

Clustering Regencies/Cities Vulnerable to Air Pollution in the Java Island: Fuzzy Geographically Weighted Clustering

ARYA CANDRA KUSUMA, ARIE WAHYU WIJAYANTO, ARISTA, VICKA KHARISMA BAHAR, TIFANI HUSNA SIREGAR

Adding MSNBURR-IIa Distribution to MultiBUGS

ELIANA PUTRI RAMADANI, ACHMAD SYAHRUL CHOIR, ANINDYA APRILIYANTI PRAVITASARI, JOYNABEL PARAGUISON

Implementation of Twofold HB Beta SAE Model to Estimate Out-of-School Children with Disabilities in Indonesia

AISHA MAHARANI, AZKA UBAIDILLAH, ADHI KURNIAWAN

Spatial Heterogeneity of Food Security in Indonesia: Unpacking the Roles of Technology and Democracy Index

DITTO SATRIO WICAKSONO, NOVI HIDAYAT PUSPONEGORO, ARBI SETIYAWAN

Forecasting Farmer Exchange Rate (FER) in Southeast Sulawesi Province Using Cheng's Fuzzy Time Series Method

RASTINA, LILIS LAOME, BAHRIDDIN ABAPIHI, GUSTI NGURAH ADHI WIBAWA, MUKHSAR, MAKKULAU, GAMA PUTRA DANU SOHIBIEN, SUKIM, FATHURRAHMAN YAHYASATRIO

Satellite Imagery for Classification Analysis of Abrasion Areas on Panaitan, Banten

GHAFFAR ISMAIL, ROBERT KURNIAWAN, SILVIA NI'MATUL MAULA

Estimating Economic Activity Using Geospatial Big Data in East Java, Indonesia: Relative Spatial GDP Index Approach

RIFQI RAMADHAN, I MADE SATRIA AMBARA PUTRA, TAUFIQ AGUNG KURNIAWAN, FITRI KARTIASIH, RADEN MUAZ MUNIM, SOMETHEA BUOY

Integrating Multi-Criteria Decision Analysis and Machine Learning for Fine-Scale Mapping of Safe Drinking Water Access in Province of Bengkulu

ANDREW MARULI TUA TAMPUBOLON, BONY PARULIAN JOSAPHAT, ASRIADI SAKKA, YOHANES WAHYU TRIO PRAMONO



Jurnal Aplikasi Statistika & Komputasi Statistik (JASKS) contains scientific papers on research findings and theoretical studies of statistics and computational statistics applied in fields, that is published twice a year in June and December. This journal is published by Politeknik Statistika STIS.

Editor-in-Chief:	Setia Pramana
Managing Editor:	Fitri Kartiasih
Editor:	Hardius Usman
	Rani Nooraeni
	Arie Wahyu Wijayanto
	Lutfi Rahmatuti Maghfiroh
	Firman M. Firmansyah
	Muhammad Rausyan Fikri
	Dani Prasetyawan
	Ika Yuni Wulansari
	Ahmad R. Pratama
Copyeditor:	Christiana Anggraeni Putri
IT:	Salwa Rizqina Putri
	Alif Wira Bayu
Administrasi:	Ary Wahyuni

Editorial Address:
Politeknik Statistika STIS
Jl. Otto Iskandardinata 64C
Jakarta Timur 13330
Telp. 021-8191437

The editorial accepts scientific papers or research articles on theoretical studies of statistics and computational statistics in fields. The editorial has the right to edit writings without changing the substance of the writing. The contents of the Jurnal Aplikasi Statistika & Komputasi Statistik are cited by referring to the source material.

Editorial Foreword

Jurnal Aplikasi Statistika & Komputasi Statistik (JASKS) Volume 17 Number 2 December 2025 Edition has undergone transformations such as writing articles in English, establishing a journal logo, establishing the Politeknik Statistika STIS publisher logo, changing paper template designs, updating Author Guidelines, etc. The aim of this transformation is to improve the journal's performance and expand the reach of JASKS readers.

This issue consists of 8 articles contributed by 37 authors affiliated with various institutions from Indonesia and abroad. The contributing institutions include Politeknik Statistika STIS, BPS-Statistics from various regional offices, Universitas Halu Oleo, Universitas Padjajaran, Indiana University Bloomington (USA), Asian Development Bank Institute (Japan), King Fahd University of Petroleum and Minerals (Saudi Arabia), Philippine Statistics Authority (Philippine), King's College London (UK), Jiangsu University (China), Tsinghua University (China), Hamad Bin Khalifa University (Qatar), National Institute of Statistics (Cambodia), Korea National University of Science and Technology (Korea), Victoria University of Wellington (New Zealand). The articles in this issue demonstrate diverse applications of statistical and computational methods to address environmental, economic, and social challenges in Indonesia. The studies encompass spatial and geospatial approaches, Bayesian and small area estimation methods, fuzzy time series forecasting, and satellite-based classification, as well as the integration of multi-criteria decision analysis and machine learning for fine-scale mapping and regional analysis. Collectively, these contributions highlight the practical relevance of advanced statistical techniques for evidence-based policy and regional planning.

The editorial board extends its sincere appreciation to all authors and reviewers for their valuable contributions and rigorous evaluations that have ensured the quality of this issue. We also thank the editorial team for their dedication in managing the publication process. It is hoped that this issue will enhance readers' understanding of applied statistical and computational methods, and the editorial team welcomes further scholarly contributions to support the dissemination of statistical knowledge within the community.

Jakarta, December 2025

Editor-in-Chief,

Setia Pramana

Contents

Editorial Foreword	iii
Contents	iv
Clustering Regencies/Cities Vulnerable to Air Pollution in the Java Island: Fuzzy Geographically Weighted Clustering Arya Candra Kusuma, Arie Wahyu Wijayanto, Arista, Vicka Kharisma Bahar Tifani Husna Siregar.....	94-108
Adding MSNBURR-IIa Distribution to MultiBUGS Eliana Putri Ramadani, Achmad Syahrul Choir, Anindya Apriliyanti Pravitasari, Joynabel Paraguison	109-130
Implementation of Twofold HB Beta SAE Model to Estimate Out-of-School Children with Disabilities in Indonesia Aisha Maharani, Azka Ubaidillah, Adhi Kurniawan	131-145
Spatial Heterogeneity of Food Security in Indonesia: Unpacking the Roles of Technology and Democracy Index Ditto SatrioWicaksono, Novi Hidayat Pusponegoro, Arbi Setiyawan.....	146-160
Forecasting Farmer Exchange Rate (FER) in Southeast Sulawesi Province Using Cheng's Fuzzy Time Series Method Rastina, Lilis Laome, Bahriddin Abapihi, Gusti Ngurah Adhi Wibawa, Mukhsar, Makkulau, Gama Putra Danu Sohibien, Sukim, Fathurrahman Yahyasatrio.....	161-176
Satellite Imagery for Classification Analysis of Abrasion Areas on Panaitan, Banten Ghaffar Ismail, Robert Kurniawan, Silvia Ni'matul Maula.....	177-188
Estimating Economic Activity Using Geospatial Big Data in East Java, Indonesia: Relative Spatial GDP Index Approach Rifqi Ramadhan, I Made Satria Ambara Putra, Taufiq Agung Kurniawan, Fitri Kartiasih, Raden Muaz Munim, Somethea Buoy.....	189-210
Integrating Multi-Criteria Decision Analysis and Machine Learning for Fine-Scale Mapping of Safe Drinking Water Access in Province of Bengkulu Andrew Maruli Tua Tampubolon, Bony Parulian Josaphat, Asriadi Sakka, Yohanes Wahyu Trio Pramono.....	211-223



Clustering Regencies/Cities Vulnerable to Air Pollution in the Java Island: Fuzzy Geographically Weighted Clustering

Arya Candra Kusuma^{1*}, Arie Wahyu Wijayanto², Arista³, Vicka Kharisma Bahar⁴, Tifani Husna Siregar⁵

¹BPS-Statistics Indonesia, Jakarta, Indonesia, ²Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia, ³Indiana University Bloomington, Bloomington, Indiana, USA, ⁴Asian Development Bank Institute, Japan, ⁵King Fahd University of Petroleum and Minerals, Saudi Arabia

*Corresponding Author: E-mail address: arya.candra@bps.go.id

ARTICLE INFO

Article history:

Received 4 August, 2024

Revised 2 April, 2025

Accepted 12 July, 2025

Published 31 December, 2025

Keywords:

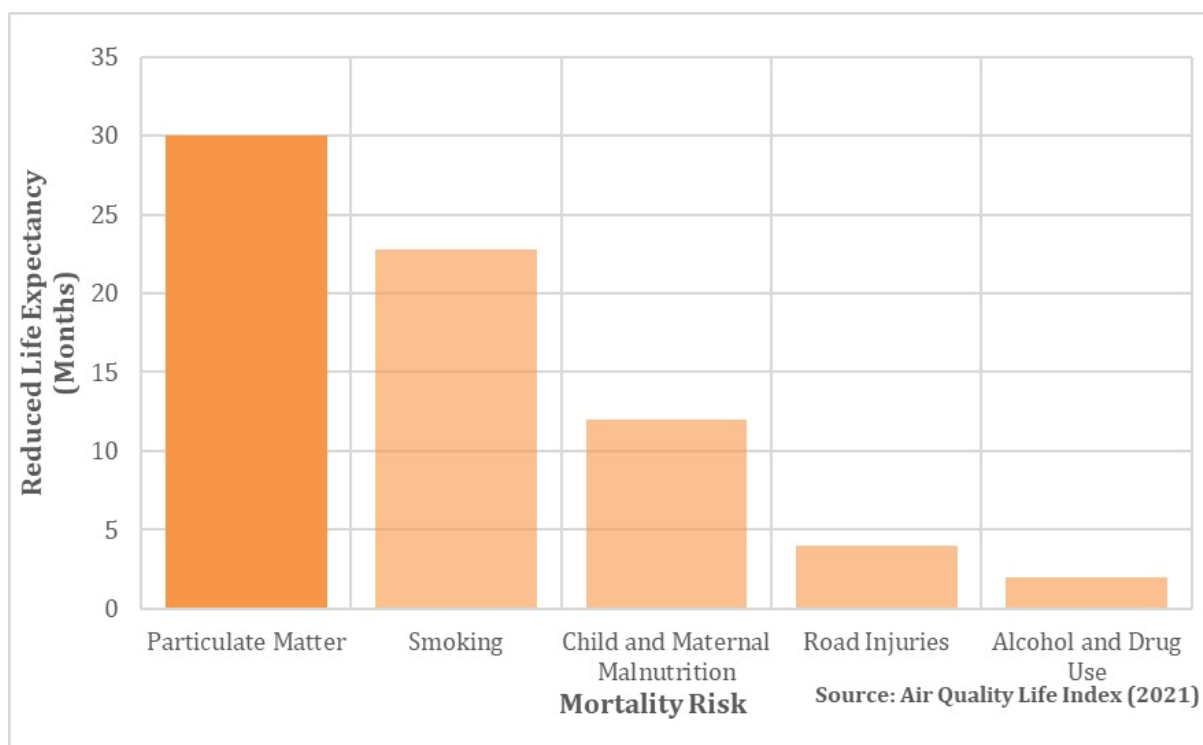
Air Pollution; Artificial Bee Colony (ABC); Environment; Fuzzy C Means (FCM); Fuzzy Geographically Weighted Clustering (FGWC)

Abstract

Introduction/Main Objectives: Air pollution has become a critical global concern with substantial effects on human health and the environment. **Background Problems:** Java Island in Indonesia, recognized for its high population density and industrial activities, necessitates focused effort in resolving this issue. **Novelty:** While air pollution research has been enormous, there has been no effort to cluster regencies or cities on Java Island utilizing spatially-based data. This research seeks to cluster regencies and cities on Java Island according to air pollution levels and to compare geodemographic and non-geodemographic clustering methodologies. **Research Methods:** This study employs secondary data regarding air pollution, obtained from the Openweather API. This study employs a geodemographic clustering technique, namely fuzzy geographically weighted clustering (FGWC), optimized by the artificial bee colony (ABC) algorithm. **Finding/Results:** The study findings indicate that the geodemographic clustering method ABCFGWC surpasses Fuzzy C-Means (FCM) according to the TSS (Tang-Sun-Sun) index. The data reveal that the Greater Jakarta or Jabodetabek area and its adjacent territories are more susceptible to air pollution. The findings of this study are expected to enhance the spatial planning and mapping of air pollution management strategies on Java Island.

1. Introduction

Air pollution has emerged as a global concern due to its harmful impacts on human health and the environment [1], [2], [3]. The deterioration of global air quality is attributable to anthropogenic activities that release pollutants into the atmosphere, including car emissions, industrial emissions, and fossil fuel combustion [2], [4]. The pollutants encompass particulate matter (PM_{2.5} and PM₁₀), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), carbon monoxide (CO), and several other organic chemicals. The World Health Organization (WHO) reports that 91% of the global population lives in regions where air pollution is above the established air quality guidelines [2]. In Indonesia, the average air pollution level exceeds the WHO recommended limit by three times [5]. Consequently, analyzing air pollution is essential for developing effective strategies to alleviate its harmful effects.



Source: Lee & Greenstone [5]

Figure 1. The impact of reduced life expectancy caused by mortality risk in Indonesia

Numerous scientific research papers [6], [7], [8] have provided evidence of the detrimental long-term and short-term health effects of air pollution. From Figure 1, it is evident that air pollution is one of the most significant risk factors contributing to reduced life expectancy, exceeding all other determinants [5]. PM_{2.5}, a category of fine particulate matter, is a highly harmful pollutant closely associated with respiratory and cardiovascular problems [8]. Vulnerable populations, including children, the elderly, and individuals with a history of respiratory or cardiovascular disease, are particularly susceptible to the health impacts of air pollution [9].

In addition to its effects on human health, air pollution also poses a significant threat to the ecosystem and the overall welfare of humanity [2], [4]. Sulfur dioxide (SO₂) and nitrogen oxides (NO₂) may react chemically to form acid rain, which has harmful consequences for the ozone layer and ultimately contributes to global warming [4]. In terms of human well-being, air pollution has a significant impact on both labor productivity and the cognitive capacities of students [10], [11]. Therefore, resolving air pollution problems will not only tackle health concerns but also address environmental concerns [2].

To alleviate the adverse effects of air pollution, several governments and organizations across different nations have enacted a series of measures, including the establishment of emission regulations, the adoption of renewable energy sources, and the promotion of sustainable transportation [2], [13], [14]. Furthermore, the European Union is implementing the Carbon Border Adjustment Mechanism (CBAM), a scheme intended to charge fees on carbon-intensive goods that will be fully enacted in 2026, following a transitional phase from 2023 to 2026 [12]. In Indonesia, the urgency to address air pollution at the local level is increasing as there are different levels of air pollution in each region [15]. Java Island is a significant area for the assessment of air pollution levels attributable to its concentrated industrial and economic activities.

Several previous studies have examined air pollution using different methodologies. A comprehensive analysis incorporating spatiotemporal assessment, clustering, and correlation techniques has been applied to air pollution data in northern China [16]. An enhanced variant of the K-means algorithm has been proposed to improve accuracy and computational efficiency in determining the air quality index (AQI) [17]. Cluster analysis combined with spatial correlation methods has also been employed to examine air pollution patterns in China [18]. In addition, the K-means clustering technique has been used to classify air pollution regions in Makassar [19]. Despite extensive study on air pollution, no studies have focused on clustering areas by air pollution at the regency or city level specifically for Java Island, including geographical characteristics. Moreover, previous studies mostly focused on clustering, neglecting the relationship among distinct regions. Clustering analysis, which incorporates

spatial characteristics, might serve as a methodological approach to formulate policies that acknowledge regional variability [20]. Geodemographic clustering enhances the identification of inter-regional linkages and supports spatially explicit analysis, hence assisting in policy formulation [20], [21]. This approach has been successfully implemented in several regions [22] as well as in other domains [23], [24], [25].

Thus, this study aims to cluster regencies/cities on Java Island based on air pollution conditions. In addition, this research also aims to compare clustering results between clustering methods that consider spatial aspects and clustering methods that do not consider spatial aspects. By clustering regencies/cities, it is possible to identify pollution levels and air pollution characteristics between regions. Utilizing this knowledge, regional development strategies can be designed to mitigate air pollution by taking into account the unique characteristics of each location. This research contributes to the field of air pollution analysis and management in Java.

2. Material and Methods

2.1. Air Pollution Data

The data utilized in this research was acquired using the Application Programming Interface (API) from the website <https://openweathermap.org/>. The study utilized regencies/cities on Java Island as observation units, specifically 119 regencies/cities. The study represents air pollution in each regency/city by using the centroid point of the regency/city area. Data collection occurred daily for one week, from June 15, 2023, to June 22, 2023, at 12:00 WIB. The data collection is specifically targeted at periods of high air pollution, which are typically associated with increased population activity. The acquired data is then averaged to determine the pollution level in the regency/city. The grouping of vulnerable regencies/cities is achieved by using air pollution measurement variables, as shown in Table 1.

Table 1. Research variables

Variable	Definition	Units
CO	CO (Carbon monoxide) concentration	$\frac{\mu\text{g}}{\text{m}^3}$
NO	NO (Nitrogen monoxide) concentration	$\frac{\mu\text{g}}{\text{m}^3}$
NO ₂	NO ₂ (Nitrogen dioxide) concentration	$\frac{\mu\text{g}}{\text{m}^3}$
O ₃	O ₃ (Ozone) Concentration	$\frac{\mu\text{g}}{\text{m}^3}$
SO ₂	SO ₂ (Sulfur dioxide) concentration	$\frac{\mu\text{g}}{\text{m}^3}$
PM _{2.5}	PM _{2.5} (Fine particulate matter) concentration	$\frac{\mu\text{g}}{\text{m}^3}$
PM ₁₀	PM ₁₀ (Coarse particulate matter) concentration	$\frac{\mu\text{g}}{\text{m}^3}$
NH ₃	NH ₃ (Ammonia) Concentration	$\frac{\mu\text{g}}{\text{m}^3}$

2.2. Fuzzy C Means (FCM)

FCM (Fuzzy C-Means) is one of the fuzzy clustering algorithms that assigns a membership level to each observation in each cluster, ranging from zero to one [23]. The membership level for each observation in the cluster is determined by optimizing an objective function. The most commonly used and studied objective function for FCM clustering is minimizing the least squares error, as in equation (1) [26].

$$J_m(\mathbf{U}, \mathbf{v}) = \sum_{i=1}^N \sum_{j=1}^c \mu_{ij}^m d_{ij}^2 \quad (1)$$

With: $\mathbf{Y}$ represents the set of all N observations, with each observation a vector in D -dimensional space, i.e. $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N\} \subset R^D$. c is the number of clusters in the set of all observations $\mathbf{Y}$; $2 \leq c < N - 1$, m is the fuzziness weighting exponent with a value of $1 \leq m < \infty$, μ_{ij} is the element in the i -th row and j -th column of the fuzzy membership matrix, indicating the degree of membership of observation i in cluster j , $\mathbf{v} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c)$ is the cluster centroid vector, $\mathbf{v}_j = (v_{j1}, v_{j2}, \dots, v_{jD})$ is the centroid for cluster j , d_{ij}^2 is the squared distance between observation i and cluster j . The fuzzy membership matrix $\mathbf{U} = [\mu_{ij}]$ is initialized using equation (2).

$$\mu_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{\|\mathbf{y}_i - \mathbf{v}_j\|}{\|\mathbf{y}_i - \mathbf{v}_k\|} \right)^{\frac{2}{m-1}}} \quad (2)$$

The cluster centroids, $\mathbf{v}_j$, are obtained using equation (3).

