ENHANCING FAY–HERRIOT MODEL PERFORMANCE THROUGH ENSEMBLE LEARNING FOR SMALL AREA ESTIMATION IN THE PRESENCE OF NONLINEARITY AND MULTICOLLINEARITY

ARI SHOBRI BUKHARI

STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2026

STATEMENT OF ORIGINALITY, SOURCES OF INFORMATION, AND COPYRIGHT TRANSFER

I hereby declare that the dissertation entitled "Enhancing Fay–Herriot Model Performance through Ensemble Learning for Small Area Estimation in the Presence of Nonlinearity and Multicollinearity" 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 Sources of information derived or quoted from published and unpublished works by other authors have been mentioned in the text and included in the Bibliography at the end of this dissertation.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, August 2026

Ari Shobri Bukhari
G1601211007

RINGKASAN

ARI SHOBRI BUKHARI. Peningkatan Kinerja Model Fay–Herriot melalui Pembelajaran Rakitan untuk Pendugaan Area Kecil pada Kondisi Nonlinearitas dan Multikolinearitas. Dibimbing oleh KHAIRIL ANWAR NOTODIPUTRO, INDAHWATI, ANWAR FITRIANTO, dan ANANG KURNIA.

Pendugaan area kecil (*small area estimation*/SAE) digunakan untuk meningkatkan presisi pendugaan pada domain dengan ukuran contoh terbatas. Dalam statistik resmi, model Fay–Herriot (FH) merupakan pendekatan SAE tingkat area yang paling luas digunakan karena menggabungkan informasi survei langsung dan peubah penyerta dalam kerangka model linear campuran. Namun, kinerjanya sangat bergantung pada asumsi linearitas, sedangkan data sosial ekonomi sering menunjukkan hubungan nonlinear, heterogenitas, dan multikolinearitas. Akibatnya, model FH standar menjadi kurang fleksibel dan dapat menghasilkan pendugaan yang kurang presisi.

Perkembangan pembelajaran mesin (*machine learning*) membuka peluang pengembangan SAE yang lebih adaptif karena mampu menangkap pola nonlinear tanpa spesifikasi fungsi tertentu. Meskipun demikian, integrasi pembelajaran mesin dalam SAE tingkat area masih relatif terbatas, terutama yang tetap mempertahankan struktur inferensial Fay–Herriot. Selain itu, model yang mampu menangani nonlinearitas, multikolinearitas, dan peubah respons berbatas (yang umum dijumpai pada indikator SDG's) secara simultan masih jarang ditemukan dalam literatur.

Disertasi ini mengembangkan dua model SAE berbasis pembelajaran rakitan (*ensemble learning*), yaitu model Fay-Herriot dengan hutan acak (*Fay–Herriot with Random Forest*/FHRF) dan model Fay-Herriot dengan hutan rotasi (*Fay–Herriot with Rotation Forest*/FHRoF). Masing-masing dikembangkan melalui pendekatan koreksi (*correction*/C) dan penggantian (*replacement*/R). Pendekatan koreksi mempertahankan struktur dasar FH dan memodelkan sisaan menggunakan pembelajaran mesin, sedangkan pendekatan penggantian mengganti komponen pengaruh tetap dengan model rakitan. Pengembangan ini ditujukan untuk meningkatkan fleksibilitas model dalam menghadapi nonlinearitas, multikolinearitas tinggi, dan selanjutnya dikembangkan untuk menangani peubah respons berbentuk proporsi melalui transformasi *additive log-ratio* (alr).

Kajian pertama mengembangkan SAE–FHRF untuk mengatasi pelanggaran asumsi linearitas pada model FH. Hasil simulasi menunjukkan bahwa SAE–FHRF secara konsisten mengungguli FH–EM pada kondisi nonlinear melalui penurunan galat akar kuadrat tengah relatif (*relative root mean squared error*/RRMSE) dan bias relatif (*relative bias*/RB), dengan keunggulan yang semakin meningkat seiring bertambahnya jumlah area. Sebaliknya, pada kondisi linear, FH–EM tetap memberikan kinerja terbaik karena sesuai dengan asumsi model yang mendasarinya.

