapter 5 is a study of the development of GMET and GMERF for a response variable that follows a Four Parameter Beta distribution. The GLMM model will be used to estimate paddy productivity at the farmer level if the relationship between the mean and the independent variables is linear. Meanwhile, if the relationship between the independent variables is complex, the Four Parameter Beta GMET and GMERF model will be more sufficient. These models will then be applied to predict paddy productivity data, which is measured through the CCE conducted by SI. The independent variables used in the models will be derived from farmer surveys conducted during the CCE, as well as from Sentinel 2A Satellite Imagery Data.

@Hak cipta milik IPB University

IPB University

At the end of Chapter 5, We have also selected the most accurate paddy productivity predictions that have been obtained from Chapter 4 and Chapter 5. We will use the most accurate paddy prediction to estimate premiums and risks in the development of the Area Yield Index (AYI) insurance policy. The estimations were based on Bootstrap simulations.

5. Chapter 6 is the general discussion. The material in this chapter is a summary of the findings from the previous chapters.
6. Chapter 7 is the conclusion and suggestions section.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

II. LITERATURE REVIEW

2.1. Beta Distribution and Four Parameter Beta Distribution

The Beta distribution is one of the continuous variable distributions that have an interval of $(0, 1)$ and has the following Probability Density Function (PDF):

$$f(y, \omega_1, \omega_2) = \frac{\Gamma(\omega_1 + \omega_2)}{\Gamma\omega_1\Gamma\omega_2} y^{\omega_1-1} (1-y)^{\omega_2-1}, 0 < y < 1, \omega_1 > 0, \omega_2 > 0 \quad (2.1)$$

where ω_1 is the location parameter and $\omega_1 - \omega_2$ is the scale parameter. Meanwhile, the Four Parameter Beta distribution has the following PDF:

$$f(y, \omega_1, \omega_2, a, b) = \frac{\Gamma(\omega_1 + \omega_2)}{\Gamma\omega_1\Gamma\omega_2} \frac{(y-a)^{\omega_1-1} (b-y)^{\omega_2-1}}{(b-a)^{\omega_1+\omega_2-1}}, a \leq y \leq b, \omega_1 > 0, \omega_2 > 0 \quad (2.2)$$

Figure 2.1 illustrates the difference between a standard Beta distribution and a Four Parameter Beta distribution when the shape is defined by the parameters $\omega_1 = 2$ and $\omega_2 = 5$. The Four Parameter Beta distribution's range extends from -2 to 3 , while the beta distribution range is limited from 0 to 1 . This affects the height of the probability density, where the probability density curve of the Four Parameter Beta distribution is shorter compared to the standard beta distributions, especially near the boundaries of the range

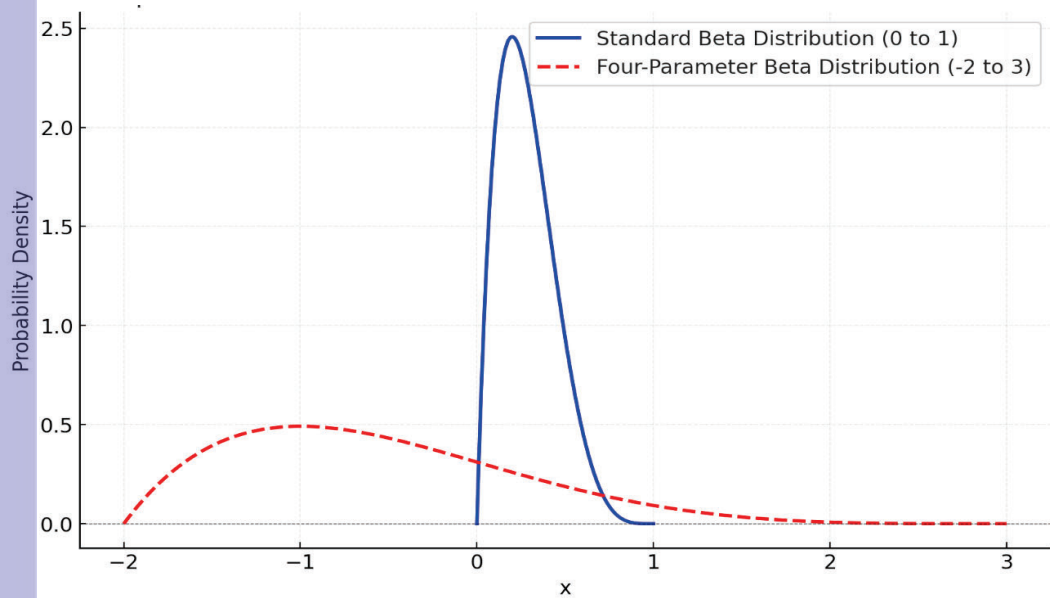


Figure 2. 1 Comparison of a Beta Distribution and a Four Parameter Beta Distribution with $\omega_1 = 2$ and $\omega_2 = 5$

Figure 2.1 illustrates the difference between a standard Beta distribution and a Four Parameter Beta distribution when the shape is defined by the parameters $\omega_1 = 2$ and 5 and $\omega_2 = 2$ and 5 . The shape parameters ω_1 and ω_2 change the skewness and symmetry of the distributions: For a condition of $\omega_1 = 2$ and $\omega_2 = 5$ or a condition where $\omega_1 < \omega_2$ the distribution is skewed to right. Meanwhile when $\omega_1 = 5$ and $\omega_2 = 2$ or a condition where $\omega_1 > \omega_2$, the distribution is skewed to the left. This behavior is consistent for both the standard beta and the Four Parameter Beta distributions

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya atau tinjauan masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University
2. Dilarang mengumunkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Table 2. 1 Formula of Mean, Variance, and Skewness of the Four Parameter Beta Distribution

@Hak cipta milik IPB University		
Parameter	Beta Distribution (*)	Four Parameter Beta Distribution
Relationship Between Distributions		
Mean (μ)	$\mu^* = \frac{\omega_1}{\omega_1 + \omega_2}$	$\mu = a + \frac{\omega_1}{\omega_1 + \omega_2} (b - a)$
Variance (ϕ)	$\phi^* = \frac{\omega_1 \omega_2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 - 1)}$	$\phi = \frac{\omega_1 \omega_2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 - 1)} (b - a)^2$
Skewness (S_k)	$S_k^* = 2 \frac{\omega_2 - \omega_1}{(\omega_1 + \omega_2 + 2)} \sqrt{\frac{\omega_1 + \omega_2 + 1}{\omega_1 \omega_2}}$	$S_k^* = S_k$

2.2. Beta Generalized Linear Model (GLM)

Generalized Linear Models (GLM) models various types of response variables belonging to the exponential family, as introduced by Nelder and Wedderburn (1972). GLMs differ from common linear regression models in that the distribution of the response variable comes from the exponential family, and a transformation process is applied to the expected value of the response variable, which is then linearly linked to the independent variables. Typically, distributions from the exponential family modelled with GLM models include binomial, Poisson, and Gamma distributions.

McCullagh and Nelder (1989) stated that in general GLMs consist of three main components:

- Random Component: A response variable Y , where $E(Y)=\mu$.
- Systematic Component: A linear predictor $\boldsymbol{\eta} = \mathbf{X}^T \boldsymbol{\beta}$, where \mathbf{X} represents the independent variables matrix, and $\boldsymbol{\beta}$ represents the vector of fixed effect coefficients.
- Link Function: A function g that links the expected value $E(Y)$ to the linear predictor $\boldsymbol{\eta}$ such that $g(\mu) = \boldsymbol{\eta}$.

The link function must be invertible and differentiable, and its inverse is referred to as the mean function or denoted as $\mu = g^{-1}(\boldsymbol{\eta})$. In a linear regression, the identity link function $g(\mu) = \mu$ is used.

Ferrari and Cribari-Neto (2004) proposed a regression model for continuous response variables whose values lie within the interval (0,1). It assumes the response variable follows a Beta distribution, giving rise to the Beta regression model. If it is known that $Y_{ij} \sim B(\mu_{ij}, \phi)$, where $i = 1, \dots, q$ and $j = 1, 2, \dots, n_i$ then the Beta GLM is formulated as

$$\boldsymbol{\eta}_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} \quad (2.3)$$

Where $g(\cdot)$ is a link function and $h(\cdot) = g^{-1}(\cdot)$ is the inverse link function. Therefore, the researcher can model the expected value of Y_i as $E(Y_{ij}) = h(\boldsymbol{\eta}_{ij}) = \mu$. Then, the predicted value of Y_{ij} is equivalent to

$$Y_{ij} = h(\boldsymbol{\eta}_{ij}) + \varepsilon_{ij}$$

where ε is the residuals and assumed as $\varepsilon_{ij} \sim N(\mu, \sigma^2)$.

To include an intercept in the model, $X_{i1} = 1$ is commonly set for all i . A frequently used link function in GLM is the logit function, where:

$$g(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) \quad (2.4)$$

Other link functions, such as probit, log-log, and reciprocal, can also be applied.

2.3. Beta Generalized Linear Mixed Model (GLMM)

It has been shown that the mean and variance of a beta distribution can be written as $\mu = \frac{\omega_1}{\omega_1 + \omega_2}$. We can express ω_1 and ω_2 in terms of μ as shown below:

First, we can express ω_1 in terms of μ and ω_2

$$\mu = \frac{\omega_1}{\omega_1 + \omega_2}, \text{ therefore } \omega_1 = \mu(\omega_1 + \omega_2)$$

If we set a precision parameter $\phi = \omega_1 + \omega_2$, then $\omega_1 = \mu\phi$

Next, rearranging $\phi = \omega_1 + \omega_2$ to solve for ω_2 , we will have:

$$\omega_2 = \phi - \omega_1 = \phi - \mu\phi$$

$\omega_2 = (1 - \mu)\phi$
As a conclusion we have shown that $\omega_1 = \mu\phi$ and $\omega_2 = (1 - \mu)\phi$. Furthermore, the variance of the beta distribution can be expressed as follows:

$$\phi = \frac{\omega_1 \omega_2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 - 1)}$$

Substituting with $\omega_1 = \mu\phi$ and $\omega_2 = (1 - \mu)\phi$, we will have:

$$\phi = \frac{(\mu\phi)((1 - \mu)\phi)}{(\phi)^2 (\phi - 1)}$$

by simplify the numerator and canceling out ϕ^2 we will have

$$\phi = \frac{\mu(1 - \mu)\phi^2}{\phi^2 (\phi - 1)} = \frac{\mu(1 - \mu)}{(\phi - 1)}$$

Hence, based on the above beta distribution PDF can be parametrized in terms of mean and precision parameters (ϕ) and it can be written as:

$$f(y|\mu, \phi) = \frac{\Gamma\phi}{\Gamma(\mu\phi)\Gamma(1-\mu)\phi} (y)^{\mu\phi-1} (1-y)^{(1-\mu)\phi-1}, \quad 0 < y < 1$$

The parameters of the probability density function are the average value and the precision value, where $\Gamma(\cdot)$ is the Gamma function, the average value is in the interval $0 < \mu < 1$, $\phi > 0$ is the precision parameter, $E(Y) = \mu = \frac{\omega_1}{\omega_1 + \omega_2}$, $V(Y) = \frac{\mu(1-\mu)}{(1+\phi)}$, and the greater the precision parameter the lesser the variance of Y

Suppose the response variable (Y_{ij}) represent an observation for unit $j = 1, 2, \dots, n_i$ measured in a particular group or area $i = 1, 2, \dots, q$. Furthermore, it is assumed that the Y_{ij} variable depends on K independent variables $X_{ij} = (X_{1ij}, X_{2ij}, \dots, X_{Kij})$ and it is also influenced by groups/area level variability (McCullagh and Nelder 1989). In the context of this study, Y_{ij} represents the paddy productivity of farmer j that is cultivating a paddy field in a certain district or sub district i. It can be assumed that paddy productivity can be influenced by district/subdistrict and farmer variability. District or subdistrict variations are commonly due to unobserved contextual factors, such as differences in agroecological conditions, infrastructure, and support services. These factors are commonly experienced by all farmers in a district/subdistrict, even if not directly measured. In general, a GLMM can be formulated as

$$\eta_{ij} = X_{ij}^T \beta + Z_{ij}^T b_i \quad (2.5)$$

Where X_{ij}^T and Z_{ij}^T are fixed effect matrix design and random effect matrix design. The fixed effect estimators of the independent variables to be used are denoted as β and b_i are the random effect estimators in a particular group/area. Fixed effects account for population-level effects that are consistent across all groups. While a random effect captures group/area specific deviations from the fixed effects, introducing correlations within group/area. The link function that will be used in GLMM is the logit function.

In designing the random effects, Z_{ij} can correspond either to an intercept-only random effect or to the independent variables effect varying by group or known as random slopes. For the intercept-only random effect GLMM model, the random effect b_i affects all individuals within group/area i equally, meaning that the effect does not vary across the independent variable. Since there is only one random effect per group/area (the intercept), the random effects are generally assumed to follow the distribution $b_i \sim \text{iid } N(0, \sigma_i^2)$. Where σ_i^2 represents the variance of the random intercepts. Meanwhile for the random intercept and slope models, both the intercept and slopes are allowed to vary by group/area. In contrast, for models that include both random intercepts and slopes, both the intercept and slope are allowed to vary by group/area. More generally, when the slopes are designed to vary by group, the random intercepts and slopes are commonly modeled as a multivariate normal distribution, $[b_i | \Sigma] \sim$

$N(0, \Sigma)$ where Σ represents the covariance matrix that captures both the variances and the correlation between the random intercept and slope.

Additionally, when we design \mathbf{Z}_{ij} we can also consider nested, crossed, or interaction random effects. In this study we can also model hierarchical levels (districts and farmers). We can consider that within a district often farmers have different experience, skills, motivation, or land quality that will introduce additional variation, known as that individual farmer level random effect (\mathbf{v}_{ij}). Therefore, the GLMM model can be denoted as

$$\eta_{ij} = \mathbf{X}_{ij}^T \boldsymbol{\beta} + \mathbf{u}_i + \mathbf{v}_{ij} \quad (2.5a)$$

Here, \mathbf{u}_i is the random effect for district i and \mathbf{v}_{ij} is the random effect for farmer j in district i . This model is often easier to understand and can be interpreted directly. Nevertheless, it is suitable when we don't consider random slopes.

Vise versa, we can also determine a more simplified model if we are just interested in modelling the response variable aggregated at the group or area level and the random effects are only associated at the group/area level.

$$\eta_i = \mathbf{X}_i^T \boldsymbol{\beta} + \mathbf{b}_i \quad (2.5b)$$

This model assumes that each group has one observation, or we are modeling just the group means (e.g., average productivity in district i).

Selecting on which model to use depends on the data structure and the objective of the analysis. For instance, if we apply the model in the context of developing predictions for crop insurance, we can use the general model in Equation 2.5 if we have individual-farmer level data and we want to predict individual predictions such as productivity, loss, or claim severity. Here, there is also an indication of district or sub district area variation. Next, if we are tailoring crop insurance to specific farmers and setting personalized premiums. Then, we can use the hierarchical model shown in Equation 2.5a. Last, if we have aggregate data (e.g., average loss or average productivity) at a group/area level and predictions are needed at the group/area level then we can use the model shown in Equation 2.5b. This is most suitable for Area-based insurance schemes where payouts are triggered by average group outcomes, not individuals.

After deciding on which GLMM model structure to develop, we can now discuss on how the parameters were estimated. Considering that \mathbf{b}_i is a q -dimensional vector of random effects and assuming that the responses Y_{ij} are conditionally independent given $\mathbf{b}_i, \boldsymbol{\beta}$, and φ , thus the conditional density is defined as

$$f(\mathbf{y}_{ij} | \mathbf{b}_i, \boldsymbol{\beta}, \varphi) = \frac{\Gamma \varphi}{\Gamma(\mu_{ij} \varphi) \Gamma(1 - \mu_{ij} \varphi) (y_{ij})^{\mu_{ij} \varphi - 1} (1 - y_{ij})^{(1 - \mu_{ij} \varphi) - 1}}, \quad 0 < y_{ij} < 1 \quad (2.6)$$

Model parameters in Equation 2.5 can be estimated by maximizing the marginal likelihood function obtained by integrating the combined probability density function $[Y, \mathbf{b}]$. The contribution of the likelihood for the i -th group/individual is

$$f_i(\mathbf{y}_{ij} | \boldsymbol{\beta}, \boldsymbol{\Sigma}, \varphi) = \int \prod_{j=1}^{n_i} f(\mathbf{y}_{ij} | \mathbf{b}_i, \boldsymbol{\beta}, \varphi) f(\mathbf{b}_i | \boldsymbol{\Sigma}) d\mathbf{b}_i \quad (2.7)$$

If it is assumed that groups/areas are mutually independent, then the likelihood function can be simplified as follows

$$L(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \varphi) = \prod_{i=1}^q f_i(\mathbf{y}_{ij} | \boldsymbol{\beta}, \boldsymbol{\Sigma}, \varphi) \quad (2.8)$$

Evaluation of the equation above requires solving the integral q times. For the simpler model with a single random effect the integrals are one dimensional. More generally, the dimension equals the number of random effects in the model which imposes practical limits to numerical methods and approximations required to evaluate the likelihood. The integrals in the example given are up to five dimensions. Thus, the marginal likelihood is maximized by using a sampling-based data cloning algorithm (Bonat *et al.* 2015).

2.4. Four Parameter Beta Regression Model

Modeling data that has a Four Parameter Beta distribution is not commonly found. Previously, most references suggested to conduct a transformation of the response variable from a Four Parameter Beta distribution into a beta distribution (Gray and Alava, 2018) and apply a Beta GLMM (Bonat *et al.* 2015) that has been formulated in Equation 2.3 Nevertheless, transformation process tends to lead to bias in parameter estimates and it is also sometimes complicated to interpret the results. We will explain

Zhou and Huang (2022) show that if it is known that the response variable has a Four Parameter Beta distribution or $Y \sim B(\omega_1, \omega_2, a, b)$, then $W = \frac{Y-a}{b-a}$ will also have a beta distribution with shape parameters ω_1 and ω_2 . Let μ_W and μ_Y denote the mean for W and Y , thus $\mu_W = \frac{\omega_1}{\omega_1 + \omega_2}$ and $\mu_Y = a + \frac{\omega_1}{\omega_1 + \omega_2}(b-a) = a + \mu_W(b-a)$. Next, we can set $\omega_1 = \phi m$ and $\omega_2 = \phi(1-m)$, for $0 < m < 1$ and $\phi > 0$, where ϕ is the precision parameter. A larger precision value will lead to a lower variance of the Four Parameter Beta distribution. The parameters ω_1 and ω_2 in the beta distribution control the shape of the distribution. By introducing m , you can directly link m to the mean of the Four Parameter Beta distribution because this reparameterization process will set the value of μ_W to be equivalent to m . Through this reparameterization we will also directly link ϕ to a scaling factor of the Four Parameter Beta distribution. In most econometric applications it will be more meaningful to express the Four Parameter beta distribution in terms of mean. Furthermore, in analyzing the results of the model, it will be more interpretable to analyze the mean of the distribution rather than analyzing the distribution solely in terms of ω_1 and ω_2 .

Zhou and Huang (2022) then developed the mean Four Parameter Beta mean regression model that can be formulated as follows:

$$g\left[\frac{\text{Mean}[Y_{ij}|X_{ij}]-a}{b-a}\right] = X_{ij}^T \beta \quad (2.9)$$

The left side of the equation can be interpreted as the quantile score for the position of the Mean $[Y_{ij}|X_{ij}]$ in the range (a, b) . As known previously, X_{ij}^T and β are fixed effect matrix design and fixed effect parameters, were $\beta = (\beta_0, \beta_1, \dots, \beta_K)$. The link functions used are Logit, Probit, and log-log functions. In addition to the above approach, researchers can also derive a canonical link function from the developed model. A canonical link function $g(\cdot)$ can be described from the natural parameter θ on an exponential family distribution

2.5. Generalized Mixed Effect Tree (GMET)

The GMET model was first introduced because the GLMM model is not always appropriate for modelling multilevel data, especially if the relationship between the response variables and the independent variables is not linear. GMET general model formulation is as follows:

$$\begin{aligned} \mu_{ij} &= E[Y_{ij}|b_i] \text{ where } i = 1, \dots, q \text{ and } j = 1, 2, \dots, n_i \text{ with } g(\mu_{ij}) = \eta_{ij} \\ \eta_{ij} &= f(X_{ij}) + Z_i b_i \text{ dan } b_i \sim \text{iid } N(0, \sigma_i^2) \end{aligned} \quad (2.10)$$

In this model, the random effect structure is simplified into a random intercept model, and the response variables can be adjusted. So, the model formulation for the Y_i observation can be written as follows:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \text{ with } i = 1, \dots, q \text{ and } j = 1, 2, \dots, n_i \text{ then } p_{ij} = E[Y_{ij}|b_i] \\ \text{logit}(p_{ij}) &= f(X_{ij}) + b_i \text{ with } b_i \sim \text{iid } N(0, \sigma_i^2) \end{aligned} \quad (2.11)$$

Where i is the group/area and q is the total number of group/area, n_i is the number of observations within the i -th group and $\sum_{i=1}^q n_i = J$. In addition, η_{ij} is the n_i -dimensional linear

predictor vector and $g(\cdot)$ is the link function. Finally, \mathbf{X}_{ij} is the $n_i \times (K + 1)$ matrix of fixed-effects regressors of observations in group i , \mathbf{Z}_{ij} is the $n_i \times q$ matrix of regressors for the random effects, \mathbf{b}_i is the $(q + 1)$ -dimensional vector of their coefficients and Σ_i is the $q \times q$ within group covariance matrix of the random effects. As in a GLMM, \mathbf{b}_i and $\mathbf{b}_{i'}$ are independent for $i \neq i'$. Fixed effects are identified by a regression tree model associated with the entire population, whereas random effects are identified by group-specific parameters.

The dataset used in Fontana *et al.* (2021) consist of PoliMi data, with information on 18,612 undergraduate students who enrolled at the Politecnico di Milano in Italy between the academic years 2010/2011 and 2013/2014. These students are nested in 19 different departments. The response variable used has a Bernoulli distribution, indicating whether students either dropped out or successfully graduated.

The algorithm used in GMET model can be done based on the following steps (Fontana *et al.* 2021):

1. Initialize the random effect estimate (\mathbf{b}_i) to zero.
2. Estimate the target variable (μ_{ij}) using the GLM model given that the fixed effect independent variables are $\mathbf{X}_{ij} = (X_{1ij}, X_{2ij}, \dots, X_{Kij})^T$ for $i=1, \dots, q$ and $j = 1, 2, \dots, n_i$. We will obtain an estimate of $\hat{\mu}_{ij}$.
3. Construct a regression tree that will approximate the function $f(\mathbf{X}_{ij})$ in Equation 2.8 by using $\hat{\mu}_{ij}$ as the response variable and $\mathbf{X}_{ij} = (X_{1ij}, \dots, X_{Kij})^T$ as the independent variable. We will then obtain L terminal nodes (\mathbf{R}_ℓ), where $\ell = 1, 2, \dots, L$. For each observation i , will belong to one of the terminal nodes, that will be described by a certain set of independent variables \mathbf{X}'_{ij} . Hence, we can define a set of indicator $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ for $\ell = 1, 2, \dots, L$. Then, $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ will be given the value of 1 if observation i , belongs to the ℓ -th terminal node and 0 otherwise. Therefore, the indicator $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ is in the form of a dummy variable.
4. Fit the mixed effects model in Equation 2.8, using Y_i as the response variable and the set of indicator variables $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ as the fixed effect independent variables. For $i = 1, \dots, q$ and $j = 1, 2, \dots, n_i$ we will have $g(\mu_{ij}) = I(\mathbf{X}_i \in \mathbf{R}_\ell)\beta_\ell + \mathbf{Z}_{ij}^T \mathbf{b}_i$. Here, β_ℓ represents the fixed effects for each terminal node and $\mathbf{Z}_{ij}^T \mathbf{b}_i$ captures the random effects. So, we obtain $\hat{\mathbf{b}}_i$ from the presumed model.
5. Replace the predicted response values at each terminal node \mathbf{R}_ℓ of the regression tree with the predicted response values from the mixed effects model performed in step 4. Thus, we obtain

$$g(\mu_{ij}) = g(\hat{\gamma}_\ell) + \mathbf{Z}_{ij}^T \mathbf{b}_i$$

$$\hat{\gamma}_{ij} = g^{-1}(\mu_{ij}) = g^{-1}(g(\hat{\gamma}_\ell) + \mathbf{Z}_{ij}^T \mathbf{b}_i) \quad (2.12)$$

As the updated prediction of the response variable, where $g(\hat{\gamma}_\ell)$ is the estimated value of the fixed effects at the terminal node \mathbf{R}_ℓ and $\mathbf{Z}_{ij}^T \mathbf{b}_i$ accounts for the subject-specific random effects

2.6. Generalized Mixed Effect Random Forest (GMERF)

The Generalized Linear Mixed Effect Random Forest (GMERF) model was first introduced by Pellagatti *et al.* (2021). GMERF extends the GMET model by using Random Forest (RF) instead of regression trees to estimate fixed effects in the mixed-effect models. The core concept of GMERF is to train multiple GMET models, each using a different dataset generated by the bootstrap method from the original data. Each resulting tree is then tested using

selected independent variables, and the final Random Forest prediction is an aggregate of each individual tree's predictions.

In GMERF, fixed effects are estimated by the Random Forest model, while random effects are identified through random parameter estimators specific to each group or area, estimated by GLMM. When applying this method, as both fixed and random effects are generally unknown, researchers employ an iterative, alternating method until convergence is achieved for both the Random Forest and random effect estimations. The Random Forest method used is based on Breiman's (2001) algorithm, where the algorithm can be summarized in few steps:

- i. Generate a bootstrap sample (i) of dimension n from the training set .
- ii. For each $i=1 \dots m$ grows the i -th decision tree by randomly selecting q features without repetition and then splitting the node based on the best feature. Supposing that the Feature set contains Q features, a commonly used value for setting the dimension of the features' subset is $q = \sqrt{Q}$
- iii. Combining the prediction by a suitable combination function (such as Majority vote for classification problems and by averaging prediction in regression problems)

In general, GMERF model is an extension of GMET model in Equation 2.10 and can be formulated as follows:

$$\mu_{ij} = E[Y_{ij} | \mathbf{b}_i] \text{ where } i = 1, \dots, q \text{ and } j = 1, 2, \dots, n_i$$

$$\begin{aligned} g(\mu_{ij}) &= \eta_{ij} \\ \eta_{ij} &= f^*(X_{ij}) + \mathbf{Z}_{ij}^T \mathbf{b}_i \end{aligned} \quad (2.13)$$

where $f^*(X_{ij})$ represents the relationship between X and Y , which is not linear but assumed to follow some tree form structure, built through the ensemble-based Random Forest (RF) method (Breiman, 2001). Similar to GMET, GMERF assigns local fixed effect estimates (β) to each terminal node, while the random effects (\mathbf{b}_i) remain global, as they capture group-level deviations that apply consistently across all trees and iterations in the random forest process. Thus, determining the values of the random effects can be challenging due to the large number of values generated across the numerous iterations in the random forest process. Strategies for summarizing or aggregating these values are essential to ensure interpretability and effective utilization of the model outputs. Nevertheless, this method's merit is that it retains the capacity to model hierarchical data while satisfying the flexibility and predictive power typical of RF.

The case study used in Pellagatti *et al.* (2021) employs the same dataset as the one used in Fontana *et al.* (2021) for the development of the GMET model. Authors compared the predictive power of both models using the test dataset and reveals a slight accuracy advantage for GMERF (0.908) over GMET (0.878), aligning with GMERF's goal to improve GMET's predictive accuracy.

2.7. Mean Square Error (MSE) of Parameter Estimates

Inference for estimating the Four Parameter Beta GLMM model can be done using Mean Square Error (MSE). MSE calculations can be done analytically or empirically using bootstrap simulation. However, the analytical MSE calculation process is generally quite tricky for the GLMM model. In the GLMM model, the analytical form of MSE generally cannot be calculated explicitly. The problem often encountered when calculating MSE analytically is the existence of nonlinearity in the model, namely the existence of a non-linear relationship between the average value of the response variable and its linear predictor. In addition, it is often found that there are violations of the assumption of normality. Therefore, MSE calculations will be performed using approximation methods, such as Bootstrap. The bootstrap steps for estimating MSE are as follows

Step 1 Determine the initial values of Σ_i and β , where Σ_i represents the covariance matrix of the random effects for each group, and β represents the fixed effect parameters.

Step 2 Estimate $\hat{\theta} = \{\hat{\beta}_x, \hat{\beta}_z, \hat{\sigma}_v^2\}$

Step 3 Generate bootstrap samples $u^*_i \sim N_q(0, \Sigma_i)$, where u^*_i are random effects sampled from a multivariate normal distribution with mean 0 and covariance matrix Σ_i . In bootstrapping, u^*_i represents a sample from the random effect distribution that is used to generate new data points for the purpose of estimating uncertainty.

Step 4 Assemble the bootstrap samples to generate the bootstrap response values and prediction where $\mu_{ij}^* = \mathbf{x}_{ij}^T \hat{\beta}_x^* + \mathbf{z}_{ij}^T \hat{\beta}_z^* + u^*_i$ and $y_{ij}^* = \mu_{ij}^* + u^*_i$

Here, μ_{ij}^* is the predicted value for the response variable, incorporating both fixed effects and random effects, while y_{ij}^* is the final observed response variable after incorporating the random effect u^*_i

Step 5 Estimate the parameters $\hat{\beta}_x^*, \hat{\beta}_z^*, \hat{\sigma}_v^{2*}$

Step 6 Calculate the value of $\hat{\mu}_{ij}^*$ using the data $y_{ij}^*, \mathbf{x}_{ij}, \mathbf{z}_{ij}$

Step 7 Repeat Steps 3.–6 B times to generate B bootstrap replications.

Step 8 Calculate the MSE parametric bootstrap estimator for $i = 1, 2, \dots, m$ with the formula

$$\text{mse}_i^{\text{PB}} [\hat{\mu}_{ij}^*] = \frac{1}{B} \sum_{b=1}^B (\hat{\mu}_{ij}^* - \mu_{ij}^*)^2 \quad (2.14)$$

2.8. Paddy Productivity Prediction for Area Yield Index (AYI) Crop Insurance

Area Yield Index Insurance (AYI) is productivity-based agricultural insurance. Generally, the rules for paying compensation for AYI are when the contract is determined, and the insurance company will estimate the productivity target value or Critical Yield Index (y_c) of an area. If the average productivity of a sub-area (\bar{y}_i) is less than the y_c value, then all farmers in that area will receive compensation (Skees *et al.* 1997). Therefore, the indemnity function of AYI can be formulated as follows:

$$\text{Indm} = \max(y_c - \bar{y}_i, 0) \cdot \text{SI} \cdot L_{ij} \quad (2.15)$$

Where L_{ij} is the amount of land cultivated by a farmer j in a sub area i . For simplicity we can set L_{ij} at 1 hectare (Ha). While **SI** is the sum insured fixed at 6,00,000 IDR per Ha. This is the minimum amount needed to continue farming in the next season if risks occur (Kusumaningrum *et al.* 2021).

The results of previous simulation studies showed that AYI has the potential to overcome many of the weaknesses commonly found in MPCI (Kusumaningrum *et al.* 2021) and has been successfully adopted in various countries such as India, the Philippines and Thailand. The challenge faced in developing AYI insurance contracts is the need for an accurate paddy productivity prediction model at the area and sub-area levels. The prediction of paddy productivity in an area needs to be determined before the planting process so that the insurance company can determine the value of y_c , which will be used to determine the amount of the AYI insurance premium.

Forecasting of paddy productivity, in general, has been carried out by many researchers before, which can be seen in Figure 2. 3. More specifically, productivity forecasting for AYI insurance contracts is generally carried out based on official statistical data issued by ministries or other trusted institutions, and the forecasting method used is the smoothing method. Or time series data modelling, such as Double Exponential Smoothing and ARIMA (Skees *et al.* 1997). The first problem is that this approach only considers the effect of time and does not consider the influence of other variables, so it is not appropriate when applied to the heterogeneous conditions of agricultural land in Indonesia (Haryastuti *et al.* 2021). Therefore, it is necessary

to predict productivity by considering other independent variables and considering the effect of variation between areas.

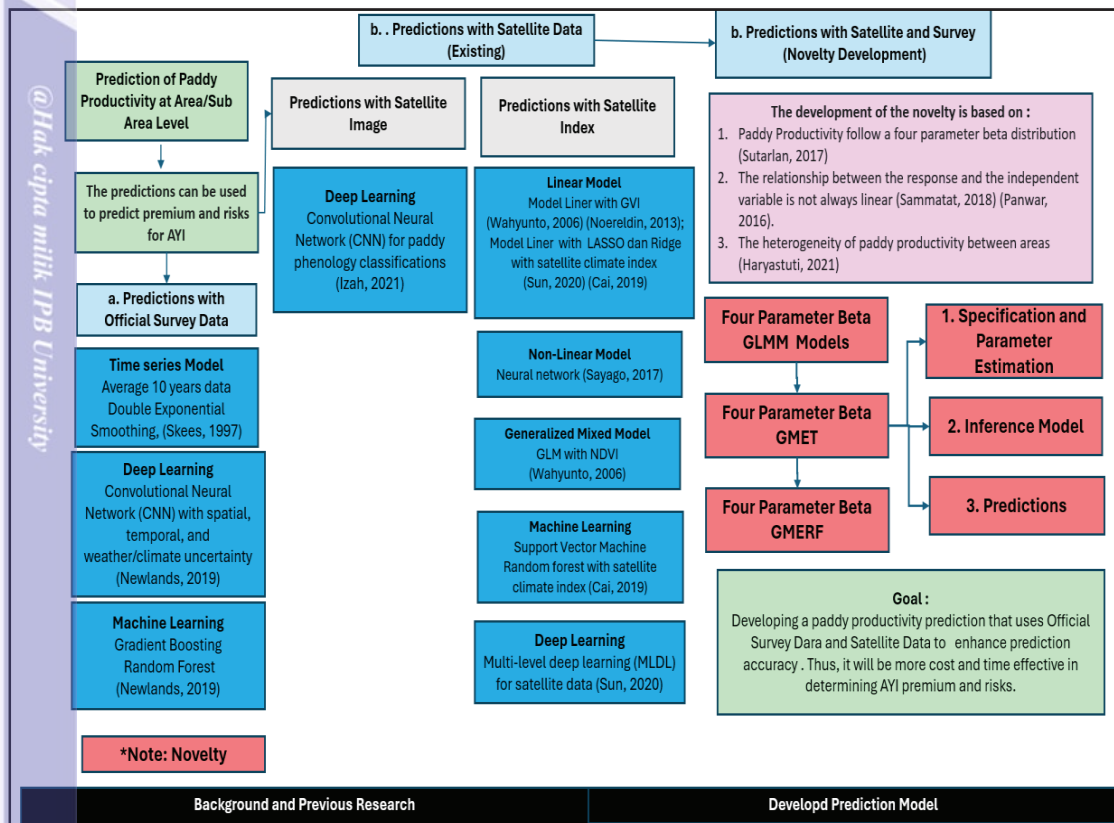


Figure 2. 3 Research Road Map of Paddy Productivity Prediction

2.9. Crop Cutting Experiment (CCE)

The CCE is a routine activity conducted by SI aimed at collecting data on the productivity of food crops, particularly paddy. This survey is conducted simultaneously within the harvesting period of farmers. The data collection method for productivity in this survey involves direct measurement on selected plots (2.5 m x 2.5 m) and interviews with sampled farmers. Indonesian's productivity official figures related to production and productivity are derived from the CCE. Typical CCEs are conducted annually in three phases: Sub round I (January–April), Sub round II (June–August) and Sub round III (October - December). The unit level of observation is an area plot of ready-to-harvest paddy.

Starting from sub-round 3 (September–December) in 2018, the CCE survey for paddy was conducted using the Area Sampling Frame (ASF) approach. ASF is a new method developed by SI in collaboration with various institutions to improve the data collection method for harvested area, which was previously based only on eye estimates. This method is conducted in a more objective and modern manner by involving technological devices, resulting in more accurate data collection (<https://rembangkab.bps.go.id/>). The CCE produces an average paddy productivity value in numerical form, which will be used as the response variable in this study.

2.10. Farmer Survey Data

Farmer surveys were also conducted during CCE. Besides coordinates of the plot area there were also thirty-seven questions asked, which include the type of land, the planting method, planting system, the type of activities that increased productivity, the quantity and variety of seeds used, the application of pesticides, the amount of fertilizer applied, information on government supports, and furthermore. These variables will be considered as independent/predictor variables in our developed model and it can be seen in Table 2. 2. Variables in the Farmer Survey Data.

