PERFORMANCE OF THE XGBOOST METHOD IN MEASURING UNCERTAINTY OF SMALL AREA ESTIMATION

FIDA FARIHA AMATULLAH

STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
BOGOR AGRICULTURAL INSTITUTE
BOGOR
2026

STATEMENT REGARDING THE THESIS, INFORMATION SOURCES, AND COPYRIGHT ASSIGNMENT

I, Fida Fariha Amatullah, with this declare that the thesis entitled "Performance of The XGBoost Method in Measuring Uncertainty of Small Area Estimation" is my original work with the direction of the supervisors and has not been submitted in any form to any other higher education institution. The information sources derived or quoted from the published or unpublished works of the other authors have been mentioned in the text and listed in the references at the end of this thesis.

In compiling this work, the author used the help of artificial intelligence Claude, ChatGPT, SciSpace, Perplexity, DeepL, and Grammarly to help the author analyze errors in the code, review journals, and improve writing in English. After using these tools/services, the author reviews and edits the content as needed and is fully responsible for the content of this final project.

I assign the copyright of my writing to IPB University.

Bogor, August 2026

Fida Fariha Amatullah
G1501231064

RINGKASAN

FIDA FARIHA AMATULLAH. Evaluasi Kinerja Metode XGBoost untuk Mengukur Ketidakpastian dalam Pendugaan Area Kecil. Dibimbing oleh ANWAR FITRIANTO dan KHAIRIL ANWAR NOTODIPUTRO.

Survei sampel merupakan salah satu sumber utama dalam menghasilkan indikator sosial ekonomi, termasuk rata-rata pengeluaran per kapita. Namun, pada tingkat area kecil seperti kecamatan, ukuran sampel yang terbatas sering menyebabkan pendugaan langsung memiliki ragam yang besar dan tingkat ketelitian yang rendah. *Small Area Estimation* (SAE) menjadi salah satu solusi untuk meningkatkan presisi pendugaan dengan memanfaatkan informasi tambahan dari peubah penyerta. Metode Fay–Herriot (FH) merupakan pendekatan SAE yang paling banyak digunakan, tetapi memiliki keterbatasan karena mengandalkan asumsi linearitas, normalitas, dan independensi komponen model. Di sisi lain, perkembangan *machine learning* menghadirkan metode *Extreme Gradient Boosting* (XGBoost) yang mampu memodelkan hubungan *non-linear* dan interaksi kompleks tanpa memerlukan asumsi parametrik yang ketat. Penelitian ini bertujuan mengevaluasi kinerja metode XGBoost dalam menduga rata-rata pengeluaran per kapita tingkat kecamatan di Provinsi Jambi, serta mengevaluasi ketidakpastian pendugaan yang dihasilkan.

Data yang digunakan berasal dari Survei Sosial Ekonomi Nasional (SUSENAS) dan Potensi Desa (PODES) Provinsi Jambi tahun 2021. Sebelum pemodelan, dilakukan pembersihan data pengeluaran rumah tangga dengan menghilangkan komponen pengeluaran yang bersifat tidak rutin, kemudian dilakukan penanganan pencilan menggunakan metode *winsorization* pada persentil ke-5 dan ke-95. Hasil eksplorasi menunjukkan bahwa pendugaan langsung memiliki distribusi yang menjulur dan ukuran sampel antar kecamatan yang tidak merata, sehingga mendukung penggunaan pendekatan SAE. Pada model EBLUP-FH dilakukan transformasi logaritma untuk memperbaiki distribusi peubah respons dan memenuhi asumsi model. Hasil pengujian menunjukkan bahwa setelah transformasi, komponen efek acak area dan galat *sampling* menjadi lebih mendekati normal dengan nilai *p-value* masing-masing sebesar 0,5263 dan 0,05009. Namun demikian, asumsi independensi antar komponen belum terpenuhi karena ditemukan korelasi yang cukup tinggi sebesar 0,8457.