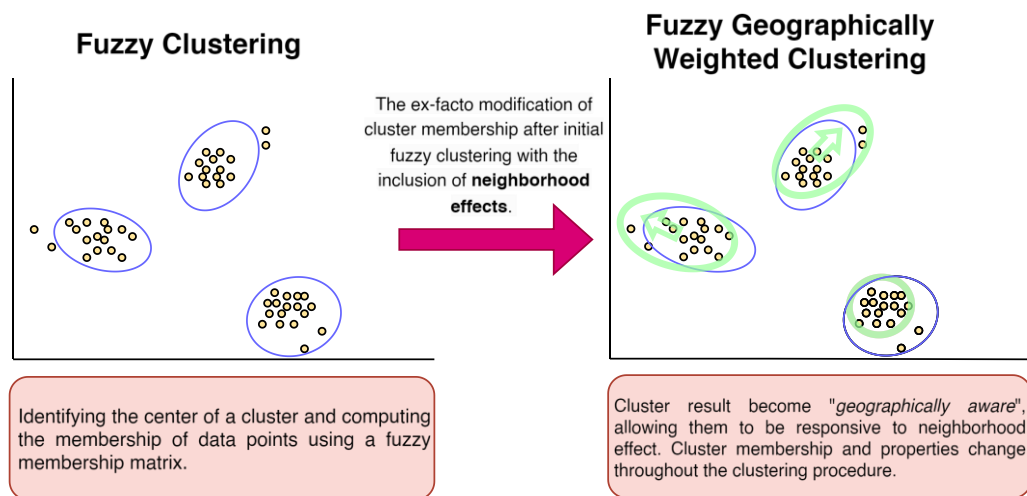
$$\mathbf{v}_j = \frac{\sum_{i=1}^N \mu_{ij}^m \mathbf{y}_i}{\sum_{i=1}^N \mu_{ij}^m} \quad (3)$$

The optimal form of fuzzy clustering for dataset $\mathbf{Y}$ is defined as the pair of values $(\mathbf{U}, \mathbf{v})$ that minimizes the objective function of equation (1). The optimal pair is obtained through an iteration process. The iteration process will stop when the condition in equation (4), which indicates the achievement of the local optimum [27]. With: ε is the stopping criterion for the iteration, ranging from 0 to 1, t is the iteration index.

$$\|\mu_{ij}^{t+1} - \mu_{ij}^t\| < \varepsilon \quad (4)$$

2.3. Fuzzy Geographically Weighted Clustering (FGWC)

The fuzzy c-means (FCM) algorithm was further developed through the integration of spatial features [28]. The resulting algorithm is called Fuzzy Geographically Weighted Clustering (FGWC), which is able to combine the effects of population and distance in cluster analysis in a geographical context so as to enrich the existing fuzzy C-means algorithm [29]. The addition of spatial elements is based on the understanding that cluster centers depend not only on variable values, but also on fuzzy values associated with cluster parameters [28]. Therefore, the spatial component is added by giving weights to the cluster membership formed so as to produce clusters that also consider spatial aspects [28]. FGWC has been applied to various studies that consider spatial aspects, particularly social-environmental studies [23], [24], [25], [29].



Source: Mason & Jacobson [28]

Figure 2. Illustrative comparison of FGWC and fuzzy clustering

Figure 2 shows an illustration of FGWC. In the process of cluster formation, FGWC adopts an approach that considers the influence of a region on other regions through population factors and the distance separating them [28]. The mathematical equation for FGWC can be described as follows. For a new cluster membership μ'_i of area i , it can be expressed as a linear combination of the previous cluster membership μ_i and the membership weight w_{ij} with respect to the weighting factors α and β . Mathematically, the FGWC formulation can be written as in equation (5) [28].

$$\mu'_i = \alpha\mu_i + \beta \frac{1}{A} \sum_{j=1}^n w_{ij}\mu_j \quad (5)$$

Where μ'_i is the new cluster membership for area i , μ_i is the previous cluster membership for area i , α and β are the scale parameters of the proportion between the previous cluster membership and the membership weight. The values of α and β must fulfill the requirement $\alpha + \beta = 1$, A is a scale factor calculated for all clusters such that the total membership of an area in each cluster remains 1. In addition, the regional weight w_{ij} in equation (5) can be defined as in equation (6). Where: m_i and m_j represent the population of areas i and j respectively. d_{ij} is the distance between areas i and j , a are parameters defined by the researcher.

$$w_{ij} = \frac{(m_i m_j)^b}{d_{ij}^a} \quad (6)$$

2.4. Artificial Bee Colony (ABC)

FGWC has the disadvantage of being prone to getting trapped in local optima [23], [25], [29]. Artificial Bee Colony (ABC) is one of the optimization approaches that can be used to minimize this limitation [30], [31]. ABC is an optimization algorithm inspired by the behavior of a group of bees in search of optimal food sources [32], [33]. The search for the optimal value in ABC is analogous to the search for the optimal food source by bees in the search space [32]. The ABC algorithm consists of three stages involving three types of bees, namely employed bees, onlooker bees, and scout bees. The stages of the ABC algorithm can be explained as follows [32].

First, employed bees travel back and forth between the nest and the food source location. Each employed bee is associated with a specific food source and remembers its location. They not only return to the hive with the food source, but also conduct local exploration to find better food sources.

Second, onlooker bees observe the behavior of employed bees returning from the food source search process. Onlooker bees are not associated with a particular food source location, and they probabilistically observe the behavior of employed bees to determine the best food source location. This selection probability can be calculated using equation (7) [34]. Where fit_i is the fitness value of the i -th solution that is directly proportional to the quality of the food source at the i -th location, and n_{emp} is the number of employed bees.

$$p_i = \frac{fit_i}{\sum_{q=1}^{n_{emp}} fit_q} \quad (7)$$

Third, scout bees observe the behavior of employed bees when returning to the nest. Scout bees are not associated with a specific food source location, and their job is to find new food sources. If scout bees observe that employed bees do not find progressively better food sources, they randomly search for new food source locations in the search space. In the ABC algorithm, the search for a new food source is expressed in equation (8) [34].

$$x_{ij}^* = x_{ij} + \phi_{ij}(x_{ij} - x_{lj}) \quad (8)$$

Where x_{ij}^* is a new candidate food source location calculated by comparing the previous food source x_{ij} with a random food source location, $l \in \{1, 2, \dots, n_{emp}\}$, and $j \in \{1, 2, \dots, D\}$. The notation ϕ_{ij} represents a number in the range $[-1, 1]$ that is randomly generated.

2.5. Analysis Flow

This study employs FGWC parameters $m = 2$, $\alpha = 0.7$, $\beta = 0.3$, $a = 1$, $b = 1$, and $\varepsilon = 10^{-5}$, which are the default settings in the *naspaclust* package [35]. Meanwhile, for the ABC algorithm, the parameters $n_{food} = 10$, $n_{onlk} = 5$, and $limit = 4$ were used. The analysis process was carried out using Rstudio 4.2.1 software by utilizing several statistical analysis packages and data visualization. The flow of analysis in this study can be described as follows.

1. First, the data was retrieved through the openweather API by determining the regency/city center point as the area to be retrieved from the API. The data is then preprocessed using Rstudio.
2. Next, cluster analysis was conducted by forming 2 to 6 clusters using both clustering methods. The range of the number of clusters formed up to 6 refers to the categorization of AQI, which also amounts to 6 categories [36]. The determination of the optimal number of clusters in both methods is done using the cluster validity index, the TSS (Tang-Sun-Sun) Index [37]. The TSS index is a development of the Xie-Beni Index and Kwon Index to overcome the problem of monotonically increasing or monotonically decreasing index values by imposing a penalty on the calculation of index validity, formulated in equation (9) [37].

$$I_{TSS} = \frac{1}{\min_{j \neq k} \|\mathbf{v}_j - \mathbf{v}_k\|^2 + \frac{1}{c}} \left(\sum_{j=1}^c \sum_{i=1}^n \mu_{ij}^2 \|\mathbf{x}_i - \mathbf{v}_j\|^2 + \frac{1}{c(c-1)} \sum_{j=1}^c \sum_{k=1, k \neq j}^c \|\mathbf{v}_j - \mathbf{v}_k\|^2 \right) \quad (9)$$

Where μ_{ij} is the degree of membership of the i -th observation to the j -th cluster, $\mathbf{x}_i$ is a vector indicating the value for the i -th observation, $\mathbf{v}_j$ and $\mathbf{v}_k$ are vectors indicating the centroid of the j -th and k -th clusters, n indicates the number of observations, and c is the number of clusters. The lower the I_{TSS} value, the better the cluster formed [37].

3. After the clusters are formed, profiling is done on the results of the clusters formed. Profiling is done by interpreting the cluster centroid results and showing biplots to support clustering.
4. The comparison between clustering methods is then carried out using the TSS index [37] and the Rand index [38]. The Rand index is a measure of the similarity of cluster results from two different clustering methods on the same data [38]. The Rand index is formulated in equation (10) [38].

$$c(Y, Y') = \frac{\binom{N}{2} - \frac{1}{2} \{ \sum_i (\sum_j n_{ij})^2 + \sum_j (\sum_i n_{ij})^2 \} - \sum_i \sum_j n_{ij}^2}{\binom{N}{2}} \quad (10)$$

Where: n_{ij} is the number of data grouped into the i -th cluster in the clustering method Y and grouped into the j -th cluster in the clustering method Y' . In addition, the notation N represents the number of clustered data. The Rand index value is between 0 and 1 with 1 denoting that the cluster results between the two methods are identical [38]. The better clustering method is then interpreted and discussed further.

3. Results and Discussion

This section explains the results and discusses cluster formation in the case of air pollution in Java. To compare the two cluster formation methods, the optimal number of clusters for each method is first determined. The determination of the optimal number of clusters is done using the TSS index. Clusters will be created using both methods, with the TSS index determining the optimal number of clusters. Next, the cluster results will be compared. After obtaining the best method, the better cluster method is interpreted, and the cluster results are discussed to analyze the case of air pollution on Java Island.

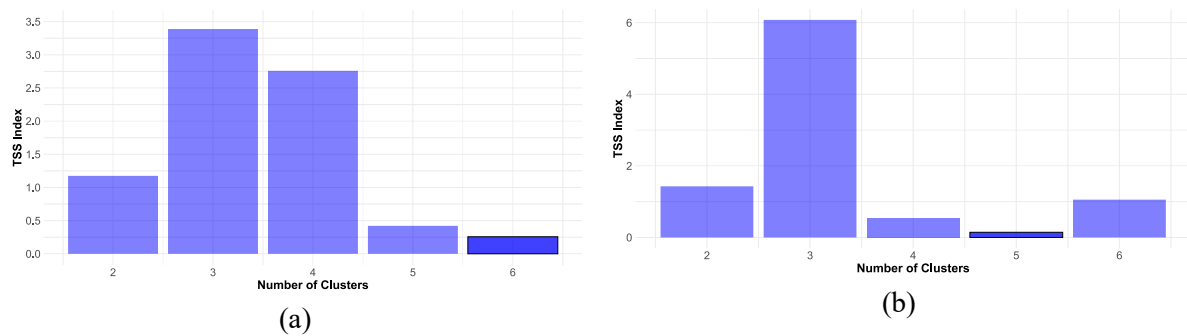


Figure 3. Determining the number of clusters with TSS index in clustering methods (a) Fuzzy C Means; (b) ABCFGWC

3.1. Optimal Number of Clusters

Figure 3 shows the TSS index graph for clusters formed using the FCM and ABCFGWC methods, with cluster numbers ranging from 2 to 6. In the FCM method, it can be seen that the lowest TSS index value in the cluster range of 2 to 6 is the TSS index for 6 clusters. As a result, the optimal number of clusters using FCM is six. Thus, six clusters were formed using the FCM method. On the other hand, in the ABCFGWC method, it can be seen that the lowest TSS index value for the range of clusters 2 to 6 is the TSS index for 5 clusters. As a result, using the ABCFGWC method, the optimal number of clusters is five. Thus, five clusters were formed using the ABCFGWC method.

3.2. Fuzzy C Means (FCM) Method

Based on the TSS index, a total of six clusters were formed using FCM. Interpretation of the cluster results can be done using the center of the cluster results. Table 2 displays the cluster centers of the 6 clusters formed using the FCM method.

Table 2. Centroid fuzzy c means (FCM) clustering result

Cluster	CO	NO	NO ₂	O ₃	SO ₂	PM _{2.5}	PM ₁₀	NH ₃
1	432.94	0.32	4.07	76.64	5.29	18.77	26.78	5.33
2	1416.28	0.92	21.35	179.71	20.95	104.10	124.82	16.96
3	3282.25	2.85	63.72	281.21	49.14	243.95	275.12	5.33
4	7809.41	3.46	116.40	639.47	69.80	721.95	832.20	1.53
5	833.50	0.43	9.34	100.76	8.50	52.69	69.99	13.18
6	2077.22	0.99	2.36	193.18	32.79	174.00	200.78	12.95

Source: Authors' Calculation

According to Table 2, Cluster 4 tends to have the highest air pollution compared to other clusters, except for the NH₃ pollutant. The next cluster that has the highest air pollution after Cluster 4 is Cluster 3. In contrast to Cluster 4, Cluster 3 tends to have lower air pollution values for all air pollutants except for the NH₃ pollutant. When it comes to the NH₃ pollutant, Cluster 3 actually has a higher average value than Cluster 4. Meanwhile, the other clusters exhibit their own unique patterns. For example, Cluster 6 has high pollution levels, ranking just below clusters 3 and 4 in most pollutants but has the lowest NO₂ pollutant among all clusters. Another example is Cluster 2, which has the highest NH₃ pollutant but is not very prominent in other pollutants.

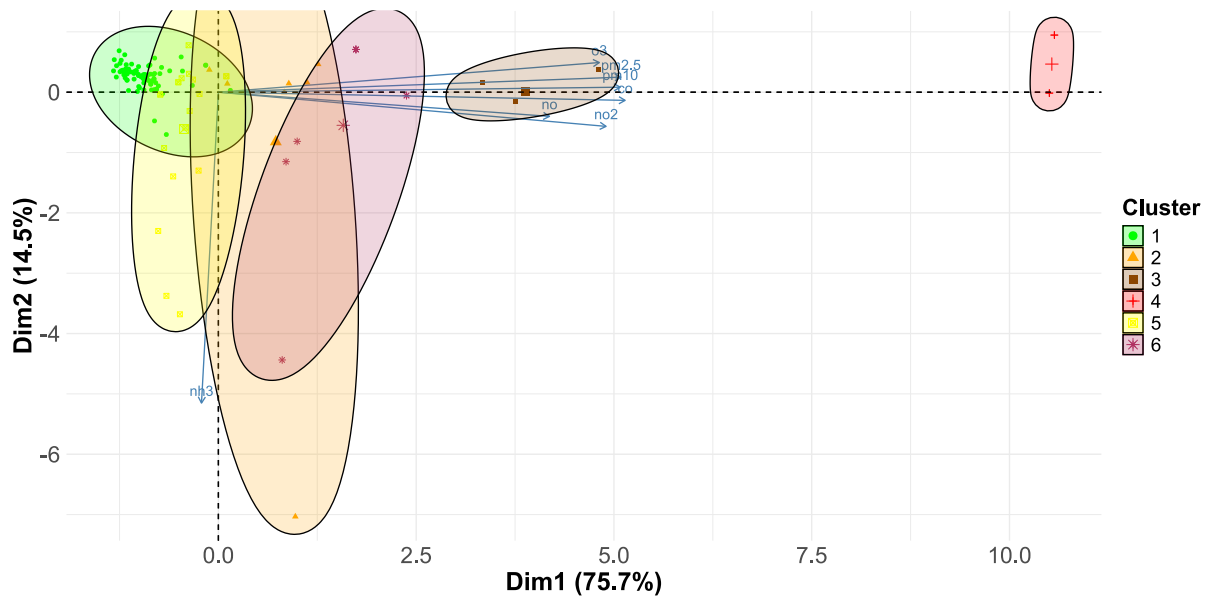


Figure 4. Biplot of principal component analysis (PCA) fuzzy c means (FCM) clustering results

In addition to using cluster centroids, biplots can be used to support profiling and show clustering in the clusters formed. Figure 4 shows a biplot using the PCA technique based on the cluster results obtained with the FCM method. Using PCA, dimension 1 is able to explain 75.7% of the data variation, while dimension 2 is able to explain 14.5% of the data variation. Thus, using only these two dimensions, the data has been well explained because it has explained more than 90% of the data variation. The majority of air pollution is explained by dimension 1 in the PCA results, except for NH_3 pollutants.

Examining Dimension 1 yields similar results, indicating that Cluster 4 typically has the highest air pollution compared to the other clusters. Figure 4 also supports Table 3 by showing that Cluster 3 is the most vulnerable cluster to air pollution after Cluster 4. On the other hand, Cluster 1 has the lowest air pollution compared to other clusters. Cluster 1, with its cluster center closest to the origin, demonstrates this result. When considering dimension 2, it can also be seen that Cluster 2, Cluster 5, and Cluster 6 are clusters characterized by high NH_3 pollutants. In addition, Cluster 2 also appears to have relatively higher NH_3 pollutants than the other clusters, as indicated by Cluster 2's elongated ellipse in dimension 2. Thus, the biplot shows similar results to the cluster centroid in Table 3.

3.3. Artificial Bee Colony (ABC) Fuzzy Geographically Weighted Clustering (FGWC) Method

The next clustering method implemented in this research is ABCFGWC. Based on the TSS index, a total of 5 clusters were formed using ABCFGWC. Similar to FCM, the interpretation of the cluster results is done using the center of the cluster results. Table 3 displays the cluster centers of the 5 clusters formed using the ABCFGWC method.

Table 3. Centroid artificial bee colony (ABC) fuzzy geographically weighted clustering (FGWC) result

Cluster	CO	NO	NO ₂	O ₃	SO ₂	PM _{2.5}	PM ₁₀	NH ₃
1	1962.00	1.02	27.74	191.95	29.46	159.75	186.28	16.00
2	900.00	0.44	10.15	109.25	9.61	59.53	77.31	13.12
3	3264.53	2.78	62.34	282.00	48.61	243.32	275.23	5.42
4	7692.77	3.46	115.38	628.55	69.42	709.67	817.87	1.61
5	460.99	0.35	4.58	78.13	5.65	20.82	29.44	5.88

Source: Authors' Calculation

Table 3 displays the cluster centers of the 5 clusters formed using the ABCFGWC method. According to Table 3, Cluster 4 tends to have the highest air pollution compared to other clusters, except for NH_3 pollution. The next cluster that has the highest air pollution after Cluster 4 is Cluster 3. In

contrast to Cluster 4, Cluster 3 tends to have high air pollution values for all air pollutants, except for the NH_3 pollutant. When it comes to the NH_3 pollutant, Cluster 3 actually has a higher average value than Cluster 4. The ABCFGWC method yields similar results to the FCM method. Other clusters exhibit characteristics unique to each cluster.