Table 2. 2. Variables in the Farmer Survey Data (a)

Variable	Variable Name	Unit Measurement/ Category
Paddy Productivity	y	Tons/Ha
Pest Attacks This Year	x_1	1 = Heavy, 2 = Medium, 3 = Light 4 = Not Affected
Pest Attacks Last Year	x_2	1 = Heavy, 2 = Medium, 3 = Light 4 = Not Affected
Impact of Climate Change This Year	x_3	1 = Affected, 2 = Not Affected
Impact of Climate Change Last Year	x_4	1 = Affected, 2 = Not Affected
Water Sufficiency This Year	x_5	1 = Not Enough, 2 = Sufficient, 3 = More than Enough
Water Sufficiency Last Year	x_6	1 = Not Enough, 2 = Sufficient, 3 = More than Enough
How to Handle Pest	x_7	0 = No Actions, 1 = Agronomist, 2 = Mechanical, 3 = Biological, 4 = Chemical
Food Plant Type	x_8	1 = Irrigated Lowland Rice, 2 = Non-Irrigated Lowland Rice
District	x_9	1 = West Kotawaringin, 2 = East Kotawaringin, 3 = Kapuas, 4 = Sukamara, 5 = Lamandau, 6 = Seruyan, 7 = Katingan, 8 = Pulang Pisau, 9 = Gunung Mas, 10 = East Barito
Strata	x_{10}	1 = S1, 2 = S2, 3 = S3
Sub Districts	x_{11}	29 Sub Districts
Sub Round	x_{12}	1 = Irrigated Lowland Rice, 2 = Non Irrigated Lowland Rice, 3 = Paddy Field
Land Type	x_{13}	1 = Irrigated Rice Field, 2 = Rain-Fed Rice Fields, 3 = Tidal Swamp Rice Fields, 4 = Rawa Lebak Rice Fields, 5 = Not a Rice Field
How to Plant	x_{14}	1 = Monoculture, 2 = Mixed
Efforts to Increase Production	x_{15}	1 = Government Assistance, 2 = Non-Government Assistance
Plant Variety	x_{16}	1 = Hybrid, 2 = Inbred
Join Farm Group	x_{17}	1 = Yes, 2 = No
Last Year Productivity	x_{18}	1 = Different Plants, 2 = Lower, 3 = No Change, 4 = Higher

Table 2. 3 Variables in the Farmer Survey Data Continued (b)

Variable	Variable Name	Unit Measurement/ Category
Last Year Productivity in Area	x_{19}	1= Lower, 2 = No Change, 3 = Higher
Number of seeds used/area of similar plants	x_{20}	Plants/Ha
Urea	x_{21}	Kg/Ha
TSP/SP36	x_{22}	Kg/Ha
KCL	x_{23}	Kg/Ha
Solid Organic Fertilizer /Compost	x_{24}	Kg/Ha
Liquid Organic Fertilizer	x_{25}	Kg/Ha
ZA	x_{26}	Kg/Ha
Seed Assistance	x_{27}	1= Yes, 2 =No
Source of Seed Assistance	x_{28}	0 = None, 1= Central Gov, 2= Local Gov, 3 = Private Sector/BUMN, 4 = Individual, 5 = Others
Seed Assistance In Line with Planting Season	x_{29}	0 = None, 1= Yes, 2 =No
Seed Assistance in Line With Plant Type	x_{30}	0 = None, 1= Yes, 2 =No
Seed Assistance Planted	x_{31}	0 = NA, 1 = All, 2 = Some, 3 = No
Fertilizer Assistance	x_{32}	1= Free, 2 = Subsidized, 3 = None
Origin of Fertilizer Assistance	x_{33}	0 = None, 1= Central Gov, 2= Local Gov, 3 = Private Sector/BUMN, 4 = Individual, 5 = Others
Fertilizer Assistance On Time	x_{34}	0 = None, 1= Yes, 2 =No
Farmer Group's Machine Assistance	x_{35}	0 = None, 1= Yes, 2 =No
Type of Machine Assistance	x_{36}	0 = None, 1= Water Pump, 2=Tractor, 3=Harvesting Tool, 4=Others
Machine Assistance Useful	x_{37}	0 = None, 1= Yes, 2 =No

2.11. Sentinel 2A Satellite Data

Sentinel 2 satellite data consist of a two-satellite constellation, Sentinel 2A and Sentinel 2B, which orbit the poles on a sun-synchronous orbit at an altitude of 786 km. First launch in 2015, these two identical satellites operate on opposite sides of the orbit and positioned 180 degrees apart from each other (Figure 2. 4). These two identical satellites operate simultaneously to achieve frequent revisits of five days at the equator. The satellites follow a Sun-synchronous orbit at 786 km altitude with a 10:30 a.m. descending node, balancing minimal cloud cover with optimal sun illumination. This schedule aligns with Landsat and SPOT, facilitating the integration of Sentinel-2 data with historical images for long-term time series. Sentinel-2 systematically captures observations over land and coastal areas from -56° to 84° latitude, including islands larger than 100 km², smaller islands within 20 km of coastlines, inland water bodies, and closed seas. Additional observations are made over specific calibration sites like Antarctica's DOME-C (Drusch *et al.* 2012)

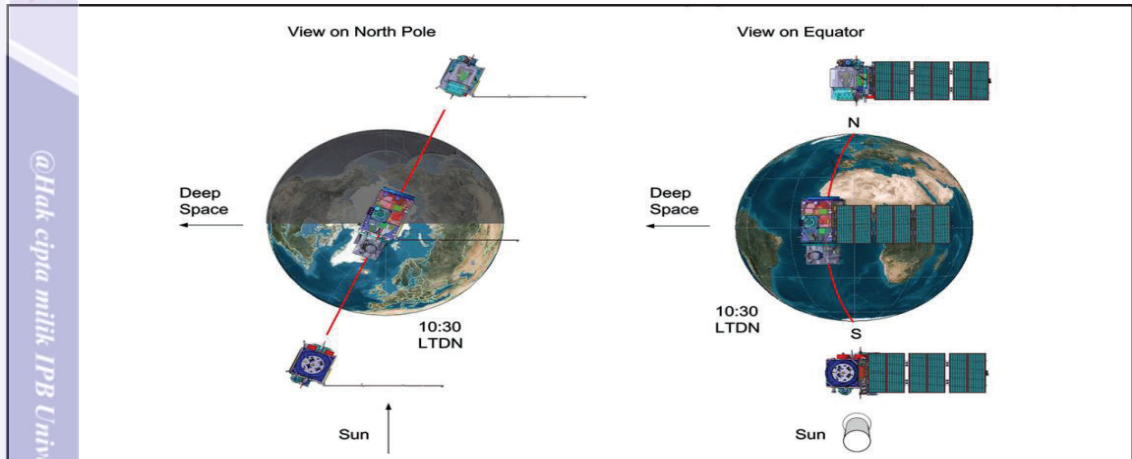


Figure 2. 4 Sentinel-2 Satellite Orbital Configuration (Drusch *et al.* 2012)

Sentinel 2 satellites are considered as medium-resolution satellites with a temporal resolution of 10 days for one satellite or 5 days with both satellites. The two satellites are nearly identical in design, they operate together to achieve the mission's goals more effectively. Sentinel 2A satellite focuses on systematically covering land and coastal areas to collect multispectral data. While Sentinel 2B satellite is a complementary satellite to enhance revisit frequency. It works in tandem with Sentinel-2A to halve the revisit time of 5 days at the equator. Both are equipped with the Multispectral Instrument (MSI) that has thirteen spectral bands ranging from visible to shortwave infrared. In general, these satellites can be used for operational observations, such as land cover mapping, land change detection mapping, and geophysical variable monitoring.

Table 2. 3 Spectral Wavelength Characteristics and Spatial Resolution of Sentinel 2 Bands

Band	Spectrum	Wavelength(μm)	Spatial Resolution (m)
1	Costal Aerosol	0.433 – 0.453	60
2	Blue	0.458 – 0.523	10
3	Green	0.543 – 0.578	10
4	Red	0.650 – 0.680	10
5	Vegetation Red Edge 1	0.698 – 0.713	20
6	Vegetation Red Edge 2	0.733- 0.748	20
7	Vegetation Red Edge 3	0.765 – 0.785	20
8	NIR	0.785 – 0.900	10
8a	Vegetation Red Edge 4	0.855 – 0.875	20
9	Water Vapor	0.855 – 0.875	60
10	SWIR Cirrus	1.365 – 1.385	60
11	SWIR1	1.565 – 1.655	20
12	SWIR2	2.100 – 2.280	20

Source : ESA (2015)

Sentinel 2A has thirteen bands, including 4 bands with a 10 m resolution (B2, B3, B4, B8), 6 bands with a 20 m resolution (B5, B6, B7, B8A, B11, B12), and 3 bands with a 60 m

spatial resolution (B1, B9, B10), with a radius width of 290 km. More detail information of these bands can be seen in Table 2.3.

In addition to using single-band analysis on Band 4 (the red band) and Band 8 (the near-infrared band), the analysis of Sentinel 2A imagery can also be conducted by calculating vegetation indices. The vegetation indices analyzed in this study include the Normalized Difference Vegetation Index (NDVI) that can be calculated as

$$NDVI = \frac{NIR - Red}{NIR + Red} \quad (2.16)$$

Where, NIR and Red correspond to the near-infrared band (Band 8) and the red band (Band 4), respectively, with the background brightness correction factor ($L = 0.5$).

In this research Sentinel-2 Satellite data mentioned above is obtained by writing specific scripts on Google Earth Engine (GEE) at <https://code.earthengine.google.com/>. The steps needed to download the Sentinel-2 Satellite data are as follows:

1. Write the sample coordinate points of the selected CCEs in CSV format.
2. Convert the CSV file from Step 1 into a shapefile using ArcGIS or QGIS.
3. Upload the shapefile from Step 2 to the Assets section of a GEE account.
4. Write commands to acquire imagery based on the sample coordinates within the desired time range.
5. Perform cloud masking to improve data quality.
6. Extract imagery values into bands Band 4 and Band 8.
7. Calculate spectral index values using formulas provided in Equation 2.16.

The script code used to extract the Sentinel 2A data used in this research can be seen in Appendix 1,

Sentinel-2A satellite offers significant potential for monitoring and estimating paddy productivity. It is highly suitable for estimating paddy productivity due to its key features, which are : (1) Its high spatial resolution (10 m, 20 m, and 60 m) allows precise monitoring at the plot level, (2) while its frequent revisit time of 10 days, reduced to 5 days when paired with Sentinel-2B, ensures timely updates for tracking crop growth stages, (3) the satellite's spectral bands, such as Red (B4), Near Infrared (B8), and vegetation index calculations such as NDVI provide insights into crop health and biomass. Last (4) the wide coverage for large areas, Sentinel-2A effectively monitors extensive paddy fields and detects water bodies, soil moisture, and vegetation stress, aiding in yield estimation and early issue identification.

III. Modeling the Four Parameter Beta Distribution Response Variable Through a Transformation Process

3.1. Introduction

Modeling response variables bounded between specific intervals, or also known as Four Parameter Beta distribution, is becoming a challenge in various fields. The Four Parameter Beta distribution is characterized by its flexibility in capturing a wide range of shapes, making it ideal for modeling proportions, rates, and other bound continuous variables. Despite this flexibility, directly modeling the Four Parameter Beta distributed response variables introduces complexities. GLMM models for a Four Parameter Beta distribution response variable have never been applied in previous research. Therefore, in this chapter, we address these challenges by applying a transformation process to simplify the modelling of the Four Parameter Beta distributed response variables. Following the transformation, we can apply the Beta GLM or Beta GLMM model that has been introduced in sub chapter 2.3 and 2.4. This model was developed by Bonat *et al.* (2015) and it is particularly well-suited for data with hierarchical structures and accommodates both fixed and random effects, enabling us to account for group/area-specific variability. By utilizing this approach, we aim to leverage prediction accuracy, ensuring more reliable results. However, it is also crucial to recognize potential limitations, such as bias introduced during the transformation process, which will be analytically assessed to ensure the validity of the model's outcomes. Besides analytically assessing the potential bias of the model transformation process we have also applied the approach for predicting paddy productivity.

Paddy is a critical crop in Indonesia, essential for food security, economic planning, and agricultural sustainability. Accurate paddy productivity predictions can mitigate harvest shortfalls, support crop insurance policies, optimize resource use, and stabilize the economy. Predicting paddy productivity is also crucial in developing AYI crop insurance policies, which is a productivity-based index agricultural insurance policy, which will be further discussed in Chapter 5. Current estimates rely on the Crop Cutting Experiment (CCE) conducted by Statistics Indonesia using a tiling tool on randomly selected plots provided by a regional sampling framework (Ardiansyah *et al.* 2021). The weight of the tile yields is recorded in the questionnaire, followed by interviews with the farmers to find other matters related to the respondent's farming business. However, there is often a significant gap between CCE forecasts and actual productivity, with discrepancies of up to 2.8 tons/hectare reported in 2022 (Marnawati, 2023).

To improve paddy productivity prediction accuracy, studies have incorporated satellite data, such as the Vegetation Index from Sentinel 2A imagery data (Said *et al.* 2015). Vitasari *et al.* (2017) succeeded in estimating paddy harvest and vegetation index. The accuracy of predictions is 73.01%. Previous studies also show that climate change has an impact on productivity and the growth of big data has accelerated further development of potential prediction models (Cedric *et al.* 2022). Thus, we incorporate both farmer survey data and satellite Sentinel 2A data as independent variables in the prediction model to increase prediction accuracy.

Until recently, prediction models applied methods such as regression (Son *et al.* 2013), non-linear regression (Son *et al.* 2013; Sammatat and Lekdee 2018), time series models (Skees *et al.* 19797; Sari and Sukojo, 2015), image classification (Vitasari *et al.* 2017; Said *et al.* 2015), and machine learning (Sun *et al.* 2020; Clauss *et al.* 2018; Newlands *et al.* 2019; Cedric *et al.* 2022; Klompenburg 2020). However, most of these approaches didn't take into consideration

the distribution of paddy productivity and variability of paddy productivity across different areas. Thus, in this study we aim to tackle these limitations.

It is known that the distribution of paddy productivity is commonly skewed and bounded to a certain minimum and maximum value, typically following a four-parameter beta distribution. If we don't apply the correct distribution, prediction results may not be accurate and applicable to the case. For example, in Indonesia, the maximum paddy productivity is around 7.8 tons per hectare (Sulaeman *et al.* 2024). Thus, when we use prediction models based on the assumption of normal distributed conditions, predictions can exceed the maximum paddy productivity value. Furthermore, it is also found that the agricultural land conditions in Indonesia are heterogeneous (Haryastuti *et al.* 2021), a condition often observed in other countries as well. Therefore, we need to consider the effect of variation between areas and incorporate a random effect in our prediction model. Therefore, we have proposed a prediction model based on the Four Parameter Beta Generalized Linear Mixed Model (GLMM) model developed through a transformation process.

To evaluate the performance of the Four Parameter Beta GLMM model developed through a transformation process in predicting paddy productivity, an empirical study in Central Kalimantan Province and Karawang Regency, Indonesia was done. We have used the CCEs, farmer survey data, and Sentinel 2A satellite data of 2020 in Central Kalimantan and Karawang to test the model. Central Kalimantan was selected because it has many areas of peatland, which can lead to variations in paddy productivity. Meanwhile, Karawang has a more stable high level of rice productivity, often reaching more than 6 tons per hectare in one planting season. This makes Karawang one of the most productive rice-producing regions in Indonesia, significantly contributing to national food security. Early studies also point out that the distribution of paddy productivity has a four-parameter beta distribution. Therefore, Central Kalimantan and Karawang are suitable areas and further on the purposed model can be upscaled in other areas. In summary, the contributions of this chapter are:

1. Increase the accuracy of paddy productivity by developing a Four Parameter Beta GLMM based on a transformation process.
2. Apply the Four Parameter Beta GLMM model based on a transformation process on farmer survey and Sentinel 2A satellite data to predict paddy productivity.
3. Evaluate the goodness of fit model and prediction accuracy of the four-parameter beta GLMM to find the most appropriate paddy productivity prediction model.
4. Analyze the best paddy productivity prediction model and deliver comprehensive recommendations.
5. Analytically analyze potential parameter bias and prediction bias caused by the transformation process

Therefore, the remaining sub chapters are structured and described as: First, in sub chapter 3.2, which is the Proposed Method section, we provide a detailed explanation about the transformation process needed before developing the model and how to calculate the prediction values. Next, in sub chapter 3.3 we will introduce the empirical case study dataset, and we also elaborated how the Four Parameter Beta GLMM models were developed based on the transformation process using the, variable selections, key performance indicators, and prediction process. Sub chapter 3.3 shows the Results and Discussion section, presents the findings and an in-depth discussion of the applied studies. Third, sub chapter 3.4, which is the Potential Transformation Bias section, analytically shows and explains the effect of the transformation process. Fourth, in sub chapter 3.5, we will discuss the model's limitations theoretical and application wise. We also give suggestions for further research development. Last, in sub chapter 3.6, which is the Conclusion and Recommendation section, we will conclude and summarize the overall findings and potential areas for further developments.

3.2. Method

3.2.1. Transformation Process for a Four Beta Parameter Distribution

The four-parameter beta distribution is a probability distribution that is defined within the interval (a, b) , where a is the minimum value and b is the maximum value. It is a flexible family of distributions that can model a wide range of shapes, making it useful in various applications (Zhou and Huang 2022). The four-parameter beta distribution has a probability density function that can be seen in Equation 2.2. Meanwhile, the Beta distribution is also a continuous variable distributions that has an interval of $(0, 1)$ and the probability density function is shown in Equation 2.1.

Zhou and Huang (2022) showed that if Y follows a four-parameter beta distribution or denoted as $Y \sim B(\omega_1, \omega_2, a, b)$, then we can use the transformation formula below:

$$Y^* = \frac{(Y-a)}{(b-a)} \quad (3.1)$$

Thus, Y^* will follow a Beta distribution with shape parameters ω_1 and ω_2 . Therefore, for modeling the Four Parameter Beta GLMM, we can apply this transformation process by using Equation 3.1. This transformation maps the actual response variable y that has an interval $[a, b]$ to Y^* with the interval $(0,1)$.

3.2.2. Predicted Values

Transforming the four-parameter beta distribution to a standard beta distribution in a sample data set helps us work with a simpler form. This transformation process enables us to model and predict data that has a Four Parameter Beta distribution by applying GLM or GLMM models that have been explained in sections 2.2 and 2.3. We can use a GLMM model when the response variable is measured in a particular group/area ($i=1, 2, \dots, q$) and $j = 1, 2, \dots, n_i$ or apply a GLM model if the data structure is more straightforward.

In a beta GLM/GLMM, the response variable y_{ij}^* lies within $(0,1)$, so the inverse link function $g^{-1}(\eta_{ij})$ maps the linear predictor η_{ij} to the predicted mean μ_{ij} of the beta distribution or denoted as

$$\mu_{ij} = g^{-1}(\eta_{ij})$$

when using the logit link function, then

$$\mu_{ij} = \frac{1}{1 + e^{-\eta_{ij}}}$$

The mean μ_i derived from the inverse link function above represents the expected value of y_{ij}^* . Therefore, the predicted value can be interpreted as:

$$\widehat{y}_{ij}^* = \mu_{ij}$$

In a beta GLMM, the beta distribution includes a dispersion parameter ϕ , which influences the variance of y_{ij}^* , thus $V(\widehat{y}_{ij}^*) = \frac{(\mu_{ij}(1 - \mu_{ij}))}{(1 + \phi)}$. While this does not directly affect the predicted mean μ_i , it is important for understanding the uncertainty of predictions.

After we have obtained the predicted value \widehat{y}_i^* we need to conduct a back transformation process where

$$\widehat{y}_j = \widehat{y}_j^*(b - a) + a$$

Where \widehat{y}_j is the predicted response variable that will follow a Four Parameter Beta distribution, this transforming process has advantages in terms of simplifying calculations, providing uniformity, normalizing data, and simplifying theoretical properties. However, it also comes with disadvantages such as bias. The bias is caused by the fact that a Four Parameter Beta distribution is a more flexible distribution that can fit a wider range of data shapes. In particular,

the four-parameter beta distribution allows for skewness and kurtosis to vary independently, whereas the Beta distribution assumes that skewness and kurtosis are fixed. One way to evaluate this potential bias is by examining the accuracy of the fitted values of the developed model, which will be discussed further on in the Key Performance Indicator section. Additionally, a systematic review on the effect of the transformation process can be seen in sub chapter 3.4.

3.3. Empirical Case Study

3.3.1. Empirical Case Study Dataset

In this research we have used the paddy productivity data obtained from 534 CCEs in Central Kalimantan Province and 363 CCEs in Karawang Regency, West Java (2020) collected by Statistics Indonesia (SI), which can be seen in Figure 3. 1. SI conducted this survey all year round in these two selected areas using a 2.5 x 2.5-meter tile tool on randomly selected plots (Adriansyah *et al.* 2021). The weight of the tile yields is recorded in the provided questionnaire, followed by interviews with the respondent farmers to find other matters related to the respondent's farming business. However, as stated above, it is known that there is a large gap between the prediction based on CCEs and the actual average national productivity. In 2022, the gap is estimated to reach at least 2.8 tons/ha (Marnawati , 2022). Therefore, to improve predictions, we will combine the farmer survey data and incorporate satellite Sentinel 2A data in this research.

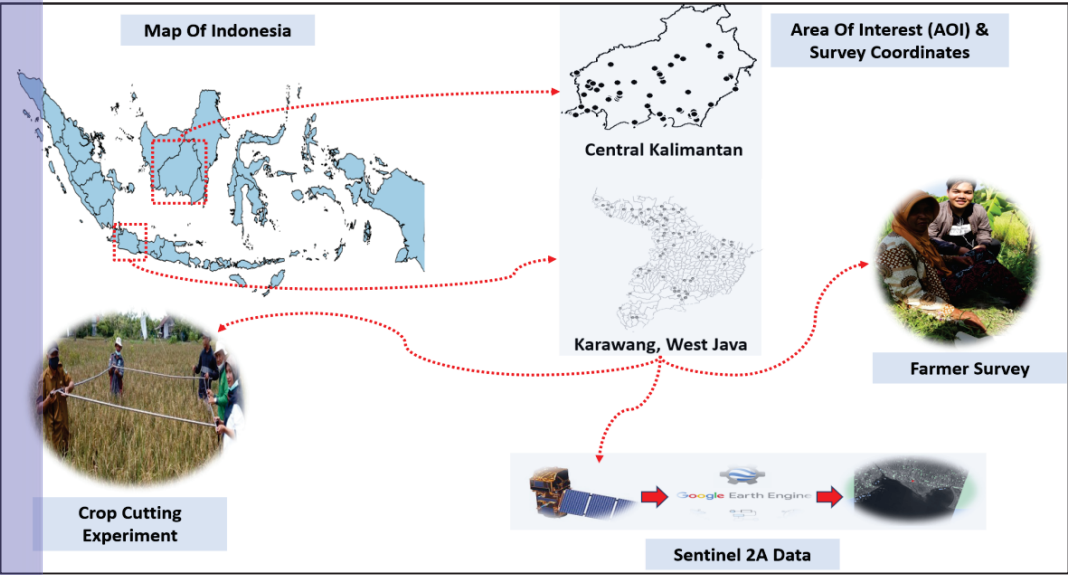


Figure 3. 1 Distribution of Survey Plots and Data Used in Central Kalimantan and Karawang (2020)

The list of all independent variables or predictors collected from the farmers survey can be seen in Table 2. 2. The farmer survey has given insights of many key factors such as pest attacks, pest management, impact of climate change, and water sufficiency will be used as the explanatory variables in the model. In addition, sentinel 2A satellite data of selected index for each plot coordinate surveyed have been included alongside its twelve-month period lag. Out of the thirteen bands recorded by the Sentinel 2A satellite, researchers have decided to use band 4 (red band), band 8 (infra-red band), and a calculated Normalized Difference Vegetation Index (NDVI) as predictor variables in the model.

The Sentinel 2A satellite is used in this research because it has good spatial, temporal, and spectral resolutions. It has a multispectral instrument (MSI) with thirteen spectral bands. The spectral bands cover the electromagnetic spectrum's visible, near infrared, and shortwave infrared regions. For this research, we have only used Band 4, Band 8, and NDVI as explanatory variables for the model. Band 4 corresponds to the red spectral region, with a wavelength range of 630-680 nm, while Band 8 corresponds to the near-infrared (NIR) spectral region, with a wavelength range of 780-900 nm. NDVI, on the other hand, is a vegetation index used to assess the health and vigor of vegetation. It is also one of the most popular indices extracted from spectral bands of satellite imagery and used for predicting crop yield (Kusumaningrum, 2022). The formula used to calculate NDVI can be seen in Equation 2.14. Such as for its temporal resolutions, the visible and near-infrared spectral bands have a spatial resolution of 10 meters, while its shortwave infrared spectral bands have a spatial resolution of 20 meters. Meanwhile, in terms of temporal resolution, Sentinel 2A has a revisit time of five days at the equator. This enables frequent monitoring of changes on the Earth's surface, such as vegetation growth and changes in land use.

3.3.2. Data Collecting, Processing, and Modelling for Predicting Paddy Productivity

This study aims to predict paddy productivity based on the prior knowledge that paddy productivity has a Four Parameter Beta distribution (Sutarlan *et al.* 2017). It is known that studies of modeling data that follow a Four Parameter Beta distribution is uncommon. Therefore, in the first study we used a transformation approach and applied the beta GLM shown in Equation 2.3 and beta GLMM model shown in Equation 2.5. This approach will be applied and evaluated in an empirical case study located in Central Kalimantan and Karawang. In general, the research process that has been conducted can be seen in Figure 3.2.

More specifically, the stages of this research are as follows:

1. Preliminary data collection, which includes:
 - a. Collecting farmer survey data mentioned in Table 2. 2 from SI
 - b. Collecting lagged Sentinel 2A satellite data by
 - In gathering the Satellite Sentinel 2A data, we must first define the Area of Interest (AOI) based on the coordinates of surveyed data.
 - Select the bands that will be used in this research, which are band 4, band 8, and NDVI. These bands were selected based on previous research (Chen *et al.* 2004).
 - Select the time period. We have collected the Sentinel 2A satellite data starting from 12 months in advance before the paddy plots were harvested
 - Afterwards we adjust the program in Google Earth Engine (GEE) to conduct cloud masking to identify and remove clouds, which will enhance more clear and accurate information.
 - The GEE syntax used to collect Sentinel 2A satellite data can be seen in Appendix 1
2. For all data sets, we have determined the proportion of the amount of training data (80%) and testing data (20%) to be used. Thus, we divide all the data into training and testing data.
3. Carry out preliminary data exploration and check the distribution of paddy productivity based on the CCEs.
4. If the paddy productivity has a beta for parameter distribution or $y_{ij} \sim B(\omega_1, \omega_2, a, b)$ then we conduct a transformation of the response variable into a beta distribution by using Equation 3.1. Therefore, we can denote that $y_{ij}^* \sim B(\omega_1, \omega_2)$. This transformation is done for both training and testing data sets.

5. For farmer survey data and the satellite data gathered, we have conducted data preprocessing that will be carried out by:

- Eliminating errors in the farmer survey data
- Outlier detection and handling
- Data exploration

6. A variable selection process was done based on previous data explorations and conducting a pre-variable selection using LASSO (Least Absolute Shrinkage and Selection Operator) applied to a multiple linear regression model. LASSO includes an L1-norm (absolute value) penalty that reduces as many parameters estimates to zero as possible (Sun *et al.* 2020). Variables with non-zero coefficients can thus be essential predictors for the response variable. The LASSO solution can be formulated as:

$$\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^q \sum_{j=1}^{n_i} (y_{ij} - \sum_{k=1}^K x_{ijk} \beta_k)^2 + \lambda \sum_{k=1}^p |\beta_k| \right\} \quad (3.2)$$

It is known that y_{ij} is the response for observation j ($j = 1, 2, \dots, n_i$) in group/area i ($i = 1, 2, \dots, q$), and x_{ijk} is the k -th independent variables for that observation, and β_k are the coefficients to estimate. Next, N is the total number of observations, p is the number of parameters ($K+1$). The penalty's size is determined by λ , which can be found using cross-validation to minimize prediction error. Alternatively, a firmer threshold for λ at one standard error above the minimum prediction error can be chosen (Sun *et al.* 2020). Although Lasso can find a multicollinear solution, Lasso commonly selects one variable arbitrarily and sets the others to zero (Newlands *et al.* 2019)

7. Define the models that will be built, which include:

- Models that only use independent variables obtained from sentinel 2 A satellite data.
- Model that only uses independent variables obtained from farmer surveys.
- Model that only uses sentinel 2A satellite data and farmer surveys data
- In addition, authors also consider models with no random effect and models with random effect for models' a-c mentioned above.
- The random intercepts trial and error include farmers, sub district, district, and nested random effects.

f. For each model a-e there will be a full model and a variable selection process

Thus, in total there are eighteen models tested on the training and testing data set of each area to find the most suitable prediction model for paddy productivity. These models consist of Four Parameter Beta GLM/GLMM models

8. Using the training data set, apply the Four Parameter Beta GLM and GLMM model explained in detail in sub chapters 2.2 and 2.3. We will use the glmmTMB package in R for modelling purposes because its most appealing features are the combination of speed, flexibility, and its interface's (Brooks *et al.* 2017)

9. Set key performance indicators to obtain not just the most suitable model but also the most accurate predictions of paddy productivity at each observed paddy field surveyed in the CCEs. First, the goodness-of-fit model will be evaluated based on Akaike Information Criteria (AIC) values that are formalized as

$$AIC = 2K - 2 \ln(L) \quad (3.3)$$

Where, K is the number of independent variables used in the model, and L is the log-likelihood estimate. Lower AIC values indicate a better-fit model.

10. Use the best-fit selected models, shown by the lowest AIC value, for further predicting paddy productivity on the testing data set. Thus, we will obtain \hat{y}_{ij}^* and it is known that $\hat{y}_{ij}^* \sim B(\omega_1, \omega_2)$

11. Conduct back transformation process on the prediction results (\hat{y}_{ij}) by using the formula

$$\hat{y}_{ij} = \hat{y}_{ij}^* (b - a) + a \quad (3.4)$$

Where the predicted paddy productivity $\widehat{y}_{ij} \sim B(\omega_1, \omega_2, a, b)$

12. Evaluate prediction accuracy based on the second key performance indicators, which is the Relative Root Mean Square Error of Prediction (RRMSEP). RMSEP is the standard deviation of the prediction errors and formulated as

$$RMSEP = \sqrt{\frac{\sum_{i=1}^q \sum_{j=1}^{n_i} (\widehat{y}_{ij} - y_{ij})^2}{N}} \quad (3.5)$$

Here, \widehat{y}_{ij} denotes the predicted paddy productivity value, y_{ij} represents the actual paddy productivity value, and n is the total number of observations or data points. We can finally obtain RRMSEP by dividing RMSEP by the means of actual paddy productivity values (\bar{y}). Lower RRMSEP values will indicate better predictions of the selected models.

13. The best model will then be chosen based on the smallest RMSEP and RRMSEP values.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

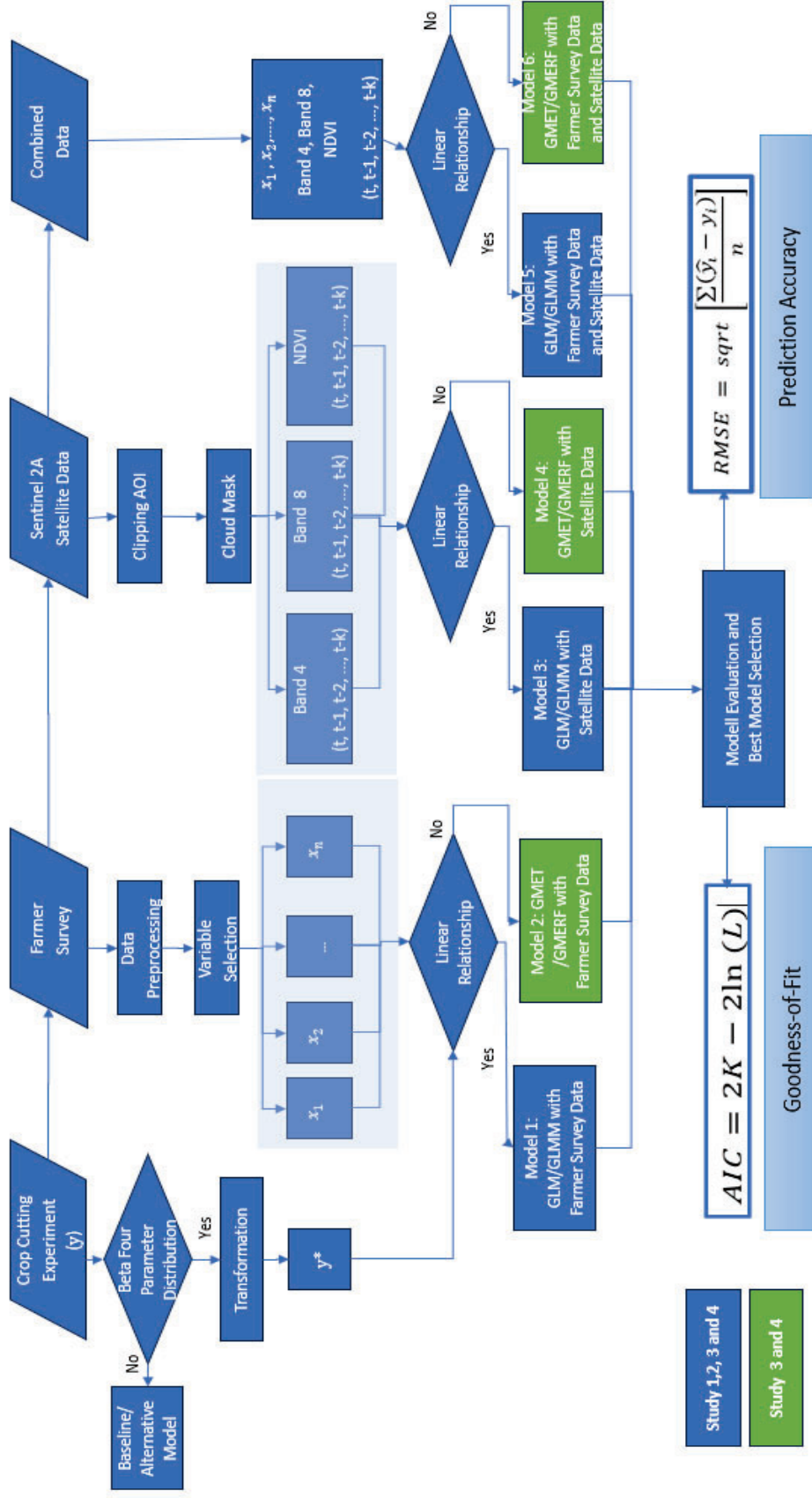


Figure 3. 2 Research Process Flowchart of the Empirical Study

3.3.3. Variable Selection

As mentioned in the methodology, we have first performed a variable selection process by using LASSO. Through this process, seven independent variables were selected. Variables such as pest attacks, pest management, impact of climate change, and water sufficiency were selected and will be used as the explanatory variables in the model. In addition, sentinel 2A satellite data of selected index for each plot coordinate surveyed have been included alongside its twelve-month period lag. Out of the thirteen bands recorded by the Sentinel 2A satellite, researchers have decided to use band 4 (red band), band 8 (infra-red band), and a calculated NDVI as predictor variables in the model. Thus, the list of all selected independent variables or predictors used in this research can be seen in Table 3. 1.

Table 3. 1 Independent Variables Used in the Prediction Model

Variable	Variable Name	Unit Category	Measurement/ Category	Data Source
Paddy Productivity	y	Tons/Ha		CCE (SI)
Pest Attacks This Year	x_1	1 = Heavy, 2 = Medium, 3 = Light, 4 = Not Affected		Farmer Survey (SI)
Pest Attacks Last Year	x_2	1 = Heavy, 2 = Medium, 3 = Light, 4 = Not Affected		Farmer Survey (SI)
Impact of Climate Change This Year	x_3	1 = Affected, 2 = Not Affected		Farmer Survey (SI)
Impact of Climate Change Last Year	x_4	1 = Affected, 2 = Not Affected		Farmer Survey (SI)
Water Sufficiency This Year	x_5	1 = Not Enough, 2 = Sufficient, 3 = More than Enough		Farmer Survey (SI)
Water Sufficiency Last Year	x_6	1 = Not Enough, 2 = Sufficient, 3 = More than Enough		Farmer Survey (SI)
How to Handle Pest	x_7	0 = No Actions, 1 = Agronomist, 2 = Mechanical, 3 = Biological, 4 = Chemical		Farmer Survey (SI)
Band 4	Lag	Mm		Sentinel-2A Satellite
Band 8	rlag	Mm		Sentinel-2A Satellite
NDVI	Nlag	Index		Sentinel-2A Satellite

3.3.4. Paddy Productivity Distribution

Paddy productivity in Central Kalimantan and Karawang have a minimum and maximum value that slightly changes for each season. The productivity of paddy in Central Kalimantan ranges from 1.020 to 5.890 tons per hectare. Slightly higher, in Karawang paddy productivity ranges from 0.500 up to 6.900 tons per hectare. A distribution fit test in Table 3. 2 shows that the Four Parameter Beta distribution is most suitable compared to the normal and Laplace distribution in both areas because it has the lowest AIC value. Based on the Numerical Maximum Likelihood parameter estimates (R Software - ExtDist Library) shown in Table 3.2, it is estimated that the average paddy productivity in Central Kalimantan is 3.031 tons per hectare with a variance of 1.070 tons per hectare. Meanwhile, in Karawang the conditions of paddy productivity are more uniform, which is indicated by the small variance of 0.206 tons per

hectare. The average paddy productivity in Karawang is also higher, reaching up to almost 5 tons per hectare. Nevertheless, the productivity of paddy in Central Kalimantan and Karawang are still below the national paddy productivity, which is around 5.23 tons per hectare.