Hasil evaluasi menunjukkan bahwa model EBLUP-FH menghasilkan *Mean Square Error* (MSE) sebesar $1{,}81\times10^{9}$ dan meningkat menjadi $2{,}14\times10^{9}$ setelah *bootstrap*, dengan nilai *Mean Absolute Percentage Error* (MAPE) masing-masing sebesar 1,89% and 2,93%. Nilai MAPE yang sangat rendah menunjukkan bahwa hasil pendugaan EBLUP-FH memiliki tingkat kedekatan yang tinggi terhadap pendugaan langsung. Selain itu, nilai *shrinkage factor* yang sebagian besar berada pada kisaran 0,80–0,95 menunjukkan bahwa EBLUP-FH secara aktif memanfaatkan mekanisme *borrowing strength* untuk meningkatkan stabilitas pendugaan pada kecamatan dengan ukuran sampel kecil. Sementara itu, model XGBoost menunjukkan kemampuan yang baik dalam menangkap pola hubungan yang kompleks antar peubah, namun menghasilkan tingkat ketidakpastian yang lebih besar dibandingkan EBLUP-FH berdasarkan hasil *bootstrap*.

Perbandingan *standard error* menunjukkan bahwa metode EBLUP-FH secara umum menghasilkan pendugaan yang lebih stabil dibandingkan pendugaan langsung. XGBoost mampu menghasilkan estimasi yang kompetitif dan lebih fleksibel terhadap struktur data yang kompleks, namun variasi hasil *bootstrap* menunjukkan bahwa ketidakpastian model masih relatif lebih tinggi. Temuan ini menunjukkan bahwa pada data pengeluaran per kapita tingkat kecamatan di Provinsi Jambi, EBLUP-FH masih menjadi pendekatan yang sangat efektif karena mampu memanfaatkan informasi tambahan melalui *borrowing strength* sehingga menghasilkan estimasi yang presisi dan stabil. Namun demikian, pelanggaran asumsi independensi komponen model mengindikasikan bahwa pendekatan *machine learning* seperti XGBoost tetap memiliki potensi sebagai alternatif atau pelengkap model SAE parametrik pada data sosial ekonomi yang kompleks.

Implikasi dari penelitian ini adalah bahwa penggunaan SAE dapat meningkatkan kualitas indikator pengeluaran per kapita pada wilayah dengan ukuran sampel kecil sehingga mendukung penyusunan kebijakan yang lebih tepat sasaran. Hasil penelitian juga menunjukkan bahwa integrasi metode statistik klasik dan *machine learning* membuka peluang pengembangan SAE yang lebih fleksibel untuk menghadapi karakteristik data sosial ekonomi *modern* yang semakin kompleks dan heterogen.

Kata kunci: *Bootstrap*, EBLUP-FH, SAE, *Winsorization*, XGBoost

SUMMARY

FIDA FARIHA AMATULLAH. Performance of The XGBoost Method in Measuring Uncertainty of Small Area Estimation. Advised by ANWAR FITRIANTO and KHAIRIL ANWAR NOTODIPUTRO.

Sample surveys are one of the primary sources for producing socioeconomic indicators, including average per capita expenditure. However, at small-area levels such as sub-districts, limited sample sizes often lead to direct estimates with large variances and low precision. Small Area Estimation (SAE) provides an effective solution by incorporating auxiliary information to improve estimation accuracy. Among SAE approaches, the Fay–Herriot (FH) model is the most widely used; nevertheless, it relies on assumptions of linearity, normality, and independence of model components. In contrast, recent advances in machine learning have introduced methods such as Extreme Gradient Boosting (XGBoost), which can capture non-linear relationships and complex interactions without requiring strict parametric assumptions. This study aims to evaluate the performance of XGBoost in estimating average per capita expenditure at the sub-district level in Jambi Province and to assess the associated uncertainty.

The study used data from the 2021 National Socioeconomic Survey (SUSENAS) and the 2021 Village Potential Statistics (PODES) of Jambi Province. Prior to modeling, household expenditure data were cleaned by removing non-recurring expenditure components, followed by outlier treatment using Winsorization at the 5th and 95th percentiles. Exploratory analysis revealed a skewed distribution of direct estimates and substantial variation in sample sizes across sub-districts, supporting the application of SAE methods. For the EBLUP-FH model, a logarithmic transformation was applied to the response variable to improve its distributional characteristics and satisfy model assumptions. The results showed that, after transformation, both the area random effect and the sampling error became closer to normality, with p-values of 0.5263 and 0.05009, respectively. However, the independence assumption was not fully satisfied, as a relatively strong correlation of 0.8457 was observed between the two model components.