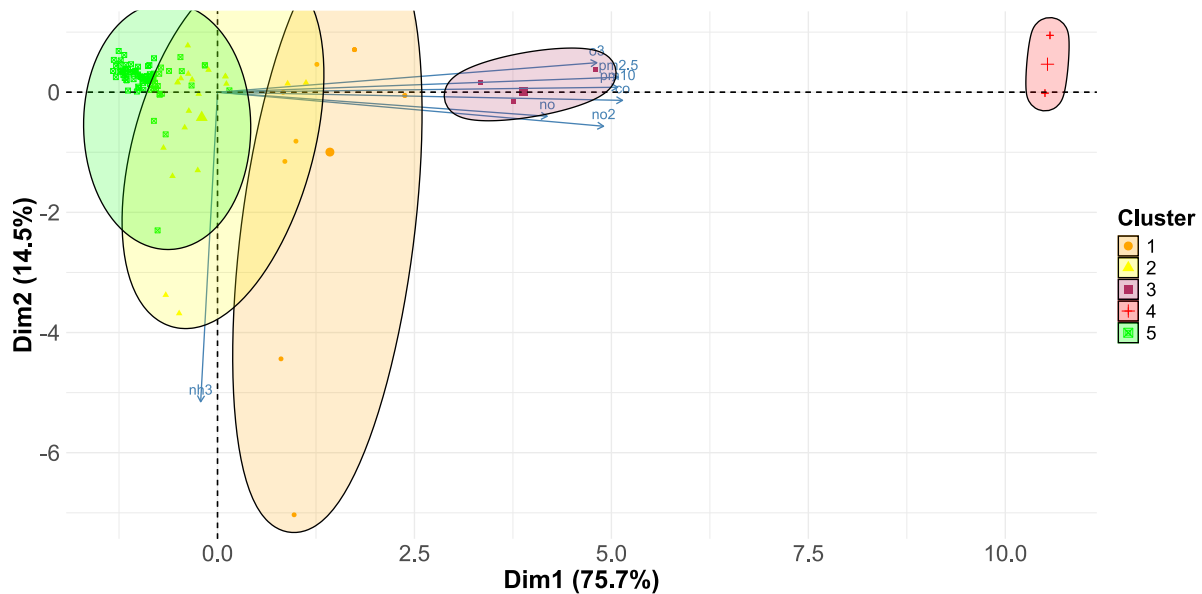


Figure 5. Biplot of principal component analysis (PCA) ABCFGWC (Artificial Bee Colony Fuzzy Geographically Weighted Clustering) results

Figure 5 illustrates the biplot using the PCA technique, based on the cluster results obtained through the ABCFGWC method. Using PCA, dimension 1 can explain 75.7% of the data variation, while dimension 2 can explain 14.5% of the data variation. Similar to the previous findings, the majority of air pollution is explained by dimension 1 in the PCA results, except for NH_3 pollutants. Examining dimension 1 yields similar results once more, suggesting that Cluster 4 typically experiences the highest levels of air pollution. Figure 5 also supports Table 3 by showing that Cluster 3 is the most vulnerable cluster to air pollution after Cluster 4. The results with the ABCFGWC method are similar to those obtained with the FGWC method, but there is a difference: in the ABCFGWC method, the separation between clusters is more visible. In addition, it should be noted that the numbering between the two clustering methods is the result of random initialization.

3.4. Comparison of the Two Clustering Methods

A comparison between the two clustering methods is done using the TSS index and Rand index values. The TSS index value is used to determine the suitability between the clusters formed with a better clustering method that will produce a smaller TSS index value. According to Table 4, the lowest TSS index values for the FCM and ABCFGWC methods are 0.257 and 0.145, respectively, in the range of clusters 2 to 6. As a result, the clusters formed by the ABCFGWC method outperform FCM based on the TSS index.

Table 4. Comparison of clustering methods by TSS index and rand index

Clustering Method	TSS Index for Number of Clusters					Rand Index
	2	3	4	5	6	
FCM	1.175	3.389	2.759	0.422	0.257	0.945
ABCFGWC	1.429	6.078	0.547	0.145	1.056	

Source: Authors' Calculation

However, the cluster membership formed using this method does not show a significant difference. The Rand index can be used to determine the suitability of members between the two clusters that formed. The Rand Index has a range of values between 0 and 1, with the higher values indicating more identical clusters [38]. Using the Rand Index, as many as six clusters generated using FCM are compared with as many as five clusters generated using ABCFGWC. The resulting Rand Index value is close to

1, indicating that these two clusters are quite similar despite having a different number of optimal clusters.

However, the Rand Index method measures the suitability of cluster members without considering the fuzzy membership degree but directly uses the cluster membership results [38]. Meanwhile, the FCM and ABCFGWC methods have different TSS index values, even showing a difference in the optimal number of clusters between the two methods. With the TSS index, the degree of cluster membership is considered for each observation. Therefore, the difference in TSS index value shows that there is a difference in the degree of membership between the two clustering methods. The lower TSS index value in the ABCFGWC method indicates that the results of clustering air pollution improve when the neighborhood effect is taken into account and the fuzzy membership degree is adjusted accordingly. Previous researchers found similar results, indicating that FGWC clustering outperforms FCM in the presence of regional aspects [39].

3.5. Discussion of Cluster Results

The clustering results of these two methods are depicted in a map visualization. Figure 6 depicts the visualization map of the clustering outcomes obtained from the FCM method. Meanwhile, Figure 7 depicts the visualization map of the clustering results obtained from the ABCFGWC method. It is evident that the two maps in Figure 6 and Figure 7 show a similar pattern of results. It should be noted, however, that the numbering between the two clustering methods is the result of random initialization. The difference in results between the two methods can be seen in a small number of regencies/cities in the eastern and western parts of Java Island. The absence of striking differences between the two clustering methods also supports the results of the rand index calculation.

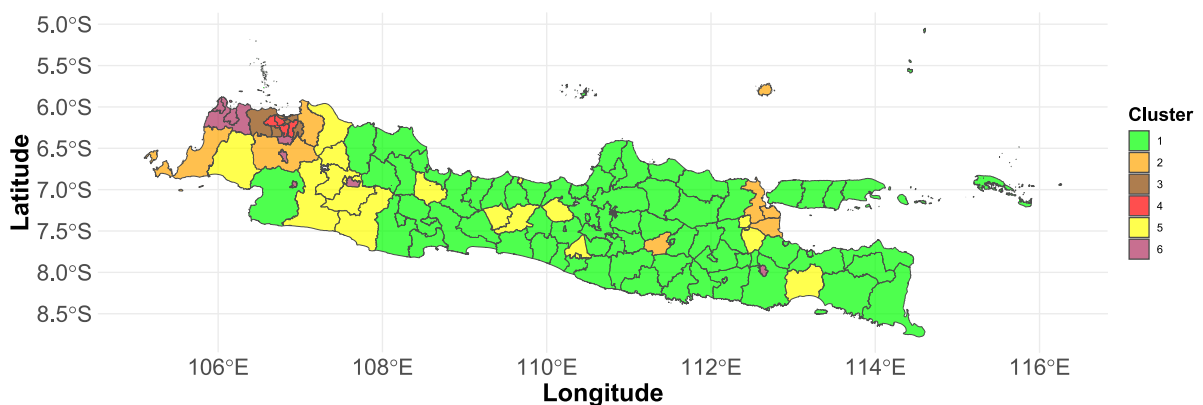


Figure 6. Map of cluster results of regency/city in Indonesia based on air pollution by fuzzy c means (FCM) approach

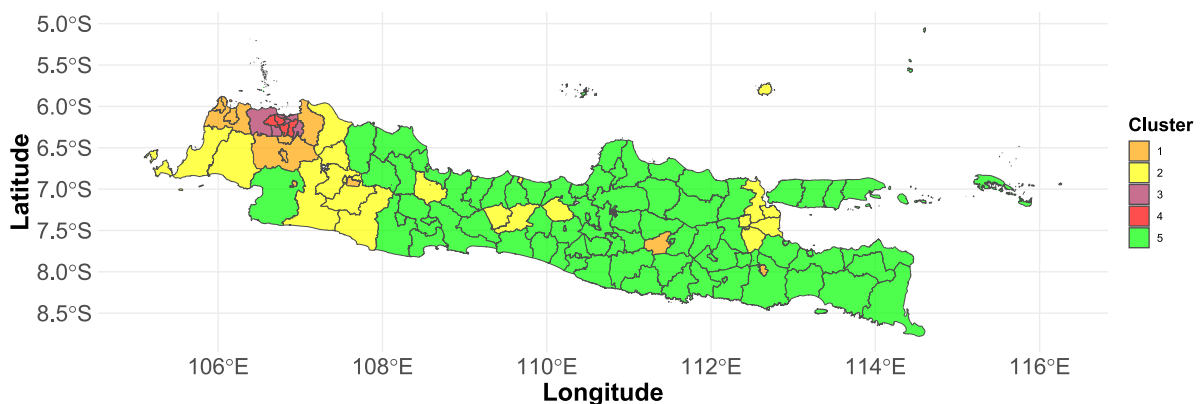


Figure 7. Map of cluster results of regency/city in Indonesia based on air pollution by artificial bee colony (ABC) fuzzy geographically weighted clustering (FGWC) approach

The FCM method, as shown in Figure 6, creates clusters solely based on data values. This approach leads to some differences in results for certain areas in the eastern and western parts of Java Island compared to the ABCFGWC method depicted in Figure 7. Meanwhile, the ABCFGWC method forms clusters by considering the distance between each region. This method makes it appear as though the regencies and cities in the eastern and western parts of Java Island are grouped together. In addition, the TSS Index result in the ABCFGWC method is lower, so the interpretation and discussion of the cluster results will be carried out in the ABCFGWC clustering method. This result is supported by the nature of air pollution, which has inter-regional linkages that are considered in the ABCFGWC method but not considered in the FCM method.

Figure 7 shows the distribution of regencies/cities on Java Island based on air pollution using ABCFGWC. The Jabodetabek area is classified as a member of Cluster 4, which, according to the profiling results, is a regency/city with highly vulnerable air pollution characteristics. Greater Jakarta or Jabodetabek, comprises Jakarta Special Capital Region and the surrounding regencies/cities of Bogor, Depok, Tangerang, and Bekasi. With dense economic and government activities, the Jabodetabek area is classified as a cluster that has high air pollution, which was also found in a related study [40]. This result aligns with Tian et al. [16], who found that air pollution is more closely linked to industrial structure than to geographical and socioeconomic conditions. However, measures to address high air pollution in industrial areas cannot be done by directly halting the industrial process. Reducing industrial and transportation activities will help reduce air pollution, but such action is not sufficient if the area is experiencing critical levels of pollution [41].

The results reveal that, in addition to the Jabodetabek area, several regencies/cities in the vicinity also fall into air pollution-vulnerable clusters, albeit not as vulnerable as Jabodetabek. In general, the eastern part of Java Island does not have any regions or cities that are susceptible to air pollution. Regencies/cities in the eastern part of Java Island mostly belong to Cluster 1, and only a small number belong to clusters apart from Cluster 1. Regencies/cities in the eastern part of Java Island tend to cluster close to the city of Surabaya, which is influenced by the dense economic and business activities. However, Figure 7 shows that the western part of Java Island is highly vulnerable to air pollution, especially the Jabodetabek area. The relationship between pollution and income typically follows an inverted U-shaped curve [42]. It starts in areas with low income and minimal pollution. When investments flow into such underdeveloped regions, it often happens because these areas have regulations allowing industry to be free to pollute. As a result, both income and pollution increase until they reach a peak. At this stage, the region becomes wealthy enough to prioritize environmental concerns. Over time, continued economic growth leads to reduced pollution as sustainable practices and cleaner technologies are adopted.

In Indonesia, particularly in Java, the region is still in the developing phase. This means the relationship between economic development and pollution has not yet reached its peak. Jakarta and Surabaya, which have higher economic output and GDP, also experience significantly higher pollution levels compared to other regions.

Table 5. Details of air quality index (AQI) and pollutant concentration range

AQI Category	Range of Pollutant Concentrations ($\frac{\mu g}{m^3}$)							
	CO	NO	NO ₂	O ₃	SO ₂	PM _{2.5}	PM ₁₀	NH ₃
Good	0-1,000	0-40	0-40	0-50	0-40	0-30	0-50	0-200
Satisfactory	1,000-2,000	41-80	41-80	51-100	41-80	31-60	51-100	201-400
Moderately Polluted	2,000-10,000	81-180	81-180	101-168	81-380	61-90	101-250	401-800
Poor	10,000-17,000	181-280	181-280	169-208	381-800	91-120	251-350	801-1200
Very Poor	17,000-34,000	281-400	281-400	209-748	801-1,600	121-250	351-430	1,201-1,800
Severe	34,000+	400+	400+	748+	1,600+	250+	430+	1,800+

Source: Kumar & Yadav [36]

Pollution in the Jabodetabek area can be said to be quite serious when compared to the AQI (Air Quality Index) settings for each pollutant. Table 5 shows the settings for the air quality index for each pollutant [36]. By examining the ABCFGWC cluster centroid results in Table 3 and comparing them with each pollutant, it can be seen that some pollutants in Cluster 4 are rated as moderately polluted or worse. On the other hand, NO, SO₂, and NH₃ pollutants in Cluster 4 are classified as satisfactory or good. Of all pollutants, the pollutants that need to be the main concern in Cluster 4 are PM_{2.5} and PM₁₀,

which are classified as severe. These two pollutants (particulate matter) are frequently mentioned by several sources as one of the most dangerous because of their carcinogenic nature and links to diseases of several organ systems [43]. As a result, policies to reduce air pollution must be applied to clusters of areas with high pollutant values, namely Cluster 4 and Cluster 3, with a particular emphasis on PM_{2.5} and PM₁₀ pollutants.

Some regions have implemented policies to reduce air pollution. For example, countries that have implemented policy measures to address air pollution are China and India. In 2014, China began implementing the "war against pollution" program and has successfully curtailed particulate pollution by 30% [13]. China established a policy of reducing air pollution through green technology reform and the use of green energy [14]. On the other hand, India implemented the National Clean Air Program (NCAP) in the form of city-specific air pollution policies along with a national strategy [44]. In addition to policies, the WHO proposes whole-community collaborative measures to reduce air pollution. WHO [2] conducts training for health workers to have the skills to implement policies and strategies for reducing air pollution. These health workers are then expected to become trainers and community mobilizers, ready to conduct innovative programs to address air pollution; this approach is called train-the-trainer [2].

Indonesia has implemented several measures to address air pollution, particularly particulate matter. Most of the policies in Indonesia focus on the transportation sector, as this sector is the biggest source of air pollution [45]; for example, the Indonesian government requires Euro-4 standard fuel from September 2018 [5]. Additionally, Indonesia has addressed air pollution by resolving peatland and forest fires and implementing coal regulations [5]. Nonetheless, an innovative step needs to be implemented to further accelerate the process of addressing air pollution problems in Indonesia. This innovative step can be achieved by involving all community groups in an air pollution reduction measure [2]. For instance, previous studies [3], [46] have proposed the integration of air pollution reduction and environmental sustainability into education. In addition, other steps can also be taken by transforming the transportation sector in Indonesia, for example, using renewable energy transportation [47].

4. Conclusion

We conducted a clustering analysis of regencies and cities on Java Island using the fuzzy clustering approach. This study further compares the performance of the fuzzy c-means (FCM) method with the ABCFGWC method, which incorporates regional factors. The findings indicate that the ABCFGWC method, due to its consideration of regional aspects, performs better in addressing air pollution cases than the FCM method. The ABCFGWC approach identified five clusters, with the Jabodetabek region classified as part of a cluster highly vulnerable to air pollution, exceeding the Air Quality Index (AQI) threshold. This work is intended to provide stakeholders with a basis for developing regional strategies to mitigate air pollution on Java Island.

There remains significant potential for further development of this study in terms of data, methodologies, and substantive focus. Future work can be advanced by employing panel data to examine the temporal dynamics of air pollution. Additionally, subsequent studies could broaden the geographic scope to include all regencies and cities across Indonesia or explore alternative clustering methods. Future studies may also investigate the interplay between air pollution issues and various socioeconomic challenges.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Arya Candra Kusuma: Conceptualization, Methodology, Data Collection, Data Analysis, Writing—Original Draft. **Arie Wahyu Wijayanto:** Conceptualization, Methodology, Manuscript Review, Research Advisor. **Arista:** Editing, Writing—Review. **Vicka Kharisma Bahar:** Editing, Writing—Review. **Tifani Husna Siregar:** Editing, Writing—Review.

References

- [1] G. Shaddick, M. L. Thomas, P. Mudu, G. Ruggeri, and S. Gumy, “Half the world’s population are exposed to increasing air pollution,” *NPJ Clim Atmos Sci*, vol. 3, no. 1, Dec. 2020, doi: 10.1038/s41612-020-0124-2.
- [2] WHO, “Mapping opportunities for training in air pollution and health for the health workforce,” 2021.
- [3] Y.-J. Huo, K.-T. Shih, and C.-J. Lin, “The study on integrating air pollution environmental education into the teaching personal and social responsibility model in physical education,” *IOP Conf Ser Earth Environ Sci*, vol. 576, no. 1, p. 012006, Nov. 2020, doi: 10.1088/1755-1315/576/1/012006.
- [4] A. A. Almetwally, M. Bin-Jumah, and A. A. Allam, “Ambient air pollution and its influence on human health and welfare: an overview,” *Environmental Science and Pollution Research*, vol. 27, no. 20, pp. 24815–24830, Jul. 2020, doi: 10.1007/s11356-020-09042-2.
- [5] K. Lee and M. Greenstone, “Polusi udara indonesia dan dampaknya terhadap usia harapan hidup [Indonesia’s air pollution and its impact on life expectancy],” 2021. Accessed: Jun. 04, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp-content/uploads/2021/09/AQLI_IndonesiaReport-2021_IND-version9.7.pdf
- [6] D. A. Glencross, T. R. Ho, N. Camiña, C. M. Hawrylowicz, and P. E. Pfeffer, “Air pollution and its effects on the immune system,” *Free Radic Biol Med*, vol. 151, pp. 56–68, May 2020, doi: 10.1016/j.freeradbiomed.2020.01.179.
- [7] A. J. White, J. P. Keller, S. Zhao, R. Carroll, J. D. Kaufman, and D. P. Sandler, “Air pollution, clustering of particulate matter components, and breast cancer in the sister study: a U.S.-wide cohort,” *Environ Health Perspect*, vol. 127, no. 10, Oct. 2019, doi: 10.1289/EHP5131.
- [8] H. Kim *et al.*, “Cardiovascular effects of long-term exposure to air pollution: a population-based study with 900 845 person-years of follow-up,” *J Am Heart Assoc*, vol. 6, no. 11, Nov. 2017, doi: 10.1161/JAHA.117.007170.
- [9] L. A. Rodriguez-Villamizar, A. Magico Bsc, A. Osornio-Vargas, and B. H. Rowe, “The effects of outdoor air pollution on the respiratory health of Canadian children: a systematic review of epidemiological studies,” *Can Respir J*, vol. 22, no. 5, pp. 282–292, 2015, doi: 10.1155/2015/263427.
- [10] A. Ebenstein, V. Lavy, and S. Roth, “The long-run economic consequences of high- stakes examinations: evidence from transitory variation in pollution,” *Am Econ J Appl Econ*, vol. 8, no. 4, pp. 36–65, 2016, doi: 10.1257/app.20150213.
- [11] R. Hanna and P. Oliva, “The effect of pollution on labor supply: evidence from a natural experiment in Mexico City,” *J Public Econ*, vol. 122, pp. 68–79, Feb. 2015, doi: 10.1016/j.jpubeco.2014.10.004.