It is known that there are several factors that can support paddy productivity growth, such as selecting proper soil management techniques and fertilizers. Apart from that, there are also several government programs that can assist farmers in Central Kalimantan and Karawang to increase their paddy productivity. Based on the models developed, we can identify these factors and provide recommendations that can be used to improve the level of paddy productivity in Central Kalimantan and Karawang.

Table 3. 2 Distribution Fit Test and Parameter Estimates

Area	AIC of Distribution Fit Test			
	Four	Laplace	Normal	
	Parameter Beta			
	Central Kalimantan	335.208	336.688	338.168
	Karawang	113.299	139.975	127.154
Parameter Estimates				
	ω_1	ω_2	a	b
Central Kalimantan	3.067	6.091	0.691	7.676
Karawang	3.099	1.382	3.186	5.488

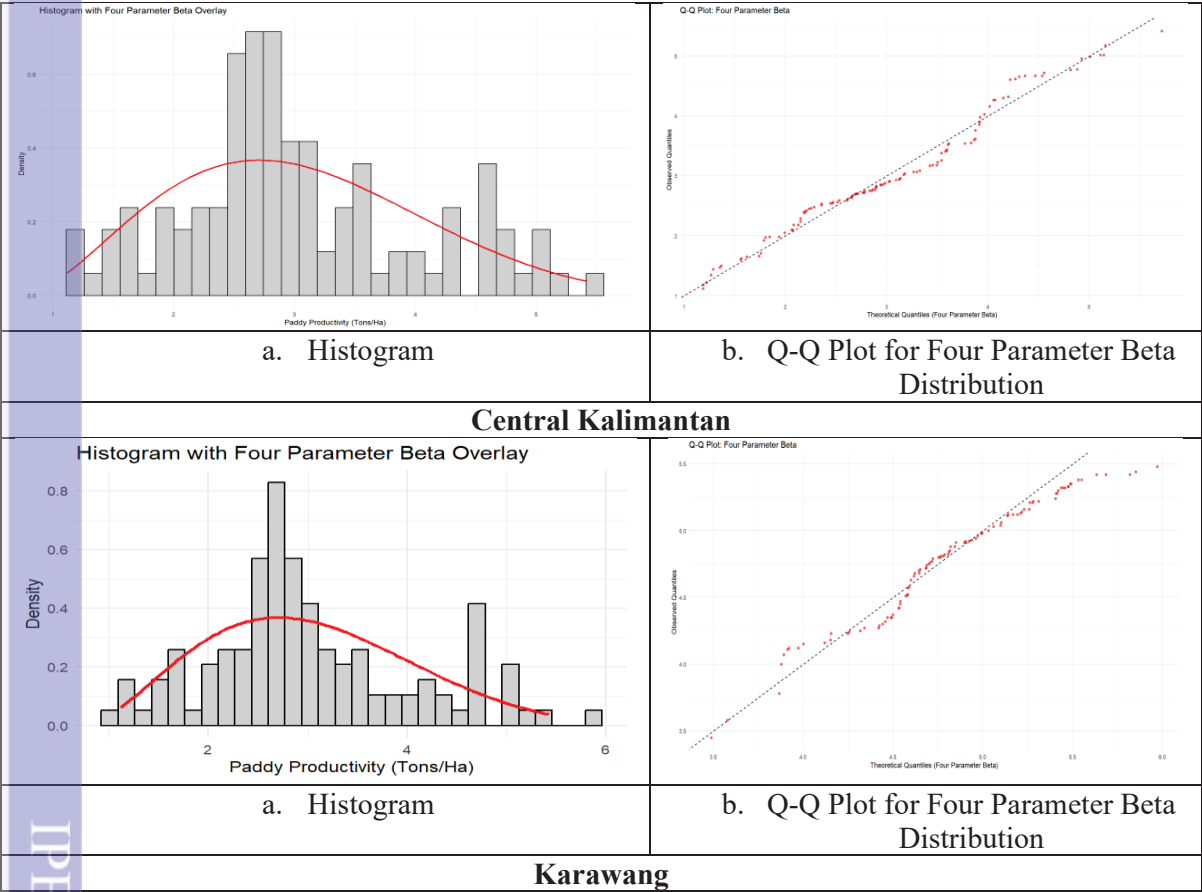


Figure 3. 3 Paddy Productivity Distribution

Figure 3. 3 shows that paddy productivity in both areas tends to skew to the left. Additionally, it is shown that paddy productivity in both areas are bounded within a specific minimum and maximum value. The QQ – Plot constructed based on a Four Parameter Beta distribution further supports the suitability of this distribution for modeling the data. Most of the observed points lie closely along the reference line, indicating a good overall fit. A few deviations are observed at the distribution's tail, which may suggest the presence of outliers or extreme values. These characteristics further justify the use of the Four Parameter Beta distribution, given its flexibility in handling skewed and bounded data.

3.3.5. Model Evaluations based on Key Performance Indicators

In total, we have evaluated eighteen GLM and GLMM models for each area. A GLM model has no random effects but only considers fixed effects. Therefore, the model tends to be simpler. Meanwhile, GLMM models incorporate random effects and fixed effects. The GLMM model is usually used when researchers have data with more than one source of variance, such as modeling data that has different areas (cross-section data) or clustered data. Therefore, in modelling paddy productivity in Central Kalimantan, incorporating suitable random effects is crucial to account for hierarchical area structures. The area-level effects (e.g., districts and subdistricts) is needed to capture the variability in environmental and management practices. While farm-level random effects can address differences in individual farming practices and interaction effects between area-level effects can model a certain unique behaviour to specific regions. By integrating these random effects, the GLMM models can better handle the complex, dynamic, and heterogeneous nature of agricultural data, leading to improved predictive accuracy and more generalizable findings.

In our study, the influence of the different districts and subdistricts on the predictor was modeled through a random effect b_i . Besides that, we have also considered the influence of farmers, sub district, district, and nested random effects in the model (Kusumaningrum *et al.* 2022). After trial-and-error process it turns out that the farmers, farmers nested in subdistricts, and subdistrict random effects has been a better fit in developing a GLMM model for predicting productivity. Therefore, in general, the best model can be written as

$$Y_{ijk}^* \sim B(\mu_{ijk}, \varphi) \text{ and } \eta_{ijk} = X_{ijk}^T \beta + u_i + w_j + v_{j(k)} \quad (3.6)$$

$$u_i \sim iid N(0, \sigma_{Farmer}^2), w_j \sim iid N(0, \sigma_{Subdistrict}^2), v_{j(k)} \sim iid N(0, \sigma_{Subdistrict(District)}^2)$$

where Y_{ijk}^* is denoted as the transformed paddy harvest productivity for farmers i located in subdistrict j , which is within district k . Next, $\eta_{ijk} = g(E(y_{ijk})) = g(\mu_{ijk}) = \log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right)$ for the logit link function. Fixed effect estimates of the selected variables are denoted by β and $u_i, v_{j(i)}$, and w_j are the random effects of farmer, farmer nested in subdistrict, and subdistrict.

It is known that the Four Parameter Beta GLMM is better than the four-parameter beta GLM because it has the lowest AIC value (Figure 3. 4). This condition is relevant in each type of independent variable used, indicating that random effects and fixed effects are needed for predicting paddy productivity. The result was in line with previous systematic literature review written by Klompenburg et. al (2020), where it was shown that tailor models to specific areas increase the relevance and accuracy of yield predictions Klompenburg et. al (2020). In our studies, the most suitable random effects for both areas include farmers, farmers nested in sub districts, and subdistricts. Out of the eighteen models examined for each area, the best two GLMM models are the models that incorporate both farmer survey and satellite data or the models that just incorporate farmer survey data. In Central Kalimantan, the best fit model is the model that uses both farmer survey and satellite which has an AIC of -124.117. While in

Karawang the best fit model is the model that just uses farmer survey data, which has an AIC of -228.000.

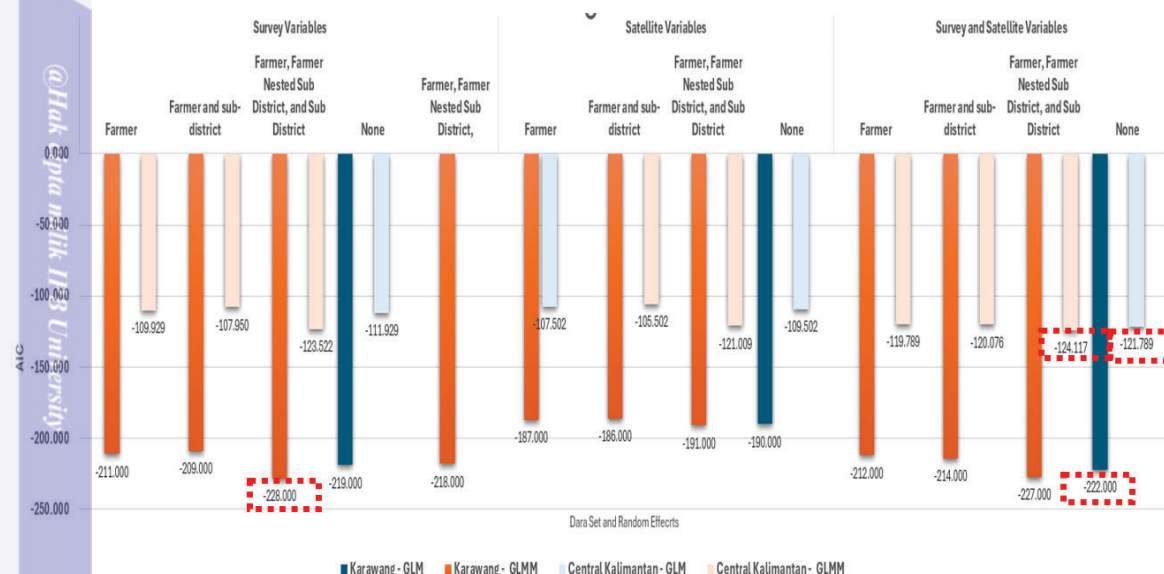


Figure 3. 4 Four Parameter Beta GLM and GLMM Goodness of Fit Model

Next, we will also examine the prediction accuracy of the selected Four Parameter Beta GLMM models. Final predictions of paddy productivity were made through three selected models (Table 3. 3) applied on the testing data set. These models are the best fit selected models when we use only farmer survey independent variables, only satellite independent variables, and a combination of both farmer survey and satellite independent variables. When comparing the three model's prediction accuracy on the testing data set, we must first do a back transformation based on Equation 3.4 to have the actual paddy productivity values. It was found that in Central Kalimantan, the model with the highest prediction accuracy is the model that uses both farmer survey and satellite. While in Karawang, the model with the highest prediction accuracy is the model that just uses farmer survey data. This is in line with the results of the goodness of fit model showed early on. Thus, incorporating survey data and satellite data increases prediction accuracy, especially in areas such as Central Kalimantan.

Table 3. 3 GLMM Model's Prediction Accuracy Based on RRMSEP

Model	Training Data Set		Testing Data Set	
	Central Kalimantan	Karawang	Central Kalimantan	Karawang
GLMM with Farmer Survey Data	0.123	0.120	0.261	0.110
GLMM with Satellite Data	0.121	0.110	0.283	0.150
GLMM with Farmer Survey and Satellite Data	0.117	0.110	0.247	0.170

3.3.6. Analyzing the Best Prediction Models

The accuracy of prediction between the three models is similar, which can be an advantage because the accuracy of prediction performs well when using farmer survey data or satellite data. Nonetheless, the models with a combination of farmer survey and satellite data slightly outperform other models. Thus, in the future, the use of satellite data for prediction purposes will be more beneficial because satellite data covers large areas, making it ideal for predicting trends and patterns over large regions. In contrast, survey data is typically limited to a specific location or a small sample size, which may only be representative of part of the region. Satellite data can also be available in near-real-time, allowing for quick and timely predictions of events. Survey data, on the other hand, can take weeks or months to collect and analyze. Finally, collecting survey data, in general, can be expensive, particularly for large samples. Meanwhile, satellite data is often readily available and can be obtained at a relatively low cost.

Based on the best models, the independent variables that have a significant effect on paddy productivity at a 10% significance level in Central Kalimantan are Pest Attacks Last Year (x_2), How to Handle Pest (x_7), Band 4 (Lag 0 and Lag 5), Band 8 (Lag 0), and NDVI (Lag 3, Lag 5). Meanwhile, in Karawang besides Pest Attacks Last Year (x_2) and the current and lagged Band 4, Band 8, and NDVI values, Impact of Climate Change Last Year (x_4) also has a significant effect on paddy productivity. The full output of the best GLM/GLMM models in Karawang and Central Kalimantan can be seen in Appendix 2.

We will conduct further exploration on the independent variables that have a significant effect on paddy productivity. First, starting with exploring patterns of the satellite data. In Indonesia, paddy productivity is influenced by three main growing seasons: Sub Round 1 (January–March), Sub Round 2 (April–June), and Sub Round 3 (August–December). Productivity peaks during Sub Round 1 and declines during Sub Round 3, primarily due to differences in irrigation and growing conditions. Satellite data, particularly NDVI and Band 8 (Near-Infrared), show a clear relationship with paddy growth stages. These indices increase during Sub Round 1 and Sub Round 2, indicating healthier, denser vegetation and higher productivity, but decline in Sub Round 3, reflecting lower productivity. NDVI captures key growth factors like leaf area, chlorophyll content, and canopy structure, while Band 8 strongly correlates with healthy plant reflectance. Furthermore, historical patterns of NDVI and Band 8 from previous growing periods (Sub Round 1 and Sub Round 2) align with current productivity trends, as shown in Figure 3. 5

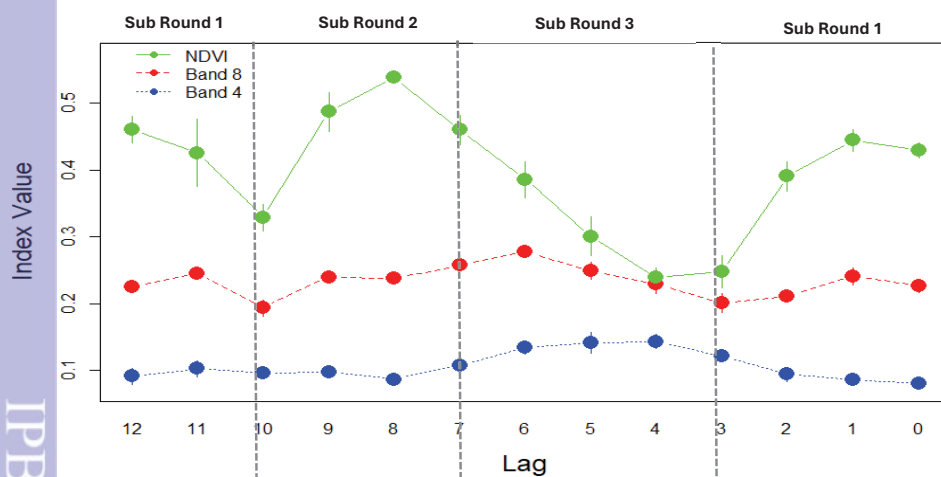


Figure 3. 5 Band 4, Band 8, and NDVI Index Trend

These findings suggest that historical satellite data can effectively serve as one of the main independent variables for predicting paddy productivity. The consistency in these patterns validates the accuracy developed model in incorporating historical satellite information for productivity assessments. These findings were also in line with Debake et al (2023). Thus, enhancing the utilization of satellite data will give stakeholders more opportunity to make productivity estimates for large areas where they might not have access to individual farmers' survey data. This information can also be used to monitor and maintain crop health during critical growth stages, which can influence higher productivity. Based on our model, the monitoring process can even start 5 months in advance.

Next, we will explore two main independent variables that had an influence in either Central Kalimantan or Karawang. In both areas, pests were a major issue found. More than 95% of the farmers have faced pest attacks in the current year and previous year (Figure 3. 6). Most farmers experience mild to moderate pest attacks. The condition is slightly more crucial in Central Kalimantan where at least 25% of farmers must handle severe pest attacks. Contrarywise, in Karawang there were no farmers facing severe pest attacks. Farmers who suffer from more severe pest attack conditions tend to have lower paddy productivity (Figure 11).

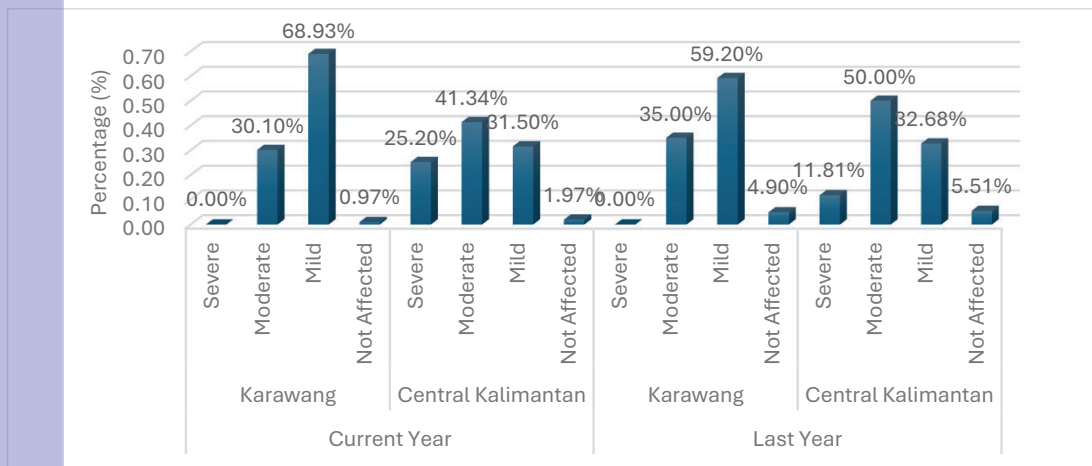


Figure 3. 6 Pest Attack Severity in Central Kalimantan and Karawang

Besides pests, climate change also is another issue found in both areas. Climate change will cause changes in temperature, precipitation patterns, or extreme weather events. There are at least 20 – 40% of farmers affected by climate change (Figure 3. 7). For the current and past year, more farmers in Central Kalimantan experience the impact of climate change compared to farmers in Karawang area. At least 25% of farmers in Central Kalimantan have experienced severe impact due to climate change. Therefore, farmers in this region tend to be more vulnerable to climate change, leading to reduced and inconsistent crop yields, making farming less reliable and more challenging. On the other hand, Karawang farmers have not faced any severe climate change impact in the current and past year. Therefore, we can perceive that Karawang farmers have advantages through a more stable climate conditions, advanced agricultural practices, better infrastructure, and better access to resources and support. These factors enable Karawang's farmers to better adapt to and mitigate the impacts of climate change, ensuring more stable and resilient agricultural productivity.

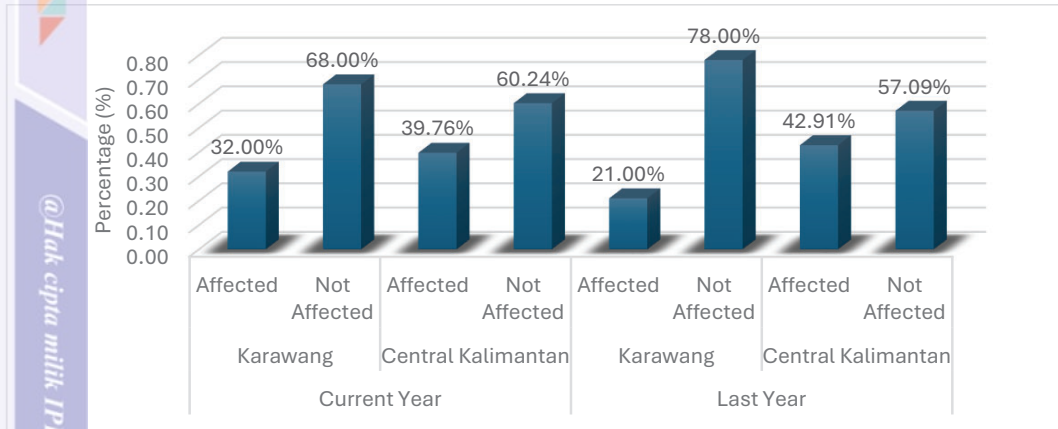


Figure 3. 7 Climate Change Impact in Central Kalimantan and Karawang

The association between climate change and pests is found to be significant and can lead to increased pest pressures on crops and ecosystems (Subedi *et al.* 2023). Climate change can influence pest populations, distributions, and impacts through changes in temperature, precipitation patterns, and ecological interactions. Therefore, effective pest management in controlling climate change requires adaptive strategies. In Central Kalimantan, it was shown that farmers who handle pests by using an agronomist approach tend to have higher average paddy productivity (3.11 tons/ha). To manage pests sustainably, agronomists usually use cultural, biological, chemical, and preventive management techniques. Their goal is to minimize threats to human health and the environment while maintaining pest populations below levels that would cause less harm through the application of Integrated Pest Management (IPM) techniques. This all-encompassing strategy aids in maximizing paddy productivity. Meanwhile, farmers who do not use the agronomist approach combined with a condition of having more or sufficient water and having severe to mild pest Attacks last year and this year tend to have the lowest average paddy productivity (2.28 tons/ha). In this case we can see that farmers who do not use the agronomist approach tend to be more vulnerable to pests, even though having sufficient water. This emphasizes the fact that it is crucial to implement IPM techniques that are suited to the local environment. By addressing these issues with better pest control techniques, paddy farmers' livelihoods can be improved, productivity can be increased, and yield variability can be decreased.

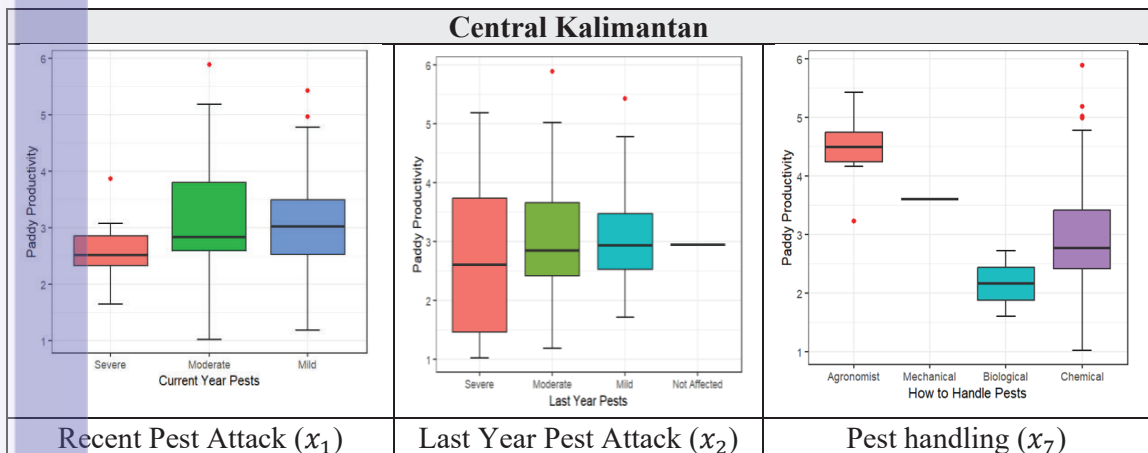


Figure 3. 8 Clustered Boxplots of Farmer Survey Data and Paddy Productivity (a)

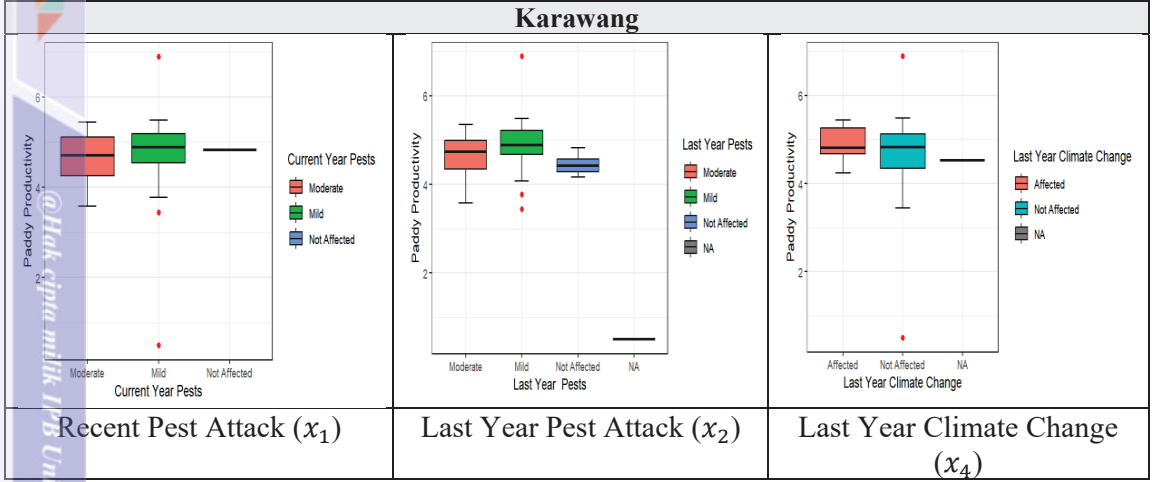


Figure 3. 9 Clustered Boxplots of Farmer Survey Data and Paddy Productivity (b)

Eventhough, the proposed approach model's fit and prediction is promising, nevertheless we must keep in mind that the transformation process can cause bias in parameter estimates and complications in the interpretation of coefficient values. In this study, we are more focused on accuracy of the prediction results. Therefore, the drawbacks will not be apparent. Nevertheless, in the next section we will indicate the potential bias as a result of the transformation process.

3.4.Effect of the Transformation Process

To understand the transformation from a Four Parameter Beta distribution to a standard beta distribution, we first have already defined the PDF of the Four Parameter Beta distribution in Equation 2.2 and the PDF of the Beta in Equation 2.1. It has already been shown that the transformation process conducted is based on Equation 3.1.

Due to the transformation process, we need to transform the PDF of the Four Parameter Beta distribution accordingly. Using the change of variables, we find the PDF of y as follows:

$$f_{Y^*}(y^*) = f_Y(y) \left| \frac{dy}{dy^*} \right|$$

where $f_Y(y)$ is the PDF of the Four Parameter Beta distribution and $\left| \frac{dy}{dy^*} \right|$ is the absolute value of the Jacobian of the transformation. Then we will calculate the Jacobian by calculating the derivative of y with respect to y^* as:

$$\frac{dy}{dy^*} = (b - a)$$

Thus, the absolute value of the Jacobian is $|b - a| = b - a$. Next, substitute $y = a + y^*(b - a)$ and the Jacobian into the PDF of y which is $f(y, \omega_1, \omega_2, a, b) =$

$\frac{\Gamma(\omega_1 + \omega_2)}{\Gamma(\omega_1)\Gamma(\omega_2)} \frac{(y-a)^{\omega_1-1}(b-y)^{\omega_2-1}}{(b-a)^{\omega_1+\omega_2-1}}$ than we have

$$f_{Y^*}(y^*) = \frac{(a + y^*(b - a) - a)^{\omega_1-1} (b - (a + y^*(b - a)))^{\omega_2-1}}{(b - a)^{\omega_1+\omega_2-1} B(\omega_1, \omega_2)} (b - a)$$

Where $B(\omega_1, \omega_2) = \frac{\Gamma(\omega_1 + \omega_2)}{\Gamma\omega_1 \Gamma\omega_2}$ is the Beta function and $f_{Y^*}(y^*)$ defining the PDF as a function of the specific value y^*

Simplify the terms:

$$a + y^*(b - a) - a = y^*(b - a) \text{ and } b - (a + y^*(b - a)) = (b - a)(1 - y^*)$$

We will have

$$f_{Y^*}(y^*) = \frac{(y^*(b - a))^{\omega_1 - 1} ((b - a)(1 - y^*))^{\omega_2 - 1}}{(b - a)^{\omega_1 + \omega_2 - 1} B(\omega_1, \omega_2)} (b - a)$$

$$f_{Y^*}(y^*) = \frac{y^{*\omega_1 - 1} (1 - y^*)^{\omega_2 - 1} (b - a)^{\omega_1 - 1} (b - a)^{\omega_2 - 1}}{(b - a)^{\omega_1 + \omega_2 - 1} B(\omega_1, \omega_2)} (b - a)$$

$$f_{Y^*}(y^*) = \frac{y^{*\omega_1 - 1} (1 - y^*)^{\omega_2 - 1} (b - a)^{\omega_1 + \omega_2 - 1}}{(b - a)^{\omega_1 + \omega_2 - 1} B(\omega_1, \omega_2)}$$

$$f_{Y^*}(y^*) = \frac{y^{*\omega_1 - 1} (1 - y^*)^{\omega_2 - 1}}{B(\omega_1, \omega_2)}$$

This will be the PDF of the standard beta distribution with shape parameters ω_1 and ω_2 .

Hence, it is shown that the transformation from a Four Parameter Beta distribution to a standard beta distribution involves rescaling the random variable Y from the interval (a, b) to Y^* with the interval $(0, 1)$. By using the transformation formula (Equation 3.1) it has been shown that this process obtains a distribution of y^* that has a standard beta distribution with the same shape parameters ω_1 and ω_2 .

Next, to analyze potential bias, first we will compare the mean and variance before and after the transformation. It is known that the Four Parameter Beta distribution mean or $E[Y] = a + \frac{\omega_1}{(\omega_1 + \omega_2)}(b - a)$. Meanwhile the beta distribution mean $E[Y^*] = \frac{\omega_1}{(\omega_1 + \omega_2)}$. When conducting a back transforming to the original scale (Y'), then we have:

$$Y' = a + Y^*(b - a), \text{ thus}$$

$$E[Y'] = a + E[Y^*](b - a) = a + \frac{\omega_1}{(\omega_1 + \omega_2)}(b - a) = E[Y]$$

Thus, the mean of the transformed variable, when back transformed, is equivalent the original mean. Thus, the transformation does not introduce bias in the mean of the distribution.

As for the variance, it is known that the Four Parameter Beta distribution variance or $\varphi = \frac{\omega_1 \omega_2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 + 1)} (b - a)^2$. Meanwhile the beta distribution variance $\varphi^* = \frac{\omega_1 \omega_2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 + 1)}$. When conducting a back transforming to the original scale, we will have:

$$\varphi(Y') = \varphi^* (b - a)^2 = \frac{\omega_1 \omega_2}{(\omega_1 + \omega_2)^2 (\omega_1 + \omega_2 + 1)} (b - a)^2 = \varphi$$

Such as the mean, the variance of the transformed variable, when back transformed, is equivalent the original variance. Thus, the transformation does not introduce bias in the variance.

Analytically, it has been shown that the linear transformation from a Four Parameter Beta distribution to a standard beta distribution preserves the mean and variance, indicating no bias in these aspects. However, potential bias could happen from non-linear relationships, changes in higher moments like skewness and kurtosis, and the handling of outliers and extreme values. It is crucial to consider these factors when applying such transformations, especially in complex real-case studies.

Analyzing further, let's consider fitting a GLMM with a Four Parameter Beta distributed response variable Y than has been formulated in Equation (2.2). The transformation to Y^* affects the GLMM components based on the steps below:

1. Define the Link Function and Linear Predictor

For the original variable Y , suppose the link function is $g(\mu_Y)$, where $\mu_Y = E[Y]$ and the linear predictor is $\eta = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$. For the transformed variable Y^* , the mean of Y^* is μ_{Y^*} . Thus, the link function is now $g(\mu_{Y^*})$

2. Define the Relationship Between μ_Y and μ_{Y^*}

It has been stated above that $E[Y] = \mu_Y = a + \frac{\omega_1}{\omega_1 + \omega_2}(b - a)$ and $E[Y^*] = \mu_{Y^*} = \frac{\omega_1}{(\omega_1 + \omega_2)}$. Then $\mu_Y = a + \mu_{Y^*}(b - a)$. Thus, we will have $\mu_{Y^*} = \frac{\mu_Y - a}{b - a}$

3. Potential Bias when using an Identity Link Function

For a GLMM of Y , we have $g(\mu_Y) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$ and for the transformed GLMM for Y^* we will have $g(\mu_{Y^*}) = g\left(\frac{\mu_Y - a}{b - a}\right) = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{Z}\mathbf{b}^*$. This is still in the same linear form, but now everything is on the scale of μ_{Y^*} . Because the transformation itself is linear, the identity link preserves this structure.

If the link function is the identity than $g(Y) = Y$ and applied on a certain data set so we will have

$$\mu_Y = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}} \text{ and } \mu_{Y^*} = \mathbf{X}\widehat{\boldsymbol{\beta}}^* + \mathbf{Z}\widehat{\mathbf{b}}^*$$

by using $\mu_Y = a + \mu_{Y^*}(b - a)$, than we will have

$$\mu_Y = a + (\mathbf{X}\widehat{\boldsymbol{\beta}}^* + \mathbf{Z}\widehat{\mathbf{b}}^*)(b - a)$$

There are two factors that can cause bias. First is a that reflects a shift in the response variable. Next is $b - a$ that reflects a rescaling of the response variable.

4. Bias Estimation Caused by the Rescaling Factor when using an Identity Link Function

First, we evaluate the impact of the rescaling factor by equating the terms below

$$a + (\mathbf{X}\widehat{\boldsymbol{\beta}}^* + \mathbf{Z}\widehat{\mathbf{b}}^*)(b - a) = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}}$$

$$(\mathbf{X}\widehat{\boldsymbol{\beta}}^* + \mathbf{Z}\widehat{\mathbf{b}}^*)(b - a) = \mathbf{X}\widehat{\boldsymbol{\beta}} + \mathbf{Z}\widehat{\mathbf{b}} - a$$

When just taking into consideration the rescaling factor $(b - a)$, then

$$(X\widehat{\beta}^* + Z\widehat{b}^*)(b - a) = X\widehat{\beta} + Z\widehat{b}$$

$$X\widehat{\beta}^* + Z\widehat{b}^* = \frac{X\widehat{\beta} + Z\widehat{b}}{b-a}$$

Thus, the transformed parameters estimates $\widehat{\beta}^*$ and \widehat{b}^* are related to the original parameters estimates ($\widehat{\beta}$ and \widehat{b}) as:

$$\widehat{\beta}^* = \frac{\widehat{\beta}}{(b-a)} \text{ and } \widehat{b}^* = \frac{\widehat{b}}{(b-a)}$$

Now, let's recall the definition of a bias of a parameter estimate β , which is defined as:

$$\text{Bias}(\widehat{\beta}) = E[\widehat{\beta}] - \beta$$

Thus, if the transformation process is unbiased, then the bias of the parameter estimates from the transformation process should be equal to β and b or it can be written as

$$E[\widehat{\beta}^*] = \beta \text{ and } E[\widehat{b}^*] = b$$

or it can also be shown that

$$\text{Bias}(\widehat{\beta}^*) = E[\widehat{\beta}^*] - \beta = 0 \text{ and } \text{Bias}(\widehat{b}^*) = E[\widehat{b}^*] - b = 0$$

Now, let's calculate the bias

$$\text{Bias}(\widehat{\beta}^*) = E[\widehat{\beta}^*] - \beta \text{ and } \text{Bias}(\widehat{b}^*) = E[\widehat{b}^*] - b$$

We know that $\widehat{\beta}^* = \frac{\widehat{\beta}}{(b-a)}$ and $\widehat{b}^* = \frac{\widehat{b}}{(b-a)}$. Therefore we will obtain

$$\text{Bias}(\widehat{\beta}^*) = E\left[\frac{\widehat{\beta}}{(b-a)}\right] - \beta \text{ and } \text{Bias}(\widehat{b}^*) = E\left[\frac{\widehat{b}}{(b-a)}\right] - b$$

$$\text{Bias}(\widehat{\beta}^*) = \frac{E[\widehat{\beta}]}{(b-a)} - \beta \text{ and } \text{Bias}(\widehat{b}^*) = \frac{E[\widehat{b}]}{(b-a)} - b$$

assuming that $\widehat{\beta}$ is unbiased or $E[\widehat{\beta}] = \beta$ and $E[\widehat{b}] = b$ then we will have

$$\text{Bias}(\widehat{\beta}^*) = \frac{\beta}{(b-a)} - \beta \text{ and } \text{Bias}(\widehat{b}^*) = \frac{b}{(b-a)} - b$$

$$\text{Bias}(\widehat{\beta}^*) = \beta \left(1 - \frac{1}{(b-a)}\right) \text{ and } \text{Bias}(\widehat{b}^*) = b \left(1 - \frac{1}{(b-a)}\right)$$

If a and b are chosen such that $(b - a) = 1$, then the bias reduces to 0. Otherwise, the scaling factor impacts the estimates systematically. If $b - a > 1$, the bias is negative, indicating $\widehat{\beta}^*$ and \widehat{b}^* underestimates β and b . Meanwhile If $b - a < 1$, the bias is positive, indicating $\widehat{\beta}^*$ and \widehat{b}^* overestimates β and b .