Aplikasi empiris pada pendugaan pengeluaran per kapita rumah tangga pertanian di Indonesia menunjukkan bahwa penduga langsung memiliki keterbatasan pada area dengan ukuran contoh kecil dan galat baku relatif (*relative standard error*/RSE) tinggi. Diagnostik model mengindikasikan adanya nonlinearitas, kemenjuluran (*skewness*), dan multikolinearitas. Dalam kondisi tersebut, SAE–FHRF, khususnya FHRF-R, mampu menurunkan RSE secara signifikan dibandingkan estimator langsung maupun FH–EM.

Kajian kedua mengembangkan SAE–FHRoF melalui integrasi *Rotation Forest* berbasis Analisis Komponen Utama (*Principal Component Analysis*/PCA) ke dalam kerangka Fay–Herriot untuk menangani multikolinearitas tinggi dan hubungan semi-linear. Hasil simulasi menunjukkan bahwa FHRoF-C memberikan kinerja terbaik pada kondisi semi-linear dan meningkatkan stabilitas pendugaan ketika multikolinearitas tinggi. *Rotation Forest* mampu mempertahankan struktur hubungan yang mendekati linear sekaligus mengurangi dampak korelasi antarpeubah penyerta. Namun, pada hubungan nonlinear yang lebih kompleks, SAE–FHRF tetap lebih unggul karena *Random Forest* lebih fleksibel dalam menangkap interaksi dan pola nonlinear. Dengan demikian, FHRF lebih adaptif pada nonlinearitas kompleks, sedangkan FHRoF lebih efisien pada hubungan mendekati linear dengan korelasi tinggi antarpeubah penyerta.

Hasil evaluasi berdasarkan data empiris menunjukkan bahwa SAE–FHRoF juga mampu meningkatkan presisi pendugaan dibandingkan estimator langsung dan FH–EM, terutama pada area dengan RSE tinggi. Secara statistik, kinerjanya relatif setara dengan SAE–FHRF sehingga kedua pendekatan dapat dipandang sebagai metode yang saling melengkapi.

Kajian ketiga memperluas SAE–FHRF dan SAE–FHRoF untuk peubah respons berbentuk proporsi melalui transformasi *log-ratio* agar hasil pendugaan tetap ada pada rentang nol hingga satu. Pendugaan parameter dilakukan menggunakan kerangka *compositional mixed model* untuk mengakomodasi bias akibat transformasi. Hasil simulasi menunjukkan pola yang konsisten dengan kajian sebelumnya: SAE–FHRF lebih unggul pada hubungan nonlinear kompleks, sedangkan SAE–FHRoF lebih stabil pada hubungan linear dan semi-linear. Selain itu, pendekatan koreksi (C) secara konsisten menghasilkan kinerja yang lebih baik dibandingkan pendekatan penggantian (R).

Dalam evaluasi secara empiris, indikator perkawinan anak (SDG,s Tujuan 5), kedua model berbasis pembelajaran rakitan berhasil menurunkan RSE secara nyata dibandingkan penduga langsung dan FH–EM, terutama untuk kabupaten/kota dengan ukuran contoh kecil dan variabilitas tinggi. Temuan ini menunjukkan bahwa integrasi pembelajaran mesin dalam SAE tingkat area tetap efektif meskipun peubah respons memiliki keterbatasan sebaran.