- [12] M. Dobranschi, D. Nerudová, V. Solilová, and K. Stadler, "Carbon border adjustment mechanism challenges and implications: the case of Visegrád countries," *Heliyon*, vol. 10, no. 10, p. e30976, May 2024, doi: 10.1016/j.heliyon.2024.e30976.
- [13] M. Greenstone and C. Fan, "Is China winning its war on pollution?," 2020. Accessed: Jun. 04, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp-content/uploads/2021/02/China-Report_2020updateGlobal.pdf
- [14] P. Wang, "China's air pollution policies: progress and challenges," *Curr Opin Environ Sci Health*, vol. 19, Feb. 2021, doi: 10.1016/j.coesh.2020.100227.
- [15] M. Greenstone and Q. Fan, "Indonesia's worsening air quality and its impact on life expectancy," 2019. Accessed: Jun. 08, 2023. [Online]. Available: <https://aqli.epic.uchicago.edu/wp-content/uploads/2019/03/Indonesia-Report.pdf>
- [16] D. Tian *et al.*, "Characteristic and spatiotemporal variation of air pollution in Northern China based on correlation analysis and clustering analysis of five air pollutants," *Journal of Geophysical Research: Atmospheres*, vol. 125, no. 8, Apr. 2020, doi: 10.1029/2019JD031931.
- [17] G. R. Kingsy, R. Manimegalai, D. M. S. Geetha, S. Rajathi, K. Usha, and B. N. Raabiathul, "Air pollution analysis using enhanced k-means clustering algorithm for real time sensor data," in *2016 IEEE Region 10 Conference (TENCON)*, Singapore: Institute of Electrical and Electronics Engineers (IEEE), Feb. 2017, pp. 1945–1949. doi: 10.1109/TENCON.2016.7848362.
- [18] Z. Xu *et al.*, "Classification of urban pollution levels based on clustering and spatial statistics," *Atmosphere (Basel)*, vol. 13, no. 3, Mar. 2022, doi: 10.3390/atmos13030494.
- [19] S. Annas, U. Uca, I. Irwan, R. H. Safei, and Z. Rais, "Using k-means and self organizing maps in clustering air pollution distribution in Makassar City, Indonesia," *Jambura Journal of Mathematics*, vol. 4, no. 1, pp. 167–176, Jan. 2022, doi: 10.34312/jjom.v4i1.11883.
- [20] A. B. Santoso, A. C. Candra, R. Nooraeni, and A. W. Wijayanto, "Development of a hybrid fuzzy geographically weighted k-prototype clustering and genetic algorithm for enhanced spatial analysis: application to rural development mapping," *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 16, no. 2, pp. 122–139, Dec. 2024, doi: 10.34123/jurnalasks.v16i2.789.
- [21] A. Páez, M. Trépanier, and C. Morency, "Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards," *Transp Res Part A Policy Pract*, vol. 45, no. 7, pp. 640–652, 2011, doi: 10.1016/j.tra.2011.04.002.
- [22] E. Austin, B. A. Coull, A. Zanobetti, and P. Koutrakis, "A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition," *Environ Int*, vol. 59, pp. 244–254, Sep. 2013, doi: 10.1016/j.envint.2013.06.003.
- [23] B. I. Nasution, R. Kurniawan, T. H. Siagian, and A. Fudholi, "Revisiting social vulnerability analysis in Indonesia: an optimized spatial fuzzy clustering approach," *International Journal of Disaster Risk Reduction*, vol. 51, Dec. 2020, doi: 10.1016/j.ijdrr.2020.101801.
- [24] B. I. Nasution, F. M. Saputra, R. Kurniawan, A. N. Ridwan, A. Fudholi, and B. Sumargo, "Urban vulnerability to floods investigation in Jakarta, Indonesia: a hybrid optimized fuzzy spatial clustering and news media analysis approach," *International Journal of Disaster Risk Reduction*, vol. 83, p. 103407, Dec. 2022, doi: 10.1016/j.ijdrr.2022.103407.
- [25] R. E. Caraka *et al.*, "Micro, small, and medium enterprises' business vulnerability cluster in Indonesia: an analysis using optimized fuzzy geodemographic clustering," *Sustainability*, vol. 13, no. 14, Jul. 2021, doi: 10.3390/su13147807.
- [26] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput Geosci*, vol. 10, no. 2–3, pp. 191–203, 1984, doi: 10.1016/0098-3004(84)90020-7.
- [27] G. Grekousis, "Local fuzzy geographically weighted clustering: a new method for geodemographic segmentation," *International Journal of Geographical Information Science*, vol. 35, no. 1, pp. 152–174, 2021, doi: 10.1080/13658816.2020.1808221.
- [28] G. A. Mason and R. D. Jacobson, "Fuzzy Geographically Weighted Clustering," 2006, Accessed: Aug. 04, 2023. [Online]. Available: <https://www.researchgate.net/publication/255650957>
- [29] A. W. Wijayanto, A. Purwarianti, and L. H. Son, "Fuzzy geographically weighted clustering using artificial bee colony: an efficient geo-demographic analysis algorithm and applications to the analysis of crime behavior in population," *Applied Intelligence*, vol. 44, no. 2, pp. 377–398, Mar. 2016, doi: 10.1007/s10489-015-0705-7.
- [30] A. W. Wijayanto, S. Mariyah, and A. Purwarianti, "Enhancing clustering quality of fuzzy geographically weighted clustering using ant colony optimization," in *4th International Conference on Data and Software Engineering (ICoDSE 2017)*, Palembang, Indonesia: institute of electrical and electronics engineers (IEEE), 2018. doi: 10.1109/ICODSE.2017.8285858.

- [31] A. W. Wijayanto and A. Purwarianti, "Improvement design of fuzzy geo-demographic clustering using artificial bee colony optimization," in *2014 International Conference on Cyber and IT Service Management (CITSM)*, 2014, pp. 69–74. doi: 10.1109/CITSM.2014.7042178.
- [32] D. Simon, *Evolutionary optimization algorithms*, 1st ed. Wiley, 2013.
- [33] D. Karaboga and B. Basturk, "On the performance of artificial bee colony (ABC) algorithm," *Applied Soft Computing Journal*, vol. 8, no. 1, pp. 687–697, Jan. 2008, doi: 10.1016/j.asoc.2007.05.007.
- [34] F. De Rango, N. Palmieri, and M. Tropea, "Multirobot coordination through bio-inspired strategies," in *Nature-Inspired Computation and Swarm Intelligence*, 1st ed., X.-S. Yang, Ed., Academic Press, 2020, ch. 19, pp. 361–390. doi: 10.1016/b978-0-12-819714-1.00030-0.
- [35] B. I. Nasution, R. Kurniawan, and R. E. Caraka, "naspaclust: Nature-inspired spatial clustering," 2021. [Online]. Available: <https://cran.r-project.org/package=naspaclust>
- [36] S. Kumar and S. Yadav, "Air quality index and criteria pollutants in ambient atmosphere over selected sites: impact and lessons to learn from covid-19," in *Environmental Resilience and Transformation in Times of Covid-19*, 1st ed., Elsevier, 2021, ch. 15, pp. 153–162. doi: 10.1016/c2020-0-02703-9.
- [37] Y. Tang, F. Sun, and Z. Sun, "Improved validation index for fuzzy clustering," in *Proceedings of the American Control Conference*, 2005, pp. 1120–1125. doi: 10.1109/acc.2005.1470111.
- [38] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *J Am Stat Assoc*, vol. 66, no. 336, pp. 846–850, 1971, doi: 10.2307/2284239.
- [39] A. W. Wijayanto, "Improvement of fuzzy geo-demographic clustering using metaheuristic optimization on Indonesia population census," 2015. Accessed: Jun. 12, 2023. [Online]. Available: <https://www.researchgate.net/publication/279977406>
- [40] S. Kautsar, R. R. Ridwansyah, and N. Priscilla, "Formulation of greenhouse gas emission index and strategy towards net zero emission for cities on the java island in support of golden Indonesia 2045," *Seminar Nasional Official Statistics*, vol. 2024, no. 1, pp. 591–602, Nov. 2024, doi: 10.34123/semnasoffstat.v2024i1.2119.
- [41] P. Wang, K. Chen, S. Zhu, P. Wang, and H. Zhang, "Severe air pollution events not avoided by reduced anthropogenic activities during COVID-19 outbreak," *Resour Conserv Recycl*, vol. 158, Jul. 2020, doi: 10.1016/j.resconrec.2020.104814.
- [42] B. Özcan and I. Öztürk, *Environmental Kuznets Curve (EKC)*. Elsevier, 2019. doi: 10.1016/C2018-0-00657-X.
- [43] P. J. Landrigan *et al.*, "The lancet commission on pollution and health," *The Lancet Commission*, vol. 391, no. 10119, pp. 462–512, Feb. 2018, doi: 10.1016/S0140-6736(17)32345-0.
- [44] M. Greenstone and Q. Fan, "India's 'war against pollution': an opportunity for longer lives," 2019. Accessed: Jun. 08, 2023. [Online]. Available: https://aqli.epic.uchicago.edu/wp-content/uploads/2019/01/India.NCAP_Global.English.pdf
- [45] P. Lestari, M. K. Arrohman, S. Damayanti, and Z. Klimont, "Emissions and spatial distribution of air pollutants from anthropogenic sources in Jakarta," *Atmos Pollut Res*, vol. 13, no. 9, p. 101521, Sep. 2022, doi: 10.1016/j.apr.2022.101521.
- [46] E. Sáez de Cámara, I. Fernández, and N. Castillo-Eguskita, "A holistic approach to integrate and evaluate sustainable development in higher education. the case study of the university of the basque country," *Sustainability*, vol. 13, no. 1, p. 392, Jan. 2021, doi: 10.3390/su13010392.
- [47] R. Kurniawan, A. C. Kusuma, B. Sumargo, P. U. Gio, S. K. Wongsonadi, and K. Sasmita, "The role of renewable energy and foreign direct investment toward environmental degradation convergence to achieve sustainability: evidence from ASEAN countries," *International Journal of Energy Sector Management*, May 2024, doi: 10.1108/IJESM-02-2024-0012.



Adding MSNBURR-IIa Distribution to MultiBUGS

Eliana Putri Ramadani¹, Achmad Syahrul Choir^{2*}, Anindya Apriliyanti Pravitarsari³, Joynabel Paraguison⁴

¹BPS-Statistics Kapuas Hulu Regency, Putussibau, Indonesia, ²Politeknik Statistika STIS, Jakarta, Indonesia,

³Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Bandung, Indonesia, ⁴Phillippine Statistics Authority, Phillippine

*Corresponding Author: E-mail address: madsyair@stis.ac.id

ARTICLE INFO

Article history:

Received 12 September, 2024

Revised 13 February, 2025

Accepted 12 July, 2025

Published 31 December, 2025

Keywords:

Bayesian; MCMC; MSNBurr-IIa; MultiBUGS; Neo-normal

Abstract

Introduction/Main Objectives: The MSNBurr-IIa distribution is a neo-normal distribution designed to fit right-skewed data better. This article aims to integrate the MSNBurr-IIa distribution into MultiBUGS, thereby enabling Bayesian estimation of its parameters. **Background Problems:** Markov Chain Monte Carlo (MCMC) is a popular method for Bayesian computations, although its implementation is frequently challenging. MultiBUGS, a statistical tool that uses the BUGS language, is used to make this easier. **Novelty:** This paper details integrating the MSNBurr-IIa distribution into MultiBUGS, allowing for estimating its parameters. The module's effectiveness is demonstrated through its application on both simulated data and regional economic growth data of Indonesian districts/cities in 2021. **Research Methods:** The MSNBurr-IIa module was developed using five steps: requirement, design, development, testing, and implementation in simulation and real-world data. It was built with Blackbox Component Builder, an integrated development environment (IDE) for the Component Pascal programming language. **Finding/Results:** The findings confirm that MultiBUGS, with the MSNBurr-IIa module, successfully estimates the distribution's parameters across various datasets.

1. Introduction

The normal or Gaussian distribution is one of the most widely used statistical distributions. However, many real-world datasets exhibit skewness or fat tails, often resulting from outliers or other underlying factors. The symmetrical and mesokurtic nature of the normal distribution may not always align with the characteristics of real-world data, particularly those exhibiting skewness or non-normal kurtosis. Applying the normal distribution to non-normally distributed data, such as those with skewness or differing kurtosis, can lead to inaccurate inferences, as it fails to capture the underlying characteristics of the data adequately [1].

Modifying the normal distribution can overcome its limitations, yielding a more flexible model capable of accommodating diverse data characteristics [2]. Several methods exist for achieving this modification, including directly altering the probability density function (PDF) of the normal distribution [2], [3], applying Tukey's-*gh* transformation [4], [5], and employing the compounding method [6], [7]. These modifications can alter the distribution's symmetry and kurtosis, controlled by additional parameters introduced into the model [1]. For instance, in the skew-normal distribution, an

added skewness parameter allows for the control of skewness, with changes to this parameter directly altering the distribution's asymmetry [3].

Amemiya [8] proposed that the logistic distribution serves as an approximation of the standard normal distribution. Building upon this foundation, Iriawan [9] extended the concept to model skewed non-normal data by utilizing the Burr II distribution [10], which is a more flexible generalization of the logistic distribution. These distributions were termed 'neo-normal' by Iriawan [9].

Neo-normal distributions represent a class of distributions that either are inherently normal or can approximate normality under specific conditions [1], [9], [11]. For instance, the MSNBurr distribution can closely approximate normality when symmetric, the skew-normal distribution converges to the standard normal distribution when its skewness parameter is zero (indicating symmetry), and the exponential power distribution becomes identical to the normal distribution when its shape parameter results in a kurtosis of 3 (the kurtosis of a normal distribution, also known as mesokurtic).

The MSNBurr-IIa distribution extends the MSNBurr distribution to handle cases with opposite skewness characteristics [1]. Derived from the Burr-IIa distribution [12], the MSNBurr-IIa distribution is better suited for accommodating right-skewed data than left-skewed data. MSNBurr-IIa can estimate both symmetrical and skewed data distributions [1]. The parameters represent the shape, location, and scale parameters of the MSNBurr-IIa distribution α , μ , and σ , respectively. The shape parameter is related to variations in its skewness. Currently, the specific values of each parameter for the MSNBurr-IIa distribution are unknown. Consequently, estimation is required to determine these parameter values.

There are two primary methods for estimating parameters in statistics: the frequentist and Bayesian approaches [13]. The Bayesian approach is preferred in this study because it offers more flexible results than the frequentist method, particularly in handling complex models and incorporating uncertainty. This approach incorporates prior knowledge about the parameters, combining it with observed data to produce a probability distribution of the parameters after the data are observed, known as posterior distribution. While standard Bayesian calculations can become more complex and time-consuming when dealing with a larger number of parameters and data [14], advancements in user-friendly software, such as MultiBUGS, driven by scientific and technological progress, have made the Bayesian approach readily accessible.

Bayesian inference using Gibbs Sampling (BUGS) is a project developed to facilitate the implementation of Bayesian inference. The advantages of BUGS compared to other Bayesian software include ease of use and speed. It also has a model visualization feature called Doodle, which simplifies model understanding and analysis [15]. The latest and fourth version of the BUGS program is MultiBUGS. MultiBUGS was developed to address computational problems in previous software by employing an MCMC parallelization strategy. MultiBUGS can run on Linux and is fully implemented within other statistical programs, such as R [14], [16].

Currently, only 29 theoretical distributions are available in MultiBUGS, although the program includes many specialized distributions, such as those for temporal, spatial, and reliability modeling. However, because MultiBUGS is open-source, users can modify the application as needed. Extensions to MultiBUGS can be implemented through the BlackBox Components Builder, an open-source Pascal-based Integrated Development Environment (IDE). BlackBox provides tools for creating module-based documents, assembling and running modules, and developing user interfaces, enabling users to customize MultiBUGS according to their specific requirements.

To address the absence of the MSNBurr-IIa distribution in MultiBUGS, we developed an MSNBurr-IIa distribution module. This addition aims to simplify the parameter estimation process for the MSNBurr-IIa distribution using the Bayesian approach, making it more accessible and user-friendly. The module underwent rigorous testing to ensure its functionality and accuracy. An example application involving real-world data with symmetry issues, for which the normal distribution is unsuitable, demonstrates the use of the module.

The addition of the MSNBurr-IIa distribution in MultiBUGS provides significant advantages for modeling heavy-tailed data, offering flexibility in handling both symmetric and asymmetric distributions. Compared to traditional distributions like the normal or logistic, MSNBurr-IIa allows better modeling of extreme values and heteroskedasticity, making it particularly useful for real-world data that exhibit high variability.

Beyond MultiBUGS, this distribution has also been integrated into other Bayesian software. Specifically, the R package *neodistr* extends support for Stan, while the R package *neojags* adds it to JAGS. These implementations enable Bayesian practitioners to leverage MSNBurr-IIa across multiple platforms, ensuring robust statistical modeling for diverse data structures.

2. Material and Methods

2.1. Bayesian Method

The Bayesian method is an inference method derived from Bayes' Theorem, where parameter values are unknown and treated as random variables [14]. These random variables are expressed using probability distributions: $f(\theta)$ represents the prior distribution and $f(\theta|x)$ represents the posterior distribution. The prior distribution reflects existing knowledge or beliefs about the parameters before analyzing the data (x) [8]. The posterior distribution is obtained by updating the prior distribution with observed data using Bayes' Theorem. The equation for the posterior distribution is

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \propto f(x|\theta)f(\theta). \quad (1)$$

2.2. MSNBurr-IIa Distribution

The probability density function of the MSNBurr-IIa distribution is given by

$$f(x|\mu, \sigma, \alpha) = \frac{\omega}{\sigma} \exp\left(\frac{\omega}{\sigma}(x - \mu)\right) \left(1 + \frac{\exp\left(\frac{\omega}{\sigma}(x - \mu)\right)}{\alpha}\right)^{-(\alpha+1)} \quad (2)$$

where $-\infty < x < \infty$, $-\infty < \mu < \infty$, $\alpha > 0$, $\sigma > 0$, and

$$\omega = \frac{\left(1 + \frac{1}{\alpha}\right)^{(\alpha+1)}}{\sqrt{2\pi}} \quad (3)$$

which α is a shape parameter, μ is location parameter, σ and is a scale parameter. If $\alpha > 1$ then data is left-skewed, if $\alpha = 1$ then data is symmetric, and if $0 < \alpha < 1$ then data is right-skewed. The mean, variance, mode, skewness, excess kurtosis, and quantile of the MSNBurr-IIa distribution can be computed using formulas (4)-(9), respectively.

$$E(x) = \mu + \frac{\sigma}{\omega(\alpha)} \{\psi^0(1) + \ln(\alpha) - \psi^0(\alpha)\} \quad (4)$$

$$Var(x) = \frac{\sigma^2}{\omega(\alpha)^2} \{\psi^1(\alpha) + \psi^1(1)\} \quad (5)$$

$$Mode = \mu \quad (6)$$

$$Skew(x) = \frac{\psi^2(1) - \psi^2(\alpha)}{\{\psi^1(\alpha) + \psi^1(1)\}^{\frac{3}{2}}} \quad (7)$$

$$K(x) = \frac{(\psi^3(\alpha) - \psi^3(1))}{(\psi^1(\alpha) + \psi^1(1))^2} \quad (8)$$

$$q(u) = \mu + \frac{\sigma}{\omega} \left(\ln(\alpha) + \ln\left((1-u)^{-\frac{1}{\alpha}} - 1\right) \right) \quad (9)$$

where $0 < u < 1$, ω follows equation (3), and $\psi^0(\cdot)$ is $\psi^1(\cdot)$, $\psi^2(\cdot)$, $\psi^3(\cdot)$ are digamma, trigamma, tetragamma, and pentagamma functions, respectively [1].

2.3. Deviance Information Criterion (DIC)

The Deviance Information Criterion (DIC), a widely used Bayesian model selection criterion, assesses model validity and goodness of fit. It balances model complexity and goodness of fit to the observed data. In the BUGS program, DIC is automatically calculated and can be used to compare different models. The model with the smallest DIC value is preferred, as it indicates the best fit to the data. DIC is calculated as

$$DIC = \bar{D} + P_D \quad (10)$$

where $\bar{D}$ is the posterior mean of the deviance and P_D is the adequate number of parameters. (See [17] for further details).