5. Bias Estimation Caused by the Shifting Factor when using an Identity Link Function

Next, we will evaluate the impact of the shifting factor a

$$X\widehat{\beta} + Z\widehat{b} - a = (X\widehat{\beta}^* + Z\widehat{b}^*)(b - a)$$

$$\frac{X\widehat{\beta} + Z\widehat{b}}{(b - a)} - \frac{a}{(b - a)} = X\widehat{\beta}^* + Z\widehat{b}^*$$

$$\frac{\mu_Y - a}{b - a} = X\widehat{\beta}^* + Z\widehat{b}^*$$

The shifting factor a does not depend on \mathbf{X} or \mathbf{Z} and acts uniformly across all observations. It will affect, the intercept within the fixed-effects parameter β , This result is obtained by considering that the fixed-effects design matrix \mathbf{X} includes an intercept term, β_0 . Then, the model can be explicitly written as:

$$\mu_Y = \widehat{\beta}_0 + X'\widehat{\beta} + Z\widehat{b} \text{ and } \mu_{Y^*} = \widehat{\beta}_0^* + X'\widehat{\beta}^* + Z\widehat{b}^*$$

where \mathbf{X}' is the matrix of fixed-effect predictors excluding the intercept. Next, rewriting μ_Y and μ_{Y^*} in terms of the original GLMM

$$\frac{\widehat{\beta}_0 + X'\widehat{\beta} + Z\widehat{b} - a}{(b - a)} = \widehat{\beta}_0^* + X'\widehat{\beta}^* + Z\widehat{b}^*$$

Next, if we focus on the intercept-related terms, we will have :

$$\frac{\widehat{\beta}_0 - a}{(b - a)} = \widehat{\beta}_0^* \text{ or } \widehat{\beta}_0^* = \frac{\widehat{\beta}_0}{(b - a)} - \frac{a}{(b - a)}$$

Regarding that $\widehat{\beta}_0$ has already been scaled by $\frac{1}{(b - a)}$ at the previous step, then we will have

$$\widehat{\beta}_0^* = \widehat{\beta}_0 - a$$

Now, let's calculate the bias of a parameter estimate $\widehat{\beta}_0^*$. If the transformation process is unbiased, then the bias of the parameter estimates from the transformation process should be equal to β_0 or it can be written as

$$E[\widehat{\beta}_0^*] = \beta_0$$

or it can also be shown that

$$\text{Bias}(\widehat{\beta}_0^*) = E[\widehat{\beta}_0^*] - \beta_0 = 0$$

$$\text{Bias}(\widehat{\beta}_0^*) = E[\widehat{\beta}_0 - a] - \beta_0$$

$$\text{Bias}(\widehat{\beta}_0^*) = E[\widehat{\beta}_0] - E[a] - \beta_0$$

assuming that $\widehat{\beta}_0$ is unbiased or $E[\widehat{\beta}_0] = \beta_0$, then we will have

$$\text{Bias}(\widehat{\beta}_0^*) = \beta_0 - a - \beta_0 = -a$$

This means that the fixed-effects predictions are shifted downward (if $a > 0$) or upward (if $a < 0$) proportionally to a . This relationship shows how the transformation shifts the baseline of the fixed-effects component of the model.

6. Bias Estimation of the Predicted Values when using an Identity Link Function

Next, for the prediction values, we will have $\widehat{\mu}_Y = g^{-1}(\mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b})$. While, for the transformed model predictions, we will have $\widehat{\mu}_{Y^*} = g^{-1}\left(\frac{\widehat{\mu}_Y - a}{b - a}\right) = g^{-1}\left(\frac{\mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b} - a}{b - a}\right)$. To examine potential bias, consider the implications of transforming both the response variable and the predictions back and forth. When back transformed in order to obtain predictions on the original scale we must set $\widehat{\mu}_Y' = a + \widehat{\mu}_{Y^*}(b - a)$ then we will have

$$\widehat{\mu}_Y' = a + g^{-1}\left(\frac{\mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b} - a}{b - a}\right)(b - a)$$

If g is a linear identity function, then there is no bias because:

$$\widehat{\mu}_Y' = \mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b}$$

Henceforward, we can conclude that using the identity link, it doesn't add any extra complexity. It directly models the response. This means the relationship between the predictors (\mathbf{X} and \mathbf{Z}) and the transformed response (Y^*) is still straightforward and linear. The transformation doesn't change the form of the equation—it just rescales the response variable. The predictors (\mathbf{X}) and their coefficients (β) still influence Y^* in the same way as they influenced Y , but now everything is on the rescaled range of 0 to 1.

7. Potential Bias when using a Logit Link Function

In a GLMM where the logit link function is used, it can be defined as:

$$g(\mu_Y) = \text{logit}(\mu_Y) = \log\left(\frac{\mu_Y}{1 - \mu_Y}\right) = \mathbf{X}\beta + \mathbf{Z}b$$

This relates the mean of the response variable μ_Y to the linear predictor $\mathbf{X}\beta + \mathbf{Z}b$. Equivalently, solving for μ_Y , we will have

$$\mu_Y = \frac{\exp(\mathbf{X}\beta + \mathbf{Z}b)}{1 + \exp(\mathbf{X}\beta + \mathbf{Z}b)}$$

Through the transformation process applied on a certain data set we will have

$$g(\mu_{Y^*}) = \text{logit}(\mu_{Y^*}) = \log\left(\frac{\mu_{Y^*}}{1 - \mu_{Y^*}}\right) = \mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b}$$

Substituting $\mu_{Y^*} = \frac{\mu_Y - a}{b - a}$, then we have

$$g(\mu_{Y^*}) = \log\left(\frac{\mu_{Y^*}}{1 - \mu_{Y^*}}\right) = \log\left(\frac{\frac{\mu_Y - a}{b - a}}{1 - \frac{\mu_Y - a}{b - a}}\right) = \mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b}$$

We can simplify the denominator by $1 - \frac{\mu_Y - a}{b - a} = \frac{b - a - \mu_Y + a}{b - a} = \frac{b - \mu_Y}{b - a}$, therefore the transformed GLMM now becomes

$$g(\mu_{Y^*}) = \log\left(\frac{\mu_{Y^*}}{1 - \mu_{Y^*}}\right) = \log\left(\frac{\frac{\mu_Y - a}{b - a}}{\frac{b - \mu_Y}{b - a}}\right) = \log\left(\frac{\mu_Y - a}{b - \mu_Y}\right) = \mathbf{X}\widehat{\beta} + \mathbf{Z}\widehat{b}$$

So, if we compare the two link functions, we have the transformed model link function $\log\left(\frac{\mu_Y - a}{b - \mu_Y}\right)$ and for the untransformed model, the link function is $\log\left(\frac{\mu_Y}{1 - \mu_Y}\right)$. In general, such as in the point above, there will be a shift and rescaling factor. The term $\mu_Y - a$ reflects a shift in the response variable. This shift affects the intercept of the model, as the baseline (reference point) for μ_Y as it has been altered by subtracting a . Meanwhile the

term $b - \mu_Y$ reflects a rescaling of the response variable. This scaling modifies the range of μ_Y , compressing or stretching the influence of $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$ on the mean response.

To be more specific, we can rewrite the transformed model logit above, then we will have:

$$\log\left(\frac{\mu_Y - a}{b - \mu_Y}\right) = \log\left(\frac{\mu_Y}{1 - \mu_Y}\right) + \log\left(\frac{1 - \mu_Y}{b - \mu_Y}\right) - \log\left(\frac{\mu_Y}{\mu_Y - a}\right)$$

If we set $\Delta_{(\mu_Y, a, b)} = \log\left(\frac{1 - \mu_Y}{b - \mu_Y}\right) - \log\left(\frac{\mu_Y}{\mu_Y - a}\right)$, then

$$\mu_Y^* = \mu_Y + \Delta_{(\mu_Y, a, b)}, \text{ or}$$

$$\mathbf{X}\hat{\boldsymbol{\beta}}^* + \mathbf{Z}\hat{\mathbf{b}}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}} + \Delta_{(\mu_Y, a, b)}.$$

The term $\Delta_{(\mu_Y, a, b)}$ introduces bias in $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{b}}^*$, as it shifts the relationship between the linear predictor $\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}$ and the response μ_Y , such as mentioned before. Furthermore, let's go back to $\Delta_{(\mu_Y, a, b)}$, where it can be written as

$$\Delta_{(\mu_Y, a, b)} = \log(1 - \mu_Y) - \log(b - \mu_Y) - \log(\mu_Y) + \log(\mu_Y - a)$$

To simplify calculations, let $\Delta_{(\mu_Y, a)_1} = \log(\mu_Y - a) - \log(\mu_Y)$ and $\Delta_{(\mu_Y, b)_2} = \log(1 - \mu_Y) - \log(b - \mu_Y)$. For the first component, $\Delta_{(\mu_Y, a)_1} = \log\left(1 - \frac{a}{\mu_Y}\right)$. By

using the Taylor expansion, where $\log(1 + x) \approx x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$, it is known that for small values of x , higher-order terms (x^2, x^3, \dots) become negligible, so $\log(1 + x) \approx x$. Hence for a small value a , $\Delta_{(\mu_Y, a)_1} \approx -\frac{a}{\mu_Y}$. For the second component if we have large

b the approximate of $b - \mu_Y \approx \frac{1}{b}$, so $\Delta_{(\mu_Y, b)_2} = \log(1 - \mu_Y) - \log(b)$. Further on, for small μ_Y (near 0), $\log(1 - \mu_Y) \approx -\mu_Y$. Thus, $\Delta_{(\mu_Y, b)_2} \approx -\frac{\mu_Y}{b}$. Therefore, combining the two approximations, we will have $\Delta_{(\mu_Y, a, b)} \approx -\frac{a}{\mu_Y} - \frac{\mu_Y}{b}$. The additional term $\Delta_{(\mu_Y, a, b)}$ will affect the linear predictor, introducing bias in $\boldsymbol{\beta}^*$ and \mathbf{b}^*

Moving forward, the transformed parameters estimate $\hat{\boldsymbol{\beta}}^*$ and $\hat{\mathbf{b}}^*$ are related to the original parameters estimates ($\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$) as:

$$\mathbf{X}\hat{\boldsymbol{\beta}}^* + \mathbf{Z}\hat{\mathbf{b}}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}} - \frac{a}{\mu_Y} - \frac{\mu_Y}{b}$$

$$\mathbf{X}\hat{\boldsymbol{\beta}}^* + \mathbf{Z}\hat{\mathbf{b}}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}} - \frac{a}{\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}} - \frac{\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{b}}}{b}$$

Since *both* $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ contribute to μ_Y let us isolate their respective effects to estimate the bias of each parameter estimates more easily. Split the contributions of *both* $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{b}}$ into terms scaled by \mathbf{X} and \mathbf{Z} :

When $\mathbf{Z}\hat{\mathbf{b}}^* = \mathbf{0}$ and $\mathbf{Z}\hat{\mathbf{b}} = \mathbf{0}$, this simplifies the equation to:

$$\mathbf{X}\hat{\boldsymbol{\beta}}^* = \mathbf{X}\hat{\boldsymbol{\beta}} - \frac{a}{\mathbf{X}\hat{\boldsymbol{\beta}}} - \frac{\mathbf{X}\hat{\boldsymbol{\beta}}}{b}$$

Assuming \mathbf{X} is invertible, divide through \mathbf{X} to isolate $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\beta}}$, we will have

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - \frac{a}{\hat{\boldsymbol{\beta}}} - \frac{\hat{\boldsymbol{\beta}}}{b}$$

Now, when $\mathbf{X}\hat{\boldsymbol{\beta}}^* = \mathbf{0}$ and $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$, this simplifies the equation to:

$$\mathbf{Z}\hat{\mathbf{b}}^* = \mathbf{Z}\hat{\mathbf{b}} - \frac{a}{\mathbf{Z}\hat{\mathbf{b}}} - \frac{\mathbf{Z}\hat{\mathbf{b}}}{b}$$

Assuming \mathbf{Z} is invertible, divide through \mathbf{Z} to isolate $\hat{\mathbf{b}}^*$ and $\hat{\mathbf{b}}$, we will have

$$\widehat{b}^* = \widehat{b} - \frac{a}{\widehat{b}} - \frac{\widehat{b}}{b}$$

Such as explained above, if the the trnasformation process using a logit link function is unbiased, then the bias of the parameter estimates from the transformation process should be equal to β and b or it can be writteh as

$$E[\widehat{\beta}^*] = \beta \text{ and } E[\widehat{b}^*] = b$$

or it can also be shown that

$$\text{Bias}(\widehat{\beta}^*) = E[\widehat{\beta}^*] - \beta = 0 \text{ and } \text{Bias}(\widehat{b}^*) = E[\widehat{b}^*] - b = 0$$

Now, let's calculate the bias

$$\text{Bias}(\widehat{\beta}^*) = E[\widehat{\beta}^*] - \beta \text{ and } \text{Bias}(\widehat{b}^*) = E[\widehat{b}^*] - b$$

$$\text{We know that } \widehat{\beta}^* = \widehat{\beta} - \frac{a}{\widehat{\beta}} - \frac{\widehat{\beta}}{b} \text{ and } \widehat{b}^* = \widehat{b} - \frac{a}{\widehat{b}} - \frac{\widehat{b}}{b}$$

Therefore we will obtain the bias first for $\widehat{\beta}^*$

$$\text{Bias}(\widehat{\beta}^*) = E\left[\widehat{\beta} - \frac{a}{\widehat{\beta}} - \frac{\widehat{\beta}}{b}\right] - \beta$$

$$\text{Bias}(\widehat{\beta}^*) = E[\widehat{\beta}] - E\left[\frac{a}{\widehat{\beta}}\right] - E\left[\frac{\widehat{\beta}}{b}\right] - \beta$$

Assuming that $\widehat{\beta}$ is unbiased or $E[\widehat{\beta}] = \beta$, then we will have

$$\text{Bias}(\widehat{\beta}^*) = \beta - \frac{1}{\beta}a - \beta\frac{1}{b} - \beta$$

$$\text{Bias}(\widehat{\beta}^*) = -\frac{1}{\beta}a - \beta\frac{1}{b}$$

$$\text{Bias}(\widehat{\beta}^*) = -\frac{ab+\beta^2}{\beta b}$$

Next, we will obtain the bias for \widehat{b}^*

$$\text{Bias}(\widehat{b}^*) = E[\widehat{b}^*] - b$$

$$\text{Bias}(\widehat{b}^*) = E\left[\widehat{b} - \frac{a}{\widehat{b}} - \frac{\widehat{b}}{b}\right] - b$$

$$\text{Bias}(\widehat{b}^*) = E[\widehat{b}] - E\left[\frac{a}{\widehat{b}}\right] - E\left[\frac{\widehat{b}}{b}\right] - b$$

assuming that \widehat{b} is unbiased or $E[\widehat{b}] = b$, then we will have

$$\text{Bias}(\widehat{b}^*) = b - \frac{1}{b}a - b\frac{1}{b} - b$$

$$\text{Bias}(\widehat{b}^*) = -\frac{1}{b}a - b\frac{1}{b}$$

$$\text{Bias}(\widehat{b}^*) = -\frac{ab+b^2}{bb}$$

8. Bias Estimation of the Predicted Values when using a Logit Link Function

Next, when we take the exponential of both sides to remove the logarithm and estimate the odds ratio

$$\frac{\mu_Y - a}{b - \mu_Y} = \exp(\mathbf{X}\beta + \mathbf{Z}b)$$

$$\mu_Y - a = \exp(\mathbf{X}\beta + \mathbf{Z}b) (b - \mu_Y)$$

$$\mu_Y = \exp(\mathbf{X}\beta + \mathbf{Z}b) b - \mu_Y(\exp(\mathbf{X}\beta + \mathbf{Z}b)) + a$$

$$\mu_Y + \mu_Y(\exp(\mathbf{X}\beta + \mathbf{Z}b)) = \exp(\mathbf{X}\beta + \mathbf{Z}b) b + a$$

$$\mu_Y(1 + \exp(\mathbf{X}\beta + \mathbf{Z}b)) = \exp(\mathbf{X}\beta + \mathbf{Z}b) b + a$$

$$\mu_Y = \frac{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})b + a}{(1 + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}))}$$

The solution gives the transformed mean response μ_Y as a function of the predictors $\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{Z}\mathbf{b}$, along with the parameters a and b . The term $\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})b$ represents the scaled contribution of the predictors. While a indicates the shift introduced by the transformation. When $a = 0$ and $b = 1$ we will get the standard logistic function $\mu_Y = \frac{\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})}{1 + \exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})}$. Nevertheless if $a \neq 0$ and $b \neq 1$, the response μ_Y will also be rescaled and shifted by a and b .

Next, in more detail for potential bias in the predictions, let the true mean of Y be μ_Y and the estimated mean be $\widehat{\mu}_Y = \mu_Y + \epsilon$, where ϵ is a small estimation error.

Then

$$\text{logit}(\widehat{\mu}_{Y^*}) = \log\left(\frac{\widehat{\mu}_{Y^*}}{1 - \widehat{\mu}_{Y^*}}\right) = \log\left(\frac{\widehat{\mu}_Y - a}{b - \widehat{\mu}_Y}\right)$$

Substituting $\widehat{\mu}_Y = \mu_Y + \epsilon$

$$\text{logit}(\widehat{\mu}_{Y^*}) = \log\left(\frac{(\mu_Y + \epsilon) - a}{b - (\mu_Y + \epsilon)}\right) = \log\left(\frac{\mu_Y + \epsilon - a}{b - \mu_Y - \epsilon}\right)$$

For small ϵ , this can be approximated as:

$$\text{logit}(\widehat{\mu}_{Y^*}) = \log\left(\frac{\mu_Y - a}{b - \mu_Y}\right) + \log\left(1 + \frac{\epsilon}{\mu_Y - a}\right) - \log\left(1 + \frac{\epsilon}{b - \mu_Y}\right)$$

Using the Taylor expansion $\log(1 + x) \approx x$ for small x , we get:

$$\text{logit}(\widehat{\mu}_{Y^*}) \approx \log\left(\frac{\mu_Y - a}{b - \mu_Y}\right) + \left(\frac{\epsilon}{\mu_Y - a}\right) - \left(\frac{\epsilon}{b - \mu_Y}\right)$$

So the bias introduced by the estimation error ϵ is:

$$\text{Bias}(\text{logit}(\widehat{\mu}_{Y^*})) \approx \left(\frac{\epsilon}{\mu_Y - a}\right) - \left(\frac{\epsilon}{b - \mu_Y}\right)$$

This demonstrates that small errors in estimating μ_Y can introduce a bias in the logit-transformed mean μ_Y . The degree of bias depends on the relative values of μ_Y , a , and b .

The random effects \mathbf{b} add complexity. If \mathbf{b} is normally distributed, the transformation might alter the distribution of the random effects.

Consider the random effect $\mathbf{b}_j \sim N(\mathbf{0}, \sigma^2_j)$

$$g(\mu_Y) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}$$

$$\mu_Y = g^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b})$$

Through transformation

$$\mu_{Y^*} = \frac{\mu_Y - a}{b - a}$$

If g is non-linear, the transformation of the random effect might not preserve its original properties, leading to bias.

In summary, the transformation from a Four Parameter Beta distribution to a standard beta distribution in the context of a GLMM can introduce bias, particularly when non-linear link functions are used. The linear predictor, which includes both fixed and random effects, might not be preserved accurately under the transformation, leading to systematic differences in

parameter estimates and predictions. To mitigate this bias, one should carefully consider the link function and the impact of the transformation on both fixed and random effects.

3.5. Model Limitations and Further Development

Although the proposed approach demonstrates promising model fit and prediction accuracy. Nonetheless, we have shown that transforming a Four Parameter Beta distribution into a standard beta distribution within the framework of a GLMM can potentially introduce bias. Further studies should be done on developing a method for bias corrections. As an alternative solution, the researchers also suggest the development of a four-parameter beta mean and mode GLMM model based on Bayesian approaches following Zou *et al.* (2022).

Next, the model should also be improved by taking into consideration complex linear and non-linear relationships that might occur between the response and independent variables when w use various data sets. Such as in this study when we used satellite data in Central Kalimantan, the absolute correlation value between paddy productivity and the lagged NDVI values vary from 0.01 up to 0.4 (Figure 3.10). Indicating week to moderate linear relationships.

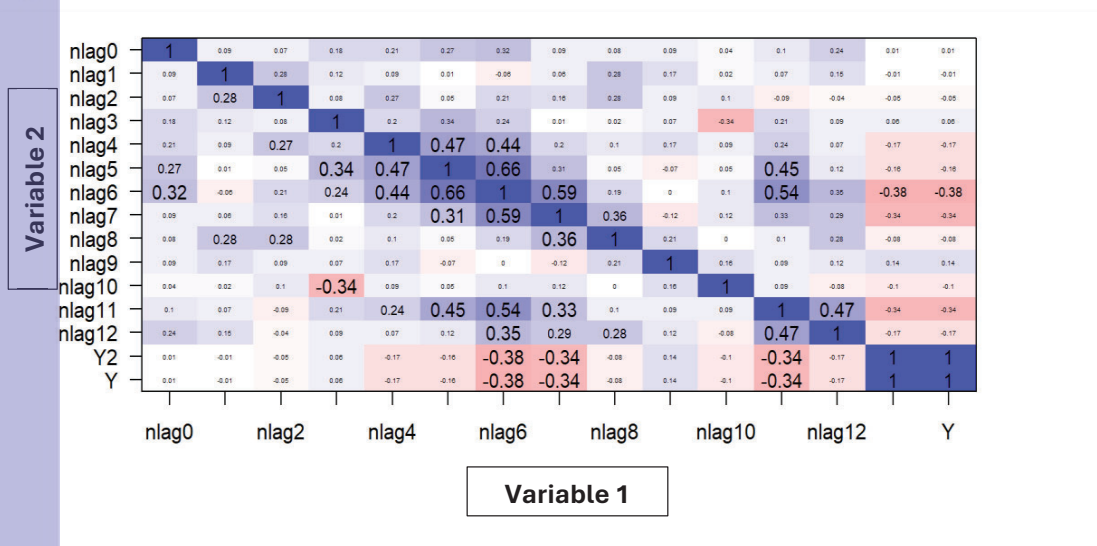


Figure 3.10 Correlation Plot of Lagged NDVI and Paddy Productivity

Figure 3.11 is a scatterplot matrix that displays relationships between paddy productivity, current and up to four lags of NDVI. We can see that paddy productivity and the lag four of NDVI have a negative weak linear relationship as indicated by the green regression lines. As for non linear relationships are indicated between paddy productivity and lag 1 up to lag 3 of NDVI. This is in line with the correlation plot above. More data exploration that displays relationships between paddy productivity and lags five up to lag twelve of NDVI can be seen in Appendix 3. This appendix also provides scatterplot matrix that displays relationships between paddy productivity, current, and up to twelve lags of Band 4 and Band 8 values. In general the patterns are similar of what has been presented here.

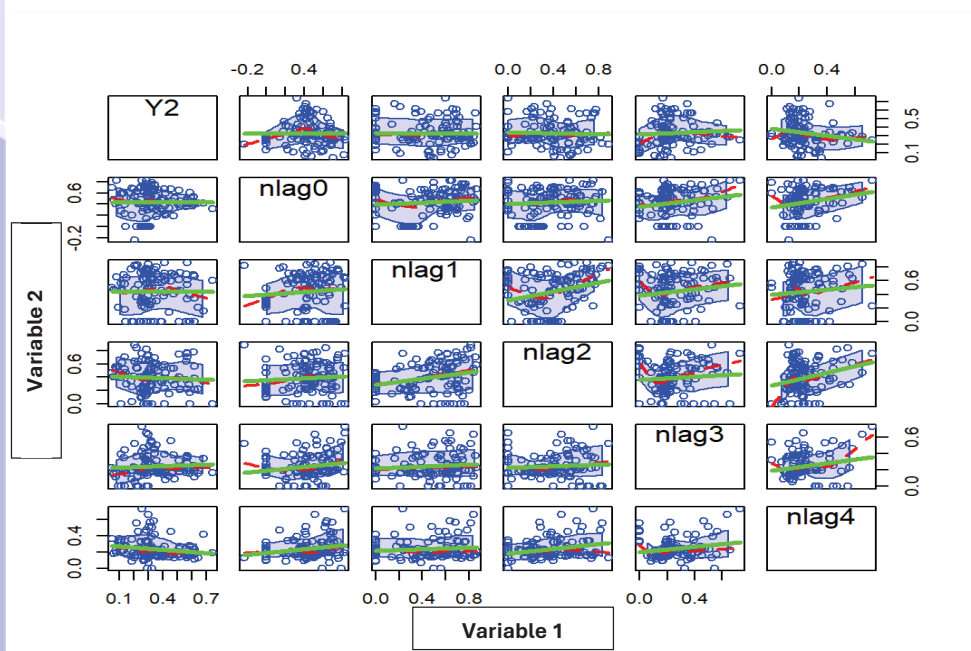


Figure 3.11 Scatterplot Between the Paddy Productivity, Current NDVI, and Lag One to Lag Four NDVI Values

Aside from the variables above, there are also other key variables that is used in this study taken from the farmer survey data. These variables are assumed to have an effect on paddy productivity, which include paddy variety, last year's paddy productivity, number of seeds used, and the amount of fertilizer components used. The relationship between the paddy productivity and these variables can be seen in Figure 3. 11. Among these variables, the amount of Solid Organic Fertilizer /Compost shows the strongest positive linear relationship with paddy productivity. Meaning that there is a tendency that with proper dosage, more use of this substance leads to higher paddy productivity. This fact is also supported by the value of correlation between these variables, which is 0.52. The dosage of Urea also shows similar patterns, but having lower impacts while others tend to show weak relationship with paddy productivity.

Therefore further developing a flexible model that takes into account (1) the Four Parameter Beta distribution, (2) area variability, and (3) complex linear and non linear relationship between the independent and dependent variables should be done. Therefore, we can consider further developing a Four Parameter Beta Generalized Mixed Effect Tree (GMET), which incorporates the Four Parameter Beta GLMM and regression trees (Fontana *et al.* 2021). Another advancement is incorporates the Four Parameter Beta GLMM and random forest, thus we will have a Four Parameter Beta Generalized Mixed Effect Random Forest (GMERF) that was first introduced by Pellagatti *et al.* (2021).

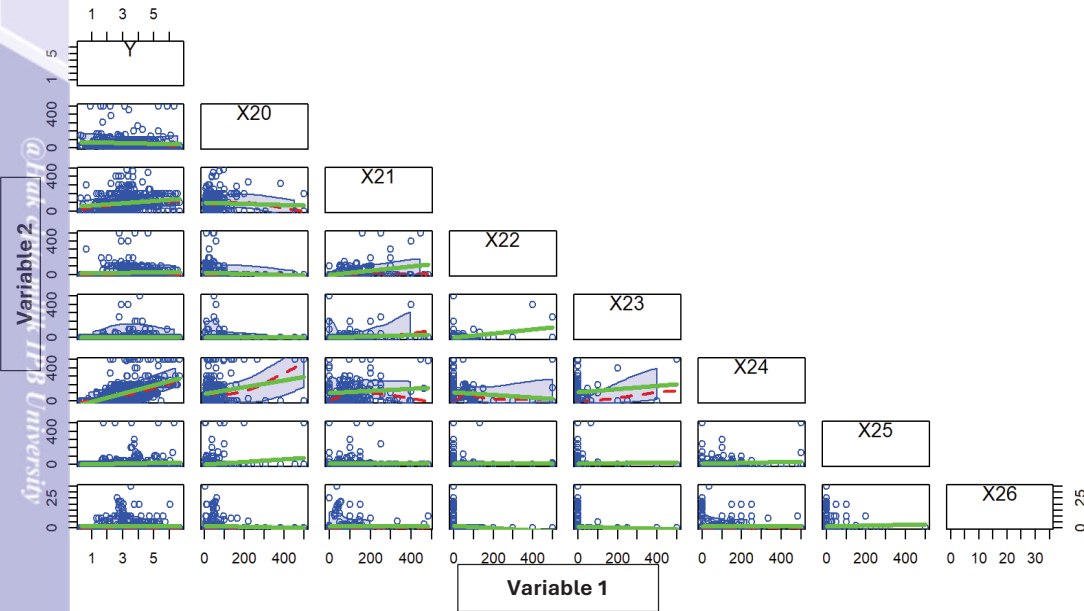


Figure 3.12. Scatterplot Between the Paddy Productivity, Number of Seeds Used, and the Amount of Fertilizer Components Used

Applicationwise, in further developing this approach for predicting paddy productivity predictions there are some other limitations that need to be taken into consideration., such as:

1. This study has used two areas, Central Kalimantan and Karawang, which tend to have different agricultural conditions. Further on we must also apply this approach in other areas that have different characteristics compared to Central Kalimantan and Karawang to increase the variation of the input dataset. Hence, enhances model generalization, robustness of the parameter, and increases prediction accuracy.
2. We should also consider the reliance on historical satellite data which may not fully capture current conditions. It is known that satellite data has certain temporal and spatial precision limitations depending on the period it was taken, and which type of satellite was used. Satellite data also can suffer from quality issues, such as cloud cover, sensor malfunctions, or calibration errors. These issues can introduce noise or gaps in the data. Thus, it is crucial to use up to date satellite data and evaluate models against current circumstances to overcome these limitations.
3. Consider other independent variables in the model including phenology, climate, farming practices, and geographical information data to enhance rice yield predictions under climate change conditions. Phenological variables play a crucial role in rice yields as they are closely related to the growth and development stages of the crop. Phenology can be measured by using UAVs (unmanned aerial vehicles) equipped with multi-spectral and hyperspectral sensors that have been proven to be a highly effective tool (Guo *et al.* 2021, 2022, 2023). Climate data include temperature, humidity, rainfall, precipitation, etc. Next, incorporating farming practices, particularly the use of fertilizers, is critical. Fertilizers play a significant role in providing the nutrients necessary for optimal crop growth, especially in rice farming, where nutrient demands can vary significantly depending on soil quality, water availability, and climatic conditions. Meanwhile, geographical information factors such as latitude, longitude, and altitude remain constant over time for a given location can be further explored by

considering spatial weights. These factors were proven to play a crucial role in rice productivity predictions (Guo *et al.* 2021)

Despite these limitations, the study provides valuable insights into the use of a transformation process in modeling a GLMM for a response variable following a Four Parameter Beta distribution for paddy productivity prediction. Our proposed four-parameter beta GMET and GLMM model have addressed the following advantages:

1. Obtain predictions for response variables that have a Four Parameter Beta distribution that varies between areas, skewed, and bounded within a certain minimum and maximum range through a transformation process.
2. Through this transformation process, the GLMM models has been able to increase prediction accuracy compared to a GLM model in areas that have high response variables variability.
3. Visualization of the relationships between the response and independent variables that were considered most important is an additional step that will help analyst and decision makers to deliver in-depth analysis and comprehensive recommendations.

3.6. Conclusion and Policy Recommendations

One of the main goals of this research is to find a model that enhances paddy productivity predictions. The goal was accomplished by applying a transformation approach for a response variable that follows a Four Parameter Beta distribution and applying the beta GLMM. This model is much more suitable for predicting paddy productivity compared to a GLM model. It was also shown that the use of both short-term and long-term satellite data is significant. This pointed out that it is possible to predict paddy productivity using Sentinel 2A satellite data along with survey data analyzed with the sufficient prediction model. As a result, it will be more time and cost efficient compared to just relying on survey data.

We have shown that this model is quite promising due to its high prediction accuracy. Thus, the Ministry of Agriculture, NGOs, Insurance companies, Farmers and other related stakeholders can benefit to further develop and apply paddy predictions based on this model. Listed below are some recommendations for policy makers and practitioners on how to effectively use this model in practice:

- Formulate policies and field action plans that support sustainable agricultural practices based on model predictions. For example, developing policies and action plans promoting crop diversification in areas with low predicted paddy productivity can help mitigate risks.
- Use predictions to identify and protect regions that are impacted by climate change by implementing suitable mitigation measures.
- Establish early warning systems for early warning systems of food insecurity and a reference for the food self-sufficiency program.
- We should also train field staff and farmers on how to interpret and use productivity predictions for their benefit. Therefore, enhance the implementation of precision farming techniques by combining GLMM predictions with GPS and IoT technologies for better field management.
- Regularly monitor the accuracy of predictions and adjust model and data accordingly. This involves comparing predicted productivity with actual outcomes to evaluate the model's performance.
- Scale up model especially in areas where there might not be any access to individual farmers' surveys and CCEs. This model is not only limited to be used to predict paddy productivity but it can also be applied to other case studies that have similar conditions mentioned in this study.



Besides the recommendation given above, further research development can be done in developing a Four Parameter Beta Mean and Mode GLMM, GMET, and GMERF model based on Bayesian approaches to prevent possibilities of bias in parameter estimates and complications in the interpretation of coefficient values when transformation process is being performed. It is also suggested to apply and evaluate the model in other various areas to increase the generalization, robustness, and prediction accuracy. We also recommend that the model be developed by adding relevant phenology, climate, and geographical information variables.

- Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber :
 - a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
 - b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
 2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

IV. The Four Parameter Beta GLMM Based on Bayesian Approach

4.1. Introduction

The Four Parameter Beta distribution is a distribution having a specific maximum and minimum value. Although it is not as common as the beta distribution, the Four Parameter Beta distribution is more flexible because it is not restricted to values between 0 and 1. It can also accommodate a wide range of shapes, skewness and even heavy tails, allowing econometric or predictive models to capture different patterns in the data. This flexibility is valuable when dealing with different economic phenomena that have different characteristics. Therefore, a model developed based on a Four Parameter Beta distribution can be used in many economic and business domains, such as modeling inflation, growth, index, rates, loss, etc. For example, the average growth of gross domestic product (GDP) in a country or region over the last ten years has a minimum value and can only reach a maximum value. In Indonesia, for example, the average GDP growth from 2009 to 2022 is 4.71 percent, reaching a maximum of 6.38 percent in 2010 and a minimum of -2.07 percent in 2020. Another example that we will discuss in more detail is in the context of Area Yield Insurance (AYI) in the agricultural sector.

The Four Parameter Beta distribution can also be used to model the productivity distribution of a particular crop in each region. We know that crop productivity is inherently bound to a certain minimum value (often greater than zero) and a certain maximum value determined by agronomic, climatic, and technological factors. These upper and lower values can vary significantly across regions due to differences in soil fertility, weather patterns, irrigation infrastructure, and farming practices. Moreover, productivity data is rarely symmetrically distributed. Instead, it often shows skewness because of natural limits on yield, higher variability on the downside (due to shocks and failures), and uneven access to resources or technology across farmers and regions. In this context, developing GLMM models based on the Four Parameter Beta distribution can be especially valuable, as they allow for the incorporation of both fixed and random effects while accommodating the bounded and skewed nature of productivity data. The predicted productivity value of this model are expected to be more accurate and reliable, thereby supporting their application in designing Area Yield Index (AYI) insurance schemes. By more precisely capturing the variability and distributional features of regional crop productivity, the resulting estimates can better capture regional yield variability and risks. These models will contribute to improved risk assessment, fair premium setting, and the design of payout structures that align closely with the actual conditions experienced by farmers.

In contrast to unbounded data, it is known that for bounded data, central tendency measures, skewness, and other characteristics of the underlying distribution are reliant on the distribution's support or range of values for which the Probability Density Function (PDF) is non-zero. As a result, greater attention must be taken when drawing conclusions for these features based on bounded data, especially when the support is unknown (Zhou and Huang, 2022). Nevertheless, existing approaches for analyzing the Four Parameter Beta distribution are limited. Most references have opted to transform the data into a beta distribution that has a prefixed support of (0, 1) and applied a beta mean regression (Ferrari and Neto, 2004). Furthermore, the model has been expanded to permit covariates to affect the precision parameter (Smithson and Verkuilen, 2006; Ferrari *et al.* 2011). On CRAN, you can find the *betareg* R package (Cribari-Neto and Zeileis, 2010; Grün *et al.* 2012) that can be used to fit the beta mean regression model with different levels of precision and perform model diagnostics. However, we must keep in mind that in many cases, the transformation process can lead to biases in the parameter estimates and complications in the interpretation of coefficient values



(Kusumaningrum, 2024). We have also shown the potential bias of transforming a response variable that follows a Four Parameter Beta to a beta distribution and apply the beta GLM and GLMM models in sub section 3.4. Thus, it was suggested in sub section 3.5 that the researchers should develop a Four Parameter Beta mean and mode GLMM model based on Bayesian approaches following Zou and Huang (2022).