Secara keseluruhan, disertasi ini menunjukkan bahwa integrasi hutan acak dan hutan rotasi dalam kerangka Fay–Herriot mampu meningkatkan fleksibilitas, akurasi, dan stabilitas pendugaan pada berbagai karakteristik data. Model SAE–FHRF dan SAE–FHRoF tidak dimaksudkan untuk menggantikan Fay–Herriot klasik, melainkan sebagai pendekatan komplementer berbasis diagnostik. FH–EM tetap optimal ketika asumsi linearitas terpenuhi, sedangkan pendekatan berbasis pembelajaran rakitan menjadi alternatif yang lebih efektif pada kondisi nonlinear. Dengan demikian, penelitian ini memberikan kontribusi metodologis bagi pengembangan SAE modern sekaligus memperluas pemanfaatan pembelajaran mesin dalam statistik resmi.

Kata kunci: Fay–Herriot, pembelajaran rakitan, pembelajaran mesin, pendugaan area kecil, hutan acak, hutan rotasi, SDG’s.

SUMMARY

ARI SHOBRI BUKHARI. Enhancing Fay–Herriot Model Performance through Ensemble Learning for Small Area Estimation in the Presence of Nonlinearity and Multicollinearity. Supervised by KHAIRIL ANWAR NOTODIPUTRO, INDAHWATI, ANWAR FITRIANTO, and ANANG KURNIA.

Small area estimation (SAE) is widely used to improve the precision of estimates for domains with small sample sizes. In official statistics, the Fay–Herriot (FH) model is the most commonly applied area-level SAE approach because it combines direct survey information with auxiliary variables within a linear mixed model framework. However, the performance of the FH model strongly depends on the linearity assumption, whereas socioeconomic data often exhibit nonlinear relationships, heterogeneity, and multicollinearity among auxiliary variables. These conditions reduce the flexibility of the standard FH model and may lead to lower estimation precision.

Recent developments in machine learning provide opportunities for more flexible SAE methods because they can capture nonlinear relationships without requiring explicit functional form specification. Nevertheless, the integration of machine learning into area-level SAE remains limited, particularly approaches that preserve the inferential structure of the Fay–Herriot model. Moreover, the development of models capable of simultaneously handling nonlinearity, multicollinearity, and bounded response variables common in SDG indicators is still limited in the literature.

This dissertation develops ensemble learning-based Fay–Herriot models through the integration of Random Forest and Rotation Forest algorithms into the area-level SAE framework. The proposed models consist of Fay–Herriot with Random Forest (FHRF) and Fay–Herriot with Rotation Forest (FHRoF), each implemented using correction (C) and replacement (R) approaches. The correction approach preserves the basic FH structure and models the residual component using machine learning, whereas the replacement approach substitutes the fixed-effects component with ensemble learning models. These developments aim to improve model flexibility in handling nonlinearity, high multicollinearity, and proportion-type response variables through additive log-ratio (alr) transformation.

The first study focuses on the development of SAE–FHRF to address violations of the linearity assumption in the standard FH model. Evaluation was conducted through simulation and empirical studies by considering relationship forms, levels of multicollinearity, and between-area variances. Simulation results show that SAE–FHRF consistently outperformed FH–EM under nonlinear conditions through reductions in relative root mean squared error (RRMSE) and relative bias (RB). The superiority increased as the number of areas increased. However, under linear conditions, FH–EM remained the best model because it is optimal within the linear mixed model framework.

An empirical application using per capita expenditure data of agricultural households in Indonesia showed that the direct estimator had limitations in areas with small sample sizes and high relative standard errors (RSE). Diagnostic analyses also indicated nonlinearity, skewness, and multicollinearity. Under these conditions, SAE–FHRF, particularly FHRF-R, substantially reduced RSE

compared with both the direct estimator and FH–EM, especially in areas with high variability and small sample sizes.