2.4. MSNBurr-IIa Distribution Module Development

The research began by analyzing the requirements for a functional MSNBurr-IIa distribution module within MultiBUGS. This analysis informed the module's design, which was then implemented using Pascal and the "UnivariateTemp1.odc" template file. The following adjustments were made to the template to accommodate the MSNBurr-IIa distribution: (1) Module initiation adjustments, (2) Parameter adjustments in all procedures, (3) Equation changes in several procedures, including ClassifyPrior, DevianceUnivariate, DiffLogLikelihood, DiffLogPrior, LogLikelihood-Univariate, LogPrior, Cumulative, and Sample, (4) Addition of procedures for generating random values from the MSNBurr-IIa distribution in "Randnum.odc" file, (5) Updating "The 'Cumulative.odc' file with a new procedure to compute the cumulative distribution function of the MSNBurr-IIa distribution, (6) Updating The 'External.odc', 'Make.odc', and 'Linking.odc' files to incorporate 'GraphMSNBurr2a', allowing the MSNBurr-IIa distribution module to be accessed and utilized, and (7) Modification of the 'Strings.odc' file to include the parameters and default values for the MSNBurr-IIa distribution, enabling its graphical representation within the software.

The next step involved rigorous testing and validation of the developed module to ensure its functionality and accuracy for practical use. This testing process consisted of two main steps: First, the MSNBurr-IIa distribution module was tested to ensure proper compilation and the absence of coding errors. This was achieved by running the module through Doodle, MultiBUGS's graphical interface for model visualization. The absence of error messages indicated a successful compilation. Second, the probability values generated by the module were validated by comparing them with equivalent calculations performed in the R programming language, using identical input parameters. This aimed to ensure the accuracy of the probability calculations within MultiBUGS. The validation used the absolute difference in cumulative distribution function (CDF) and probability density function (PDF) values, following the approach used by Annis et al. [18] for adding distributions to the Stan program. More minor absolute differences indicate higher accuracy in the MultiBUGS.

After confirming the module's functionality and accuracy, the next step focused on its implementation for simulated and real-world data. Simulation data was used to assess the performance of the MSNBurr-IIa distribution in accurately modeling data with various characteristics, including right-skewed, symmetrical, and left-skewed data. Normal distribution simulations served as a benchmark for comparison. Simulation data is generated from the distribution of MSNBurr-IIa for three scenarios with varying degrees of skewness. Scenario 1 for right-skewed data is MSNBurr-IIa(0,1,0.1), Scenario 2 for symmetrical data is MSNBurr-IIa (0,1,1), and Scenario 3 for left-skewed data is MSNBurr-IIa(0,1,3). Additionally, Scenario 4 is normal(0,1). Each scenario comprised 1,000 simulated samples.

For real-world application, the module analyzed the economic growth rate of districts/cities in Indonesia in 2021. This data was used because it has a non-normal distribution, thus demonstrating the advantages of the MSNBurr-IIa distribution. This application involved several steps: First, the economic growth rate data was examined using the Shapiro-Wilk test for normality and the skewness coefficient to understand its characteristics and determine the degree of skewness. Second, the parameters of the MSNBurr-IIa distribution were estimated using the developed module in MultiBUGS. Convergence of the estimation process was assessed through trace plots, which visually display the sampled values of the parameters over time. Finally, the model was validated by comparing the characteristics of the fitted MSNBurr-IIa distribution (e.g., mean, variance, and skewness) with the observed characteristics of the economic growth rate data. This comparison, which included visual assessments of histograms, evaluated the model's goodness of fit and ability to represent the real-world data accurately.

3. Results and Discussion

3.1. Requirements

The necessary conditions in MultiBUGS can be identified by examining the functions and conditions in existing distribution modules, such as the normal and logistics distribution modules, which served as references for developing the MSNBurr-IIa distribution module. To create a new module in MultiBUGS, the following are required:

1. The MultiBUGS source code, including a template for new distribution modules and necessary modules (see Figure 3)
2. BlackBox component builder v1.7 from Oberon Microsystems (<http://www.oberon.ch>).
3. Mathematical formulas, including formulas for the log-likelihood, its derivatives, random variate generation, and the cumulative distribution function.

3.2. Module Design

The MSNBurr-IIa module in MultiBUGS is developed using the 'univariatetemp.odc' file template. To adhere to the MultiBUGS architecture, modifications are made not only to the 'Univariatetemp.odc' file but also involve adding or modifying code in several other files, including 'Randnum.odc', 'Cumulative.odc', 'External.odc', 'Make.odc', 'Linking.odc', 'Strings.odc'. The files modified within the architecture are highlighted in the diagram in Figure 1. Green indicates folders, while purple highlights indicate the files requiring modification.

For consistency, we change the parameter σ to τ , where $\tau = \frac{1}{\sigma}$. This change affects the probability density function (pdf). The pdf of the MSNBurr-IIa distribution after parameter changes is as follows.

$$f(x|\omega, \mu, \tau, \alpha) = \omega \tau \exp(\omega \tau (x - \mu)) \left(1 + \frac{\exp(\omega \tau (x - \mu))}{\alpha}\right)^{-(\alpha+1)} \quad (11)$$

where ω follow equation (3). Changes in the PDF will cause changes to the other equations used, namely, as follows.

1. The natural logarithm of the probability density function

$$\ln(f(x)) = \ln(\omega) + \log(\tau) + \omega \tau (x - \mu) - (1 - \alpha) \log\left(1 + \frac{\exp(\omega \tau (x - \mu))}{\alpha}\right) \quad (12)$$

2. The natural logarithm of the unnormalized probability density function

$$\ln(p(x)) = \omega \tau (x - \mu) - (1 - \alpha) \log\left(1 + \frac{\exp(\omega \tau (x - \mu))}{\alpha}\right) \quad (13)$$

3. The natural logarithmic derivative of the probability density function for variable X

$$\frac{d}{dx} \log(f(x)) = -\frac{\alpha \omega \tau (\exp(\omega \tau (x - \mu)) - 1)}{\exp(\omega \tau (x - \mu)) + \alpha} \quad (14)$$

4. Partial derivative of parameter α

$$\frac{\partial}{\partial \alpha} \log(f(x)) = -\frac{\exp(\omega \tau (x - \mu))(-1 - \alpha)}{\left(\frac{\exp(\omega \tau (x - \mu))}{\alpha} + 1\right)\alpha^2} - \log\left(\frac{\exp(\omega \tau (x - \mu))}{\alpha} + 1\right) \quad (15)$$

5. Partial derivative of parameter μ

$$\frac{\partial}{\partial \mu} \log(f(x)) = \frac{\alpha \omega \tau (\exp(\omega \tau (x - \mu)) - 1)}{\exp(\omega \tau (x - \mu)) + \alpha} \quad (16)$$

6. Partial derivative of parameter

$$\frac{\partial}{\partial \tau} \log(f(x)) = \frac{\omega(-1 - \alpha)(x - \mu) \exp(\omega \tau (x - \mu))}{\alpha \left(\frac{\exp(\omega \tau (x - \mu))}{\alpha} + 1\right)} + \frac{1}{\tau} + \omega(x - \mu) \quad (17)$$

7. Quantile

$$x = \mu + \frac{1}{\omega \tau} \left(\log(\alpha) + \log\left((1 - u)^{-\frac{1}{\alpha}} - 1\right) \right) \quad (18)$$

8. Cumulative distribution function

$$F(x) = 1 - \left(1 - \frac{\exp(\omega \tau (x - \mu))}{\alpha}\right)^{-\alpha} \quad (19)$$

Making doodles the MSNBurr-IIa distribution in MultiBUGS requires a list of parameters used in the MSNBurr-IIa distribution: location parameter, rate parameters, shape parameters, and the default values of each parameter. along with the default values of each. The default values used are 0 for the location parameter, 1.0×10^{-3} for the rate parameter and 0.1 for the shape parameter. The diagram illustrating the relationships between modules in creating the MSNBurr-IIa distribution module is visualized in Figure 1.

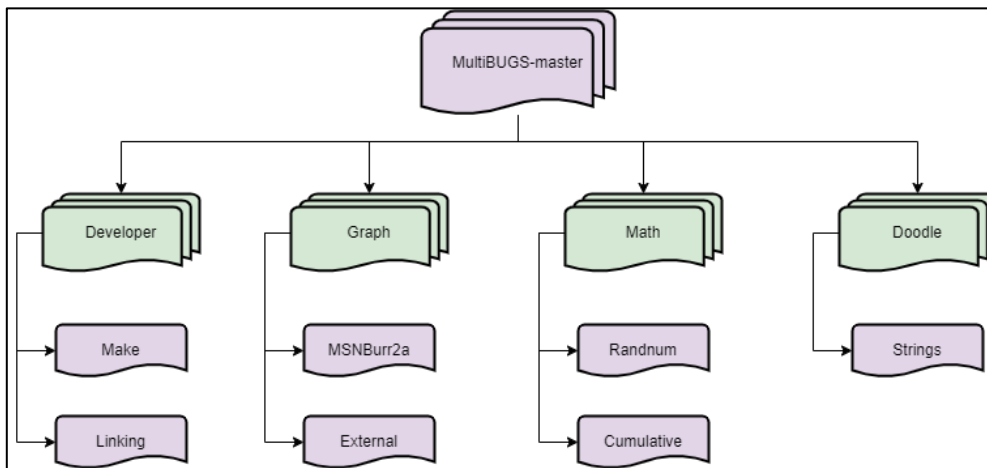


Figure 1. Diagram of the relationship between modules in the creation of the MSNBurr-IIa distribution module

3.3. Module Development

During the module creation stage, the module code will be written according to the design and equations prepared during the module design stage. The module is created using the "Univaritetempl.odc" template in the MultiBUGS-master file. The code template is written in Pascal and displayed in black, red, and green colors. Lines of code in green and enclosed by (* ... *) are not executed by the program, as they are comments. Lines of code in black should not be changed except for definitions added. Lines of code in red can be modified according to the user's needs.

The MSNBurr-IIa distribution module-file will be written in a new file named "MSNBurr2a.odc". An example of the beginning code in the file "MSNBurr2a.odc" can be seen in Figure 2. This file is saved in the "Graph\mod" folder

The MSNBurr-IIa distribution module is initiated at the beginning of the file, first by naming the module "GraphMSNBurr2a". This name is chosen because the module will be stored in the Graph folder for the MSNBurr-IIa distribution. After naming the module, the folders used in the MSNBurr-IIa distribution module are specified: the Math folder, which stores mathematical function modules, and the Stores folder for storage in the IMPORT function. The TYPE function declares the data types of the nodes and parameters that will be used. Next, the Factory function is an abstract class to create nodes in the graphical model. A constant is also declared, specifically ϵ (eps), with a value of 1.0×10^{-10} . Finally, variables are declared, namely fact, version, and maintainer.

```

MODULE GraphMSNBurr2a;

IMPORT
  Math, Stores := Stores64,
  GraphNodes, GraphRules, GraphStochastic, GraphUnivariate,
  MathCumulative, MathFunc, MathRandnum;

TYPE
  Node = POINTER TO RECORD(GraphUnivariate.Node)
    mu, tau, alpha: GraphNodes.Node
  END;

  Factory = POINTER TO RECORD(GraphUnivariate.Factory)
  END;

CONST
  eps = 1.0E-10;

VAR
  fact: GraphUnivariate.Factory;
  version: INTEGER;
  maintainer: ARRAY 40 OF CHAR;

```

Figure 2. Code of the beginning part of the file "MSNBurr2a.odc"

The *BoundsUnivariate* procedure establishes that the distribution is a continuous univariate distribution, with a domain of $(-\infty, \infty)$. The *CheckUnivariate* procedure verifies if the distribution is univariate and checks compliance with the MSNBurr-IIa distribution. These requirements dictate that the parameters τ and α must be positive. If $\tau < -\epsilon$ and $\alpha < -\epsilon$, an error will be triggered, indicating that the second and third arguments contain non-positive parameters (which is invalid for this distribution). The code implementing the *BoundsUnivariate* and *CheckUnivariate* procedures within the "MSNBurr2a.odc" file is shown in Figure 3.

```
PROCEDURE(node: Node)CheckUnivariate(): SET;
BEGIN
  IF node.tau.value < -eps THEN
    RETURN{GraphNodes.posative,GraphNodes.arg2}
  END;
  IF node.alpha.value < -eps THEN
    RETURN{GraphNodes.posative,GraphNodes.arg3}
  END;
  RETURN{}
END CheckUnivariate;
```

Figure 3. Code of boundsunivariate and checkunivariate procedure in “MSNBurr2a.odc”

The *ClassifyLikelihoodUnivariate* procedure classifies the likelihood of a distribution by examining the parent nodes of each parameter (α , μ , and τ). Other parts of the MultiBUGS system likely use this classification to determine how to handle the distribution during calculations and statistical inference. The parent nodes of parameters are categorized by the *ClassFunction*, which is found in the "Stochastic.odc" file (Figure 4).

```
PROCEDURE(node: Node)ClassifyLikelihoodUnivariate(parent: GraphStochastic.Node): INTEGER
VAR
  density, density0, density1, f0, f1: INTEGER
BEGIN
  f0 := GraphStochastic.ClassFunction(node.mu, parent);
  f1 := GraphStochastic.ClassFunction(node.tau, parent);
  CASE f0 OF
    |GraphRules.const:
      density0 := GraphRules.unif
    |GraphRules.ident, GraphRules.prod, GraphRules.linear:
      density0 := GraphRules.logCon
    |GraphRules.other:
      density0 := GraphRules.general
  ELSE
    density0 := GraphRules.genDiff
  END;
  density1 := GraphRules.ClassifyShape(f1);
  IF density0 = GraphRules.unif THEN
    density := density1
  ELSIF density1 = GraphRules.unif THEN
    density := density0
  ELSIF (density0 # GraphRules.general) & (density1 # GraphRules.general) THEN
    density := GraphRules.genDiff
  ELSE
    density := GraphRules.general
  END;
  RETURN density
END ClassifyLikelihoodUnivariate;
```

Figure 4. Code of classifylikelihoodunivariate procedure in “MSNBurr2a.odc”

The *ClassifyPrior* procedure, as shown in Figure 5, returns a value indicating that the prior for the MSNBurr-IIa distribution is logCon (log-concave). This suggests that the parameters of this distribution are assumed to have priors belonging to the log-concave distribution family. The *Cumulative* procedure calculates the cumulative probability for the MSNBurr-IIa distribution, using Equation (19) for the cumulative distribution function.

```

PROCEDURE(node: Node) ClassifyPrior(): INTEGER;
VAR
  class : INTEGER;
BEGIN
  class := GraphRules.logCon;
  RETURN class
END ClassifyPrior;

PROCEDURE(node: Node) Cumulative (x: REAL): REAL;
VAR
  cumulative, mu, tau, alpha, omega: REAL;
BEGIN
  mu := node.mu.value;
  tau := node.tau.value;
  alpha := node.alpha.value;
  omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
  cumulative := 1 - Math.Power((1 + (Math.Exp(omega * (x-mu) * tau)) / alpha), -alpha)
  RETURN cumulative;
END Cumulative;

```

Figure 5. The code of classifyprior dan cumulative procedure in “MSNBurr2a.odc”

The DevianceUnivariate procedure (Figure 6) defines the distribution's deviance, which is calculated as negative two times the logarithm of the probability density function, as described in Equation (12). This procedure is also commonly used in calculating the Deviance Information Criterion (DIC).

```

PROCEDURE(node: Node) DevianceUnivariate(): REAL;
VAR
  logDensity, logTau, logOmega, logExp, mu, tau, alpha, x, omega: REAL;
BEGIN
  x := node.value;
  mu := node.mu.value;
  tau := node.tau.value;
  alpha := node.alpha.value;
  omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
  logTau := MathFunc.Ln(tau);
  logOmega := MathFunc.Ln(omega);
  logDensity := logOmega + logTau + (omega*(x-mu)*tau) - (1 + alpha) * Math.Ln(1.0 +
  (Math.Exp(omega*(x-mu)*tau)) / alpha);
  RETURN -2.0 * logDensity;
END DevianceUnivariate;

```

Figure 6. The code of devianceunivariate procedure in “MSNBurr2a.odc”

The *DiffLogLikelihood* procedure calculates the derivative of the log-likelihood function concerning each parameter. These derivatives are defined in Equations (15), (16), and (17). The *DiffLogPrior* procedure calculates the derivative of the log probability density function for the MSNBurr-IIa distribution, based on the random variable X . In the context of the MSNBurr-IIa distribution, this derivative refers to the log of the probability density function (log-pdf) concerning the variable X , as defined in Equation (14). The *DiffLogLikelihood* and *DiffLogPrior* procedures are shown in Figures 7 and 8, respectively.

```

PROCEDURE(node: Node)DiffLogLikelihood(x: GraphStochastic.Node): REAL;
VAR
  mu, tau, alpha, val, omega, differential, diffTau, diffMu, exp, diffAlpha: REAL;
BEGIN
  val := node.value;
  mu := node.mu.value;
  tau := node.tau.value;
  alpha := node.alpha.value;
  omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
  exp := Math.Exp(omega * (val - mu) * tau);
  IF (GraphStochastic.hint1 IN x.props) OR (GraphNodes.dataIN node.tau.props) OR (GraphNodes.data
IN node.alpha.props) THEN
    diffMu := node.mu.Diff(x);
    differential := diffMu * ((alpha * omega * tau * exp - 1) / (exp + alpha));
    ELSIF (GraphStochastic.hint1 IN x.props) OR (GraphNodes.dataIN node.mu.props) OR
(GraphNodes.dataIN node.alpha.props) THEN
    diffTau := node.tau.Diff(x);
    differential := diffTau * (((-alpha - 1) * omega * (val - mu) * exp) / (alpha * (exp / alpha + 1)) + (1/tau) +
omega * (val - mu));
    ELSIF (GraphStochastic.hint1 IN x.props) OR (GraphNodes.dataIN node.mu.props) OR
(GraphNodes.dataIN node.tau.props) THEN
    diffAlpha := node.alpha.Diff(x);
    differential := diffAlpha * (- (exp * (-1 - alpha)) / ((exp / alpha + 1) * alpha * alpha) - Math.Ln(exp /
alpha + 1));
  ELSE
    diffMu := node.mu.Diff(x);
    diffTau := node.tau.Diff(x);
    diffAlpha := node.alpha.Diff(x);
    differential := (diffMu * ((alpha * omega * tau * exp - 1) / (exp + alpha)) + (diffTau * (((-alpha - 1) *
omega * (val - mu) * exp) / (alpha * (exp / alpha + 1)) + (1/tau) + omega * (val - mu))) + (diffAlpha * (- (exp *
(-1 - alpha)) / ((exp / alpha + 1) * alpha * alpha) - Math.Ln(exp / alpha + 1))));
  END;
  RETURN differential;
END DiffLogLikelihood;

```

Figure 7. The code of the diffloglikelihood procedure in “MSNBurr2a.odc”

```

PROCEDURE(node: Node)DiffLogPrior(): REAL;
VAR
  differential, exp, mu, tau, alpha, omega, x: REAL;
BEGIN
  x := node.value;
  mu := node.mu.value;
  tau := node.tau.value;
  alpha := node.alpha.value;
  omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
  exp := Math.Exp(omega * (x - mu) * tau);
  differential := (- (alpha * omega * tau * (exp - 1)) / (exp + alpha));
  RETURN differential;
END DiffLogPrior;

```

Figure 8. The code of the difflogprior procedure in “MSNBurr2a.odc”