Zhou and Huang (2022) developed the mean and mode Four Parameter Beta regression model based on a Bayesian approach. They also provided the R package betaBayes, which estimates the parameters and provides model evaluations. Nevertheless, the inclusion of a random effects component to a model is often very important. As in econometrics or paddy productivity predictions, researchers often analyze panel data and cross-sectional data. In panel data, observations are collected over time for multiple individuals or units. Therefore, random effects can capture individual-specific effects that vary over time and unobserved heterogeneity across units. In contrast for cross sectional data, where data are collected by observing multiple units at a single point in time, the use of random effects is less common than for panel data. Nevertheless, there are many cases in which heterogeneity across units occur.

Hence, we further developed Zhou and Huang's regression model by including a random effects component and introducing the Four Parameter Beta mean GLMM. This model was developed based on a Bayesian inference by block sampling Markov Chain Monte Carlo (MCMC) approach. This gives us the advantages of (1) handling uncertainty within the specific complex model framework, (2) making inferences for all unknown parameters of the model, including the fixed effects, random effects, precision, and support bounds (minimum and maximum value), (3) obtaining probabilistic predictions, prediction intervals, and information about the distribution of possible outcomes, and (4) appropriately selecting a prior distribution for setting parameter restrictions.

To evaluate the process of developing the Four Parameter Beta mean GLMM model based on a Bayesian approach, we conducted a simulation study. The analysis of the accuracy and efficiency of the estimated parameter as well as the accuracy of the predictions are the focus of the evaluation process. Subsequently, this model is applied to predict paddy productivity in the Indonesian province of Central Kalimantan. This province was chosen because it is one of the provinces in Indonesia that has been selected for Indonesia's food estate program and the level of productivity among areas in this province also tends to be diverse. Paddy productivity itself has a bounded distribution with a certain minimum and maximum value. Paddy productivity conditions usually vary from area to area as they depend on agroclimatic conditions. Nevertheless, accurate predictions of paddy productivity in Indonesia are very important. Not only because rice is a staple food in Indonesia, but also because Indonesia has started to conduct pilot studies and introduce an alternative Area Yield Index crop insurance policy from 2022 (Sanyu Consultants Inc. and Sompom Risk Management Inc, 2023). This policy is a productivity-based index agricultural insurance policy that is highly dependent on accurate paddy productivity predictions in various areas.

The remainder of this article is organized as follows. Section 4.2 describes the proposed method, where we begin with an introduction to the Four Parameter Beta distribution and then explain the Four Parameter Beta regression and its further development. We then explain the Bayesian inference in subsection 4.2.1 to 4.2.4. Here, we have described the model specifications, the selection of priors, the estimation and evaluation of parameter models, the model selection criteria, and the predictions. Next in sub section 4.3 we explain about the simulation process and simulation results. While in subsection 4.4 we have explained about the empirical case of study beginning with the research process, results, and discussions. In the empirical study we will apply the developed model to predict paddy productivity. Section 4.5 we will explain about the model's limitation. Last section 4.6 presents the conclusions, summarizing the results of our research and outlining potential future research.

4.2. Proposed Method

The Four Parameter Beta distribution is one of the continuous variable distributions that have an interval of (a, b) and has a probability density function that can be seen in Equation 2.1. It is known that a is the minimum value and b is the maximum value. This distribution also has aslocation parameter (ω_1) and the scale parameter $(\omega_1 - \omega_2)$. Furthermore, the mean of a Four Parameter Beta distribution can be seen in Table 2. 1. The Four Parameter Beta distribution is highly flexible and can model a wide range of shapes, including symmetric and skewed distributions. This flexibility makes it suitable for fitting a variety of data patterns.

It has been known that modeling data that follow a Four Parameter Beta distribution is not commonly found. Nevtrtheless, Zhou and Huang (2022) developed the mean Four Parameter Beta mean regression model that has been explained in sub section 2. 4 and the model's equation can be seen in Equation 2.9.

Inferences of the developed model is based on Bayesian method. There are three main aspects of Bayesian approach in developing a Four Parameter Beta GLMM, which includes 1) model specification, 2) prior distribution selection, 3) parameter model estimation, 4) Predictions. Each aspect will be explained in more detail below.

4.2.1. Model Specifications

In this paper we will introduce an advancement of the Four Parameter Beta regression model, where in this model we assume that the response variable has a Four Parameter Beta distribution or $Y_{ij} \sim B(\omega_{1ij}, \omega_{2ij}, a, b)$ where object $j = 1, 2, \dots, n_i$ is measured within a certain groups/area and $i = 1, 2, \dots, q$. It is also known that, when $W_{ij} = \frac{Y_{ij} - a}{b - a}$ will also have a beta distribution with shape parameters ω_{1ij} and ω_{2ij} . Let $\mu_{W_{ij}}$ and $\mu_{Y_{ij}}$ denote the mean for W_{ij} and Y_{ij} , thus $\mu_{W_{ij}} = \frac{\omega_{1ij}}{\omega_{1ij} + \omega_{2ij}}$ and $\mu_{Y_{ij}} = a + \frac{\omega_{1ij}}{\omega_{1ij} + \omega_{2ij}}(b - a) = a + \mu_{W_{ij}}(b - a)$. Next, we can set $\omega_{1ij} = \varphi m_{ij}$ and $\omega_{2ij} = \varphi(1 - m_{ij})$, for $0 < m_{ij} < 1$ and $\varphi > 0$, where φ is the precision parameter such as explained in sub chapter 2.4.

In this model we further assumed that the response variable doesn't only depend on the independent variables $X_{ij} = (X_{1ij}, X_{2ij}, \dots, X_{Kij})$ but it is also affected by the variance between objects/groups/areas (McCullagh dan Nelder 1989). Therefore, the advanced and novel Four Parameter Beta mean GLMM can be formulated as:

$$g \left[\frac{\text{Mean}[Y_{ij}|X_{ij}] - a}{b - a} \right] = X_{ij}^T \beta + Z_{ij}^T b_i \quad (4.1)$$

The random effect matrix design is denoted by Z_{ij}^T , while b_i represents the random effect estimators in a particular entity/group/area. Following Bonat *et al.* (2015), we assumed that $b_i \sim \text{iid } N(0, \sigma_i^2)$ and applied the logit link function.

In this model a greater mean value results from a higher quantile-score. The quantile-score for the mean value increases by β_k units for every unit increase in X_k , where k is the amount of independent variables used in the model. Meanwhile, b_i represents the difference in the quantile-score for the mean value that can be attributed to entities/groups/areas that are not directly included in the model. These factors may include unmeasured individual characteristics or time-varying confounding variables that are not accounted for in the fixed effects of the model.

To estimate the parameters β, b_i , alongside the minimum (a) and maximum (b) value, and the precision value (φ) of the Four Parameter Beta mean GLMM a Bayesian approach will be used, which is specified below

$$Y_{ij} | m_{ij}, \varphi, a, b \sim B(\varphi m_{ij}, \varphi(1 - m_{ij}), a, b)$$



$$m_{ij} \equiv m(X_{ij}) = g^{-1}(X_{ij}^T \boldsymbol{\beta} + Z_{ij}^T \mathbf{b}_i) \quad (4.2)$$

$$p(\boldsymbol{\beta}, \mathbf{b}_i, \varphi, a, b) = p(\boldsymbol{\beta})p(\mathbf{b}_i)p(a)p(b)p(\varphi)$$

where $g(\cdot)$ is a link function, commonly logit, but also can be probit, and log-log. While $p(\cdot)$ represents a prior density, φ is the precision parameter, and $\omega_1 = \varphi m_{ij}$ while $\omega_2 = \varphi(1 - m_{ij})$, for $0 < m_{ij} < 1$ and $\varphi > 0$.

4.2.2. Prior Distributions

After defining the model specifications of the Four Parameter Beta mean GLMM we need to define the prior distributions for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{b}_i, a, b, \varphi)$ by specifying the parameters probability distributions (e.g. normal, uniform, gamma, etc.). The main aim of Bayesian statistics is to find an appropriate prior distribution of model parameters with a probability distribution and use the current data to update the prior to create a posterior distribution with reduced uncertainty (Lynch, 2007). Nevertheless, selecting the right choice of prior tends to be challenging.

In general, there are three categories of prior distribution, which are: non-informative, informative, and weakly informative priors. A non-informative or flat prior gives each value an equal probability and does not consider any previous information. Alternatively, informative prior typically characterizes a particular range of parameter values and expresses a stronger pre-existing belief because it is commonly derived from previous empirical research. Meanwhile the weakly informative priors aim to balance between the first two by including "weak" parameter information that covers all potential "real-world" values without assigning any value with an excessively high probability (Johnston *et al.* 2023).

Referring to Zhou and Huang (2022) it is suggested that the prior's distribution should be specified as:

- i. **Prior for Minimum (a) and Maximum Value (b)** is a uniform distribution $a \sim \text{unif}(a_1, b_1)$ and $b \sim \text{unif}(a_2, b_2)$. If we have information of previous conditions, we can set $y_{(1)} (a_1 < y_{(1)} < b_1)$ and $y_{(k)} (a_2 < y_{(k)} < b_2)$. Where $y_{(k)}$ refers to the k-th order statistic $\{y_{(1)}, \dots, y_{(k)}\}$. Nonetheless, if we do not have previous knowledge of these boundaries, we can set $a_1 = y_{(1)} - \Delta$, $b_1 = Y_{(1)} - 10^{-15}$, $a_2 = Y_{(n)} + 10^{-15}$, $b_2 = Y_{(n)} + \Delta$, where $\Delta > 10^{-15}$ or $\Delta = 2s_y$, where s_y is the sample standard deviation of y_i 's
- ii. **Prior for the Precision Parameter (φ)** can be inferred from the observed data. Nevertheless, when there is limited information, priors can be set to a gamma distribution, where $\Gamma(a_\varphi, b_\varphi)$ and $a_\varphi = b_\varphi = 0.001$ are the defaults. When doing so, we believe that the precision parameter is small and does not heavily influence the posterior.
- iii. **Prior for the Fixed Effect Parameters ($\boldsymbol{\beta}$)** must take into attention the prior information while adjusting for covariates. For this purpose, Zhou (2022) considered the g -prior (Zellner, 1986) where $\boldsymbol{\beta} \sim N_{p+1}(b\mathbf{e}_1, g\mathbf{X}_{ij}'\mathbf{X}_{ij})^{-1}$, $\mathbf{e}_1 = (1, 0, \dots, 0)'$ is of length $p + 1$, b is a prior mean for the intercept, and $g > 0$ is a scaling constant. In our purposed model we can also consider using parameter estimates of a Four Parameter Beta regression model as the prior for the fixed effect ($\boldsymbol{\beta}$) because this model is an advanced extension of the Four Parameter Beta regression
- iv. **Prior for Random Effect Parameters (\mathbf{b}_i)** is commonly based on the normal distribution. Therefore, we can set $\mathbf{b}_i \sim \text{iid } N(0, \boldsymbol{\Sigma})$ and the variance in the matrix variance covariance ($\boldsymbol{\Sigma}$) should be greater than zero.

In this paper, we have applied informative prior and weakly informative priors because often non-informative priors will lead to bias parameter estimates, especially when the sample size is

small (LeMoine, 2019). Besides that, in selecting the priors we have considered relevant research studies.

4.2.3. Bayesian Parameter Estimation

In general we know that a bayesian parameter estimate of θ can be calculated as:

$$\widehat{\theta}_B = E(\theta|Y_{ij}) = \int \theta \pi(\theta|Y_{ij}) d\theta$$

Where the posterior density function is $\pi(\theta|Y_{ij}) = \frac{f(Y_{ij}, \theta)}{f(Y_{ij})}$ and the marginal distribution function is $f(Y_{ij}) = \int f(Y_{ij}, \theta) d\theta$. Next, the joint distribution function is formulated as

$$f(Y_{ij}, \theta) = f_{Y_{ij}}(Y_{ij}|\theta)\pi(\theta)$$

Here $f_{Y_{ij}}(Y_{ij}|\theta)$ is the likelihood function and $\pi(\theta)$ is the prior.

Incorporating the priors and the likelihood function allows us to obtain Bayesian estimation of the parameters of a Four Parameter Beta mean GLMM through the posterior distributions. Given data $D = \{(y_{ij}, x_{ij}), j = 1, 2, \dots, n_i \text{ and } i = 1, 2, \dots, q\}$, the likelihood function of parameters $\theta = (\beta, b_i, a, b, \varphi)$ associated with the Four Parameter Beta mean GLMM can be given as

$$f_{Y_{ij}}(D|\theta) = L(D|\theta) = \prod_{i=1}^q \prod_{j=1}^{n_i} \frac{\Gamma(\omega_{1ij} + \omega_{2ij})(y_{ij} - a)^{\omega_{1ij}-1} (b - y_{ij})^{\omega_{2ij}-1}}{\Gamma(\omega_{1ij})\Gamma(\omega_{2ij})(b - a)^{\omega_{1ij} + \omega_{2ij}-1}} \quad (4.3)$$

Where $\omega_{1ij} = \varphi m_{ij}$ and $\omega_{2ij} = \varphi(1 - m_{ij})$, for $0 < m_{ij} < 1$ and $\varphi > 0$.

The posterior density function then can be formulated as:

$$\pi(\theta|D) = \frac{f(D, \theta)}{\int f(D, \theta) d\theta}$$

$$\text{Or if } p(\beta, b_i, \varphi, a, b|D) \propto L(D|\theta) \times \exp\left\{-\frac{1}{2gn}(\beta - be_1)'X_{ij}'X_{ij}(\beta - be_1)\right\} \times \exp\left(-\frac{1}{2}b_i\Sigma^{-1}b_i\right) \times I(a_1 < a < b_1)I(a_2 < b < b_2) \times \varphi^a\varphi^{-1} \exp(-b_\varphi\varphi) \quad (4.4)$$

$$\text{Then, the posterior density function } \pi(\theta|D) = \frac{p(\beta, b_i, \varphi, a, b|D)}{\int_\theta p(\beta, b_i, \varphi, a, b|D)}$$

To find the parameter estimates of $\widehat{\theta}_B$. We need to solve for each parameter $\beta, b_i, a, b, \varphi$

$$\widehat{\theta}_B = E(\theta|D) = \int \theta \pi(\theta|D) d\theta$$

The denominator $\int_\theta p(\beta, b_i, \varphi, a, b|D)$ is the normalizing constant (also called the evidence or marginal likelihood). Which ensures that the posterior distribution integrates to 1. However, in practice. This integral is constant for a fixed dataset D. So, when calculating the posterior mean, $E(\theta|Y)$, the proportionality constant cancels out because it does not depend on the parameters. Therefore, we have

$$\begin{aligned} \pi(\theta|D) &\propto p(\beta, b_i, \varphi, a, b|D) \\ \widehat{\theta}_B = E(\theta|D) &\propto \int \theta p(\beta, b_i, \varphi, a, b|D) d\theta \end{aligned}$$

Example for parameter β

$$\widehat{\beta} = E(\beta|D) = \int \beta \pi(\beta|D) d\beta$$

Where $\pi(\beta|D)$ is the posterior distribution of β marginalized over all other parameters b_i, φ, a, b , hence

$$\pi(\beta|D) \propto \int \int \int \int p(\beta, b_i, \varphi, a, b|D) db_i d\varphi da db$$

And for other parameters, we have

$$\pi(\mathbf{b}_i | \mathbf{D}) \propto \int \int \int \int p(\boldsymbol{\beta}, \mathbf{b}_i, \varphi, a, b | \mathbf{D}) d\boldsymbol{\beta} d\varphi da db$$

$$\pi(\varphi | \mathbf{D}) \propto \int \int \int \int p(\boldsymbol{\beta}, \mathbf{b}_i, \varphi, a, b | \mathbf{D}) d\boldsymbol{\beta} d\mathbf{b}_i da db$$

$$\pi(a | \mathbf{D}) \propto \int \int \int \int p(\boldsymbol{\beta}, \mathbf{b}_i, \varphi, a, b | \mathbf{D}) d\boldsymbol{\beta} d\mathbf{b}_i d\varphi db$$

$$\pi(b | \mathbf{D}) \propto \int \int \int \int p(\boldsymbol{\beta}, \mathbf{b}_i, \varphi, a, b | \mathbf{D}) d\boldsymbol{\beta} d\mathbf{b}_i d\varphi da$$

In many Bayesian analyses, computing the marginal posterior $\pi(\boldsymbol{\theta} | \mathbf{D})$ and the integral explicitly is challenging. Even, for a beta distribution it does not have a closed form. Instead, we can use MCMC. Where in this study the Posterior sampling is then conducted based on adaptive Metropolis samplers (Haario *et al.* 2001). There was a consideration that central tendency measures of bounded data typically entangle with the support of the underlying distribution. Thus, the posterior distribution of $\boldsymbol{\beta}$ and (a, b) are often highly correlated. Zhou and Huang (2022) have considered this problem and updated it into a single block to effectively eliminate problematic MCMC mixing. It has been well documented in the literature that block sampling can improve MCMC efficiency relative to updating each parameter independently (Liu *et al.* 1994; Roberts and Sahu, 1997; Sargent *et al.* 2000). The algorithm for posterior sampling that incorporates block sampling can be seen in Zhou and Huang (2022).

4.2.4. Predictions of the Response Variable

Prediction in the Four Parameter Beta GLMM implies estimating the response variable based on fixed and random effects while accounting for the Four Parameter Beta distribution's constraints (a and b) and the link function. As we know, the response is modeled between bounds a and b and it depends on shape parameters ω_1 and ω_2 that are derived from the predicted mean μ and a precision parameter ϕ . The stages of obtaining a prediction are:

1. We will have $\boldsymbol{\theta}^{(s)} = (\boldsymbol{\beta}^{(s)}, \mathbf{b}_i^{(s)}, a^{(s)}, b^{(s)}, \varphi^{(s)})$ after generating posterior samples using the methods explained above. Here, (s) refers to the index of posterior samples obtained from MCMC
2. Then we compute the linear predictor ($\eta_{ij}^{(s)}$) by using the posterior samples of $\boldsymbol{\beta}^{(s)}$ and $\mathbf{b}_i^{(s)}$. First, recall the modeling framework in Equation 4.2, Thus we will have

$$\eta_{ij}^{(s)} = g(\mu_{w_{ij}}^{(s)}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}^{(s)} + \mathbf{Z}_{ij}^T \mathbf{b}_i^{(s)}$$

where g is the logit link function, solve for $\mu_{w_{ij}}$, then

$$\mu_{w_{ij}}^{(s)} = g^{-1}(\mathbf{X}_{ij}^T \boldsymbol{\beta}^{(s)} + \mathbf{Z}_{ij}^T \mathbf{b}_i^{(s)})$$

3. Next, apply the inverse logit function to obtain the predicted mean of the response variable

It is known that $g^{-1}(z) = \frac{\exp(z)}{1 + \exp(z)}$, hence

$$\mu_{w_{ij}}^{(s)} = \frac{\exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}^{(s)} + \mathbf{Z}_{ij}^T \mathbf{b}_i^{(s)})}{1 + \exp(\mathbf{X}_{ij}^T \boldsymbol{\beta}^{(s)} + \mathbf{Z}_{ij}^T \mathbf{b}_i^{(s)})}$$

Due to the reparameterization described in section 4.2, expanding from Equation 4.1 for the posterior samples it is known that $\mu_{w_{ij}}^{(s)} = \frac{Mean[y_{ij}|x_{ij}]^{(s)} - a^{(s)}}{b^{(s)} - a^{(s)}}$, hence

$$\mu_{w_i}^{(s)} = \frac{Mean[y_{ij}|x_{ij}]^{(s)} - a^{(s)}}{b^{(s)} - a^{(s)}} = \frac{\exp(X_{ij}^T \beta^{(s)} + Z_{ij}^T b_i^{(s)})}{1 + \exp(X_i^T \beta^{(s)} + Z_i^T b_i^{(s)})}$$

$$\mu_{w_{ij}}^{(s)} = Mean[y_{ij}|x_{ij}]^{(s)} = \frac{a^{(s)} + (b^{(s)} - a^{(s)}) \exp(X_{ij}^T \beta^{(s)} + Z_{ij}^T b_i^{(s)})}{1 + \exp(X_i^T \beta^{(s)} + Z_i^T b_i^{(s)})} \quad (4.6)$$

This gives the predicted mean of the response variable in the Four Parameter Beta distribution for each sample (s).

4. Drive the Four Parameter Beta distribution parameters $\omega_{1ij}^{(s)}$ and $\omega_{2ij}^{(s)}$ for each MCMC sample. We have obtained the predicted mean of the response variable for each MCMC sample. Nevertheless, the distribution of the response variable also consists of the shape parameters ω_{1ij} and ω_{2ij} . These parameters indicate how skewed or peaked the distribution is and most importantly the variance of the distribution. The precision parameter (φ) indicates the variance and is related to both ω_{1ij} and ω_{2ij} . Zou and Huang (2022) has explained that $\omega_{1ij}^{(s)}$ and $\omega_{2ij}^{(s)}$ can be driven by $\omega_{1ij}^{(s)} = \varphi^{(s)} m_{ij}$ and $\omega_{2ij}^{(s)} = \varphi^{(s)} (1 - m_{ij})$, where $m_{ij} \equiv \mu_{w_{ij}}^{(s)}$ for $0 < m_{ij} < 1$ and $\varphi > 1$
5. Generate the predicted response $Y_{ij}^{(s)} \sim B(\omega_{1ij}^{(s)}, \omega_{2ij}^{(s)}, a^{(s)}, b^{(s)})$ based on the estimated $\omega_{1ij}^{(s)}$ and $\omega_{2ij}^{(s)}$ that has been obtained from step 4 and the estimated $a^{(s)}$ and $b^{(s)}$ that has been explained in step 1.
6. Summarize the predictions that have been given in step 5. We will have $\{Y_{ij}^{(s)}\}_{s=1}^S$ for all posterior samples ($s = 1, 2, \dots, S$), therefore we can calculate summary statistics, such as:

- a. The posterior predictive mean can be calculated based on

$$E[Y_{ij}] = \frac{1}{S} \sum_{s=1}^S Y_{ij}^{(s)} \quad (4.7)$$

- b. The posterior predictive median

Sort the simulated predictions: $\{Y_{ij}^{(1)}, Y_{ij}^{(2)}, \dots, Y_{ij}^{(S)}\}$. Afterwards, we can compute the median by taking the middle value based on the number of samples (S), when S is odd. If S is even the median is the average of the two middle values

- c. The posterior variance and other relevant statistical values
The posterior variance can be calculated based on

$$Var[Y_{ij}] = \frac{1}{S} \sum_{s=1}^S (Y_{ij}^{(s)} - E[Y_{ij}])^2 \quad (4.8)$$

While other relevant statistics can be also calculated based on the formulation examples given

- d. Construct credible interval (α) for the predicted responses. If $\alpha = 0.05$, then we will have the 95% credible intervals for the predicted responses. Like calculating the median, we first need to sort the simulated predictions: $\{Y_{ij}^{(1)}, Y_{ij}^{(2)}, \dots, Y_{ij}^{(S)}\}$. Next, we need to identify the lower and upper bound, where the lower bound (CI_{low}) and upper bound (CI_{high}) are formulated as

$$CI_{low} = Y_{ij}(\lceil \frac{\alpha}{2} \cdot s \rceil) \text{ and } CI_{high} = Y_{ij}(\lceil (1 - \frac{\alpha}{2}) \cdot s \rceil) \quad (4.9)$$

where $\lceil x \rceil$ is the ceiling function that rounds up to the nearest integer. The final credible interval is $[Y_{ij}^{low}, Y_{ij}^{high}] = [CI_{low}, CI_{high}]$. Suppose we $S=10,000$ and $\alpha=0.05$, then we can take the 2.5th percentiles and 97.5th percentiles of the sorted predictions. The final credible interval is $[Y_{ij}^{[250]}, Y_{ij}^{[9750]}]$

These measures summarize the posterior predictive distribution of the response variable and provide insights into its central tendency, spread, and uncertainty

4.3. Simulation Study

4.3.1. Simulation Design and Process

We construct a simulation study to evaluate the effectiveness of the Bayesian approach that was applied to develop the Four Parameter Beta GLMM. In general, there are ten main steps in the simulation, which can be seen in Figure 4. 1.

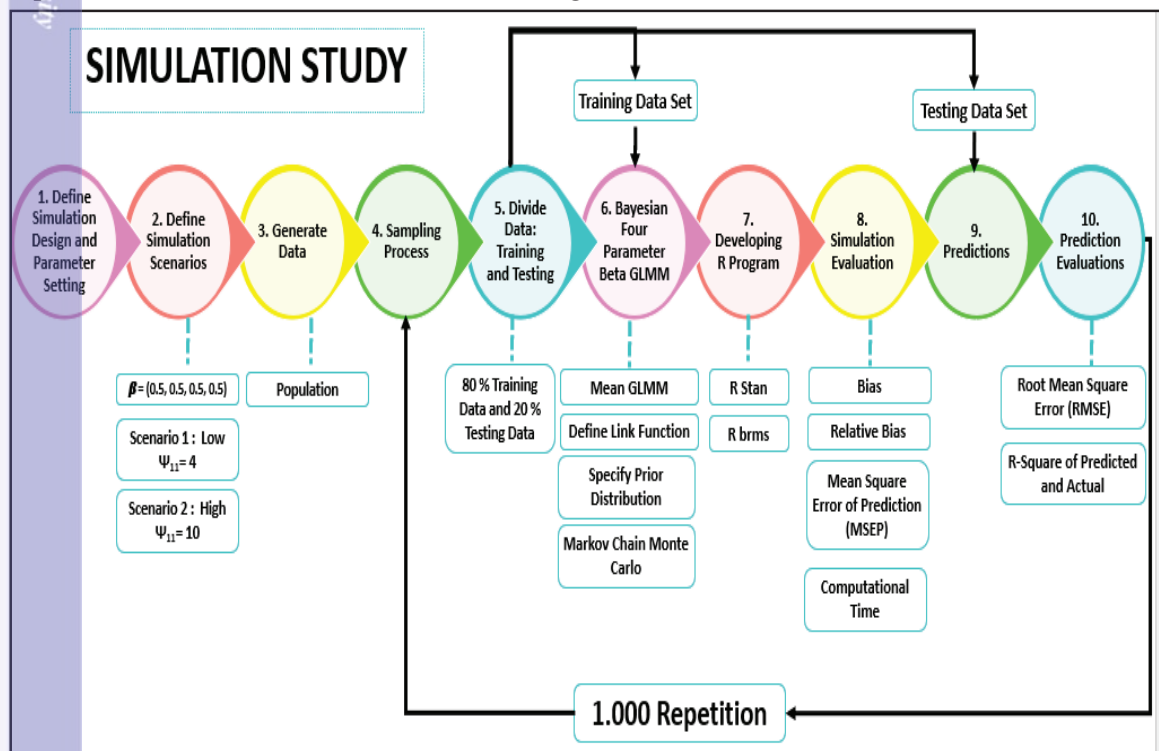


Figure 4. 1 Simulation Workflow Diagram

The pseudo population was designed to mimic the characteristics of a real population of interest. Therefore, two scenarios have been developed in the simulation. A more detailed explanation of the stages of the simulation is given below.

1. Define Simulation Design and Parameter Setting

In this step, we must define the parameters and simulation design below:

- Determine the number of areas ($q=10$) and the total population in each area ($N_i=20,000$)
- Determine K independent variables (X_1, \dots, X_K) with its distribution and the random effect with its distribution. In our study $K=3$, thus

- $X_1 \sim N(0, 5)$ is the first independent variable that has an effect on the response variable
- $X_2 \sim \text{Bin}(n, 0.5)$ is the second independent variable that has an effect on the response variable
- $X_3 \sim N(0, 5)$ is the third independent variable that doesn't have an effect on the response variable
- $\mathbf{b}_i \sim N(\mathbf{0}, \Sigma)$ is the random effect of area, where $\Sigma = \psi_{11}$ for random intercepts in the GLMM model

iii. Note that the assumptions underlying the simulation design were made based on previous relevant studies and empirical data provided by Statistics Indonesia (SI) that will also be used in the case study.

2. Define Simulation Scenarios

- Determine the fixed effect β , the random effect \mathbf{b}_i , and ψ_{11} based on the scenarios given below

Table 4. 1. Data Generating Scenarios in the Simulation Process

Scenario	Random Effect	Ψ_{11}	Fixed Effect
1	Low	4	$\beta = (0.5, 0.5, 0.5, 0.5)$
2	High	10	

- The considerations in developing the scenarios include:
 - In this study we limit the scope to only develop a Four Parameter Beta mean GLMM that has random intercepts.
 - The fixed effect parameter (β) was set to 0.5.
 - Determine the precision parameter ($\phi = 10$), minimum ($a = 0.3$), and maximum ($b = 8.4$) initial values based on obtained data of paddy productivity in Central Kalimantan.
 - While designing the random effects, researchers considered two specifications (low and high) for the covariance matrix. As we have known that the variability of productivity among areas in Indonesia is a common problem. Thus, considering different magnitude levels of inter-group variability will be a beneficiary.

3. Generate Data

- Generate the independent variables (X_1, X_2, X_3) and the random effect based on the simulation design explained above
- Generate the response variables $Y_{ij} \sim B(\omega_{1ij}, \omega_{2ij}, a, b)$, where $i=1,2,\dots,q$ and $j = 1,2, \dots, N_i$ based on the initialization and scenario above. Thus, we must determine $m_{ij}, \omega_{1ij}, \omega_{2ij}, a$, and b based on empirical data provided by Statistics Indonesia (SI). Then, following Equation 4.2 we will have:
 - $\mu_{w_{ij}} \equiv m_{ij} = g^{-1}(\mathbf{X}_{ij}^T \beta + \mathbf{Z}_{ij}^T \mathbf{b}_i)$, where g^{-1} is the inverse logit link function
 - With the known $\mu_{w_{ij}}$, we obtain ω_{1ij} and ω_{2ij} by using the formula $\omega_{1ij} = \phi m_{ij}$ and $\omega_{2ij} = \phi(1 - m_{ij})$

- c) In order to be able to generate the random variable Y_{ij} , we also need to generate $a \sim \text{unif}(a_1, b_1)$ and $b \sim \text{unif}(a_2, b_2)$, where $a_1 < b_1 < a_2 < b_2$
- iii. Obtain the response variable $Y_{ij} \sim B(\omega_{1ij}, \omega_{2ij}, a, b)$ by using the ExtDist package in R with the basic syntax shown below:

```
library(ExtDist)

set.seed(123)

shape1 =  $\omega_1$ 
shape2 =  $\omega_2$ 
a = a
b = b
n_all = 20.000

for(i in 1:n_all)
{
  y[i] = rbeta.ab(1, shape1, shape2, a, b,)
}
```

Note:

- If it is assumed that a logit link function is used, then the GLMM model equation used is

$$\mu_{w_{ij}} = g^{-1}(X_{ij}^T \beta + Z_{ij}^T b_i), \text{ or } \mu_{w_{ij}} = \frac{1}{1 + e^{-(X_{ij}^T \beta + Z_{ij}^T b_i)}}$$

- The definition of a Four Parameter Beta GLMM model requires a transformation of $\mu_{Y_{ij}}$ to $\mu_{w_{ij}}$ by applying $\mu_{w_{ij}} = \frac{\mu_{Y_{ij}} - a}{b - a}$, thus $0 < \mu_{w_{ij}} < 1$
- Therefore, we have mapped $\mu_{w_{ij}}$, that has an interval (0, 1), onto the real line by using a logit link function, which is suitable.
- It is also known from sub chapter 4.2.1 that

$$\mu_{w_{ij}} = \frac{\omega_{1ij}}{\omega_{1ij} + \omega_{2ij}}, \text{ where } 0 < \mu_{w_i} < 1 \text{ and } \phi > 0$$

$$\mu_{w_i} = \frac{\phi m_{ij}}{\phi m_{ij} + \phi(1 - m_{ij})} = \frac{\phi m_{ij}}{\phi m_{ij} + \phi - \phi m_{ij}} = \frac{\phi m_{ij}}{\phi} = m_{ij}$$

This will lead to the proof that $\mu_{w_i} \equiv m_i$

- By knowing that $\mu_{w_{ih}} \equiv m_{ih}$, we can calculate ω_{1ij} and ω_{2ij} .
It is known that in formulating the Four Parameter Beta model we have reparametrized $\omega_{1ij} = \phi m_{ij}$ and $\omega_{2ij} = \phi(1 - m_{ij})$
- Thus, we will have $Y_{ij} \sim B(\phi m_{ij}, \phi(1 - m_{ij}), a, b)$ or simply $Y_{ij} \sim B(\omega_{1ij}, \omega_{2ij}, a, b)$,

4. Sampling Process

- i. Determine the sample size (n_i) for each area $i=1, 2, \dots, q$.

- ii. The sample size for each area will be calculated based on a stratified proportional sampling and the sampling because we will take a sample for each area and the sample sizes of each area will depend on the size of the population itself.
- iii. Select data randomly for each area based on the calculated sample sizes.
- iv. Repeat the Sampling process ($B = 1000$ times).
- v. Repeat Steps 1-10 for each scenario.

Divide Data: Training and Testing

- i. Determine the proportion of training data (80%) and testing data (20%).
- ii. Divide the data into training and testing data according to predetermined proportions.
- iii. Training Data will be used to build the novel Four Parameter Beta mean GLMM model. In this step we will obtain the parameter estimates of a Four Parameter Beta mean GLMM $\theta = (\hat{\beta}, \hat{b}_i, \hat{a}, \hat{b}, \hat{\varphi})$
- iv. Testing Data will be used for predictions and evaluation of prediction results from the novel Four Parameter Beta mean GLMM. In this step we will obtain predictions of the developed model, \hat{y}_{ij}

6. Bayesian Four Parameter Beta GLMM

- i. Specify Model based on Equation 4.1.
- ii. Define the prior distributions as explained in sub section 4.2
- iii. Estimate posterior distribution based on a Bayesian Approach, where the likelihood function is shown in Equation 4.3.
- iv. Evaluate parameter estimates of the

7. Developing R Program

- i. For the computational process of developing the advanced and novel Four Parameter Beta mean GLMM and the simulation process we will use R software and have utilized
- ii. First, we will utilize R-Stan, which is a well-known and robust Bayesian analysis platform available for a variety of analytical programming languages (Stan Development Team, 2022). Stan has the benefit of strong development and support community. Stan is also the foundation for many Bayesian-oriented R packages, such as "RStan" (Stan Development Team, 2021). These packages provide comparable functionality. Although it offers the utmost flexibility, nevertheless it is not user-friendly for novice users having minimum programming skills.
- iii. Next, we will use R "brms" package (Bürkner, 2017), which enables models to be written in a standard, user-friendly formula notation. Nevertheless, it is less flexible than the RStan. As for the Four Parameter Beta regression model we can use the betaBayes package in R that has been developed by Zhou and Huang (2022).

8. Simulation Evaluation

- i. Simulation evaluation of the developed model is based on the training data set.
- ii. In the evaluation process we focus on examining two main aspects in the simulation study:
 - Accuracy and efficiency of the estimated parameters. are shown by the bias and relative bias that can be formulated as:

$$\text{Bias}(\hat{\theta}) = \theta - E(\hat{\theta}) \text{ with } E(\hat{\theta}) = \frac{\sum_{b=1}^B \hat{\theta}}{B} \quad (4.10)$$

$$\text{Relative Bias}(\hat{\theta}) = \frac{\text{Bias}(\hat{\theta})}{\theta} \quad (4.11)$$

- Computational time of R-Stan and R "brms". The computational time will be based on the average time needed to complete a parameter estimation process of 1000 time repetition.
- iii. The Bias and relative bias will be calculated for all parameters in the Four Parameter Beta mean GLMM, which are $\beta, b_i, a, b, \varphi,$.
- iv. We will do the evaluation for each scenario and draw a conclusion.

9. Predictions(\hat{y}_{ij})

- i. Predictions are conducted on the testing data set.
- ii. Prediction will be made based on the developed model.
- iii. Predictions will be made for all scenarios.

10. Prediction Evaluations

- i. The accuracy of the predictions is shown by the Mean Square Error of Prediction (MSEP) of parameter estimates as follows.

$$MSEP(\hat{y}_{ij}) = \frac{1}{B} \sum_{b=1}^B (\hat{y}_{ij} - E(\hat{y}_{ij}))^2 \quad (4.12)$$

- ii. The MSEP will be calculated for all scenarios.
- iii. We will do an evaluation of the predictions accuracy for each scenario and draw a conclusion.