The second study develops SAE–FHRoF to address high multicollinearity and semi-linear relationships through the integration of Rotation Forest based on Principal Component Analysis (PCA) into the Fay–Herriot framework. Simulation results indicate that SAE–FHRoF, particularly FHRoF-C, achieved the best performance under semi-linear conditions and improved model performance in the presence of high multicollinearity. Rotation Forest was able to preserve the linear relationship structure while improving estimation stability. In contrast, under nonlinear specification B, SAE–FHRF remained superior because Random Forest is more flexible in capturing interaction patterns and strong nonlinearities. Thus, FHRF is more adaptive for more complex nonlinear relationships, whereas FHRoF is more efficient for near-linear relationships or conditions with highly correlated auxiliary variables.

Empirical applications demonstrate that SAE–FHRoF improved estimation precision compared with the direct estimator and FH–EM, particularly in areas with high RSE. Compared with SAE–FHRF, both ensemble learning-based approaches showed relatively comparable statistical performance and complemented each other in the development of modern SAE.

The third study extends SAE–FHRF and SAE–FHRoF to proportional response variables using log-ratio transformation to ensure that estimates remain within the unit interval. Parameter estimation is conducted within a compositional mixed-model framework to accommodate transformation-induced bias. Simulation results are consistent with those of the previous studies: SAE–FHRF performs better under nonlinear specification B, whereas SAE–FHRoF provides more stable results under linear and semi-linear conditions. In addition, the correction approach consistently outperforms the replacement approach.

In the empirical application to estimating SDG 5 (Gender Equality) child marriage indicators, both ensemble learning-based models substantially reduce RSE compared with direct estimation and FH–EM, particularly in districts and municipalities with small sample sizes and high variability. These findings demonstrate that integrating machine learning into area-level SAE remains effective even when response variables are subject to distributional constraints.

Overall, this dissertation demonstrates that integrating Random Forest and Rotation Forest into the Fay–Herriot framework enhances the flexibility, accuracy, and stability of small area estimates across a wide range of data characteristics. The SAE–FHRF and SAE–FHRoF models are not intended to replace the classical Fay–Herriot model but rather to serve as diagnostic-based complementary approaches. FH–EM remains optimal when the linearity assumption holds, whereas ensemble learning-based approaches provide more effective alternatives under nonlinear conditions. Therefore, this research contributes to the methodological advancement of modern SAE and broadens the application of machine learning in official statistics.

Keywords: ensemble learning, Fay–Herriot, machine learning, random forest, rotation forest, SDG's, small area estimation.

ENHANCING FAY–HERRIOT MODEL PERFORMANCE THROUGH ENSEMBLE LEARNING FOR SMALL AREA ESTIMATION IN THE PRESENCE OF NONLINEARITY AND MULTICOLLINEARITY

ARI SHOBRI BUKHARI

Disertation
as a partial fulfillment of the requirements for a doctoral degree
in Statistics and Data Science

STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
IPB UNIVERSITY
BOGOR
2026

External Examiners on Closed Examination:

1 Dr. Kusman Sadik, S.Si., M.Si.

2 Dr. Bagus Sartono, S.Si., M.Si.

External Promotors on Doctoral Promotion:

1 Dr. Bagus Sartono, S.Si., M.Si.

2 Dr. Erni Tri Astuti, M.Math.

Dissertation Title: Enhancing Fay–Herriot Model Performance through Ensemble Learning for Small Area Estimation in the Presence of Nonlinearity and Multicollinearity

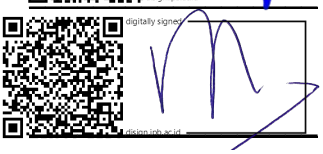
Name : Ari Shobri Bukhari

Student ID : G1601211007

Approved by

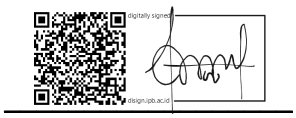
Supervisor:
Prof. Dr. Anang Kurnia, M.Si.

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



Co-Supervisor:
Dr. Ir. Indahwati, M.Si.

Co-Supervisor:
Dr. Anwar Fitrianto, M.Sc.