The *ExternalizeUnivariate* procedure, shown in Figure 9, functions to externalize or write the nodes of each parameter—namely α , μ , and τ ,—to be stored in Stores. Meanwhile, the *InternalizeUnivariate* procedure is used to internalize or read the nodes of each stored parameter. The *InitUnivariate* procedure initializes the nodes for the univariate distribution, with the initial value of each node being NIL.

```

PROCEDURE(node: Node) ExternalizeUnivariate (VAR wr: Stores.Writer)
BEGIN
  GraphNodes.Externalize(node.mu, wr);
  GraphNodes.Externalize(node.tau, wr);
  GraphNodes.Externalize(node.alpha, wr);
END ExternalizeUnivariate;

PROCEDURE(node: Node) InitUnivariate
BEGIN
  node.mu := NIL;
  node.tau := NIL;
  node.alpha := NIL;
END InitUnivariate;

PROCEDURE(node: Node) InternalizeUnivariate (VAR rd: Stores.Reader)
BEGIN
  node.mu := GraphNodes.Internalize(rd);
  node.tau := GraphNodes.Internalize(rd);
  node.alpha := GraphNodes.Internalize(rd);
END InternalizeUnivariate;

```

Figure 9. The code of externalizeunivariate, initunivariate, and internalizeunivariate procedures in “MSNBurr2a.odc”

The *Install* procedure in Figure 10 sets up the MSNBurr-IIa distribution, ensuring that the "MSNBurr2a.odc" file is automatically integrated when the "External.odc" file is compiled. The *Location** procedure retrieves the value of the location parameter, which is μ in the MSNBurr-IIa distribution.

```

PROCEDURE(node: Node) Install (OUTinstall: ARRAY OF CHAR)
BEGIN
  install := "GraphMSNBurr2a.Install"
END Install;

PROCEDURE(node: Node) Location(): REAL;
VAR
  mu: REAL;
BEGIN
  mu := node.mu.value;
RETURN mu;
END Location;

```

Figure 10. The code of install and location procedure in “MSNBurr2a.odc”

The *LogLikelihoodUnivariate* procedure in Figure 11 is used to compute the log-likelihood function of the probability density function, as specified in equation (33). This procedure ensures that the log-likelihood is correctly defined for univariate distributions, essential for accurate Bayesian computation.

```

PROCEDURE(node: Node) LogLikelihoodUnivariate(): REAL;
VAR
  logDensity, logTau, logOmega, mu, tau, alpha, x, omega: REAL;
BEGIN
  x := node.value;
  mu := node.mu.value;
  tau := node.tau.value;
  alpha := node.alpha.value;
  omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
  logTau := MathFunc.Ln(tau);
  logOmega := MathFunc.Ln(omega);
  logDensity := logOmega + logTau + omega*(x-mu)*tau - (1 + alpha) * Math.Ln(1.0+
(Math.Exp(omega*(x-mu)*tau)/ alpha));
RETURN logDensity;
END LogLikelihoodUnivariate;

```

Figure 11. The code of loglikelihoodunivariate procedure in “MSNBurr2a.odc”

The *LogPrior* procedure, as in Figure 12, defines the equation for the log-likelihood function of the probability density function, which is dependent on the variable X , as specified in equation (31). This procedure is crucial for incorporating prior distributions into the model, allowing for a

comprehensive Bayesian analysis by accounting for the influence of prior knowledge on the current data.

```

PROCEDURE(node: Node)LogPrior(): REAL;
VAR
  x, mu, tau, alpha, omega, logPrior: REAL;
BEGIN
  x := node.value;
  mu := node.mu.value;
  tau := node.tau.value;
  alpha := node.alpha.value;
  omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
  logPrior := omega*(x-mu)*tau - (1 + alpha) * Math.Ln(1.0 + (Math.Exp(omega*(x-mu)*tau)/alpha));
RETURN logPrior;
END LogPrior;

```

Figure 12. Source code prosedur logprior in “MSNBurr2a.odc”

The *ParentsUnivariate* procedure (Figure 13) adds the parents of each parameter node to a list. The *SetUnivariate* procedure is used to insert the nodes of each parameter into an array.

```

PROCEDURE(node: Node)ParentsUnivariate(all: BOOLEAN): GraphNodes.List;
VAR
  list: GraphNodes.List;
BEGIN
  list := NIL;
  node.mu.AddParent(list);
  node.tau.AddParent(list);
  node.alpha.AddParent(list);
RETURN list;
END ParentsUnivariate;

PROCEDURE(node: Node)SetUnivariate(IN args: GraphNodes.Args; OUT res: SET);
BEGIN
  res := {};
  WITH args: GraphStochastic.Args DO
    ASSERT(args.scalars[0]# NIL, 21);
    node.mu := args.scalars[0];
    ASSERT(args.scalars[1]# NIL, 21);
    node.tau := args.scalars[1];
    ASSERT(args.scalars[2]# NIL, 21);
    node.alpha := args.scalars[2];
  END;
END SetUnivariate;

```

Figure 13. The code of parentsunivariate and setunivariate procedure in “MSNBurr2a.odc”

The *Sample* procedure shown in Figure 14 defines the sampling method to be used in the estimation process. If the distribution has no constraints or domain, sampling will use the MSNBurr2a function from the "Randnum.odc" file in the “Math\mod” folder. If the distribution has constraints or a domain, sampling will use the MSNBurr2aLB, MSNBurr2aRb, or MSNBurr2aIB functions.

```

PROCEDURE(node: Node) Sample (OUTres: SET);
VAR
    mu, tau, omega, alpha, x, lower, upper: REAL;
    bounds: SET;
BEGIN
    mu := node.mu.value;
    tau := node.tau.value;
    alpha := node.alpha.value;
    bounds := node.props * {GraphStochastic.leftImposed, GraphStochastic.rightImposed};
    IF bounds = {} THEN
        x := MathRandnum.MSNBurr2a(mu, tau, alpha);
    ELSE
        node.Bounds(lower, upper);
        IF bounds = {GraphStochastic.leftImposed} THEN
            x := MathRandnum.MSNBurr2aLB(mu, tau, alpha, lower);
        ELSIF bounds = {GraphStochastic.rightImposed} THEN
            x := MathRandnum.MSNBurr2aRB(mu, tau, alpha, upper);
        ELSE
            x := MathRandnum.MSNBurr2aIB(mu, tau, alpha, lower, upper);
        END;
    END;
    node.value := x;
    res := {};
END Sample;

```

Figure 14. The code of the sample procedure in “MSNBurr2a.odc”

The New procedure (Figure 15) is responsible for creating and initializing a new instance of a Node object, which represents a key element in the probabilistic graphical model used by MultiBUGS. It allocates memory for the Node, calls its Init method to configure default properties, and returns the fully initialized object. This process is essential for defining new distributions, as each corresponds to a unique node in the model's computational graph. The Signature procedure, on the other hand, assigns a specific identifier ("sssCT") to the distribution or object, where "sss" indicates that the MSNBurr-IIa distribution has three parameters.

```

PROCEDURE(f: Factory) New (): GraphUnivariate.Node;
VAR
    node: Node;
BEGIN
    NEW(node);
    node.Init;
    RETURN node;
END New;

PROCEDURE(f: Factory) Signature (OUTsignature: ARRAY OF CHAR)
BEGIN
    signature := "sssCT";
END Signature;

```

Figure 15. The code of the new and signature procedure in “MSNBurr2a.odc”

The code in Figure 16 defines procedures for initializing and managing a module named GraphMSNBurr2a. The *Init* procedure creates a new instance of Factory, sets it to fact, and calls the Maintainer procedure to assign a version number and maintainer name. The *Install* procedure sets the fact object as the factory for GraphNodes, and the entire module begins by calling *Init* to initialize its components.

```

PROCEDUREInstall*;
BEGIN
  GraphNodes.SetFactory(fact);
END Install;

PROCEDUREMaintainer;
BEGIN
  version:= 500;
  maintainer := "Eliana P. Ramadani";
END Maintainer;

PROCEDUREInit;
VAR
  f: Factory;
BEGIN
  Maintainer;
  NEW(f);
  fact := f
END Init;

BEGIN
  Init
END GraphMSNBurr2a.

```

Figure 16. The code of procedure install and signature in “MSNBurr2a.odc”

In addition to creating the MSNBurr-IIa distribution module file, it is necessary to add and modify code in several other files to support the MSNBurr-IIa distribution module. Code additions to the "Randnum.odc" file (Figure 17) involve implementing four procedures for generating values distributed according to the MSNBurr-IIa distribution, which will then be used for sampling during estimation. The algorithm used in value generation employs the inverse transform method with the following steps:

1. Generate u from a Uniform(0,1) distribution.
2. Let $v = F(a) + [F(b) - F(a)]u$
3. The generated value is $x = F^{-1}(v)$,

where $F(x)$ is cumulative distribution function in equation (19), $F(a)$ is the cumulative probability of the function when x equals the lower bound of the distribution, $F(b)$ is the cumulative probability when x equals the upper bound, and $x = F^{-1}(v)$ represents the inverse function as defined in equation (18).

```

PROCEDUREMSNBurr2a*(mu, tau, alpha: REAL): REAL;
VAR
  u, omega: REAL;
BEGIN
  u := generator.Rand();
  omega := Math.Power((1 + 1/alpha), (1+alpha)) / Math.Sqrt(2 * Math.Pi());
  RETURN mu + (1/(omega*tau)) * (Math.Ln(alpha) + Math.Ln(Math.Power(1 - u, (-1/alpha) - 1)))
END MSNBurr2a;

```

Figure 17. The code of generate MSNBurr2a distribution in “Randnum.odc”

When a distribution is known to have a lower bound, an upper bound, or both, it will have a specific generating function for these conditions. The conditional distribution resulting from restricting the domain of certain other probability distributions is called a truncated distribution. The function for the truncated distribution is as follows:

$$f^*(x) = \begin{cases} \frac{f(x)}{F(b)-F(a)}, & a \leq x \leq b \\ 0, & \text{others} \end{cases} \quad (20)$$

To calculate $F(a)$ and $F(b)$, we need to compute the value of the CDF of MSNBurr-IIa. The code for calculating it is written in the file Cumulative.odc in the “Math\Mod” folder, as shown in Figure 18.

```

PROCEDURE MSNBurr* (mu, tau, alpha, x: REAL): REAL;
  VAR
    omega: REAL;
  BEGIN
    omega := Math.Power(1 + (1/alpha), (1 + alpha)) / Math.Sqrt(2.0 * Math.Pi());
    RETURN Math.Power(1 + (Math.Exp(-omega*(x-mu)*tau)/alpha), alpha);
  END MSNBurr;

```

Figure 18. The code of CDF MSNBurr2a distribution in “Cumulative.ode”

Therefore, the additions to the "Randnum.ode" file for truncated distributions are as follows: for truncation on the left side, truncation on the right side, and truncation on both sides (Figure 19-21).

1. The distribution is truncated on the left side

```

PROCEDURE MSNBurr2aLB* (mu, tau, alpha, lower: REAL): REAL;
  VAR
    u, pLower, omega: REAL;
  BEGIN
    pLower := MathCumulative.MSNBurr2a(alpha, mu, tau, lower);
    u := pLower + (1 - pLower) * generator.Rand();
    omega := Math.Power((1 + 1/alpha), (1+alpha)) / Math.Sqrt(2 * Math.Pi());
    RETURN mu + (1/(omega*tau)) * (Math.Ln(alpha) + Math.Ln(Math.Power(1 - u, (-1/alpha)) - 1));
  END MSNBurr2aLB;

```

Figure 19. The code of MSNBurr2aLB procedure in “Randnum.ode”

2. The distribution is truncated on the right side

```

PROCEDURE MSNBurr2aRB* (mu, tau, alpha, upper: REAL): REAL;
  VAR
    u, pUpper, omega: REAL;
  BEGIN
    pUpper := MathCumulative.MSNBurr2a(alpha, mu, tau, upper);
    u := pUpper * generator.Rand();
    omega := Math.Power((1 + 1/alpha), (1+alpha)) / Math.Sqrt(2 * Math.Pi());
    RETURN mu + (1/(omega*tau)) * (Math.Ln(alpha) + Math.Ln(Math.Power(1 - u, (-1/alpha)) - 1));
  END MSNBurr2aRB;

```

Figure 20. The code of MSNBurr2aRB procedure in “Randnum.ode”

3. The distribution is truncated on both sides, the left and right

```

PROCEDURE MSNBurr2aIB* (mu, tau, alpha, lower, upper: REAL): REAL;
  VAR
    u, pLower, pUpper, omega: REAL;
  BEGIN
    pLower := MathCumulative.MSNBurr2a(alpha, mu, tau, lower);
    pUpper := MathCumulative.MSNBurr2a(alpha, mu, tau, upper);
    u := pLower + (pUpper - pLower) * generator.Rand();
    omega := Math.Power((1 + 1/alpha), (1+alpha)) / Math.Sqrt(2 * Math.Pi());
    RETURN mu + (1/(omega*tau)) * (Math.Ln(alpha) + Math.Ln(Math.Power(1 - u, (-1/alpha)) - 1));
  END MSNBurr2aIB;

```

Figure 21. The code of MSNBurr2aIB procedure in “Randnum.ode”

Code additions to the "External.ode" file (Figure 22), located in the "Bugs\Mod" folder, are made to integrate the MSNBurr-IIa module so it can be accessed alongside other distributions. These additions are made within the *LOAD* procedure. The first addition includes the MSNBurr-IIa univariate distribution from the previously created module file. The second addition loads the function for the MSNBurr-IIa cumulative distribution function (CDF). The third addition loads the function for the MSNBurr-IIa probability density function (PDF). Finally, additional code is added to load the available deviance function.

```

Density("dmsnburr2a", "GraphMSNBurr2a.Install");
Function("pdf.msnburr2a", "GraphDensity.DensityUVInstall(dmsnburr2a)");
Function("cdf.msnburr2a", "GraphDensity.CumulativeInstall(dmsnburr2a)");
Function("dev.msnburr2a", "GraphDensity.DevianceUVInstall(dmsnburr2a)");

```

Figure 22. Adding MSNBurr-IIa code in “External.ode”

The code additions to "Linking.ode", as shown in Figure 23, are intended to ensure that the MSNBurr-IIa distribution module can be executed as part of a dynamic-link library. The compilation process depends on the operating system used, either Linux or Windows. In addition to modifying

"Linking.ode", code is also added to "Make.ode" (Figure 24) to ensure that the MSNBurr-IIa module can be compiled and linked correctly along with other programs. The modifications to both files involve including "GraphMSNBurr2a" (the name of the module) in the same section as the other distribution files. Both files are in "Developer" folder.

```
GraphRulesGraphNodes GraphGrammarGraphLogical GraphStochasticGraphLimits
GraphScalar GraphLinkfuncGraphVector GraphParamtransGraphWeight
GraphUnivariateGraphConjugateUVGraphVD GraphCensoringTruncation
GraphMultivariateGraphConjugateMVGraphChainGraphGMRFGGraphFunctional
GraphBlock GraphMessages GraphResources GraphMSNBurr2a
```

Figure 23. Adding graphMSnburr2a in "Linking.ode"

```
GraphBernGraphBinomialGraphCat GraphFounderGraphGeometric
GraphHypergeometricGraphMendelianGraphNegbin GraphPoisson GraphRecessive
GraphZipf GraphMSNBurr2a
```

Figure 24. Adding graphMSnburr2a in "Make.ode"

Code additions to the "Strings.ode" file, shown in Figure 25, are made to include a doodle for the MSNBurr-IIa distribution, facilitating easier visual modeling within MultiBUGS. This addition involves listing the parameters used in the MSNBurr-IIa module, as defined during the module design phase. Each parameter is associated with index 17, which corresponds to the assigned sequence number of the MSNBurr-IIa distribution within the system.

```
densities[17]  dmsnburr2a
param0[17]    location
param1[17]    rate
param2[17]    shape
default0[17]  0
default1[17]  1.0E-3
default2[17]  1
```

Figure 25. Adding code for MSNBurr-Iia module in "Strings.ode"

To integrate the MSNBurr-IIa distribution module, MultiBUGS needs to be recompiled. This is done by opening the "Make.ode" file and clicking the exclamation mark icon (typically used to initiate the build process). The MSNBurr-IIa distribution module will then be compiled along with other MultiBUGS components, successfully integrating it into the system. The complete procedure for compiling MultiBUGS source code is documented at: <https://github.com/MultiBUGS/MultiBUGS>.

After the module has been created and the necessary changes have been made to other related files, MultiBUGS should be recompiled using the "Make.ode" file. Once this is done, the MSNBurr-IIa distribution module will be successfully integrated into MultiBUGS and can be accessed at: <https://github.com/elianaputrir/MultiBUGS-MSNBurr2a>.

3.4. Module testing

The module testing phase consisted of five steps: testing of doodle, validation of probability values, validation of the simulation data density plot, validation of simulation data parameter estimation results, and comparison of DIC values. This testing ensured that the program ran correctly and produced the expected results without any detected errors.

Testing of doodle was conducted by running the MSNBurr-IIa distribution with the following prior distributions for each parameter: normal(0; 1.0×10^{-6}) for parameter μ , gamma (1000; 1000) for parameter τ , and uniform(0; 10) for parameters α . Figure 26 visualizes the MSNBurr-IIa distribution model. Based on this figure, the added MSNBurr-IIa distribution module functioned correctly and ran without any errors or warnings detected by MultiBUGS.

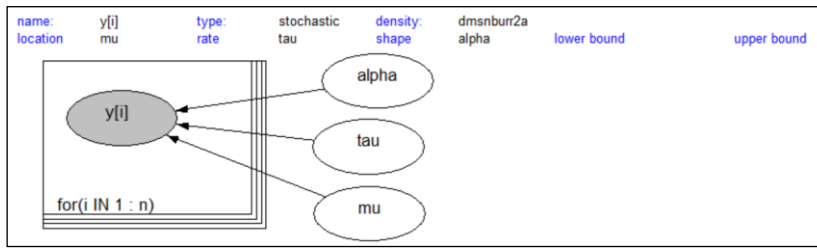


Figure 26. Visualization of trial model with doodle

Validation of the probability values was performed by comparing the probability density function (PDF) and cumulative distribution function (CDF) calculations of the MSNBurr-IIa distribution obtained from MultiBUGS [14] and neodistr package [19] in R that was published in CRAN. Ten sample points (x) were generated from the MSNBurr-IIa distribution in MultiBUGS using the parameters $\mu = 0$, $\sigma = 1$, and $\alpha = 0.1$. The PDF and CDF values were then calculated for these sample points using both MultiBUGS and the corresponding functions in the neodistr package (Table 1).

Table 1. Calculation of the probability value (PDF) and cumulative opportunity value (CDF) of the MSNBurr-IIa distribution using MultiBUGS and R functions

X	PDF			CDF		
	MultiBUGS	R	Absolute difference	MultiBUGS	R	Absolute difference
(1)	(2)	(3)	(4)	(5)	(6)	(7)
-4.750	0.03467	0.03467	0.00000	0.3191	0.3191	0.00000
-3.137	0.03750	0.03750	0.00000	0.3774	0.3774	0.00000
-10.65	0.02084	0.02085	0.00001	0.1545	0.1545	0.00000
-4.855	0.03446	0.03446	0.00000	0.3155	0.3154	0.00010
16.340	0.01020	0.01020	0.00000	0.9314	0.9313	0.00010
-8.602	0.02574	0.02574	0.00000	0.2022	0.2022	0.00000
-4.278	0.03359	0.03558	0.00001	0.3357	0.3357	0.00000
- 4.917	0.03433	0.03433	0.00000	0.3133	0.3133	0.00000
-1.452	0.03936	0.03936	0.00000	0.4423	0.4423	0.00000
11.670	0.01857	0.01857	0.00000	0.8655	0.8656	0.00010
Total			0.00002	Total		0.0003

This close agreement between the two methods indicates that the PDF and CDF values calculation in MultiBUGS is accurate. However, it's worth noting that this analysis was limited to 10 sample points. Further validation with a larger sample size could provide additional confidence in the accuracy of the MultiBUGS implementation.