The evaluation results indicate that the EBLUP-FH model achieved an MSE of 1.81×10^9, which increased to 2.14×10^9 after bootstrap resampling, while the corresponding MAPE values were 1.89% and 2.93%, respectively. These low MAPE values indicate a high degree of agreement between the EBLUP-FH estimates and the direct estimates. Furthermore, most shrinkage factor values ranged between 0.80 and 0.95, indicating that EBLUP-FH effectively employed the borrowing-strength mechanism to improve estimate stability in sub-districts with small sample sizes. Meanwhile, XGBoost demonstrated strong capability in capturing complex relationships among variables, although its bootstrap-based uncertainty measures were generally larger than those of EBLUP-FH.

A comparison of standard errors showed that EBLUP-FH generally produced more stable estimates than direct estimation. XGBoost provided competitive results and greater flexibility in handling complex data structures; however, its bootstrap results indicated relatively higher uncertainty. These findings suggest that, for estimating per capita expenditure at the sub-district level in Jambi Province, EBLUP-FH remains a highly effective approach, as it effectively utilizes

auxiliary information through the borrowing-strength mechanism, resulting in precise and stable estimates. Nevertheless, the violation of the independence assumption indicates that machine learning approaches such as XGBoost remain promising alternatives or complementary methods for SAE applications involving complex socioeconomic data.

The findings of this study highlight the importance of SAE methods in improving the quality of per capita expenditure indicators for areas with limited sample sizes, thereby supporting more targeted, evidence-based policymaking. Moreover, the results demonstrate that integrating classical statistical methods with machine learning approaches offers significant potential for developing more flexible SAE frameworks capable of addressing the increasing complexity and heterogeneity of modern socioeconomic data.

Keywords: Bootstrap, EBLUP-FH, SAE, Winsorization, XGBoost

PERFORMANCE OF THE XGBOOST METHOD IN MEASURING UNCERTAINTY OF SMALL AREA ESTIMATION

FIDA FARIHA AMATULLAH

Thesis
as a partial fulfilment for the requirements of
Masters degree
in Statistics and Data Science

STATISTICS AND DATA SCIENCE STUDY PROGRAM
SCHOOL OF DATA SCIENCE, MATHEMATICS, AND INFORMATICS
BOGOR AGRICULTURAL INSTITUTE
BOGOR
2026

External Committee Thesis Examiner:
Dr. Ir. Indahwati, M.Si.

Thesis Title : Performance of The XGBoost Method in Measuring Uncertainty of Small Area Estimation
Name : Fida Fariha Amatullah
Student ID : G1501231064

Approved by

Main Supervisor:
Dr. Anwar Fitrianto, M.Sc.

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

Known by

Head of the Study Program:
Dr. Agus Mohamad Soleh, S. Si., M.T.
NIP 197503151999031004

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

Thesis Examination Date: 8th July 2026

Date of Completion:

ACKNOWLEDGMENTS

The author would like to praise and thank Allah SWT for giving the author the strength and fluency to complete this thesis entitled "Performance of The XGBoost Method in Measuring Uncertainty of Small Area Estimation".

In completing this thesis, the author certainly experienced many difficulties, but with the guidance and encouragement of enthusiasm and motivation from everyone, finally this thesis could be completed even though it was far from perfect. Therefore, the author expresses his deepest gratitude to:

1. Dr. Anwar Fitrianto, S.Si., M.Sc. and Prof. Dr. Ir. Khairil Anwar Notodiputro, M.S. as the supervisors, for their guidance, feedback, and input for this thesis.
2. Both parents and siblings who always support the author through prayers and motivation that have been given.
3. Rizki Ananda who helped give instructions on the process of processing thesis data and explanation of small area estimation and machine learning.
4. Suci, Nadia, Tiara, Siptya, Nadine, Silvy, and author's friends from high school, who helped by listening to author's stories when author was struggling.
5. Erda, Riska, Nabila, Annisa, and other friends who have consistently provided motivation and support to the author to keep working together to complete this thesis.
6. All lecturers and educators of the IPB Statistics and Data Science Study Program who have assisted in the process of completing this scientific work.
7. The author, for having fought and worked hard to overcome all obstacles with sincerity and for never ceasing to pray.

Hopefully, this scientific work is helpful for those who need it and for the advancement of science.

Bogor, August 2026

Fida Fariha Amatullah

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES