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

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

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 Sources of information derived or quoted from published and unpublished works by other authors have been mentioned in the text and included in the Bibliography at the end of this dissertation.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

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. Dimbimbing 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, …. ,q) dan j = 1,2,… 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 $(\hat{\varphi})$. 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, …. ,q) and j = 1,2,… 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 $(\hat{\varphi})$. 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

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

External Examiners on Closed Examination:

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

External Promotors on Doctoral Promotion:

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

Disertation 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 Buono, M.Si, M.Kom
NIP 196607021993021001

Date of Closed Examination : January 30th, 2025

Graduate Date :

FOREWARD

Alhamdulillahi 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 Degreee 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. Prasetiya 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 assitance 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

LIST OF TABELS

LIST OF FIGURESS

LIST OF APPENDICES