3.5. The Implementation of simulation and real-world data

The simulated data were generated using MultiBUGS with a sample size of 1,000 across four distinct scenarios. The data of four scenarios are visualized using a density plot, as shown in Figure 27. This plot displays the distribution of the simulated data for each scenario. The x-axis represents the simulation data, and the y-axis represents the density. The shape of each curve reflects the probability density function of the MSNBurr-IIa distribution under the specific parameters of those scenarios.

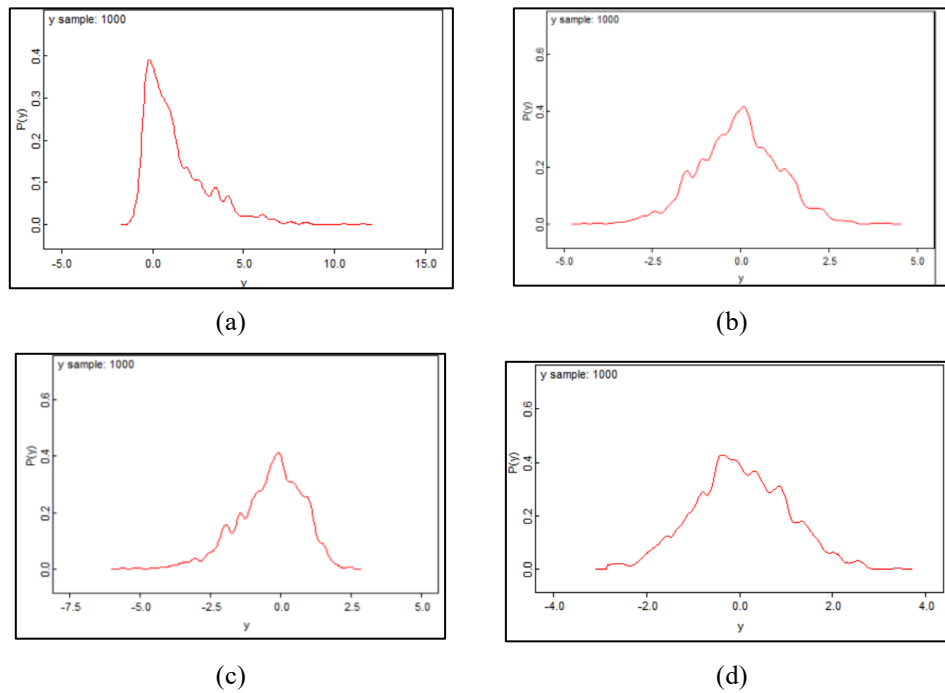


Figure 27. Density plot simulation data (a) scenario 1; (b) scenario 2; (c) scenario 3; (d) scenario 4

Figure 27 illustrates the density plots of four scenarios: three scenarios for the MSNBurr-IIa distribution with varying shape parameters (α), and one scenario as the normal distribution. Figures 27(a) through 27(d) display simulated data for right-skewed (scenario 1, $\alpha = 0.1$), symmetric (scenario 2, $\alpha = 1$), left skewed (scenario 3, $\alpha = 3$), and normal distribution, respectively. This is consistent with the theoretical characteristics of the MSNBurr-IIa distribution, that when $0 < \alpha < 1$ is a right-skewed distribution, $\alpha = 1$ is a symmetrical distribution, and $\alpha > 1$ is a left-skewed distribution. Consequently, the implemented module effectively captures and visualizes the distribution's flexible behavior

Parameter estimation value validation was conducted using four specifically designed simulated data scenarios. These scenarios encompassed various parameter combinations to assess the performance of the estimation procedure under diverse conditions. A simplified "trial model" was initially used for parameter estimation. This model focused on the core components of the MSNBurr-IIa distribution to ensure efficient convergence and accurate estimation of the key parameters. The following Markov Chain Monte Carlo (MCMC) settings were employed for parameter estimation:

Table 2. Estimated value of simulation data parameters

Scenario	Parameter	Simulation Parameter Value	Estimation MSNBurr-IIa parameter	Credible Interval	
				2.5%	97.5%
1	μ	0.0	-0.0397	-0.0976	0.0253
	τ	1.0	1.0050	0.9634	1.0490
	α	0.1	0.1054	0.0102	0.1185
2	μ	0.0	0.0154	-0.1075	0.1100
	τ	1.0	0.9738	0.9359	1.0130
	α	1.0	1.1730	0.9560	1.5210
3	μ	0.0	0.0207	-0.0839	0.1176
	τ	1.0	0.9711	0.9333	1.0090
	α	3.0	4.4510	2.5680	7.8240
4	μ	0.0	0.0598	-0.0463	0.1628
	τ	1.0	1.0450	0.9004	1.0880
	α	-	1.1530	0.8593	1.5330

1. Burn-in: 1,000 iterations. This initial period was discarded to remove samples that may not represent the target distribution.
2. Thinning: 1. Every sample from the chain was kept.
3. Refresh: 100. The chain was refreshed every 100 iterations to improve mixing.
4. Iterations: 10,000. A total of 10,000 iterations were run, resulting in 9,000 samples for analysis after the burn-in period. Table 2 Estimated value of simulation data parameters

The average time needed to estimate the four scenarios was 120.3 seconds. Table 2 shows the values of the three previously determined parameters, all falling within the 95% credible interval. This indicates that the developed MSNBurr-IIa distribution module can be used for parameter estimation and produces results as expected.

The Deviance Information Criterion (DIC) is used as an indicator for Bayesian model evaluation, such as in MultiBUGS. The model with the lowest DIC is generally preferred, as it suggests a better fit to the data. The MSNBurr-IIa distribution model's DIC will be compared to that of the normal distribution model when estimating parameters for the four simulated data scenarios.

Table 3. DIC value of MSNBurr-IIa distribution and normal distribution

Scenario	DIC value		Δ DIC
	MSNBurr-IIa	Normal	
1	3,461.0	4,202.0	741
2	3,128.0	3,140.0	12
3	3,088.0	3,169.0	81
4	2,882.0	2,858.0	-24

Table 3 shows that the Deviance Information Criterion (DIC) of the MSNBurr-IIa distribution was lower than that of the normal distribution in scenarios 1, 2, and 3. However, in scenario 4, the DIC of the MSNBurr-IIa distribution was higher than that of the normal distribution. This indicates that the MSNBurr-IIa distribution generally provided a better fit to the non-normal simulated data, with the greatest improvement in DIC observed for the right-skewed data in scenario 1.

The normal distribution outperformed the MSNBurr-IIa distribution in scenario 4 because the MSNBurr-IIa distribution tends to have heavier tails than the normal distribution, even when the overall shape is symmetric. This characteristic may lead to a slightly worse fit when the true data-generating process has lighter tails [1]. The testing process confirmed the consistency between the data characteristics and the estimation results. MultiBUGS had detected no errors in any of the testing steps. Therefore, it can be concluded that the MSNBurr-IIa distribution module created and added to MultiBUGS functions correctly and produces accurate results.

The MSNBurr-IIa distribution module and its source code have been compiled into a single MultiBUGS program available on GitHub <https://github.com/elianaputrir/MultiBUGS-MSNBurr2a/blob/main/MultiBUGS2.zip>. Users must install Microsoft MPI before running the compiled MultiBUGS program from the zip file.

The histogram of the economic growth rate data for districts/cities in 2021 reveals a right-skewed distribution (Figure 28). This indicates that most districts/cities have lower economic growth rates, with a few outliers exhibiting much higher rates. This observation is supported by the skewness test, which yielded a skewness value of 18.8663, significantly greater than 0, confirming the right-skewed nature of the distribution. Furthermore, the Anderson-Darling normality test confirms that the data is not normally distributed (p -value < 0.05). This non-normality suggests that standard statistical methods that assume normality may not be appropriate for analyzing this data, and alternative methods may need to be considered.

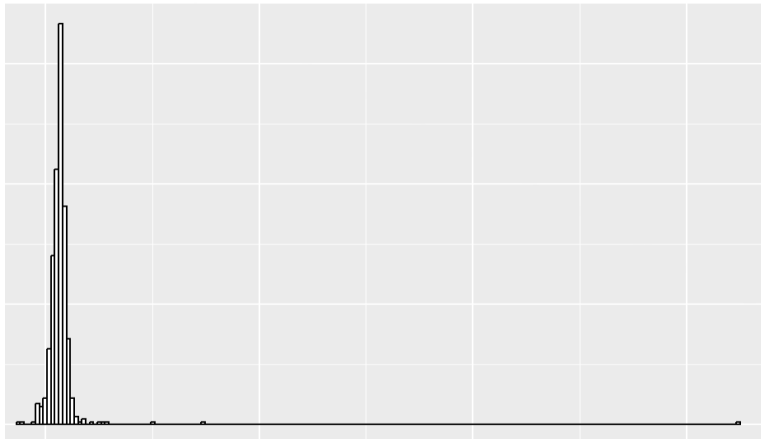


Figure 28. Histogram of district/city economic growth rate data for 2021

Parameter estimation was performed using 1,000 burn-in, 1 thin, 100 refresh, 10,000 iterations, and 9,000 samples. The prior for each parameter is normal (0; 1,0E-6) for μ , gamma (1000; 1000) for τ , and uniform (0,1; 10) for α .

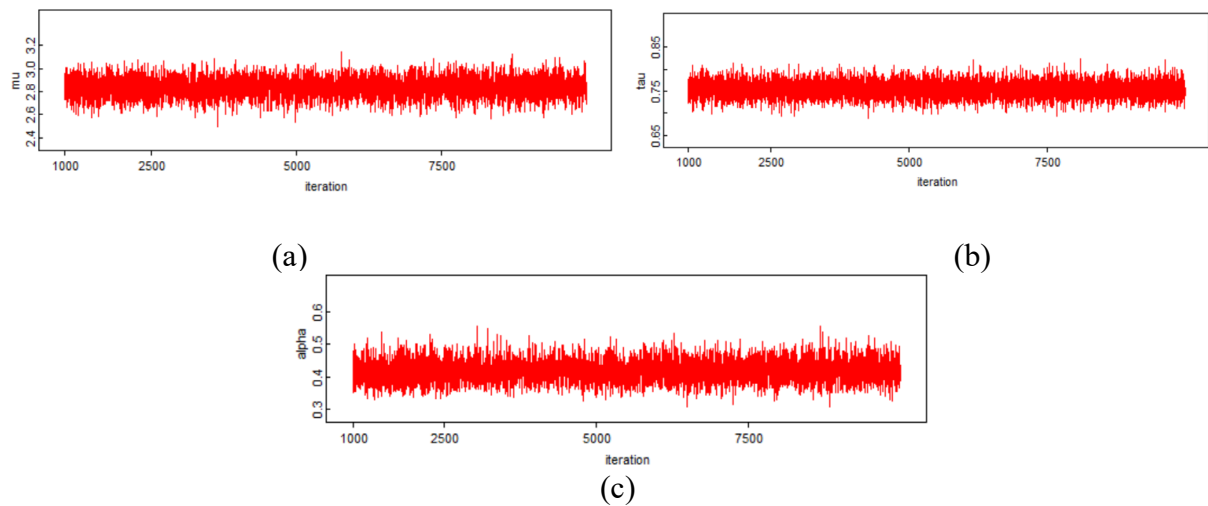


Figure 29. Trace plot of the estimated parameter of economic growth rate data for 2021: (a) μ ; (b) τ ; (c) α

Based on the trace plot of the 9,000 generated samples that were shown in Figure 29, the samples produced in the MCMC output are like fat caterpillars. It is shown that the parameter estimation process has reached a convergent condition.

Table 4. Summary of parameter estimates for the rate of economic growth data in 2021

Parameter	Estimation	Confidence Interval	
		2.5%	97.5%
μ	2.8360	2.6720	2.9940
τ	0.7541	0.7193	0.7908
α	0.4182	0.3567	0.4861

The parameters of the MSNBurr-IIa distribution for the economic growth rate data in 2021 were estimated using Bayesian inference with MCMC sampling in MultiBUGS. The estimation process took 49.938 seconds. Convergence of the MCMC chains was assessed using standard diagnostic tools to ensure reliable parameter estimates. Using the parameters estimated in Table 4, the mean, variance, skewness, and kurtosis of the MSNBurr-IIa distribution were derived and presented in Table 5. These characteristics provide a comprehensive summary of the distribution's shape and properties, which can be used to further analyze and interpret the economic growth rate data.

Table 5 shows that the estimated skewness of the fitted MSNBurr-IIa distribution is 1.0810, indicating that the distribution is right-skewed. This is consistent with the value of alpha in Table 5,

which is less than 1. This finding aligns with the right-skewed shape observed in the histogram of the economic growth rate data. The agreement between the estimated skewness and the visual representation of the data suggests that the MSNBurr-IIa distribution effectively captures the characteristics of the economic growth rate data for regencies in Indonesia.

Table 5. Characteristics of the MSNBurr-IIa distribution of economic growth data for 2021

Characteristics	Characteristic Value
Mode	3.4050
Variance	2.7789
Mode	2.8240
Skewnes	1.0810
Kurtosis	2.8590

4. Conclusion

This study successfully integrated the MSNBurr-IIa distribution into MultiBUGS by creating and compiling a dedicated distribution module. Rigorous testing confirmed the module's functionality and accuracy. The testing process included: ensuring correct computation of the CDF and PDF, confirming the module's ability to create doodle without errors, comparing the outputs with the expected characteristics of the MSNBurr-IIa distribution, assessing the model's fit against the normal distribution. The results consistently demonstrated the module's correct implementation and the accuracy of its calculations.

The MSNBurr-IIa distribution was then applied to analyze the economic growth rate data for regencies in Indonesia in 2021. The analysis revealed that the MSNBurr-IIa distribution effectively captured the characteristics of the data, particularly its right-skewness, as evidenced by the consistency between the estimated skewness, the histogram, and the skewness indicator. This successful application highlights the potential of the MSNBurr-IIa distribution for modeling and analyzing skewed data in various fields.

Future research can explore several directions to further enhance the applicability and utility of the MSNBurr-IIa distribution. One potential area is extending its integration into other Bayesian modeling platforms such as pyMC for broader accessibility. Additionally, further studies can investigate the performance of the MSNBurr-IIa distribution in different real-world datasets, particularly in financial risk modeling, survival analysis, and climate data analysis. Moreover, conducting simulations across various sample size scenarios—ranging from small-scale samples to large-scale datasets—would provide deeper insights into the model's robustness and efficiency.

Ethics approval

The study was conducted in accordance with the ethical guidelines, and informed consent was obtained from all individual participants included in the study.

Acknowledgments

In this study, we would like to express our gratitude to Politeknik Statistika STIS, and Universitas Padjadjaran for the support and resources they have provided. We would like to thank BPS-Statistics Indonesia for the publication of data.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

The research received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Eliana Putri Ramadani: Conceptualization, Methodology, Data Collection, Data Analysis, Writing–Original Draft, Software Development. **Achmad Syahrul Choir:** Conceptualization, Methodology, Manuscript Review, Research Advisor, Software Development. **Anindya Apriliyanti Pravitasari:** Editing, Writing–Review. **Joynabel Paraguison:** Editing, Writing–Review.

References

- [1] A. S. Choir, “The new neo-normal distributions and their properties”, Unpublished Dissertation, Institut Teknologi Sepuluh Nopember, Surabaya, 2020.
- [2] G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis*. Reading, Massachusetts: Addison Wesley, 1973.
- [3] A. Azzalini, “A class of distributions which includes the normal ones”, *Scand. J. Stat.*, vol. 12, no. 2, pp. 171–178, 1985.
- [4] M. Jorge and Iglewicz Boris, “Some properties of the Tukey g and h family of distributions”, *Commun. Stat. - Theory Methods*, vol. 13, no. 3, pp. 353–369, 1984, doi: 10.1080/03610928408828687.
- [5] J. A. Jiménez, V. Arunachalam, and G. M. Serna, “A generalization of tukey’s g-h family of distributions”, *J. Stat. Theory Appl.*, vol. 14, no. 1, pp. 28–44, 2017, doi: 10.2991/jsta.2015.14.1.3.
- [6] N. Eugene, C. Lee, and F. Famoye, “Beta normal distribution and its applications”, *Commun. Stat. - Theory Methods*, vol. 31, pp. 497–512, 2002.
- [7] M. A. Aljarrah, C. Lee, and F. Famoye, “On generating T-X family of distributions using quantile functions”, *J. Stat. Distrib. Appl.*, vol. 1, no. 1, p. 24, Dec. 2014, doi: 10.1186/2195-5832-1-2.
- [8] T. Amemiya, “Qualitative response models: a survey”, *J. Econ. Lit.*, vol. 19, no. 4, pp. 1483–1536, 1981.
- [9] N. Iriawan, *Computationally intensive approaches to inference in neo-normal linear models*. Perth, Australia: Disertasi Doktor, Curtin University of Technology, 2000.
- [10] Irving. W. Burr, “Cumulative frequency functions”, *Ann. Math. Stat.*, vol. 13, pp. 215–232, 1942, doi: 10.1214/aoms/1177731607.
- [11] A. Choir, N. Iriawan, B. Ulama, and M. Dokhi, “MSEPBurr distribution: properties and parameter estimation”, *Pak. J. Stat. Oper. Res.*, vol. 15, no. 1, pp. 179–193, 2019.
- [12] A. Ragab, J. R. Green, and M. C. K. Tweedie, “The burr type II distribution: properties, order statistics”, *Microelectron. Reliab.*, vol. 31, pp. 1181–1191, 1991, doi: 10.1016/0026-2714(91)90309-U.
- [13] W. M. Bolstad, *Introduction to bayesian statistics*. New York: John Wiley and Sons, 2007.
- [14] R. J. B. Goudie, R. M. Turner, D. D. Angelis, and A. Thomas, “MultiBUGS: a parallel implementation of the BUGS modeling framework for faster bayesian inference”, *J. Stat. Softw.*, vol. 95, pp. 1–20, Oct. 2020, doi: 10.18637/jss.v095.i07.
- [15] M. K. Cowles, “Review of WinBUGS 1.4”, *Am. Stat.*, vol. 58, no. 4, pp. 330–336, Nov. 2004, doi: 10.1198/000313004X8515.
- [16] E. Štrumbelj *et al.*, “Past, present and future of software for bayesian inference”, *Stat. Sci.*, vol. 39, no. 1, pp. 46–61, Feb. 2024, doi: 10.1214/23-STS907.
- [17] D. Lunn, C. Jackson, N. Best, A. Thomas, and D. Spiegelhalter, *The BUGS book: a practical introduction to bayesian analysis*. Boca Raton, FL: CRC Press, 2013.

- [18] J. Annis, B. J. Miller, and T. J. Palmeri, “Bayesian inference with Stan: a tutorial on adding custom distributions”, *Behav. Res. Methods*, vol. 49, no. 3, pp. 863–886, Jun. 2017, doi: 10.3758/s13428-016-0746-9.
- [19] A. S. Choir, A. Faoziah, and N. Iriawan, *Neodistr: neo-normal distribution*. [Online]. Available: doi:10.32614/CRAN.package.neodistr, 2024
- [20] O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesebeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, M. Osthege, R. Vieira, T. Wiecki, and R. Zinkov, “PyMC: a modern and comprehensive probabilistic programming framework in Python,” *PeerJ Computer Science*, vol. 9, p. e1516, 2023, doi: 10.7717/peerj-cs.1516.