Acknowledge by

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

Dean of School of Data Science, Mathematics and Informatics:
Prof. Dr. Ir. Agus Buono, M.Si., M.Kom.
NIP. 197807232007011001

Date of Closed Examination:
July 22, 2026

Date of Graduation::

ACKNOWLEDGMENT

Alhamdulillahi Rabbil ‘Alamin. All praise and gratitude are due to Allah Subhanahu wa Ta’ala for His countless blessings, mercy, and guidance, which have enabled the completion of this dissertation. The dissertation is entitled “Enhancing Fay–Herriot Model Performance through Ensemble Learning for Small Area Estimation in the Presence of Nonlinearity and Multicollinearity.”

Sincere appreciation is extended to the supervisory committee, Prof. Dr. Ir. Khairil Anwar Notodiputro, MS, Dr. Ir. Indahwati, M.Si, Dr. Anwar Fitrianto, M.Sc, and Prof. Dr. Anang Kurnia, M.Si, for their invaluable guidance, insightful suggestions, continuous encouragement, and unwavering support throughout this research. Gratitude is also extended to the external examiners of the closed and open doctoral examinations, namely Prof. Dr. Anang Kurnia, M.Si, Dr. Bagus Sartono, S.Si., M.Si., and Dr. Erni Tri Astuti, M.Math., for their constructive comments and valuable suggestions that contributed significantly to the improvement of this dissertation. Appreciation is likewise expressed to all teachers and lecturers who have played important roles in the author’s academic and personal development.

Profound gratitude is extended to BPS–Statistics Indonesia for providing financial support, research opportunities, and access to the data used in this study. Special appreciation is addressed to the Head of BPS, Amalia Adininggar Widyasanti, S.T., M.Si., M.Eng., Ph.D, the Deputy Head of BPS, Dr. Sonny Harry Budiutomo Harmadi S.E., M.E, and all BPS leaders for their support and the opportunity to pursue doctoral studies. Sincere thanks are also conveyed to the Education and Training Center of BPS (Pusdiklat BPS), particularly Dr. Ardi Adji, S.Si., M.Si., Yuli Cahya Ningrum, S.Sos., M.A., and their colleagues, for their assistance throughout the study leave period. The author also gratefully acknowledges the Directorate General of Higher Education, Research, and Technology, Ministry of Education, Culture, Research, and Technology of Indonesia, for supporting part of this research through the Doctoral Dissertation Research Grant (*Penelitian Disertasi Doktor – PDD*).

Appreciation is also extended to the Graduate School of IPB University and its staff members for fostering a supportive and productive academic environment. Special thanks are conveyed to colleagues and fellow doctoral students at the School of Data Science, Mathematics, and Informatics, IPB University, particularly the 2021 doctoral cohort and other cohorts, including Reza, Ade, Bayu, Yunus, Dani, Asri, Herlin, Dwi, and many others, for their friendship, collaboration, encouragement, and support throughout this academic journey.

The author’s deepest gratitude is dedicated to his beloved family. Special remembrance is devoted to his late father, Uu Yusup Buhori, whose greatest wish was to witness his son earn a doctoral degree but who passed away during this educational journey. His love, sacrifices, prayers, and aspirations remain a constant source of strength and inspiration. Heartfelt gratitude is also extended to his mother, Nanah Hamidah, his beloved wife, Fera Hermawati, his children, Hana, Arkan, and Fathiya, as well as his siblings, Rita Juita and the late Ade Mirah, for their endless love, prayers, patience, encouragement, and unwavering support throughout the completion of this doctoral study.

It is hoped that this dissertation will be beneficial to researchers, practitioners, and other interested parties while contributing to the advancement of scientific knowledge. Although this work has certain limitations, it is hoped that it may serve as a foundation for future research and development in related fields. Constructive feedback and suggestions for further improvement are always welcome.

Bogor, August 2026

Ari Shobri Bukhari

TABLE OF CONTENTS