4.3.2. Simulation Results and Discussion

As mentioned before, we constructed a simulation study to analyze the parameter estimate's performance of our proposed Beta 4 Parameter Mean GLMM with Bayesian approach with four scenarios. We examine several things in the simulation study: (1) Bias, (2). Relative Bias, and (3). MSEP. Computationally wise, we considered two approaches in developing the model. First, we have used the Stan package in R that allows more flexibility thus we can estimate the minimum and maximum value (see Four Parameter Beta GLMM (1) in Table 4.2). Second, we have used the brms package in R, that is more user friendly but does not have much flexibility. Thus, we cannot estimate the minimum and maximum value (see Four Parameter Beta GLMM (2) in Table 4.2). Utilizing an Intel core i7 desktop computer and R Server to perform the simulation, CPU time of the brms package was much faster, it took around 1.4 – 2 minutes to complete a parameter estimation process. While for the stan program, CPU time is 3 to 4 times longer. Moving forward, we can enhance the processing time by defining the priors for the model in advance. In this study we have used the results of a Beta GLM parameter estimates for the priors in our model.

The results of the estimation of fixed effect and random effect parameters of the Four Parameter Beta mean GLMM with stan and brms package can be seen in Table 4.2. Based on the table, for scenario 1 $\hat{\beta} = (0.062, 0.501, 0.502, 0.501)$. It is known that the initial value of the fixed effect parameter is $\beta = (0.5, 0.5, 0.5, 0.5)$. It appears that $\hat{\beta}$ for scenario 1 is almost unbiased, except for the intercept. This is the same case for scenario 2, where $\hat{\beta} = (0.054, 0.501, 0.503, 0.501)$ and $\beta = (0.5, 0.5, 0.5, 0.5)$. Meanwhile, the mean estimate of the random value are like the biasness of the intercept. Next, the estimates of minimum (\hat{a}) and maximum (\hat{b}) value of the Four Parameter Beta distribution for both scenarios also have minimum bias. Nonetheless, the bias, relative bias, and MSEP for the estimates of precision parameter ($\hat{\varphi}$) for Scenario 1 and Scenario 2 are the highest compared to other estimates.

Table 4. 2 Simulation Results for the Four Parameter Beta Mean GLMM with Stan (1) and the Four parameter beta Mean GLMM with BRMS (2)

a. Four Parameter Beta Mean GLMM with Stan (1)

Scenario	Parameter Estimates	Mean Estimate	Bias	Relative Bias	MSEP
Scenario 1	b ₀	0.062	0.438	0.876	0.191
	b ₁	0.501	-0.001	-0.002	0.000
	b ₂	0.502	-0.002	-0.004	0.000
	b ₃	0.501	-0.001	-0.002	0.000
	Random Effect	0.438	-0.438	-	0.198
	Phi	13.572	-3.572	-0.357	15.551
	a	0.001	-0.001	-	0.000
	b	1.962	0.038	0.019	0.001
Scenario 2	b ₀	0.054	0.446	0.892	0.198
	b ₁	0.501	-0.001	-0.002	0.000
	b ₂	0.503	-0.003	-0.006	0.000
	b ₃	0.501	-0.001	-0.003	0.000
	Random Effect	0.428	-0.428	-	0.188
	Phi	16.993	-6.993	-0.699	54.386
	a	0.000	0.000	-	0.000
	b	1.980	0.020	0.010	0.000

b. Four Parameter Beta Mean GLMM with BRMS (2)

Scenario	Parameter Estimates	Mean Estimate	Bias	Relative Bias	MSEP
Scenario 1	b ₀	0.960	-0.460	-0.920	0.212
	b ₁	0.512	-0.012	-0.024	0.000
	b ₂	0.509	-0.009	-0.018	0.000
	b ₃	0.500	0.000	0.000	0.250
	Random Effect	0.514	-0.514	-	1.944
	Phi	17.042	-7.042	-0.704	
	a				
	b				
Scenario 2	b ₀	0.883	-0.383	-0.766	0.146
	b ₁	0.506	-0.006	-0.012	0.000
	b ₂	0.504	-0.004	-0.008	0.000
	b ₃	0.500	0.000	0.000	0.250
	Random Effect	0.500	-0.500	-	
	Phi	22.247	-12.247	-1.225	
	a				
	b				

Eventhough, Four Parameter Beta GLMM (2) has faster computational processing, nevertheless Four Parameter Beta GLMM (1) in general outperforms Four Parameter Beta GLMM (2) in terms of bias, relative bias, and MSE for almost all parameters of the Four Parameter Beta mean GLMM in each scenario. The largest difference is apparent in the bias and relative bias of the precision parameter. As an example, we can see that the difference is almost double when the covariance matrix is high in Scenario 2. Not only does Four Parameter Beta GLMM (1) tend to have better parameter estimates, as mentioned earlier on, the Four Parameter Beta GLMM (1) also has an upper hand in its flexibility to estimate the minimum and maximum value. Thus, we can conclude from the simulation that Four Parameter Beta GLMM (1) based on Stan package in R is more flexible and accurate, while using Four Parameter Beta GLMM (2) based on brms package in R is faster.

4.4. Empirical Case Study

4.4.1. Research Process

In this study we have specifically applied the Four Parameter Beta GLMM to predict the paddy productivity of an area. The process consists of five main steps, starting from theory development up to defining the AYI pure premium and VaR (Figure 4. 2). We will assess the performance of the paddy productivity predictions in Central Kalimantan (2020).

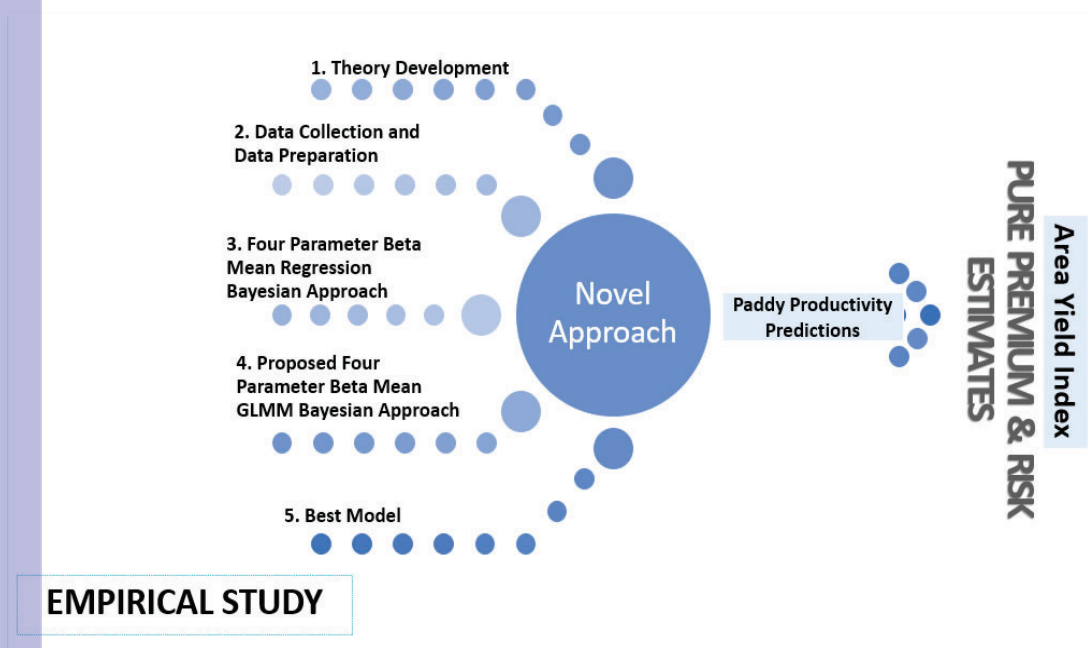


Figure 4. 2 Empirical Study Workflow Diagram

The theory development process of a Four Parameter Beta mean GLMM has been explained in detail in sub section 4.2, while the simulation process have been explained in sub section 4.3. Next, for the data collection and data preparation Statistics Indonesia (SI) has collected 534 CCEs in Central Kalimantan at that period, which indicates the actual paddy productivity at each rice field. The CCEs are also followed by farmer interviews to gather information related to climate, agriculture conditions, and farming business. In total there were 37 items asked in the farmer survey. The variable selection process was based on preliminary research done by Kusumaningrum *et al.* (2024). The seven selected independent variables

farmer survey data can also be seen in Table 3. 1. Nevertheless, as suggested in the previous research (see sub section 3.5), we have added information about farming practices. We have added three independent variables that indicete the use of fertilizers (x_{12} - x_{14}) and one independent variable that shows the amout of seeds used in the area (x_{11}). Additionally, there will be the current and a three-month period lag in the Sentinel 2A satellite imagery data for each plot coordinate surveyed. Researchers have determined that band 4 (red band), band 8 (infra-red band), and a computed Normalized Difference Vegetation Index (NDVI) will also be incorporated as the model's independent variables variables. These variables were also proven to have a significant effect on paddy productivity in Central Kalimantan (Kusumaningrum, 2024). Hence, in total there are fourteen independent variables used in the model (Table 4. 3)

Table 4. 3 Variables Used in the Proposed Four Parameter Beta Mean GLMM

Variable	Variable Name	Unit Measurement/ Category	Data Source
Paddy Productivity	y	Tons/Ha	Crop Cutting Survey (SI)
Pest Attacks This Year	x ₁	1 = Heavy 2 = Medium 3 = Light 4 = Not Affected	Farmer Survey (SI)
Pest Attacks Last Year	x ₂	1 = Heavy 2 = Medium 3 = Light 4 = Not Affected	Farmer Survey (SI)
Impact of Climate Change This Year	x ₃	1 = Affected 2 = Not Affected	Farmer Survey (SI)
Impact of Climate Change Last Year	x ₄	1 = Affected 2 = Not Affected	Farmer Survey (SI)
Water Sufficiency This Year	x ₅	1 = Not Enough 2 = Sufficient 3 = More than Enough	Farmer Survey (SI)
Water Sufficiency Last Year	x ₆	1 = Not Enough 2 = Sufficient 3 = More than Enough	Farmer Survey (SI)
How to Handle Pest	x ₇	0 = No Actions 1 = Agronomist 2 = Mechanical 3 = Biological 4 = Chemical	Farmer Survey (SI)
Band 4	x ₈	Mm	Sentinel-2A Imagery
Band 8	x ₉	Mm	Sentinel-2A Imagery
NDVI	x ₁₀	Index	Sentinel-2A Imagery
Number of seeds used/area of similar plants	x ₁₁	Plants/Ha	Farmer Survey (SI)
KCL	x ₁₂	Kg/Ha	Farmer Survey (SI)
Solid Organic Fertilizer /Compost	x ₁₃	Kg/Ha	Farmer Survey (SI)
Liquid Organic Fertilizer	x ₁₄	Kg/Ha	Farmer Survey (SI)

Such as in the simulation process, we will split the data into training data (80%) and testing data (20%). The selected variables shown in Table 4. for the training data set will be used in the Four Parameter Beta Mean Regression shown in Equation 2.9. We will also apply this data set to the proposed Four Parameter Beta Mean GLMM with the model shown in Equation 4.1. Both models will apply the Bayesian approach in the process of estimating the parameters. To compare the goodness of fit of the two different models, we will use Watanabe-Akaike information criterion (WAIC) that can be calculated as

$$WAIC = -2 \sum_{i=1}^n \log \left\{ \frac{1}{L} \sum_{l=1}^L \log L_i (D_i | \Omega^{(l)}) \right\} + 2p_w, \text{ were}$$

$$p_w = \sum_{i=1}^n \left[\frac{1}{L-1} \sum_{l=1}^L \left\{ \log L_i (\Omega^{(l)}) - \frac{1}{L} \sum_{k=1}^L \log L_i (D_i | \Omega^{(k)}) \right\}^2 \right] \quad (4.13)$$

WAIC has gained popularity in recent years due to its stability compared to DIC (Watanabe, 2010). A smaller value of WAIC indicates a better fit of the model.

After we find the best fit model, it hence becomes efficient to obtain specified predictions (\hat{y}_i) based on the testing data set. The predictions are the by-products of our MCMC sampling approach. The predictions, will than be evaluated based on RMSE value, which is the standard deviation of the prediction errors formulated in Equation 3.5.

To assess the results of parameter estimates of the selected model, we can evaluate the convergence of the posterior distribution to ensure that the parameter estimates of the Four Parameter Beta mean GLMM are accurate. Diagnostic checks that help assess for convergence include trace plots, the R ("Rhat") statistic, and Effective Sample Size (ESS) that was suggested in Reis et al (2023). A Trace plot will show whether the MCMC chain converges or not. Meanwhile consistency of model parameter estimates is proven if the Rhat value is lower than 1. Meanwhile the ESS value should be at least 100 (approximately) to show reliability of the parameter estimates (Vehtari *et al.* 2014).

4.4.2. Paddy Productivity in Central Kalimantan

In Central Kalimantan (2020), paddy productivity is bounded and varies slightly across planting seasons, ranging from 0.691 to 7.676 tons per hectare. A four-parameter beta distribution was found to best fit the data, as indicated by the lowest AIC and AICc values compared to Normal and Laplace distributions (Table 3.1). The estimated variance is 1.070 tons per hectare, with an average productivity of 3.03 tons per hectare. The distribution is skewed to the right, as shown by the shape parameters ($\omega_1=3.067$, $\omega_2=6.091$), indicating that most farmers' actual productivity tends to be lower than the average. Insurers should consider this fact when assessing risks for Area Yield Index (AYI) crop insurance, as these events may have significant financial implications for the insurance product

Given that AYI premiums and risks are season-specific, the first planting season (January–March 2020) in Central Kalimantan was selected for the case study where the productivity ranges from 0.320 up to 6.340. In general, the first planting season distribution characteristics are consistent with all year-round distribution. The Four Parameter Beta distribution also has the best fitted distribution for this planting season. The shape parameters are $\omega_1=3.915$ and $\omega_2=7.413$, which also shows that for this season $\omega_1 < \omega_2$, indicating positive skewness (Figure 4.3 (a)). Variations in productivity across sub-districts were also apparent and shown in Figure 4.3 (b). Thus, highlighting the importance of incorporating district and sub-district random effects in the models to improve prediction accuracy. Nevertheless, the estimated average (2.4

tons per hectare) and variance (0.664 tons per hectare) of paddy productivity are slightly lower compared to all year-round conditions.

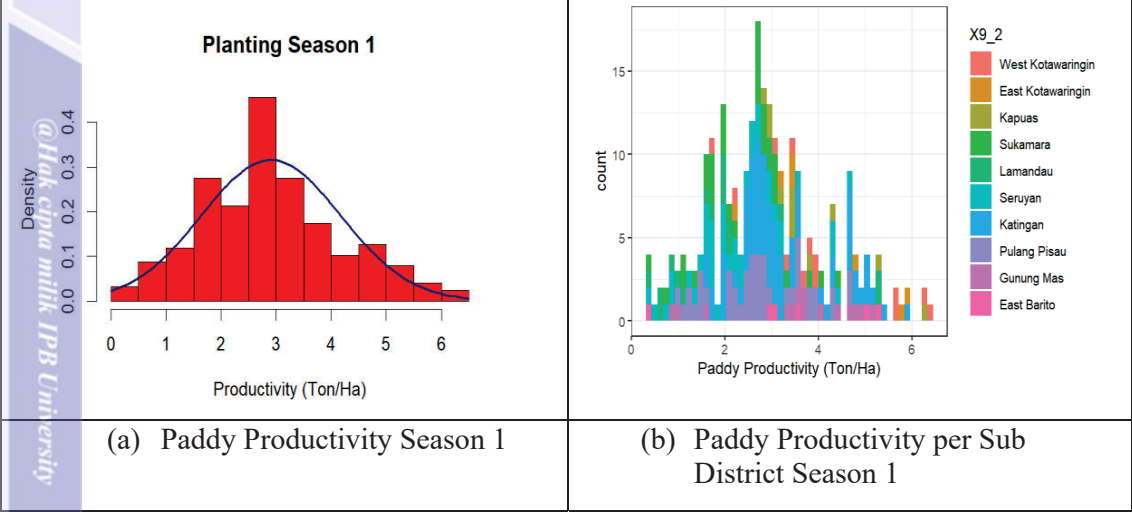


Figure 4. 3 Paddy Productivity Season 1 and Per Sub District Paddy Productivity in Central Kalimantan

4.4.3. Applying the Four Parameter Beta GLMM and Evaluations

We will apply the novel Four Parameter Beta GLMM not only to predict paddy productivity but also indicate significant factors that affect paddy productivity. We have known that predicting paddy productivity is very crucial in the calculations of the premium. Meanwhile, identifying significant variables affecting paddy productivity are important to enhance precise risk assessment associated with paddy productivity fluctuations in a specific area and develop a comprehensive risk management strategy to mitigating various risks.

In general, this process starts with data preparation and exploration to ensure data quality and relevancy. Next, benchmark models are applied to establish a baseline for comparison within the study. The researchers then apply the novel Four Parameter Beta GLMM and conduct evaluations comparing the fit of the proposed and benchmark models. In specifying the model, we begin by using (1) farmer survey data, (2) satellite data along with the 3 months lag of the satellite data, and (3) farmer survey data combined with all the satellite data with its 3 months lagged data. Thus, there will be three main conditions to choose from, allowing flexibility for policymakers and insures to select which condition is most suitable with the data provided. A full model of all the three conditions was done and variable selection was conducted subsequently. The nominated best models can be seen in Table 4.4 where assessments were made based on WAIC. Afterwards, we can identify key factors affecting paddy productivity and analyze the results of the best-chosen model. Lastly, researchers will predict paddy productivity. Further on, we will demonstrate how these predictions can be used to estimate the pure premium and VaR of the AYI crop insurance policy in Central Kalimantan.

To apply the novel Four Parameter Beta mean GLMM based on the Bayesian inference, we must first define prior distributions as follows (1). $a \sim U(-1e - 15, 0.1790452)$ dan $b \sim U(1e - 15, 1)$, (2) $\phi \sim \Gamma(0.01, 0.01)$, (3) fixed effect (β) was given by the parameter estimates of a Four Parameter Beta GLM, and (4) random effect $b_i \sim N(0, \Sigma)$. In this study, district or sub-district areas are incorporated as random effects to account for unobserved heterogeneity in paddy productivity outcomes across areas. The initial values given were based on the initial data explorations and benchmark modeling. Next, we will estimate the posterior through a Markov Chain Monte Carlo (MCMC) simulation with a Burn-in of 1000 replications, and the iterations

of simulations MCMC is set to 10.000. Aside from setting a prior distribution for the random effects (\mathbf{b}_i), the settings were the similar for conducting the four parameter beta mean regression models (Zhou and Huang, 2022).

Table 4. 4 Model Evaluation for the Four Parameter Beta Mean Regression and Four Parameter Beta Mean GLMM

Model	Name of Model	Variables	Fixed Effect	Random Effect	WAIC
Four parameter beta Mean Regression (Zou, 2023)	Model 1	All Survey Data	X_1-X_{14}	-	-126.020
	Model 2	Selected Survey Data	X_2, X_7, X_{12}, X_{13}	-	-115.000
	Model 3	Satellite Data Lag 3 Months	X_8, X_9, X_{10} and Lags of X_8, X_9, X_{10} up to the past 3 months	-	-91.620
	Model 4	Selected Survey Data and Satellite Data 3 Months	$X_2, X_7, X_{12}, X_{13}, X_8, X_9, X_{10}$ and Lags of X_8, X_9, X_{10} up to the past 3 months	-	-127.000
	Model 5	Selected Survey Data and Satellite Data Current Month	$X_2, X_7, X_{12}, X_{13}, X_8, X_9, X_{10}$	-	-124.000
Proposed Four parameter beta Mean GLMM	Model 6	Selected Survey Data	X_2, X_7, X_{12}, X_{13}	District Area	-181.666
	Model 7	Satellite Data Lag 3 Months	X_8, X_9, X_{10} and Lags of X_8, X_9, X_{10} up to the past 3 months	District Area	-203.887
	Model 8	Selected Survey Data and Satellite Data 3 Months	$X_2, X_7, X_{23}, X_{24}, X_8, X_9, X_{10}$ and Lags of X_8, X_9, X_{10} up to the past 3 month	District Area	-132.385
	Model 9	Selected Survey Data and Satellite Data Current Month	$X_2, X_7, X_{12}, X_{13}, X_8, X_9, X_{10}$	District Area	-188.180
	Model 10	Selected Survey Data	X_2, X_7, X_{12}, X_{13}	Sub District Area	-202.609
	Model 11	Satellite Data Lag 3 Months	X_8, X_9, X_{10} and Lags of X_8, X_9, X_{10} up to the past 3 months	Sub District Area	-187.735
	Model 12	Selected Survey Data and Satellite Data Current Month	$X_2, X_7, X_{12}, X_{13}, X_8, X_9, X_{10}$	Sub District Area	-206.745

Table 4. proofs that the proposed Four Parameter Beta mean GLMM has a much better model fit compared to the Four Parameter Beta mean regression model in this case study. The proposed model exceeds the benchmark in all three conditions indicating that the random effects improve the model accuracy. Hence, accounting for the variability of a district or sub district will also lead to a better understanding of the data conditions. Overall, the best fit model is the model where we both combine the farmer survey data and satellite Sentinel 2A data (Model 12). The random effect in the model was a sub district level, where the fixed effect variables included were Pest Attacks Last Year, How to Handle Pest, Band 4, Band 8, NDVI, KCL, and Solid Organic Fertilizer /Compost. Complete results of parameter estimates, and the 95% credible intervals of this model can be seen in Table 4. 5.

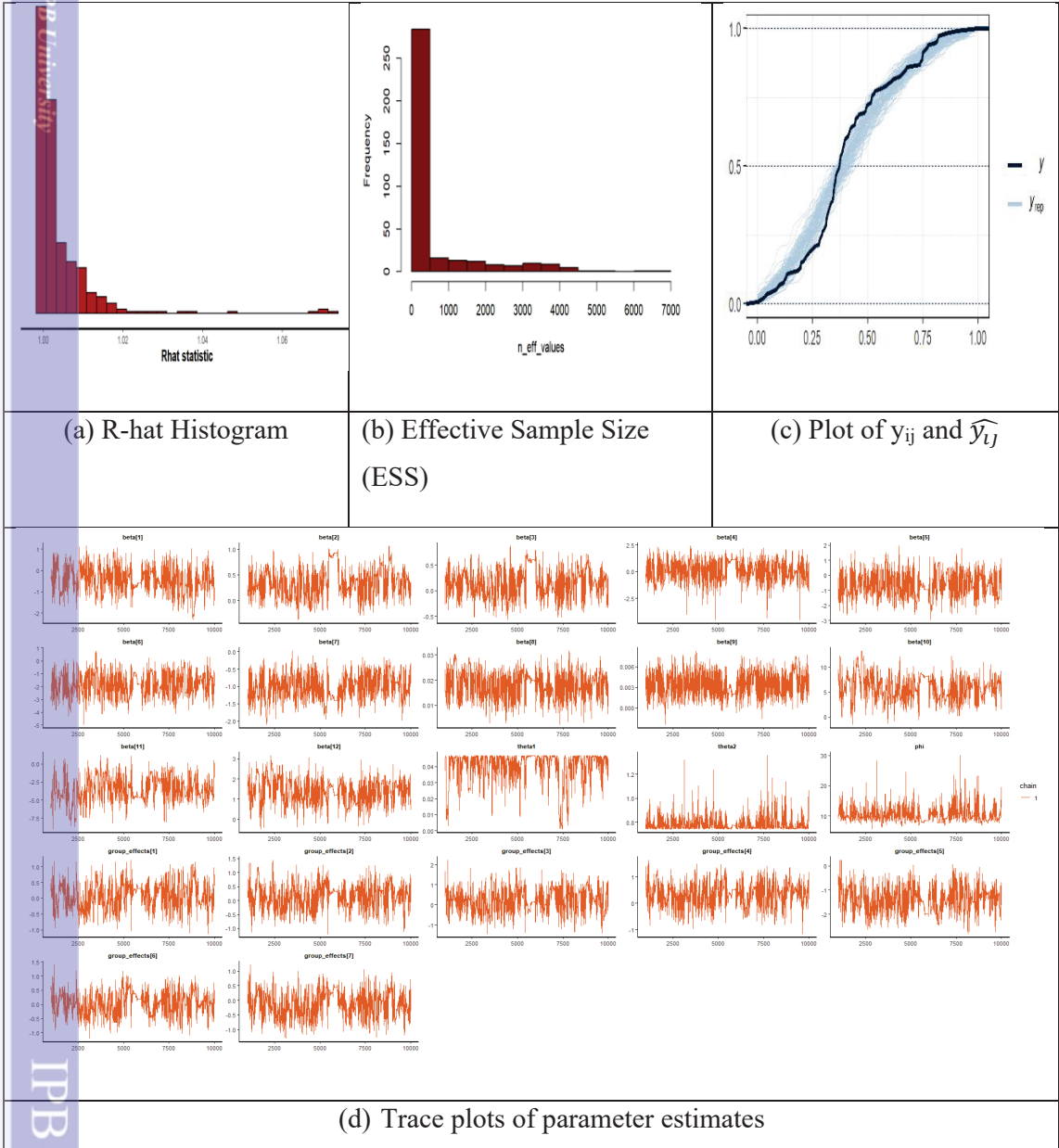


Figure 4. 4 Evaluations on the Bayesian Approach Applied for Developing the Four Parameter Beta Mean GLMM

Evaluating the Bayesian performance based on diagnostic checks of the MCMC iterations is essential. Trace plots in Figure 4. 4 (c) show that all the parameter estimates posterior distribution convergence to a certain value. The average of R-hat values is 1.004 and from the histogram of R-hat shown in Figure 4 (a), where all the values are concentrated at an R-hat value less than 1.01 ensuring that the parameter estimates are consistent. Meanwhile the mean ESS value (Figure 4. 4 (b)) is above 100, specifically the mean of ESS is 1124.0, suggesting that the parameter estimates are reliable. We have also checked the proposed Four Parameter Beta mean GLMM ability to predict paddy productivity (Figure 4. 4 (c)) and results show a good fit indicated by a high correlation of 0.797 between the actual and predicted values. Based on these diagnostic checks of the MCMC iterations and the predictions, we can conclude that the performance of Bayesian methods in estimating the parameters of the proposed Four Parameter Beta mean GLMM and predicting paddy productivity in Central Kalimantan was appropriate. Therefore, we will define the key factors affecting paddy productivity based on this model.

Table 4. 5 The Mean, Standard Deviation (Stdev), and the 95% Credible Interval Estimate for the Best Four Parameter Beta GLMM Model

Variables	Notations of the Parameter Estimate	Mean Estimate	Stdev of Estimate	The 95% Credible Interval	
				Lower Bound	Upper Bound
(Intercept)	beta[1]	-0.424	0.523	-1.521	0.601
X ₂ (Moderate)	beta[2]	0.353	0.249	-0.072	0.919
X ₂ (Mild)	beta[3]	0.145	0.237	-0.303	0.610
X ₂ (Not Affected)	beta[4]	0.061	0.822	-1.618	1.527
X ₇ (Mechanical)	beta[5]	-0.584	0.681	-1.990	0.795
X₇ (Agronomist)	beta[6]	-1.824	0.752	-3.418	-0.415
X₇ (Mechanical)	beta[7]	-1.029	0.316	-1.661	-0.443
X ₂₃	beta[8]	0.018	0.004	0.010	0.027
X ₂₄	beta[9]	0.004	0.001	0.001	0.006
X ₈	beta[10]	5.946	2.212	1.677	10.765
X ₉	beta[11]	-3.563	1.540	-7.131	-0.778
X ₁₀	beta[12]	1.394	0.533	0.372	2.519
<i>a</i>	theta1	0.040	0.010	0.008	0.047
<i>b</i>	theta2	0.786	0.054	0.745	0.955
ϕ	phi	10.414	2.110	7.709	15.626
Katingan Kuala	group_effects[1]	0.387	0.358	-0.606	0.783
Mendawai	group_effects[2]	0.429	0.377	-0.582	0.907
Tewang Sangalang	group_effects[3]				
Garing		0.641	0.501	-0.771	1.181
Pulau Malan	group_effects[4]	-0.155	0.377	-0.448	1.094
Sebangau Kuala	group_effects[5]	-0.800	0.457	-2.337	-0.516
Pandih Batu	group_effects[6]	0.325	0.355	-0.636	0.707
Maliku	group_effects[7]	0.115	0.384	-0.852	0.557

Summary statistics of Bayesian parameter estimates of Model 12 are shown in Table 4. 5. First, we have stronger indications to support our belief that the random effect will significantly deviate from the average effect when the 95% credible interval does not overlap with zero, as indicated by Sebangau Kuala subdistrict. Confidence intervals for this subdistrict are less than zero, indicating low average paddy productivity and significant variations among the districts. Variables such as How to Handle Pest (X_7), KCL (X_{12}), Solid Organic Fertilizer /Compost (X_{13}), Band 4, Band 8, and NDVI have a significant effect on paddy productivity. Pest handling through agronomist and mechanical has a potential to decrease paddy productivity. Meanwhile, using the right dosage of KCL and Solid Organic Fertilizer /Compost are in line with paddy productivity. For the upcoming season, it is estimated that the minimum value of paddy productivity predictions of the next season in Central Kalimantan is 1.037 while the maximum is 2.183. We can also see that the precision parameter is estimated to be in the interval of 7.088 and 14.902, which shows that the variability of productivity is high. Having information on these parameters is a beneficiary because it will contribute to a more comprehensive understanding of complex relationships, better predictive outcomes, and improved decision-making for both academic and policy-oriented contexts in economic and other fields.

In this model it is also shown that satellite data doesn't only have a significant effect but also enhances accuracy. Current and previous satellite data have potential for predicting paddy productivity. Model 7 and Model 11, also have low WAIC and therefore show potential for further development. The challenge will be to define the length of lag. As for the chosen Model 12, results show that the RMSE of predicted values are 0.020, which has the lowest value among all proposed four beta parameter mean GLMM. Thus, we will use the predictions results to calculate the premium and VaR of AYI.

4.5. Model Limitations and Further Development

This research has limitations, primarily due to the rare nature of studies addressing the use of Four Parameter Beta distributions in predictive modeling. Specifically, model wise, we have considered developing a Four Parameter Beta GLMM that has a random intercept, using a logit link function, and based on a mean central tendency. These options provide a good foundation for further research. Next, the model can be expanded based on random slopes, applying and evaluating other link functions, and use mode as the central tendency.

We have also not tackle the problem of weak to moderate linear relationships between paddy productivity and lagged NDVI values in Central Kalimantan that was issued in Chapter 3. Thus, in the next chapter it is crucial to develop flexible models that account for the Four Parameter Beta distribution, area variability, and complex linear and non-linear relationships. Adjusting for these advancements could enhance the model's flexibility and applicability to diverse datasets and scenarios.

Despite these limitations, the study provides a solution and advancements in modelling a GLMM for a response variable following a Four Parameter Beta distribution based on Bayesian approach. Expect for the precision parameter ($\hat{\phi}$), simulations show that the parameters estimates are considered unbiased. Also when the model is applied to predict paddy productivity, the prediction accuracy was also higher, indicated by the lower RMSE values (0.020).

4.6. Conclusion and Recommendations

In this study we focus on the development of Four Parameter Beta mean GLMM using a Bayesian approach. This model was an extension of the Four Parameter Beta regression model developed by Zhou (2021), which incorporates random effects into the model. Using a

simulation study, we showed that the developed Four Parameter Beta mean GLMM are considered relatively unbiased for the estimated parameters $\hat{\theta} = (\hat{\beta}, \hat{b}_1, \hat{a}, \hat{b}, \hat{\phi})$. However, estimates of the precision parameter ($\hat{\phi}$) was still biased. Further research is suggested to increase the precision and accuracy of this parameter. The simulation results also showed that the development of this model using Stan package in R software does not have faster computation time compared to the brms package in R software, but it is more flexible and accurate.

The empirical case study shows that the paddy productivity in Central Kalimantan has a Four Parameter Beta distribution that is skewed to the right. Hence, it makes more sense to develop the paddy productivity prediction model based on this distribution. Especially when the predictions are used as the basis to calculate the premium and Value at Risk (VaR) of the alternative Area Yield Index (AYI) crop insurance policy. Insurers need to carefully assess the risk associated with low paddy productivity as it has a significant financial implication for the insurance product. We have also shown that the proposed Four Parameter Beta mean GLMM predictions are more accurate than Four Parameter Beta regression models. The key factors affecting paddy productivity based on this model include how to manage pest (X_7), the right dosage of KCL (X_{12}) and solid organic fertilizer /compost (X_{13}), Band 4, Band 8, and NDVI. This model also demonstrates that satellite data not only has a significant effect but also improves accuracy. Satellite data from the past and present can be used to predict paddy productivity. As a result, we will calculate the premium and VaR of AYI based on the prediction results in Chapter 6.

V. Four Parameter Beta Generalized Mixed Effect Tree and Random Forest

5.1 Introduction

In this chapter, we will extend the Four Parameter Beta GLMM model developed in Chapter 4 into a Generalized Mixed Effect Tree (GMET) model by integrating GLMM with tree regression methods (Fontana *et al.* 2021). A regression tree is a well-established machine learning algorithm used to predict continuous values by partitioning all the independent variables (or features) into non-overlapping regions. By combining these two approaches, the Four Parameter Beta GMET model will offer greater flexibility, and it is expected to provide more accurate predictions of paddy productivity. Next, building on the advancements of the Four Parameter Beta GMET, we hope to further increase the model's predictive ability by incorporating random forests, leading to the development of a Four Parameter Beta Generalized Mixed Effect Random Forest model. Random forests are a powerful ensemble learning method, that combines multiple decision trees to improve prediction accuracy by averaging over many individual models. Thus, reducing overfitting and enhancing the model's robustness.

The development of the Four Parameter Beta GMET and GMERF models is driven from the limitations of the existing Four Parameter Beta GLMM. Eventhough, the Four Parameter Beta GLMM is a potential framework for modelling response data that follows a Four Parameter Beta distribution, it falls short in addressing non-linear relationships and complex interactions among variables. Consequently, it struggles to fully adapt to the dynamic and heterogeneous nature of agricultural data, as illustrated in Figure 3.9 to Figure 3.11. This limitation is particularly evident when working with datasets that integrate diverse data sources, such as satellite imagery and farmer surveys. To overcome these challenges, the Four Parameter Beta GMET and Four Parameter Beta GMERF models introduce enhanced flexibility and adaptability. Once these advanced models are developed, they will be applied to predict paddy productivity, building on the methodologies established in the previous chapter.

It has been known that predicting paddy productivity accurately is crucial for agricultural planning, as it directly impacts crop insurance policies and risk management strategies. In areas like Central Kalimantan, where productivity is influenced by a variety of factors such as climate conditions, irrigation, soil quality, and farming practices, developing reliable prediction models becomes increasingly important. These models help predict productivity, enabling effective agricultural planning, resource allocation, and risk management. Additionally, accurate predictions play a key role in crop insurance, allowing insurers to assess risks and determine premiums based on expected yields. Without precise predictions, farmers and insurers may face significant financial uncertainty, making it difficult to make informed decisions regarding crop management and insurance coverage.

However, achieving accurate paddy productivity predictions requires addressing multiple challenges. Firstly, the Four Parameter Beta distribution, identified as the best fit for the data, must be considered. Additionally, area variability plays a significant role, which led to the development of a Four Parameter Beta GLMM in Chapter 3. Despite its convenience and prediction accuracy, this transformation process may introduce potential biases in the parameter estimates and predictions. To mitigate this, a Bayesian approach to the four-parameter GLMM, extending the Four Parameter Beta regression model introduced by Zou *et al.* (2022), was developed in Chapter 4. Simulation and empirical results demonstrated advancements compared to the initial approach. Nonetheless, as highlighted in Chapters 3 and 4, it is also crucial to

account for complex relationships that may arise from combining different datasets. For instance, weak to moderate linear relationships were observed between paddy productivity and lagged NDVI values in Chapter 3. Therefore, developing flexible models that can account for the four-parameter beta distribution, area variability, and complex linear and non-linear relationships is essential.

Combining survey and satellite data has been shown to improve prediction accuracy for paddy productivity. While survey data often show linear relationships with productivity, satellite data typically exhibit non-linear patterns (Son *et al.* 2013). These differing relationships introduce complexity in identifying the most suitable prediction model. However, integrating both current and lagged satellite data with survey data has been proven effective in producing accurate predictions. This paper aims to address these complexities by proposing advanced modelling approaches that integrate diverse data types.

Hence, through the integrations of random forests and regression trees with the Four Parameter Beta GLMM models, the Four Parameter Beta GMERF and Four Parameter Beta GMET framework will not only provide a more flexible model but also increases its capability to capture complex interactions and non-linearities present in the data. This is particularly useful when developing prediction models based on various data sources, such as satellite data and survey data, where relationships between variables may vary significantly across different regions or seasons. Therefore, the Four Parameter Beta GMERF and Four Parameter Beta GMET s expected to outperform traditional prediction models, offering more reliable estimates of paddy productivity and providing valuable insights for agricultural planning, resource allocation, and crop insurance strategies.

5.2.The Proposed Method

At first, we have developed a Four Parameter Beta GLMM by using a Bayesian approach in Chapter 4 that extends the Four Parameter Beta regression model (Zou *et al.* 2022). In this model we assume that $y \sim B(\omega_1, \omega_2, a, b)$. Where, a is the minimum value, b is the maximum value, ω_1 is the location parameter and $\omega_1 - \omega_2$ is the scale parameter. Thus, the model can be seen in Equation 2.9.

To accommodate nonlinear relationships GMET has been developed based on a regression tree and GLMM (Fontana *et al.* 2021). The model first assumed a Bernoulli distribution, which can be seen in Equation 2.10. We have further we develop GMET for a four paramter beta distribution that can be formulated as:

$$Y_{ij} \sim B(\omega_{1ij}, \omega_{2ij}, a, b) \text{ and } \eta_{ij} = f(\mathbf{X}_{ij}) + \mathbf{Z}_{ij}^T \mathbf{b}_i^*, \text{ where } \mathbf{b}_i \sim iid N(0, \sigma_i^2) \quad (5.1)$$

The fixed effect $f(\mathbf{X}_{ij})$ is estimated through CART algorithm and $\mathbf{Z}_{ij}^T \mathbf{b}_i^*$ in Equation 5.1 is estimated through the Four Parameter Beta GLMM based on the Bayesian approach introduced in Chapter 4. Details on the algorithms for GMET can be seen in sub section 2.5 (Fontana *et al.* 2021). We have expanded the algorithm by applying a Four Parameter Beta regression model in step two instead of a logistic regression model. We have also used the Four Parameter Beta GLMM based on Bayesian approach in Chapter 4. Thus, the algorithm for Four Parameter Beta GMET can be written as

1. Initialize the random effect estimate (\mathbf{b}_i^*) to zero .
2. Estimate the response variable ($\hat{\mu}_{ij}^*$) using the Four Parameter Beta regression model given that the fixed effect independent variables are $\mathbf{X}_{ij} = (X_{1ij}, \dots, X_{Kij})^T$ for $i=1, \dots, q$ and $j = 1, 2, \dots, n_i$. We will obtain an estimate of $\hat{\mu}_{ij}^*$.

3. Construct a regression tree that will approximate the function $f(\mathbf{X}_{ij})$ in Equation 5.1 by using $\hat{\mu}_{ij}^*$ as the response variable and $\mathbf{X}_{ij} = (X_{1ij}, \dots, X_{Kij})^T$ as the independent variable. We will then obtain L terminal nodes (\mathbf{R}_ℓ), where $\ell = 1, 2, \dots, L$. For each observation j in group/area i and, will belong to one of the terminal nodes, that will be described by a certain set of independent variables \mathbf{X}'_{ij} . Hence, we can define a set of indicator $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ for $\ell = 1, 2, \dots, L$. Then, $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ will be given the value of 1 if observation i , belongs to the ℓ -th terminal node and 0 otherwise. Therefore, the indicator $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ is in the form of a dummy variable.
4. Fit the Four Parameter Beta GLMM model in Equation 2.9, using Y_i as the response variable and the set of indicator variables $I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)$ as the fixed effect independent variables. For $i = 1, \dots, q$ and $j = 1, 2, \dots, n_i$ we will have $g(\mu_{ij}^*) = I(\mathbf{X}_{ij} \in \mathbf{R}_\ell)\gamma_\ell + \mathbf{Z}_{ij}^T \mathbf{b}_i^*$. Here, γ_ℓ^* represents the fixed effects for each terminal node and $\mathbf{Z}_{ij}^T \mathbf{b}_i^*$ captures the random effects. So, we obtain $\hat{\mathbf{b}}_i^*$ from the presumed Four Parameter Beta GLMM model.
5. Replace the predicted response values at each terminal node R_ℓ of the regression tree with the predicted response values $g(\hat{\gamma}_\ell^*)$ from the mixed effects model performed in step 4. Thus, we obtain

$$g(\mu_{ij}^*) = g(\hat{\gamma}_\ell^*) + \mathbf{Z}_{ij}^T \hat{\mathbf{b}}_i^*$$

$$\text{thus, } \mu_{ij}^* = g^{-1}(g(\hat{\gamma}_\ell^*) + \mathbf{Z}_{ij}^T \hat{\mathbf{b}}_i^*) \quad (5.2)$$

As the updated final prediction of the response variable. Then, we can calculate predictive accuracy by using RMSEP shown in Equation 3.5 and interpret the models result.

Next, GMERF for a Four Parameter Beta distribution was also introduced as Four Parameter Beta GMERF. Such as in the Four Parameter Beta GMET, we have also apply the Four Parameter Beta GLMM to estimate $\mathbf{Z}_{ij}^T \mathbf{b}_i^*$ in Equation 5.3 below

$$\mathbf{Y}_{ij} \sim B(\omega_{1ij}, \omega_{2ij}a, b) \text{ and } \boldsymbol{\eta}_{ij} = f^*(\mathbf{X}_{ij}) + \mathbf{Z}_{ij}^T \mathbf{b}_i^*, \text{ where } \mathbf{b}_i \sim iid N(0, \sigma_i^2) \quad (5.3)$$

where $f^*(\mathbf{X}_{ij})$ represents the relationship between the independent variables (\mathbf{X}_{ij}) and the response variable (Y_{ij}), which is not linear but assumed to have a complex and unknown structure that we estimate with a tree form structure, built through the ensemble-based RF method (Breiman, 2001). Meanwhile, $\mathbf{Z}_{ij}^T \mathbf{b}_i^*$ is estimated through the Four Parameter Beta GLMM based on the Bayesian approach introduced in Chapter 4. In general, the fundamental concept of a Random Forest (RF) is to train numerous decision trees, each constructed using a distinct dataset generated through bootstrap sampling from the original data. Additionally, each tree is trained using only a randomly selected subset of the K independent variable. The final prediction of the RF is obtained by appropriately aggregating the predictions made by all the individual trees. Details on the algorithms for the Four Parameter Beta GMET have been explained above. The Four Parameter Beta GLMM based on Bayesian approach also has been explained in Chapter 4. We have expanded the algorithm by Four Parameter Beta GMET by applying a random forest algorithm.

Thus, the algorithm for Four Parameter Beta GMERF can be written as

1. Set the maximum iterations ($iter=M$)
2. Initialize the random effect estimate (\mathbf{b}_i^*) to zero.
3. Start the first iteration ($iter = 1$) by taking a random sample from the original data set

4. Estimate the response variable $(\hat{\mu}_{ij}^*)^{(iter=1)}$ using the Four Parameter Beta regression model given that the fixed effect independent variables are $\mathbf{X}_{ij} = (X_{1ij}, \dots, X_{Kij})^T$ for $i=1, \dots, q$ and $j = 1, 2, \dots, n_i$. We will obtain an estimate of $\hat{\mu}_{ij}^*$.
5. With the random sample, construct a regression tree that will approximate the function $f^*(\mathbf{X}_{ij})$ in Equation 5.3 by using $(\hat{\mu}_{ij}^*)^{(iter=1)}$ as the response variable and $(\mathbf{X}_{ij})^{(iter=1)}$ as the independent variable. We will then obtain L terminal nodes $(R_\ell)^{(iter=1)}$, where $\ell = 1, 2, \dots, L$. For each observation i , will belong to one of the terminal nodes, that will be described by a certain set of independent variables $(\mathbf{X}'_{ij})^{(iter=1)}$. Hence, we can define a set of indicator $(I(\mathbf{X}_{ij} \in R_\ell))^{(iter=1)}$ for $\ell = 1, 2, \dots, L$. Then, $(I(\mathbf{X}_{ij} \in R_\ell))^{(iter=1)}$ will be given the value of 1 if observation i with in group/area j , belongs to the ℓ -th terminal node and 0 otherwise. Therefore, the indicator $(I(\mathbf{X}_{ij} \in R_\ell))^{(iter=1)}$ is in the form of a dummy variable.
6. Fit the Four Parameter Beta GLMM model in Equation 2.9, using $(Y_i)^{(iter=1)}$ as the response variable and the set of indicator variables $(I(\mathbf{X}_{ij} \in R_\ell))^{(iter=1)}$ as the fixed effect independent variables. For $i = 1, \dots, q$ and $j = 1, 2, \dots, n_i$ we will have $(g(\mu_{ij}^*))^{(iter=1)} = (I(\mathbf{X}_{ij} \in R_\ell))^{(iter=1)} \gamma_\ell + (\mathbf{Z}_{ij}^T \mathbf{b}_i^*)^{(iter=1)}$. Here, γ_ℓ^* represents the fixed effects for each terminal node and $\mathbf{Z}_{ij}^T \mathbf{b}_i^*$ captures the random effects for each iteration. So, we obtain $(\hat{\mathbf{b}}_i^*)^{(iter=1)}$ from the presumed Four Parameter Beta GLMM model.
7. Replace the predicted response values at each terminal node $(R_\ell)^{(iter=1)}$ of the regression tree with the predicted response values from the mixed effects model performed in step 4. Thus, we obtain

$$(g(\mu_{ij}^*))^{(iter=1)} = (g(\hat{\mathbf{y}}_\ell^*))^{(iter=1)} + (\mathbf{Z}_i^T \hat{\mathbf{b}}_i^*)^{(iter=1)} \quad (5.2)$$

$$(\hat{\mathbf{y}}_{ij})^{(iter=1)} = \left(g^{-1}(\mu_{ij})\right)^{(iter=1)} = \left(g^{-1}(g(\hat{\mathbf{y}}_\ell) + \mathbf{Z}_i^T \mathbf{b}_i)\right)^{(iter=1)}$$

As the updated prediction of the response variable

8. Repeat step 3 upto step 7 upto $iter = M$
9. Aggregation of regression trees, combining predictions from all trees (e.g., by averaging) to obtain the final prediction

$$(\hat{\mathbf{y}}_{ij}) = \frac{1}{M} \sum_{iter=1}^M (\hat{\mathbf{y}}_{ij})^{(iter)}$$

5.3. Methodology

The proposed models Four Parameter Beta GMERF and Four Parameter Beta GMET will be applied to an empirical case study in Central Kalimantan's first planting season (2020). Paddy productivity was measured by CCE's and there were seven selected explanatory variables from the farmer survey. Current and a four-month period lag of Sentinel 2A satellite data for each plot will also be used. The data set used is the same as used in the previous study (See Table 4.3).

Meanwhile the research stages are as follows:

1. Data Preparation and Exploration
2. Apply the Four Parameter Beta GLM and Four Parameter Beta GLMM explained in chapter 4 as a benchmark of the study
3. Incorporate the Four Parameter Beta GLM and Four Parameter Beta GLMM with the regression tree to obtain a Four Parameter Beta GMET such as explained in subsection 5.3
4. Incorporate the Four Parameter Beta GLM and Four Parameter Beta GLMM with the random forest to obtain a Four Parameter Beta GMERF such as explained in subsection 5.3
5. For steps 2-4, we will develop models that only use farmer survey data independent variables, only satellite data independent variables, and a combination of survey and satellite data independent variables
6. Evaluate all of the models fit based on WAIC values shown in Equation 4.12. RMSE values in Equation 3.5 will be used for evaluating the prediction values.
7. Identifying key factors affecting paddy productivity and analyse the effect based on the coefficient results of the best model.

Futher on in sub chapter 5.4.4 we will futher use the best prediction results to calculate the premium and risks of AYI.

5.4.Results and Discussion

5.4.1. Model Evaluation

First, when we evaluate the models goodness of fit based on WAIC show that a Four Parameter Beta GLMM is more suitable compared to the Four Parameter Beta GLM. Hence, random effects of a sub district improve model fit and leads to a better understanding of variability of productivity among areas. Combining farmer survey and satellite data also enhances the models fit. Therefore we will incorporate the Four Parameter Beta GLMM using farmer survey and satellite data along with regresion tree to obtain the Four Parameter Beta GMET. We will also incorporate the Four Parameter Beta GLMM using farmer survey and satellite data along with random forest to obtain Four Parameter Beta GMERF.

Prediction wise, Table 5. 1 proofs that the proposed Four Parameter Beta GMERF and Four Parameter Beta GMET has a much better prediction accuracy compared to the benchmark Four Parameter Beta GLM and Four Parameter Beta GLMM models. This shows that Four Parameter Beta GMET and Four Parameter Beta GMEF there is an indictaion that these models are more able to handle complex non-linear relationships and occurrences of possible interactions between independent variables more effectively than the Four Parameter Beta GLM or Four Parameter Beta GLMM models. While Four Parameter Beta GLM and GLMM rely on predefined linear or generalized linear relationships, the tree-based structure of GMET and the ensemble approach of GMERF allow these models to capture intricate patterns in the data without requiring explicit specification of functional forms.

The observed improvements in prediction accuracy of Four Parameter Beta GMET and Four Parameter Beta GMERF highlights the practical advantages of using advanced machine learning approaches like GMET and GMERF in scenarios requiring nuanced prediction capabilities. These models offer better adaptability to real-world complexities, making them more suitable for diverse datasets and conditions.

Table 5. 1 Goodness of Fit and Prediction Evaluation for Developed Models

Model		Independent Variables Data Source	WAIC	RMSEP
@H cipa milk IPB University	Four Parameter Beta GLM	Farmer Survey Data	-115.000	0.010
		Satellite Data	-91.615	0.011
		Farmer Survey and Satellite Data	-124.000	0.011
	Four Parameter Beta GLMM	Farmer Survey Data	-202.609	0.029
		Satellite Data	-187.735	0.028
		Farmer Survey and Satellite Data	-206.745	0.020
	Four Parameter Beta GMET	Farmer Survey Data		0.003
		Satellite Data		0.004
		Farmer Survey and Satellite Data		0.026
	Four Parameter Beta GMERF	Farmer Survey Data		0.003
		Satellite Data		0.008
		Farmer Survey and Satellite Data		0.010

The Four Parameter Beta GMET and Four Parameter Beta GMERF discussed above are the results of the pruned model. Thus, it is a simpler model because we have eliminated unnecessary parameters. As an example, we will explain one Four Parameter Beta GMET that has been driven from satellite data in Central Kalimantan.

In general, the best four-parameter beta GMET model for each area (Figure 5.1) identifies the lagged values of Band 4, Band 8, and NDVI as important predictors. In Central Kalimantan, Band 8 values from 11 months prior can predict current paddy productivity. Thus, Band 4 is the most important predictor variable for defining the variability of paddy productivity conditions in Central Kalimantan. Nevertheless, further enhancements are needed by adding additional NDVI and Band 8 current and past conditions. NDVI has the most selected lag values in both areas indicating that not only the current values but also the past values of NDVI have a substantial impact on the accuracy of current predictions. These values can provide unique information about the state of vegetation that influences the current paddy productivity. It also reflects seasonal patterns that will be further analysed in the next section. Hence, we must ensure consistent and accurate collection of NDVI data over time to maintain the predictive power of the model.

5.4.2. Significant Variables for Predicting Paddy Productivity

Significant effects of lagged satellite data suggest that historical satellite data information can be used as a main predictive factor for paddy productivity. These findings were in line with Debake *et al.* (2023) where enhancing the utilization of satellite data will give stakeholders the

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

opportunity to make yield estimates for large areas where they might not have access to individual farmers' survey data. This information can also be used to monitor and maintain crop health during critical growth stages, which can influence higher productivity). Based on our model, the monitoring process can even start 7 months in advance. While gathering such survey practices requires time and money, utilizing satellite data provides a more cost-effective way to work on a wide scale.

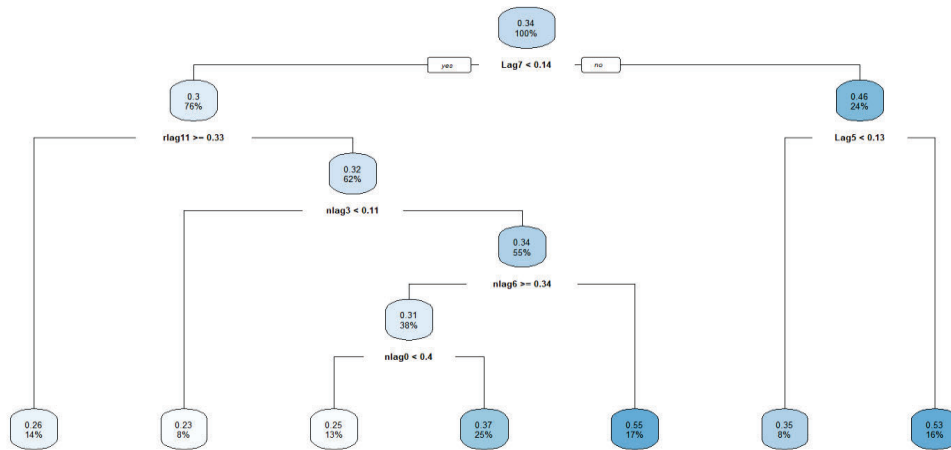


Figure 5.1. Example of a Four Parameter Beta GMET Model in Central Kalimantan

With regards to AYI, positive results were shown when utilizing satellite data for predicting paddy productivity. Variable importance pointed out that current and lagged infrared and near infrared reflectance along with current NDVI relationship with paddy productivity are apparent. Thus, it can serve as a remedy for current concerns among stakeholders regarding the insufficient availability of accurate and real-time data at both the area and sub-area levels.

In more detail, in Figure 5.2 we can analyze the relationship between Band 4, Band 8, and NDVI with productivity at current conditions and with its previous lag values. For data exploration purposes, we grouped productivity level into four groups (high, medium, low, and very low) based on its quartiles. For each group, the plot is colored based on 2D high density estimates. Darker colors of the regions will vary from light to dark based on the density, with darker colors indicating the highest density regions. In the plots we can see that there are some observations that overlap between certain groups, indicating that these observations have the possibility to be considered in two groups. Thus, further work on grouping productivity based on more advanced cluster analysis is suggested. Nevertheless, there are many interesting facts shown in Figure 5.2 First, we can see that for the current condition, where paddy is being harvested, the relationship between paddy productivity (Y) and index values of Band 4, Band 8, and NDVI are negative, except for the low group. This indicates that paddy fields are being harvested and there is less vegetation. Second, except for Sub Round 3, we can see that for each group, there is an increase shift of Band 4, Band 8, and NDVI index values in the next upcoming month/lag indicating growth of paddy. The shift is higher for NDVI, followed by Band 8 and Band 4. Third, there is also indication of a positive trend between Band 4, Band 8, and NDVI index values and productivity as the time gets closer to harvest period. This indicates that higher Band 4, Band 8, and NDVI index values will lead to higher productivity levels. The pattern is shown more clearly in Sub Round 1 and 2 that have high, medium, and very low productivity. Meanwhile if the paddy productivity is low most of the patterns are constant. Last, Sub Round 3 shows less shift and less pattern in each group of paddy productivity level due to the condition at that time is dry season. Thus, not so many farmers plant during that season and there will not

be so much plant growth captured by the Sentinel 2A satellite data. Due to the diverse conditions and trends between Band 4, Band 8, and NDVI with productivity at current conditions and with its previous lag values that can be seen in Figure 5.2, therefore four-parameter beta GMET tends to outperform the four-parameter beta GLMM model.

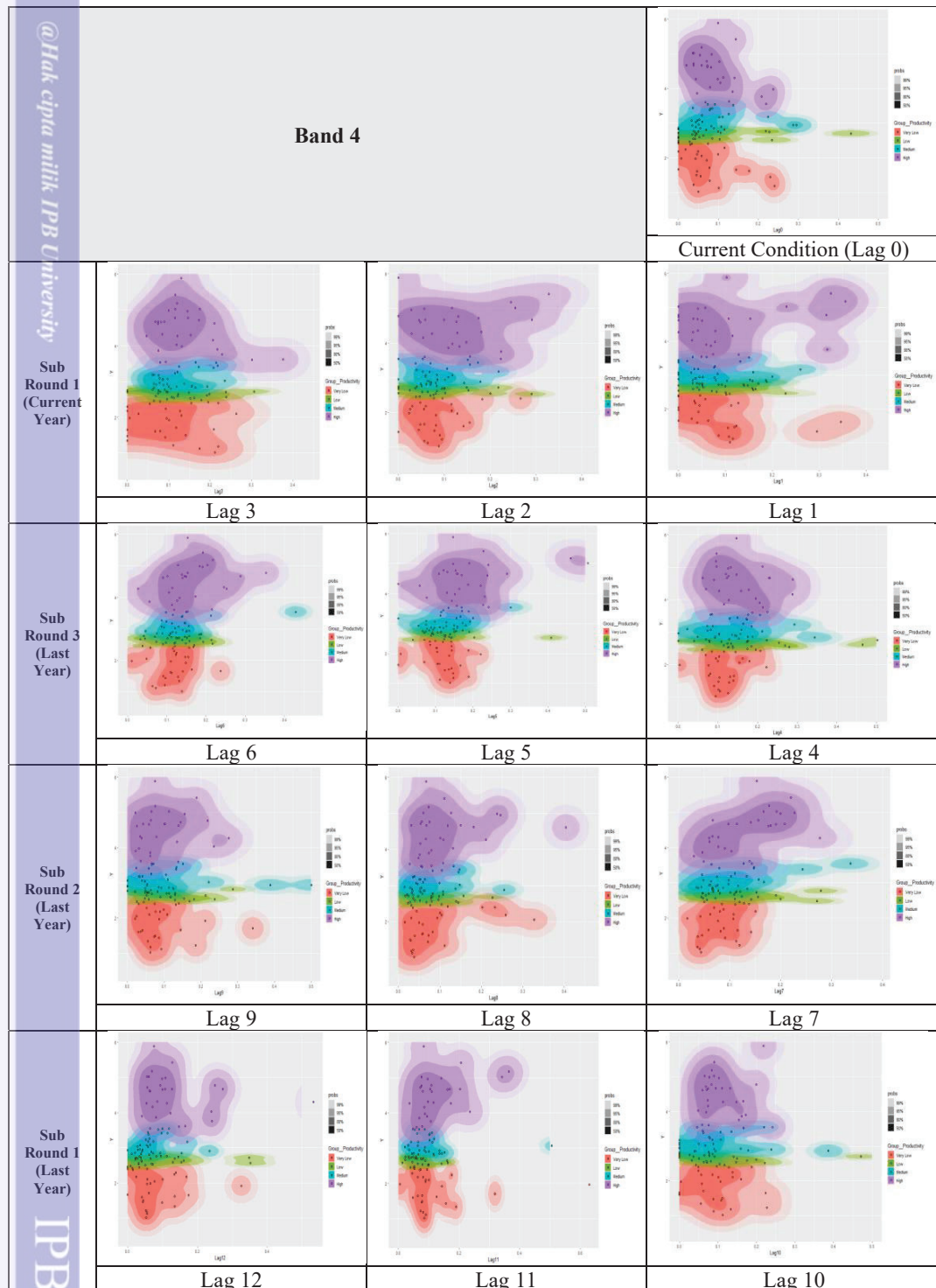


Figure 5.2 . Band 4, Band 8, and NDVI Scatterplots with Productivity Groups (a)

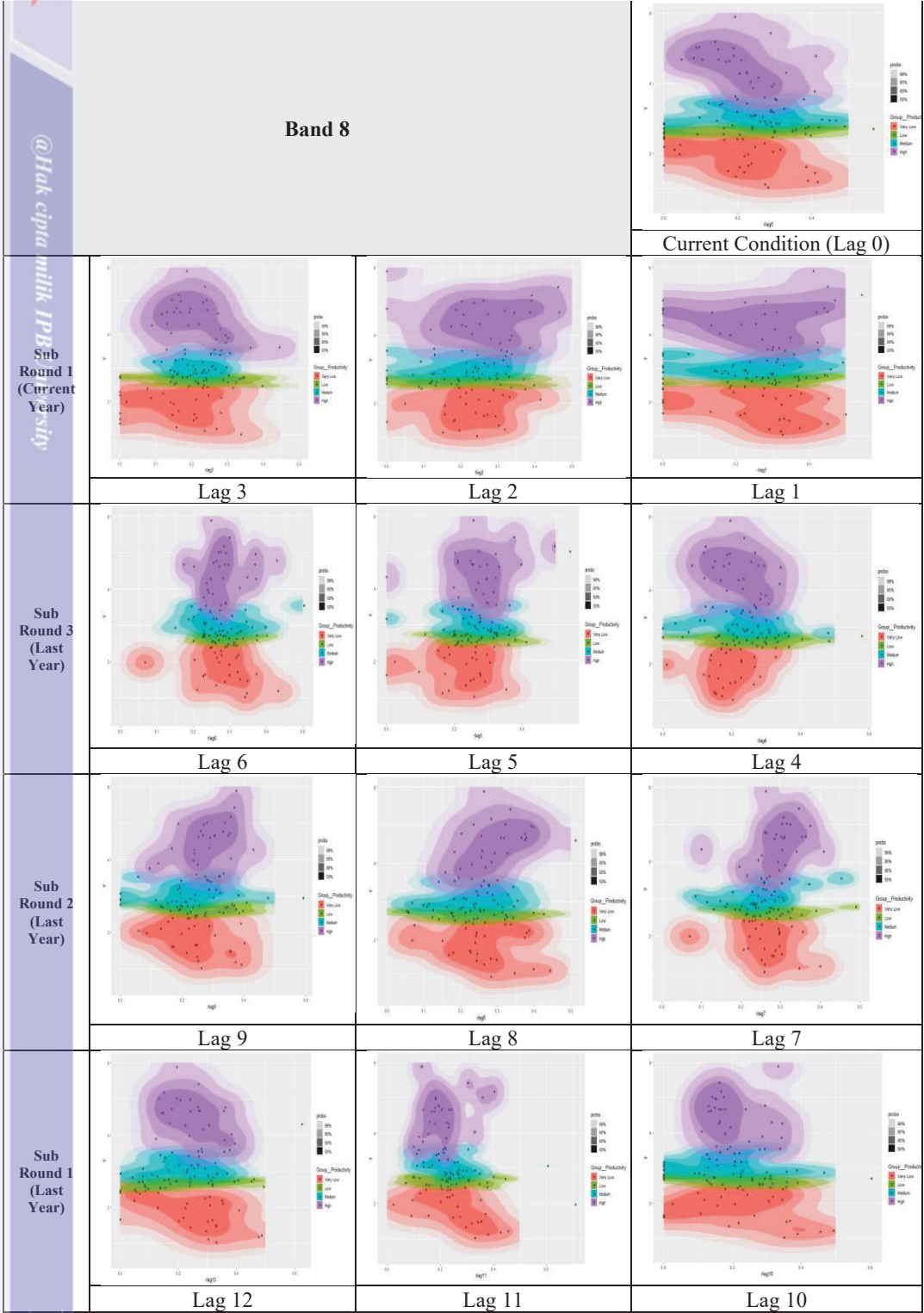


Figure 5.2 . Band 4, Band 8, and NDVI Scatterplots with Productivity Groups (b)

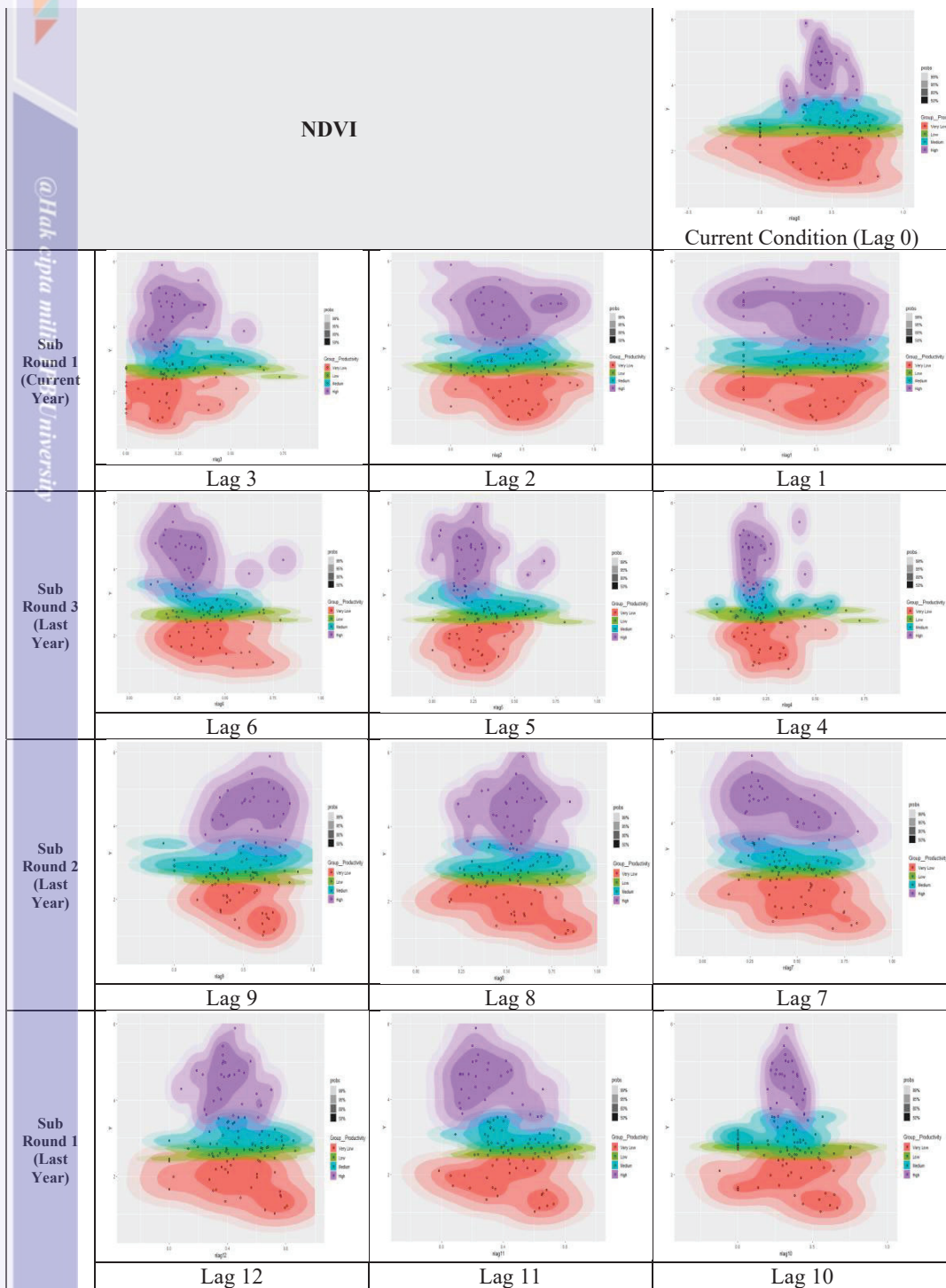


Figure 5.2 . Band 4, Band 8, and NDVI Scatterplots with Productivity Groups (c)

What differentiates Four Parameter Beta GMET from other tree-based models is that the model considers random effect value, which is calculated through GLMM models. When we have specified information on the random effect of an explanatory variable, the Four Parameter Beta GMET model can make an adjusted prediction to account for the effect and make a group-specific prediction (Fontana *et al.* 2021). The results analyzed above have taken into calculation

the random effects. The structure of the random effects were inline with the findings shown in sub chapter 3.3.5. Thus, we have used the GLMM model that can be seen in Equation 3.6, where farmers, farmers nested in subdistricts, and subdistrict random effects have been shown as a better fit in developing the GLMM model for paddy productivity. Thus, the average results of paddy productivity will differ for each farmer, or each farmer nested in a sub district, or for each district. If we want to analyze the random effect further, researchers can calculate the confidence interval of the random effects. For example, when the 95% confidence interval of the estimated random effects does not overlap with 0, we have evidence to assert that the random effect will cause a significant difference from the average effect. In our case, this was found for the subdistrict and district random effect; there are sub districts that have confidence intervals above zero, and there are sub districts that have confidence intervals below zero, which indicates underlying substantial differences between the subdistricts and districts (Figure 5.3). Meanwhile, the farmer random effects in Central Kalimantan showed that the confidence intervals stayed above or fell behind zero. Indicating that there are farmers who tend to fall behind and there are farmers that outperform.

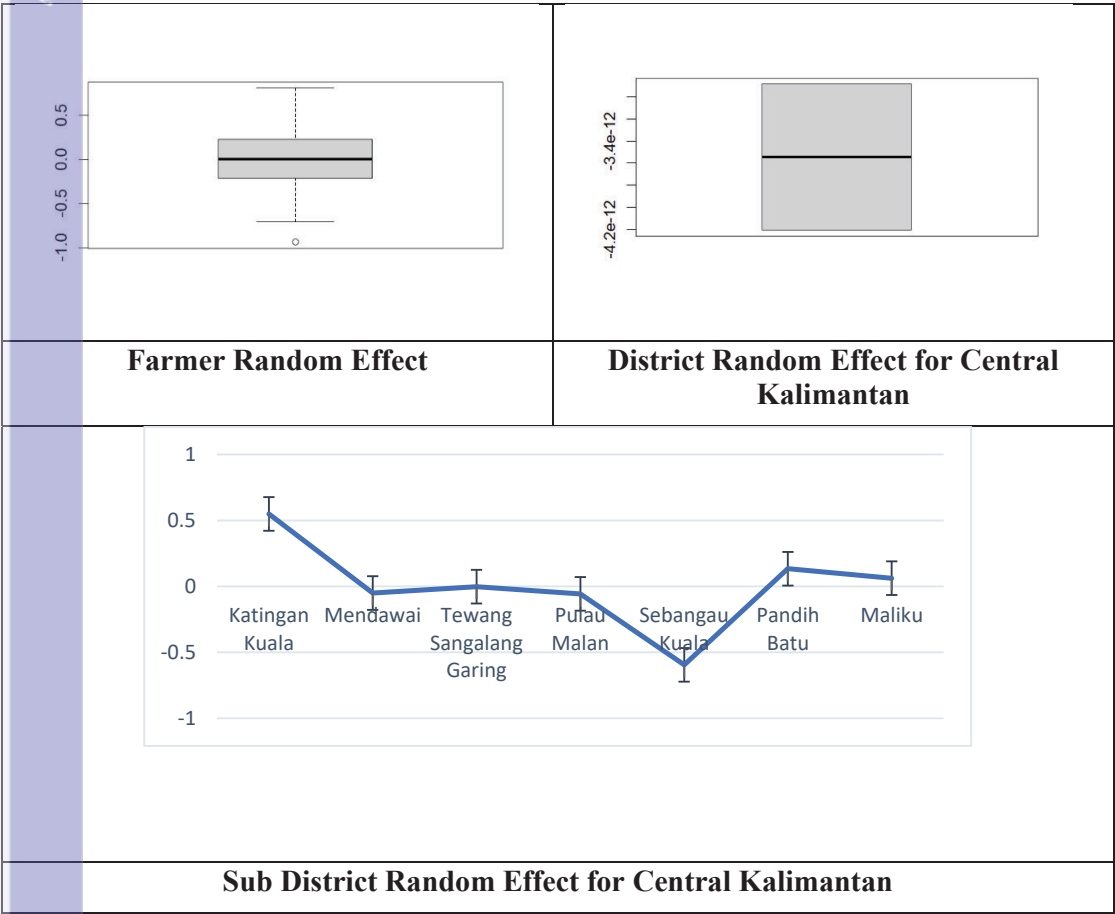


Figure 5.3. Estimated Random Effects of the Best Four Parameter Beta GMET Model

Other factors that are also considered important from the farmer survey are the severity of pest attacks of the current year and the amount of KCL and Solid Organic Fertilizer used. Lower use of KCL and Solid Organic Fertilizer tend to lead to lower productivity.

Hak Cipta Dilindungi Undang-undang

1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;

a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah

b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.

2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

5.4.3. Estimating Pure Premium and Risks of the Area Yield Index Crop Insurance

AYI crop insurance policy is a yield or productivity-based index agricultural insurance policy. National Development Planning Agency (BAPPENAS), MoA, Japan International Cooperation Agency (JICA), and other related stakeholders conducted an AYI pilot study in Karawang and Kendal districts Sanyu Consultants Inc. *et al.* (2023). Reports show that AYI exhibits many benefits such as minimal moral hazard and adverse selection. Furthermore, since the indemnity is determined by average sub district, insurers do not need to conduct loss adjustment surveys leading to lower administrative costs (Sanyu Consultants Inc. *et al.* 2023).

However, there are also obstacles found. One of the main issues is lack of reliable historical data required to calculate the benchmark yield (y_c) at an area level (province or district). Next, estimates of monthly or seasonal paddy yields (\bar{y}_l) are currently insufficient, particularly at a sub district or village level. As a response, Statistics Indonesia (SI) has developed a monthly paddy productivity estimate based on the Crop Cutting Experiments (CCEs). Nonetheless, the guidelines and data quality need to be evaluated for further applications of AYI (Shynkarenko *et al.* 2019).

Compensation paid by the insurer to all farmers in the sub area for AYI are based on indemnity that is formulated in Equation 2.15. To calculate the indemnity in Equation 2.15, it is apparent that predicting y_c is vital. BAPPENAS has used the current and average historical CCEs data in the pilot study (Adriansyah *et al.* 2021). Meanwhile, other information can also be utilized to enhance predictions accuracy, such as the farmer surveys and satellite data.

Aside from using average historical data for predicting y_c , applying Exponential Smoothing and ARIMA (Skees *et al.* 1997) are also popular. Machine Learning and Deep Learning (Sun *et al.* 2020) approaches have also been applied. Even though promising, there are several concerns that need to be accounted (1) heterogeneity of paddy productivity among areas, (2) productivity data distributions that have a certain minimum and maximum value or also known as the Four Parameter Beta distribution (Hennessy, 2009) and (3) occurrence of linear and nonlinear relationships between the response and explanatory variables.

Hence, we propose a new approach and developed a Four Parameter Beta GLMM in Chapter 3 and Chapter 4. Next we also enhanced the model by developing the Four Parameter Beta GMET and Four Parameter Beta GMERF prediction models in this chapter. We have also previously evaluated the performance of these models. Four Parameter Beta GMET and Four Parameter Beta GMERF had better prediction accuracy. For this reason, next we would like to evaluate and apply the developed model to estimate the premium and risks of AYI.

Model selection has been done based on RMSE and WAIC values in the sub chapter 5.4. Next, through a Bootstrap simulation, the predicted values of the best model will then be accounted to calculate the pure premium and VaR of AYI. The process is summarized as follows:

1. Predict paddy productivity (y_{ij}) for individual farmers based on the best fit model
2. Resample with replacement 100 bootstrap farmer samples (n) for each sub district
3. Calculate average yield for each sub district (\bar{y}_l) with $\bar{y}_l = \frac{1}{n} \sum_{i=1}^n y_{ij}$
4. Calculate benchmark yield (y_c) using average of an area, with $y_c = \frac{1}{N} \sum_{a=1}^N \bar{y}_l$. N is the total sub areas within Tan area
5. Calculate claim amounts based on Equation 2.15
6. Repeat steps 2-6 1.000.000 times and estimate the expected pure premium and VaR
7. Compare the result from simulation with actual conditions

Province and districts were used in the simulation to define a potential area level. If AYI was set at province level, the policy will have a higher pure premium and tail risks but easier to

administer. In a region where heterogeneity is apparent, setting a premium at district level is more sufficient. Therefore, preventing high basis risk of farmers and insurers. Nonetheless, administration wise it will be more challenging when in a region there are many districts. Table 6.1 also shows the estimated mean of a Four Parameter Beta distribution is much lower than the average value (y_c). Keeping in mind that paddy productivity has a Four Parameter Beta distribution, careful attention should be done in defining y_c . Setting higher average benchmark such as in this study may lead to overestimating the expected losses and cause higher premium for farmers, which may lead to decreased participation in the insurance program itself

Table 5.2 AYI Premium and VaR based on Four Parameter Beta GMERF Estimates (in Rupiah)

Area	y_c	Four Parameter Beta Mean	Pure Premium	$VaR_{95\%}$
Province	3.03	1.70	640,872.70	1,011,718.10
District 1	2.70	1.02	166,207.50	246,789.23
District 2	3.27	1.80	467,944.70	1,022,576.50

5.5.Conclusion

This paper introduces Four Parameter Beta GMET and Four Parameter Beta GMERF prediction model applied to predict paddy productivity. This model enhanced the existing models by considering (1) heterogeneity (2) appropriate Four Parameter Beta distributions, and (3) linear and nonlinear conditions. This proposed method has shown promising results compared to Four Parameter Beta GLM and Four Parameter Beta GLMM. It was also shown that the use of satellite data is more significant both in short-term and long-term compared to farmer survey data. This pointed out that it is possible to predict paddy productivity using Sentinel 2A satellite data analyzed with the right prediction model, such as the Four Parameter Beta GMET. As a result, it will be more time and cost efficient compared to using survey data. We have also shown that this model is quite promising due to its high prediction accuracy. Thus, the Ministry of Agriculture and other related stakeholders are suggested to develop paddy predictions based on this model. The model has the potential to be applied on a wider scale to predict paddy productivity where they might not have access to individual farmers' survey. It can also be used for developing an early warning system of food insecurity and a reference for the food self-sufficiency program.

We have also found that by calibrating the models to empirical data, extensive Bootstrap studies were performed to estimate the pure premium and VaR of AYI. Here, we concluded that designing AYI at district level is more appropriate when productivity among areas vary. Considerations must also be given in defining the benchmark productivity when there is proof that the distribution of paddy productivity follows a Four Parameter Beta distribution. Last, the use of satellite data in the model has proven a beneficiary. Thus, we encourage further studies in other areas and continuously improve the model.



VI. General Discussion

Advancements in developing prediction models based on the characteristics of data is important because they directly impact the accuracy, reliability, and applicability of the predictions made in various fields. By refining predictive models, researchers and practitioners can create more precise tools that better understand the nuances and patterns within data.

These advancements include developing models for data that are constrained by specific maximum and minimum values, such as a four-parameter beta distribution. This distribution offers flexibility by accommodating a wide range of shapes, skewness, and even heavy tails, enabling prediction models to capture diverse patterns in the data. This adaptability is particularly valuable when working with case studies that have varying characteristics.

Additionally, the development of prediction models should account for the complexity of relationships that may arise when applying different datasets. Understanding and modelling these complex relationships is key to improving prediction accuracy. Last it is also important to consider the inherent variability of the data, especially when working with datasets from a broad range of fields or geographical areas. Addressing this variability ensures that models are robust and applicable across different contexts.

The development of these prediction models is particularly very important for the development of Area Yield Index (AYI) crop insurance for paddy. In AYI accurate predictions of paddy productivity is essential for determining insurance premiums and risks. Paddy productivity has been proven to follow the Four Parameter Beta distribution. Therefore, the development of a Four Parameter Beta Generalized Linear Mixed Model (GLMM) is crucial. In this model, we have added the random effect to the model to account for high variability of paddy productivity across different areas. Thus, insurers can better estimate the likelihood of different productivity levels, improving the accuracy of risk assessment. Henceforward, leading to a more effective insurance product, ensuring that farmers receive fair and adequate compensation in cases of crop failure, while also maintaining the financial stability of the insurer.

Next, to develop a prediction model for AYI, integrating various sources of data, such as survey and satellite data, is essential for enhancing predictive accuracy and applicability. Survey data provides detailed, ground-level information on demographic, socioeconomic, and behavioral factors that affect paddy productivity. However, field surveys are often time consuming and constrained to sampling biases and response inaccuracies. Satellite data, on the other hand, offers large-scale, real time, continuous coverage, tracking environmental conditions like vegetation, soil moisture, and land use changes. This data type can reveal seasonal trends and extreme events, providing a temporal depth that periodic survey data lacks. Combining these sources of data with suitable prediction models allows the model to leverage both detailed local insights and expansive environmental monitoring, leading to enhanced predictive performance.

Therefore, in Chapter 3 we have developed the Four Parameter Beta GLMM model by conducting a transformation process that can be seen in Equation 2.5. This transformation maps the actual response variable y that has an interval $[a, b]$ to y^* with the interval $(0,1)$. Therefore, this process enables us to model and predict data that has a Four Parameter Beta distribution by applying beta GLM or GLMM models. We use a GLMM model when the response variable is measured in particular groups/areas ($i=1, 2, \dots, q$) or apply a GLM model if the data structure is more straightforward. Results show that the GLMM model is better than the GLM approach, indicating that random effects and fixed effects are needed for predicting paddy productivity.

It has been proven to be a more flexible and accurate prediction model for predicting paddy productivity when using satellite and farmer survey data. Hence, the Four Parameter Beta

GMET is quite promising. Out of the eighteen models examined for each area, the best two GLMM models are the models that incorporate both farmer survey and satellite data or the models that just incorporates farmer survey data.

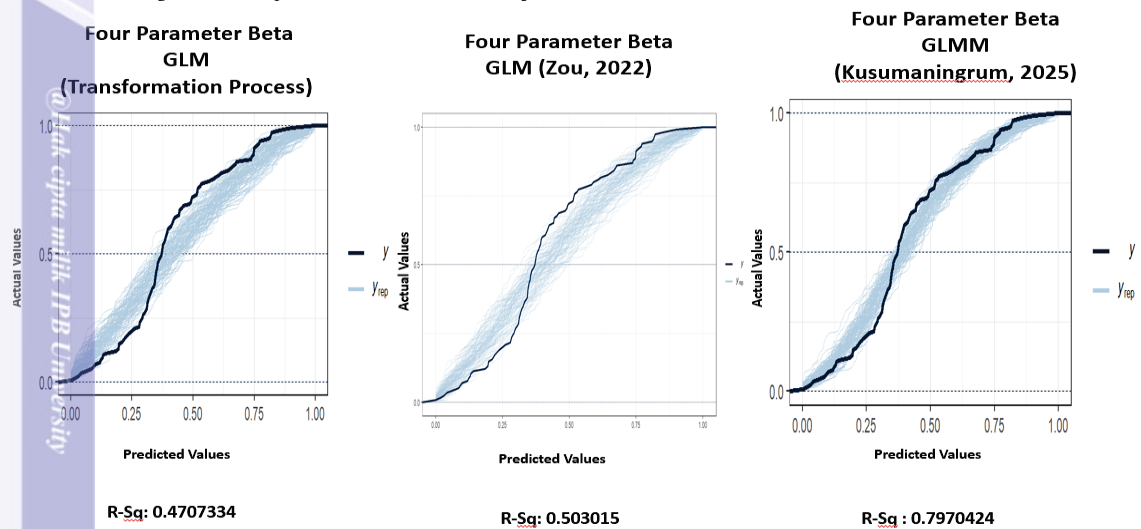


Figure 6. 1 Predicted and Actual Values Scatter Plots with R-Square of GLM and GLMM Models

Eventhough, the results of the Four Parameter Beta GLMM model are promising, the transformation process conducted in Chapter 3 can cause bias in parameter estimates and complications in the interpretation of coefficient values. Thus, we have further developed Zhou and Huang's (2022) mean Four Parameter Beta regression model in Chapter 4. This model was developed based on a Bayesian approach. The problem is that this model doesn't incorporate a random effect component. For paddy productivity predictions observations are collected over time or over multiple areas. Therefore, as mentioned above, incorporating a random effect in the model is crucial. To evaluate the developed model based on the Bayesian approach, we conducted a simulation study and applied the model in an empirical study. Evaluation processes are focused on the accuracy and efficiency of the estimated parameters and prediction results. Through the simulation study, we have shown that the developed Four Parameter Beta mean GLMM are considered relatively unbiased for the estimated parameters $\hat{\theta} = (\hat{\beta}, \hat{b}_v, \hat{a}, \hat{b}, \hat{\phi})$. However, estimates of the precision parameter ($\hat{\phi}$) was still biased. Further research is needed to increase the precision and accuracy of this parameter.

Stan package in R software was used to develop this model and it displays a more flexible and accurate result. As for the empirical study, Figure 6.1 presents a comparison of model performance across three approaches: the Four Parameter Beta GLM with transformation, the benchmark Four Parameter Beta GLM based on Zhou's (2022) model, and the developed Four Parameter Beta GLMM estimated using a Bayesian approach. Each plot shows the relationship between predicted and actual values of paddy productivity, with light blue lines representing predictive simulations and dark blue lines indicating the central predicted trend. The results highlight a clear improvement in model performance with increasing model complexity. The R-squared (R^2) values improve from 0.4707 for modelling a Four Parameter Beta GLM with transformation approach, to 0.5030 for modelling a Four Parameter Beta GLM with Zhou's (2022) benchmark model and reached 0.7970 for modelling a Four Parameter Beta GLMM using a Bayesian approach. This improvement demonstrates that incorporating random effects and adopting a Four Parameter Beta distribution in a GLMM model with Bayesian framework significantly enhances the model's predictive accuracy. The improved performance of the Four

Parameter Beta GLMM with Bayesian framework indicates its potential suitability for applications such as Area Yield Index (AYI) insurance, where accurate and reliable productivity predictions are essential for minimizing basis risk and setting fair premium rates.

Table 6. 1 Root Means Square Error Predictions (RMSEP) of the Purposed Methods

Model	Method	Independent Variables Data Source	RMSEP
Four Parameter Beta GLMM*	Four Parameter Beta GLMM Transformation Approach	Farmer Survey Data	0.120
		Satellite Data	0.110
		Farmer Survey and Satellite Data	0.110
Four Parameter Beta GLM	Four Parameter Beta GLM Bayesian Approach	Farmer Survey Data	0.010
		Satellite Data	0.011
		Farmer Survey and Satellite Data	0.011
Four Parameter Beta GLMM	Four Parameter Beta GLMM Bayesian Approach	Farmer Survey Data	0.029
		Satellite Data	0.028
		Farmer Survey and Satellite Data	0.020
Four Parameter Beta GMET	Four Parameter Beta GLMM Bayesian Approach and Tree Regression	Farmer Survey Data	0.003
		Satellite Data	0.004
		Farmer Survey and Satellite Data	0.026
Four Parameter Beta GMERF	Four Parameter Beta GLMM Bayesian Approach and Random Forest	Farmer Survey Data	0.003
		Satellite Data	0.008
		Farmer Survey and Satellite Data	0.010

@Hak cipta milik IPB University

Hak Cipta Dilindungi Undang-undang
1. Dilarang mengutip sebagian atau seluruh karya tulis ini tanpa mencantumkan dan menyebutkan sumber ;
a. Pengutipan hanya untuk kepentingan pendidikan, penelitian, penulisan karya ilmiah, penyusunan laporan, penulisan kritik atau tinjauan suatu masalah
b. Pengutipan tidak merugikan kepentingan yang wajar IPB University.
2. Dilarang mengumumkan dan memperbanyak sebagian atau seluruh karya tulis ini dalam bentuk apapun tanpa izin IPB University.

Now having better foundations on how to develop a more unbiased and accurate Four Parameter Beta GLMM prediction model based on the Bayesian problem. In chapter 5, we can further improve the prediction results for a more complex data, where linear and non linear relationships exist. Preliminary research show that GMET is more accurate in predicting the paddy productivity, especially in conditions where survey data was used alongside satellite data. Therefore we improved the algorithm to develop the Four Parameter Beta GMET by applying the Four Parameter Beta GLMM and regression tree. Further on, as an effort to enhance the predicting accuracy, we developed the Four Parameter Beta GMERF by applying the Four Parameter Beta GLMM alongside random forest. Resultwise, these proposed models predictions of paddy productivity accuracy outperformed the Four Parameter Beta GLM and GLMM models when applying farmer survey and satellite data (Table 6. 1).

In Chapter 5, apart from developing the Four Parameter Beta GLMM, GMET, and GMERF models. We have also evaluated the models prediction accuracy and selected the best model. By calibrating the models to empirical data, extensive Bootstrap studies were performed to estimate the pure premium and VaR of AYI. It was shown that designing AYI at district level is more appropriate when productivity among areas vary. Considerations must also be given in defining the benchmark productivity when there is proof that the distribution of paddy productivity follows a Four Parameter Beta distribution. The use of satellite data in the model has proven a beneficiary. Thus, we encourage further studies in other areas and continuously improve the model.

Typically, AYI premiums are calculated using only average historical productivity data. This approach lacks flexibility, as it doesn't account for variations in climate, pest outbreaks, or other dynamic factors affecting crop yields. Additionally, historical averages do not incorporate new data that may emerge after the initial calculation. By contrast, predictive models such as the Four Parameter Beta GLMM, GMET, and GMERF provide a more refined, data-driven method for estimating paddy productivity, factoring in various yield influences. This model offers improved adaptability, accuracy, and responsiveness to agricultural changes, though its effectiveness relies on the quality and relevance of training and validation data.



VII. Conclusion and Recommendation

7.1. Conclusion

Advancements in predictive modeling are crucial for enhancing accuracy and applicability across various fields. In particular, the development of models like the Four Parameter Beta GLMM that addresses the complexities and variability in data. This model takes into consideration the characteristics of paddy productivity, which follow a Four Parameter Beta distribution, and incorporates random effects to account for variability across different areas. Hence, paddy productivity predictions are more accurate. As a result, predictions can be used to calculate the premium and risks of Area Yield Index (AYI) crop insurance policy.

First, for simplicity we have developed the four-parameter beta GLMM model by conducting a transformation process. Results were quite promising. Nevertheless, there is a potential of bias in parameter estimates and complications in the interpretation of coefficient values when transformation process is being performed. The bias can also directly influence the reliability and accuracy of the model's predictions. Ignoring such biases can lead to invalid predictions, compromising the practical applications of the model, particularly in insurance premium calculation and risk assessment.

Therefore, we have further developed the model based on Zhou's (2022) Four Parameter Beta mean regression model by adding a random effect to the model. To evaluate the developed model based on the Bayesian approach, we conducted a simulation study and applied the model in an empirical study. Based on the simulation study, we have shown that the developed Four Parameter Beta mean GLMM are considered relatively unbiased for estimating parameters $\hat{\theta} = (\hat{\beta}, \hat{b}_v, \hat{a}, \hat{b}, \hat{\phi})$. However, estimates of the precision parameter ($\hat{\phi}$) was still biased. Further research is suggested to increase the precision and accuracy of this parameter. The simulation results also showed that the development of this model using Stan package in R software is more flexible and accurate.

The integration of diverse data sources, including survey and satellite data, is known to increase predictive accuracy of paddy productivity. Nonetheless, more complex linear and non linear relationship will occur. Thus, the four-parameter beta GLMM has been further developed into a Generalized Mixed Effect Tree (GMET) and a Generalized Mixed Effect Random Forest (GMERF). This model proved more suitable for predicting paddy productivity compared to the four-parameter beta GLMM when using satellite data and farmer surveys. The research also demonstrated that short-term and long-term satellite data is more significant for predictions compared to survey data. Optimizing the use of Sentinel 2A satellite data, for predicting paddy productivity is more time- and cost-efficient compared to relying solely on survey data.

Additionally, we have also shown through empirical case study that paddy productivity in Central Kalimantan and Karawang have a Four Parameter Beta distribution that is skewed to the right. Therefore, the Four Parameter Beta distribution is an appropriate basis for modeling. The proposed Four Parameter Beta mean GMET and GMERF showed higher prediction accuracy than standard regression models and proved to be a better foundation for calculating premiums and Value at Risk (VaR) for AYI crop insurance. This research emphasizes other key factors that have an influence on paddy productivity, including pest management (X7), appropriate dosage of KCL (X12), solid organic fertilizer/compost (X13), Band 4, Band 8, and NDVI from satellite data. Moreover, simulation results suggest that AYI crop insurance policies at the district level are optimal for addressing regional productivity variability and reducing basis risk.

7.2. Recommendations

This research has limitations. Model wise, we have considered developing a Four Parameter Beta GLMM that has a random intercept, using a logit link function, and based on a mean central tendency. Further research can be developed based on random slopes, applying other link functions, and use mode as the central tendency. We have also considered that the CCE plots are independent. Considering spatial effects in the developed models can be beneficial. In this research we have combined GLMM with regression tree and random forest, which show promising results for complex data set. Moving forward it is also suggested that the model be combined with other machine learning techniques that would potentially lead to better performance in terms of prediction accuracy and interpretability.

Next, simulation results show that the precision parameter is still biased, thus further research is suggested to increase the precision and accuracy of this parameter. Application wise, it is suggested to apply and evaluate the model in other various areas to increase the generalization, robustness, and prediction accuracy. We also recommend that the model be developed by adding relevant phenology, climate, and geographical information variables. For real time predictions needed to support faster claims and risk mitigations, researchers should develop a systems for real-time paddy monitoring and data integration to allow the models to be continuously updated with new information, improving responsiveness to dynamic agricultural conditions.

Last, policy wise, based on bootstrap simulations, we suggested to Design AYI policies at the district level to better accommodate productivity variability across regions. This involves recalculating AYI premiums using the proposed four-parameter beta GLMM for a more sophisticated and data-driven predictions. Farmers opting for AYI policies will benefit from lower basis risk and easier claim settlements. We should also formulate policies and action plans supporting sustainable agricultural practices based on variables that have a significant effect on paddy productivity, such as promoting pest management strategies tailored to local conditions, encourage the appropriate use of KCL (potassium chloride) and organic fertilizers, and also promote the production and distribution of affordable, high-quality compost and organic fertilizers. It was also shown that the role of satellite data is apparent in predicting paddy productivity. Therefore, we should encourage the GoI and insurance companies to invest more in satellite-based monitoring systems, emphasizing variables such as Band 4, Band 8, and NDVI to regularly assess field health and productivity. We can also utilize satellite-based insights to identify high risk areas that are impacted by climate change.

By addressing these recommendations, it is expected that we can further improve the model's precision, accuracy, generalization, and responsiveness to dynamic agricultural conditions. Hence, stakeholders such as the Ministry of Agriculture, NGOs, insurance companies, and farmers can enhance the efficiency and effectiveness of AYI crop insurance policy and contribute to improved agricultural practices overall.



References

- Agreda, D.F., Cantoni, E. 2017. Bootstrap Estimation of Uncertainty in Prediction for Generalized Linear Mixed Models. *Computational Statistics and Data Analysis* (2018), <https://doi.org/10.1016/j.csda.2018.08.006>.
- Agresti, A. 2018. *An Introduction to Categorical Data Analysis*, Wiley Hoboken, NJ, USA
- Ardiansyah, M., Kurnia, A., Sadik, K., Djuraidah, A., and Wijayanto, H. 2021. Numerical Prediction of Paddy Weight of Crop Cutting Survey using Generalized Geoadditive Linear Mixed Model. *Journal of Physics: Conference Series*, vol. 1863, pp. 1-17 (2021), doi: 10.1088/1742-6596/1863/1/012024
- Bonat, W.H., Ribeiro Jr, P.J., Zeviani, W.M. 2015. Likelihood Analysis for a Class of Beta Mixed Models. *Journal of Applied Statistics*, 42(2), 252-266. DOI: 10.1080/02664763.2014.947248
- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. 1984. *Classification and Regression Trees*, The Wadsworth Statistics and Probability Series; Wadsworth International Group: Belmont, CA, USA, p. 356.
- Breiman, L. J. H. Friedman, R. A. Alshen, and C. J. Stone. 1993. *Classification and Regression Trees*. New York: Chapman and Hall.
- Brooks, M. E., Kristensen, K., Benthem, K. J.V., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Mächler, M. and Benjamin, B.M. 2017. glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal* 9(2), pages 378-400. DOI: 10.32614/RJ-2017-066
- Butar, F. B., and Lahiri, P. 2003. On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, 112(1-2), 63-76. [https://doi.org/10.1016/S0378-3758\(02\)00323-3](https://doi.org/10.1016/S0378-3758(02)00323-3)
- Cedric, L.S., W.Y.H. Adoni, R. Aworka, J.T. Zoueu, F.K. Mutombo, M. Krichen, and C.L.M. Kimpolo. 2022. Crops Yield Prediction Based on Machine Learning Models: Case of West African Countries. *Smart Agricultural Technology* 2. <https://doi.org/10.1016/j.atech.2022.100049>.
- Chen, J., Jonsson, P., Tamura, M., Gu, Z., Matsushita, B., & Eklundh, L. (2004). A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky-Golay filter. *Remote Sensing of Environment*, 91(3-4), 332-344. <http://doi.org/10.1016/j.rse.2004.03.014>
- Clauss, K., Ottinger, M., Leinenkugel, P., and Kuenzer, C., 2018. Estimating rice production in the Mekong Delta, Vietnam, utilizing time series of Sentinel-1 SAR data, *International Journal of Applied Earth Observation and Geoinformation* 73 574–585, doi: 10.1016/j.jag.2018.07.022.
- Debacke, P., Attia, F., Champolivier, L., Dejoux, J.-F., Micheneau, A., Al Bitar, A., Trepos, R. 2023. Forecasting sunflower grain yield using remote sensing data and statistical models. *European Journal of Agronomy*, 142, Article 126677. <https://doi.org/10.1016/j.eja.2022.126677>
- Dishon, M., and G.H. Weiss., 1980. Small Sample Comparison of Estimation Methods for the Beta Distribution. *J. Stat. Comput. Sim.*, 11, 1-11, doi:10.1080/00949658008810385.
- Drusch M, Bello UD, Carlier S, Colin O, Fernandez V, Gascon F, Hoersch B, Isola C, Laberinti

P., Martimort P, Meygret A, Spoto F, Sy O, Marchese F, Bargellini P. 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment*. 120(2012): 25–36.

[ESA] European Space Agency. 2015. Sentinel-2 User Handbook. ESA Communication, Noordwijk: 1-64.

Ferrari, S.L.P., and Cribari-Neto, F. 2004. Beta Regression for Modelling Rates and Proportions.” *Journal of Applied Statistics*, 31(7), 799-815.

Fokkema, M., Smit, N., Zeileis, A., Hothorn, T., Kelderman, H. 2018. Detecting Treatment-subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees. *Behav Res* 50, 2016-2034. <https://doi.org/10.3758/s13428-017-0971>

Fontana, L., Masci, C., Ieva, F., Paganoni, A.M., 2021. Performing Learning Analytics via Generalised Mixed-Effects Trees. *MDPI Data Journal* 2021 Vol 6:74. <https://doi.org/10.3390/data6070074>

Guo, Y., Fu, Y., Hao, F., Zhang, X., Wu, W., Jin, X., Bryant, C. R., Senthilnath, J. 2021. Integrated phenology and climate in rice yields prediction using machine learning method. *Ecological Indicators*, 120, Article 106935. <https://doi.org/10.1016/j.ecolind.2020.106935>

Guo, Y., Chen, S., Li, X., Cunha, M., Jayavelu, S., Cammarano, D., Fu, Y. 2022. Machine learning-based approaches for predicting SPAD values of maize using multi-spectral images. *Remote Sensing*, 14(6), Article 1337. <https://doi.org/10.3390/rs14061337>

Guo, Y., Xiao, Y., Hao, F., Zhang, X., Chen, J., De Beurs, K. D., He, Y., & Fu, Y. H. 2023. Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images. *International Journal of Applied Earth Observation and Geoinformation*, 124, Article 103528. <https://doi.org/10.1016/j.jag.2023.103528>

Harville, D.A. 1974. Bayesian Inference for Variance Components Using Only Error Contrasts. *Biometrika*, Vol. 61 (2), 383-385. URL: <http://www.jstor.org/stable/2334370>

Haryastuti, R., Aidi, M. N., Pasaribu, S. M., Sumertajaya, I. M., Sutomo, V. A., Kusumaningrum, D., and Anisa, R. 2021. Cluster based area yield scheme for crop insurance policy in Java J. of Phys.: Conference Series 1821

Hennessy, D. A. 2009. Crop Yield Skewness and the Normal Distribution. *Journal of Agricultural and Resource Economics*, 34(1), 34-52. <http://www.jstor.org/stable/41548400>

Jiang, J., and Lahiri, P. 2001. Empirical Best Prediction for Small Area Inference with Binary Data. *Annals of the Institute of Statistical Mathematics* 53, 217-243 (2001). <https://doi.org/10.1023/A:1012410420337>

Kackar, R.N., Harville, D.1984. Approximations for Standard Errors of Estimators of Fixed and Random Effect in Mixed Linear Models. *J. Amer. Statist. Assoc.* 33 853-862.

Klompenburg, T. V., Kassahun, A., & Catal, C. 2020. Crop Yield Prediction using Machine Learning: A Systematic Literature Review. *Computers and Electronics in Agriculture*, 177. <https://doi.org/10.1016/j.compag.2020.105709>

Kudus, A. 1999. Penerapan Metode Regresi Berstruktur Pohon pada Pendugaan Masa Rawat Kelahiran Bayi (Studi Kasus Rumah Sakit Hasan Sadikin Bandung) [Tesis]. Bogor: Program Pascasarjana, Institut Pertanian Bogor.

- Kusumaningrum, D., Anisa, R., Sutomo, V.A., Tan, K.S. 2021. Alternative Area Yield Index Based Crop Insurance Policies in Indonesia. Springer Books, in: Marco Corazza & Manfred Gilli & Cira Perna & Claudio Pizzi & Marilena Sibillo (ed.), MAF Conference Series pages 285-290. Springer. DOI: 10.1007/978-3-030-78965-7_42
- Kusumaningrum, D., Sundari, M., and Kurnia, A. 2022. Bayesian Premium Calculations Of Multiperil Crop Insurance (MPCI) Based On Bayesian Beta Mixed Regression Model. AIP Conference Proceedings 2662. doi: 10.1063/5.0108843.
- Kusumaningrum D, Wijayanto H, Kurnia A, Notodiputro K A, Ardiansyah M, Islam, M.P. 2024. Beta Four Parameter Mixed Models Based on Survey and Satellite Data for Paddy Productivity Predictions". Smart Agricultural Technology 9. <https://doi.org/10.1016/j.atech.2024.100525>
- Lele, S.R., Dennis, B., Lutscher, F. 2007. Data Cloning: Easy Maximum Likelihood Estimation for Complex Ecological Models using Bayesian Markov Chain Monte Carlo Methods. Ecology Letters, 10: 551-563. <https://doi.org/10.1111/j.1461-0248.2007.01047.x>
- Manteiga, W.G., Lombardia, M.J., Molina, I., Morales, D., Santamaria, L. 2008. Analytic and Bootstrap Approximations of Prediction Errors under a Multivariate Fay-Herriot Model. Journal of Computational Statistics and Data Analysis, Volume 52(12). <https://doi.org/10.1016/j.csda.2008.04.031>
- Marnawati. 2022. Peningkatan Produktivitas Padi Lahan Kering, <https://tanamanpangan.pertanian.go.id/detil-konten/iptek/45>
- McCullagh, P. and Nelder, J.A. 1989. Generalized Linear Models. 2nd Edition. London: Chapman dan Hall.
- McCulloch, C. dan Searle, S.R. 2001. Generalized, Linear, and Mixed Models. John Wiley & Sons, Inc., New York.
- Mittelhammer, R.C., Judge, G.G., Miller, D.J. 2000. Econometric Foundations. Cambridge University Press, New York
- Mutaqin, A. K., Karyana, Y., Sunendiari, S. 2020. Pure premium calculation of rice farm insurance scheme in Indonesia based on the 4-parameter beta mixture distribution. IOP Conference Series Materials Science and Engineering 830(2):022005. DOI: 10.1088/1757-899X/830/2/022005
- Nelder J, and Wedderburn. 1972. Generalized Linear Models. Journal of the Royal Statistical Society Series A (General). 135(3): 370-384. doi: 10.2307/2344614.
- Newlands, N. K., Ghahari, A., Gel, Y. R., Lyubchich, V., and Mahdi, T. 2019. Deep Learning for Improved Agricultural Risk Management, Proceedings of the 52nd Hawaii International Conference on System Sciences.
- Nocedal, J., Wright, S.J. 1999. Numerical Optimization. Springer -Verlag. New York
- Ospina, R., Cribari-Neto, F., Vasconcellos, K.L.P. 2006. Improved Point and Interval Estimation for a Beta Regression Model. Comput. Statist. Data Anal. 51, 960-981.
- Pellagatti, M., Masci, C., Ieva, F., Paganoni, A.M. 2021. Generalized mixed-effects random forest: A flexible approach to predict university student dropout. Statistical Analysis and Data Mining: The ASA Data Science Journal, Volume 14: 3 p. 241-257 <https://doi.org/10.1002/sam.11505>
- Pregibon, D. 1980. Goodness of Link Tests for Generalized Linear Models. Journal of the Royal

Statistical Society. Series C (Applied Statistics), Vol. 29 (1), 15-23.
<https://doi.org/10.2307/2346405>

Said, H. I., Subiyanto, S., dan Yuwono, B. D. 2015. Analisis Produksi Padi dengan Penginderaan Jauh dan Sistem Informasi Geografis di Kota Pekalongan. *J. Geodesi Undip.* 4(1):1-8.

Sanyu Consultants Inc., Sompoo Risk Management Inc. 2023. *Area yield index insurance product design*. National Development Planning Agency (BAPPENAS) and Japan International Cooperation Agency (JICA)

Sammatat, S., Lekdee, K. 2018. Generalized Linear Mixed Models for Spatio-Temporal Data with an Application to Leptospirosis in Thailand. *Applied Mathematical Sciences*, Vol. 12, 2018, no. 28, 1357 - 1366 HIKARI Ltd, www.m-hikari.com

Sari, V. D. and Sukojo, B. M. 2015. Analisa Estimasi Produksi Padi Berdasarkan Fase Tumbuh Dan Model Peramalan Autoregressive Integrated Moving Average (ARIMA) Menggunakan Citra Satelit Landsat 8 (Studi Kasus: Kabupaten Bojonegoro), *Journal of Geodesy and Geomatics* 10, DOI: <http://dx.doi.org/10.12962/j24423998.v10i2.828>

Schmoor, C., Olschewski, M. and Schumacher, M. 1996. Randomized and Non-Randomized Patients in Clinical Trials: Experiences with Comprehensive Cohort Studies. *Statist. Med.* 15, 263- 271 (1996).

Sela, R.J., and Simonoff, J.S. 2012. RE-EM trees: A Data Mining Approach for Longitudinal and Clustered data. *Mach. Learn.* 86, 169-207

Sharma, S., Chatterjee, S. 2021. Winsorization for Robust Bayesian Neural Networks. *Entropy* Vol 23, 1546. <https://doi.org/10.3390/e23111546>

Shynkarenko, I., Shynkarenko, R., Krychevska, L., McConnel, R. 2019. Survey on sustainable agricultural insurance scheme in Indonesia. National Development Planning Agency (BAPPENAS) and Japan International Cooperation Agency (JICA)

Simas, A.B., Rocha, A.V. 2006, “ betareg: Beta Regression. R package version 1.2.”, URL <http://CRAN.R-project.org/src/contrib/Archive/betareg/>.

Skees, J.R., Black, J.R., Barnett, B.J. 1997. :Designing and Rating an Area Yield Crop Insurance Contract. *Am. J. Agric. Econ.* 79(2), 430-438 (1997)

Son, N. T., Chen, C. F., Chen, C. R., Chang, L. Y., Duc, H. N., & Nguyen, L. D. 2013. Prediction of rice crop yield using MODIS EVI–LAI data in the Mekong Delta, Vietnam. *International Journal of Remote Sensing*, 34(20), 7275–7292.
<https://doi.org/10.1080/01431161.2013.818258>

Subedi, B., Poudel, A., Aryal, S. 2023. The impact of climate change on insect pest biology and ecology: Implications for pest management strategies, crop production, and food security. *Journal of Agriculture and Food Research*, 14, 100733.
<https://doi.org/10.1016/j.jafr.2023.100733>

Sulaeman, Y., Aryati, V., Suprihatin, A., Santari, P., Haryati, Y., Susilawati, S., Siagian, D., Karolinoerita, V., Cahyaningrum, H., Pramono, J., Wulanningtyas, H., Fauziah, L., Raharjo, B., Syafruddin, S., Cahyana, D., Waluyo, W., Susanto, B., Purba, R., Dewi, D., ... Yasin, M. 2024. Yield gap variation in rice cultivation in Indonesia. *Open Agriculture*, 9, 20220241. <https://doi.org/10.1515/opag-2022-0241>

Sun, J., Lai, Z., Di, L., Sun, Z., Tao, J., and Shen, Y. 2020. Multilevel deep learning network

for county-level corn yield estimation in the U.S. Corn Belt, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13 5048-5060, doi: 10.1109/JSTARS.2020.3019046.

Sutarlan, N. P. B, Mutaqin, A.K, and Achmad, A.I. 2017. Pemodelan Data Produktivitas Hasil Panen Padi Menggunakan Sebaran Mixture Beta 4-Parameter. Skripsi mahasiswa Program Studi Statistika Unisba

Tierney, L., and Kadane, J.B. 1986. Accurate Approximations for Posterior Moments and Marginal Densities, Journal of American Statistics Association. Vol 81, 82-86.

Thiele, J. and Markussen, B. 2012. Potential of GLMM in Modelling Invasive Spread. CAB Reviews Perspectives in Agriculture Veterinary Science Nutrition and Natural Resources. DOI: 10.1079/PAVSNR20127016

Vitasari, W., Useng, D., Munir, A. 2017. Pendugaan Produksi dan Indeks Vegetasi Tanaman Padi Menggunakan Data Citra Platform Unmanned Aerial Vehicle (UAV) dan Data Citra Satelit Landsat-8. Jurnal AgriTechno Vol. 10(2).

Wu, C. F. J. 1986. Jackknife bootstrap and other resampling methods in regression analysis. Annals of Statistics 14, 1261-1295.

Zeileis, A., Hothorn, T., and Hornik, K. 2008. Model-based Recursive Partitioning. Journal of Computational and Graphical Statistics, 17(2), 492-514.

Zhou, H., and Huang, X. 2022. Bayesian Beta Regression for Bounded Responses with Unknown Supports. Journal of Computational Statistics and Data Analysis Vol 167. <https://doi.org/10.1016/j.csda.2021.107345>