Implementation of Twofold HB Beta SAE Model to Estimate Out-of-School Children with Disabilities in Indonesia

Aisha Maharani^{1*}, Azka Ubaidillah², Adhi Kurniawan³

^{1,2}Politeknik Statistika STIS, Jakarta, Indonesia, ³King's College London, London, United Kingdom

*Corresponding Author: E-mail address: aishaarani167@gmail.com

ARTICLE INFO

Article history:

Received 3 October, 2025

Revised 10 December, 2025

Accepted 23 December, 2025

Published 31 December, 2025

Keywords:

HB Beta; Out-of-School Children; Persons with Disabilities; SAE; Twofold Subarea-Level

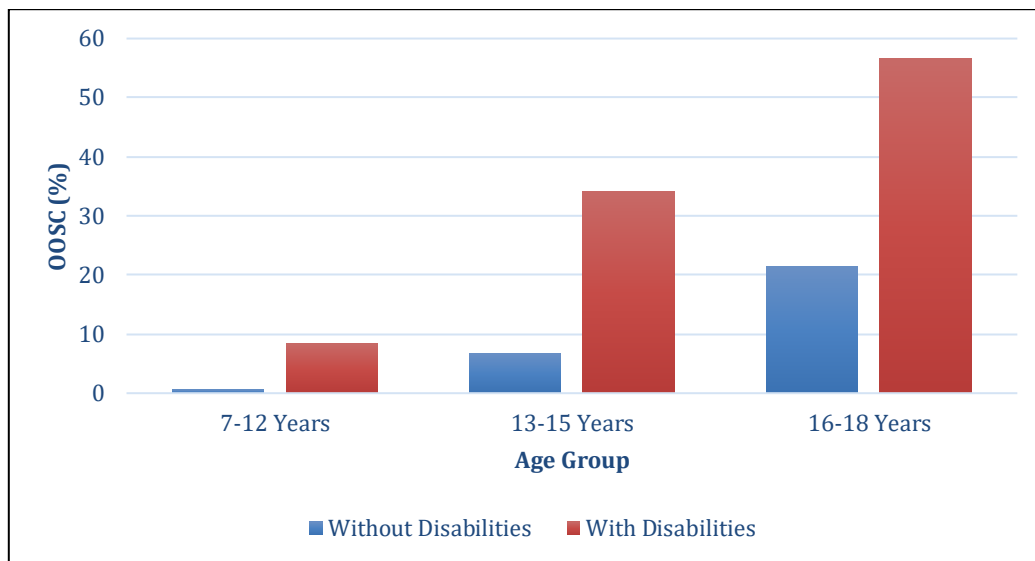
Abstract

Introduction/Main Objectives: The high percentage of out-of-school children with disabilities in Indonesia reveals a significant gap in educational participation. **Background Problems:** Due to the absence of disability-focused surveys, accurate data are only available at the national level, which is insufficient to represent regional conditions. **Novelty:** With the increasing demand for small area data, this study estimates the percentage of out-of-school children with disabilities at the provincial and district levels simultaneously, using small area estimation (SAE). **Research Methods:** This study applies SAE using a twofold subarea-level model with a Hierarchical Bayes (HB) beta approach, covering all 34 provinces and 514 districts/cities in Indonesia. This model was developed using data from the National Socio-Economic Survey (Susenas) and the Village Potential Statistics (Podes). **Finding/Results:** The twofold HB beta SAE model achieves higher precision than direct estimation, as shown by lower relative standard errors (RSE) across regions. Furthermore, spatial patterns indicate that the percentage of out-of-school children with disabilities is mostly between 35.36% and 45.34%, with clusters concentrated in Kalimantan and Papua.

1. Introduction

The rapid advancement of technology and information compels societies worldwide to adapt and innovate in response to continuous transformations [1]. Enhancing the quality of human resources is one of the key ways to address this dynamic global transformation [2]. This can be pursued, among other ways, through education [3]. Education is positioned as a primary focus within the Sustainable Development Goals (SDGs), specifically Goal 4, "Quality Education", which underscores the importance of inclusive and equitable education for all [4]. However, existing empirical evidence indicates that achieving this target remains challenging. UNESCO reported an increase of six million out-of-school children (OOSC) globally, bringing the total to 250 million in 2023 [5]. This figure accounts for 16 percent of the global child population [6]. Within this population of out-of-school children, children with disabilities constitute a particularly vulnerable group. They face greater barriers to accessing education due to physical, social, and institutional challenges. According to UNICEF, issues related to disability are not a new concern and have been a growing area of focus for more than a decade. In 2021, UNICEF released the most comprehensive global statistical analysis on children with disabilities, revealing that approximately 240 million children worldwide live with disabilities. The report highlights significant educational disparities, showing that children with disabilities are 49 percent more likely to have never attended school compared to their peers without disabilities [7]. This

evidence underscores that despite global commitments to inclusive and equitable education under SDG 4, the specific educational needs of children with disabilities remain insufficiently addressed.



Source: Education Statistics 2023 (BPS)

Figure 1. The percentage of out-of-school children by age group and characteristics

In the Indonesian context, horizontal inequality in school participation between children with and without disabilities remains a substantial issue. School participation, as an indicator of educational success, can be observed through the percentage of out-of-school children [8]. Out-of-school children refer to school-age children, generally aged 7–18 years, who have never attended school or who dropped out before completing their education [9]. Many children within the 7–18 age group in Indonesia still experience limited access to both formal and informal education [10]. Figure 1 shows the persistent gap in school participation between children with and without disabilities across all age groups in Indonesia. The percentage of out-of-school children is consistently higher among those with disabilities, indicating that school participation among children with disabilities remains significantly lower at every age level.

Persons with disabilities, as members of society, should be entitled to the same rights as others, including the right to education. Ensuring access to education for persons with disabilities is not only a constitutional mandate but also a crucial step toward achieving the fourth SDGs, which emphasizes quality, inclusive, and equitable education for all [11]. The Government of Indonesia has demonstrated a strong commitment through various policies and programs, including equivalency education, educational affirmative action programs, “Wajib Belajar 12 Tahun” program, and the implementation of inclusive and special education designed for learners with disabilities. These initiatives are expected to expand access and reduce educational disparities for persons with disabilities.

Thus, the availability of data on persons with disabilities is crucial for tracking progress toward the SDGs and other development goals. The United Nations also requires every country to collect accurate data on persons with disabilities [12]. Such data can be used to assess whether the rights of persons with disabilities, particularly the right to education, have been fulfilled. In addition, the World Health Organization (WHO) has urged countries to gather disability-related data in order to monitor their living conditions and anticipate potential gaps or inequalities [13].

In Indonesia, the availability of disability-related data is still highly restricted [14]. Data is needed not only to identify existing gaps but also as a basis for formulating evidence-based policies to meet the needs of persons with disabilities and address the inequalities they experience [15]. Data availability also enables monitoring of implemented policies so that their effectiveness in addressing existing issues can be evaluated. BPS Statistics Indonesia as the official data provider, does not conduct a survey specifically dedicated to persons with disabilities. However, disability data are collected by BPS through National Socio-Economic Survey (Susenas), which also includes information on indicators such as the percentages of out-of-school children with disabilities. However, BPS only publishes this indicator at the national level. Consequently, these figures are insufficient to accurately reflect the educational conditions of persons with disabilities across different regions. Each region naturally has its own unique conditions and challenges. Therefore, it is necessary to conduct independent estimation in order to

measure the rate of out-of-school children at smaller geographical levels. However, relying solely on Susenas data may lead to estimates with low precision due to insufficient sample sizes [16].

The problem of insufficient sample sizes can be addressed through indirect estimation methods, commonly known as Small Area Estimation (SAE). SAE is expected to enhance the effectiveness of available sample sizes, thereby producing estimates with higher precision without the need for additional sampling [17]. Generally, estimation using SAE is applied at only one area level. However, the demand for small-area data continues to increase, requiring information down to the smallest administrative units, such as districts. In this context, a model capable of producing estimates at multiple small-area levels simultaneously is required. Within SAE, such a model exists, namely the twofold subarea-level model. This model is an extension of the area-level Fay-Herriot model, which allows estimation at two small-area levels simultaneously, namely the subarea and area levels [17]. The twofold subarea-level model has been applied in several previous studies, such as Erciulescu et al. [18], Saadi & Ubaidillah [19], and Yudasena [20]. Other studies have employed the SAE method to estimate OOSC, including Eliezer et al. [21] and Trihandika et al. [22].

The main contribution of this study is to provide a more accurate indicator of OOSC among children with disabilities, which has so far been available only at the national level. Availability of data at the smallest administrative levels can support evidence-based policymaking and serve as a tool to monitor the effectiveness of government programs. Using the twofold HB beta SAE model also facilitates the estimation process, as it allows simultaneous estimation at two administrative levels. This approach helps reduce the time, effort, and costs typically required for separate estimation procedures. Furthermore, this study broadens the perspective on education issues for children with disabilities. Most previous research tends to focus on school dropout rates, whereas the OOSC indicator offers a more comprehensive view as it encompasses both children who have dropped out and those who have never attended school [23]. Therefore, this study aims to provide an overview of both direct and indirect estimates of out-of-school children with disabilities at the provincial and district/city levels in Indonesia, determine the best estimation method, and map the results.

2. Material and Methods

2.1. Persons with Disabilities

The concept of disability is highly complex and continues to evolve. The WHO has developed the International Classification of Functioning, Disability, and Health (ICF) framework as a reference for understanding and measuring health and disability [24]. This approach integrates various models of disability and acknowledges the role of environmental and personal factors in identifying disability status, as well as the relevance and impact of associated health conditions [25]. All components of disability and their interactions are illustrated in Figure 2.

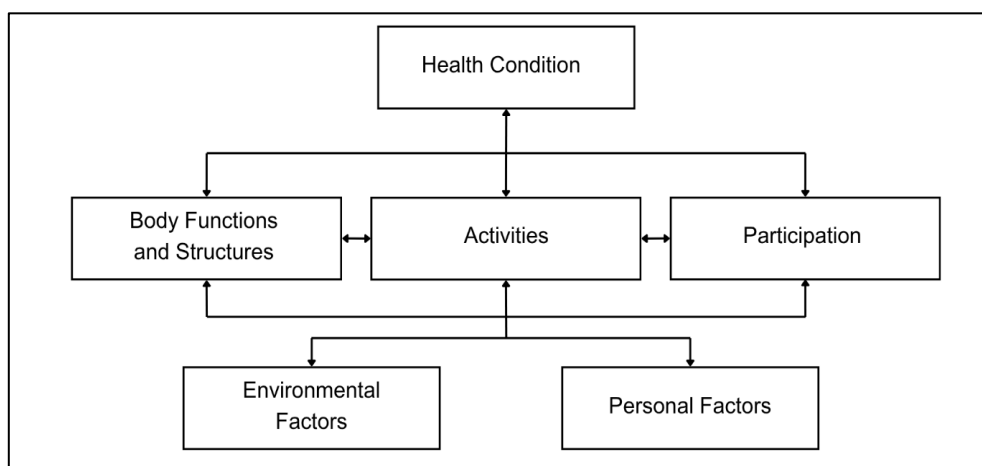


Figure 2. The ICF framework

Based on Figure 2, the ICF integrates health, individual, and social aspects, explaining the causal relationships between impairments and disabilities within a multidimensional model of functioning and health, including environmental and personal factors [26]. The concept of persons with disabilities according to BPS is aligned with the ICF framework. The emphasis in this concept is on long-term functional impairments or limitations that lead to restricted participation in society. Functional

impairments or limitations are characterized by an inability, loss, or abnormality in psychological, physiological, or anatomical structures or functions [23].

2.2. Out-of-School Children

Globally, there is no universally agreed definition of out-of-school children [27]. As part of a global initiative, the UNESCO Institute for Statistics and UNICEF define out-of-school children as school-aged children who are not enrolled in any educational setting, whether formal, nonformal, or informal [28]. This group includes children who have never attended school, those of school age who have not yet enrolled, and those who have attended school but dropped out for some reason [29]. BPS defines out-of-school children in the same way as international guidelines, namely school-aged children who are not enrolled in any educational unit [23]. BPS publishes the out-of-school children indicator to facilitate the identification of policy or program interventions specifically targeted at children who are not attending school. The calculation of out-of-school children with disabilities is carried out using the following formula.

$$OOSC \text{ with disabilities}_{7-18 \text{ years}} = \frac{\text{The number of children with disabilities aged 7-18 years who are out-of-school during a specific period}}{\text{The number of children with disabilities aged 7-18 years during the same period}} \quad (1)$$

2.3. Direct Estimation

According to Rao & Molina, direct estimation is a technique for estimating population parameters based on sample from the relevant area or domain [17]. here are two approaches to direct estimation: model-based and design-based. Generally, direct estimation uses a design-based approach, which involves estimating parameters such as means, totals, or proportions by calculating sample statistics using weights derived from the sampling design employed in the survey. However, direct estimation can also adopt a model-based approach, where population parameters are estimated from the sample without accounting for the sampling design. The model-based approach has a drawback: if the model does not align with the characteristics of the data, the estimates can be inaccurate, even with a larger sample size. Therefore, the design-based approach is more commonly used.

2.4. Small Area Estimation

Small Area Estimation (SAE) is a statistical technique used to estimate parameters in small domains. A domain is considered small if the available sample size within that domain is insufficient for producing precise direct estimates [17]. Domains can be defined as geographic areas, social or demographic groups, or other subpopulations. In this study, out-of-school children with disabilities constitute a subpopulation of the overall population of persons with disabilities. SAE can serve as a solution to address insufficient sample sizes. It is an indirect estimation model with an explicit linking model, meaning that the additional information comes not only from auxiliary variables but also from the variation between areas resulting from area-specific random effects. The random effects refer to the variation between areas that cannot be explained by the auxiliary information. Thus, there are two main concepts in SAE: the fixed effect model, which assumes that the variation in the response variable within small areas can be fully explained by auxiliary information, and the small area random effect, which accounts for variation in small areas that cannot be explained by auxiliary information, known as the area random effect. Based on the availability of auxiliary variables, there are two types of SAE models: area-level models and unit-level models [17]. Most researchers use area-level models because data are often not available at the unit level.

In the area-level model, auxiliary variables are available only at the level of the area under study. Let $\hat{\theta}_i$ represent a direct and unbiased estimate of the parameter θ_i . Since $\hat{\theta}_i$ contains sampling error, a sampling model can be formulated as follows.

$$\hat{\theta}_i = \theta_i + e_i ; i = 1, \dots, m \quad (2)$$

There is a correlation between θ_i and the auxiliary variable for each area, denoted by x_i . The linear model, serving as the linking model in the area-level model, is expressed as follows [17].

$$\theta_i = \mathbf{x}_i^T \boldsymbol{\beta} + b_i v_i ; i = 1, \dots, m \quad (3)$$

By combining the sampling model in equation (2) with the linking model in equation (3), the basic area-level model, commonly known as the Fay-Herriot model, is obtained [17].

$$\hat{\theta}_i = \mathbf{x}_i^T \boldsymbol{\beta} + b_i v_i + e_i ; i = 1, \dots, m \quad (4)$$

2.5. Twofold HB Beta Model

Essentially, the twofold subarea-level SAE model is an extension of the area-level model that can provide estimates for a given area and further divide it into smaller subareas. This model is referred to as such because auxiliary variables are available at the subarea level. Using this model, estimation can be performed simultaneously at different aggregation levels [17]. Suppose there are m areas, and each area is divided into N_i subareas with $i = 1, \dots, m$. If we aim to estimate the parameter at the area level θ_i and at the subarea level θ_{ij} where $j = 1, \dots, N_i$ and $i = 1, \dots, m$, the equation of the twofold HB beta SAE model can be written as follows [30].

$$i. \quad \hat{\theta}_{ij} | \theta_{ij} \sim \text{Beta}(a_{ij}, b_{ij}) \quad (\text{sampling model}) \quad (5)$$

$$a_{ij} = \theta_{ij} \left(\frac{n_i}{\text{def}_i} - 1 \right) \quad (6)$$

$$b_{ij} = (1 - \theta_{ij}) \left(\frac{n_i}{\text{def}_i} - 1 \right) \quad (7)$$

$$ii. \quad \text{logit}(\theta_{ij}) \sim N(\mathbf{x}_{ij}^T \boldsymbol{\beta}, \sigma_u^2) \quad (\text{linking model}) \quad (8)$$

$$iii. \quad \boldsymbol{\beta}_j \sim N(\mu_{\beta_j}, \sigma_{\beta_j}^2) \quad (\text{prior for } \boldsymbol{\beta}_j) \quad (9)$$

$$iv. \quad \sigma_u^2 \sim IG(t_1, t_2) \quad (\text{prior for } \sigma_u^2) \quad (10)$$

In this model, parameter estimation can be obtained numerically using the MCMC method with the Gibbs sampling algorithm. The Gibbs sampling estimation is performed for B iterations. The algorithm then produces a sequence of estimated parameters. Accordingly, the parameter estimates at the subarea level can be calculated as follows.

$$\hat{\theta}_{ij}^{HB} = \frac{1}{B} \sum_{k=d+1}^{d+B} \theta_{ij}^{(k)} = \theta_{ij}(\cdot) \quad (11)$$

The estimates at the area level can then be obtained by aggregating the subarea-level estimates using the weights available for each subarea, as follows.

$$\theta_i^{(k)} = \sum_{j=1}^{N_i} W_{ij} \theta_{ij}^{(k)} \quad (12)$$

$$\hat{\theta}_i^{HB} = \frac{1}{B} \sum_{k=d+1}^{d+B} \theta_i^{(k)} \quad (13)$$

2.6. Relative Standard Error

The quality of an estimate can be assessed by its total error, which reflects the difference between the estimated value and the true value. The smaller the total error, the more accurate the estimate, as it deviates less from the true value. Assuming bias is relatively small, accuracy can be approximated by precision. In this context, Relative Standard Error (RSE) is a measure of the precision of an estimate relative to its estimated value, calculated by comparing the standard error to the estimate and expressed as a percentage [23]. RSE can be calculated using the following equation.

$$RSE(\hat{\theta}) = \frac{SE(\hat{\theta})}{\hat{\theta}} \times 100\% \quad (14)$$

where $\hat{\theta}$ is the estimated value for a specific domain and $SE(\hat{\theta})$ is the measure of precision represented by standard error.

Sampling errors in the estimation results need to be considered. BPS sets RSE thresholds to assess the accuracy of an estimate: an estimate with $RSE \leq 25\%$ is considered accurate, an estimate with $RSE > 25\%$ and $\leq 50\%$ is considered moderately accurate and should be used with caution, and an estimate with $RSE > 50\%$ is regarded as highly inaccurate [23].

2.7. Analysis Method

This study was conducted to obtain estimates of out-of-school children with disabilities in Indonesia. This is motivated by the fact that BPS, as the data provider for out-of-school children with disabilities in Indonesia, is only able to publish data at the national level. Estimating out-of-school children with disabilities could be done using direct estimates from the March 2023 Susenas data. However, obtaining reliable estimates for the population of children with disabilities remains challenging. This is because the survey's minimum sample size was not designed for the disability group, so the small sample size results in estimates with relatively large errors.

One of the solution to address inaccurate estimates due to high relative standard errors (RSE) is to apply indirect estimation methods. Indirect estimation, which utilizes auxiliary information, is known as small area estimation (SAE). As the demand for small-area data continues to increase, requiring information down to the smallest administrative units, this study estimates out-of-school children with disabilities simultaneously at two levels of aggregation: provincial and district/city levels. Accordingly, the proposed model is a twofold subarea-level model using the Hierarchical Bayes (HB) estimation method with a beta distribution. After performing indirect estimation, the results will be evaluated by comparing the RSE values obtained. The estimation method that produces the lowest RSE will be selected as the best method for estimating out-of-school children with disabilities in Indonesia. Based on the best method, the estimated results will then be mapped.

The auxiliary variables must be sourced from census data or other datasets that are free from sampling error. At the time of the research, the only available and most recent source meeting this criterion was the 2021 Village Potential Statistics (Podes). There are 20 candidate auxiliary variables obtained from Podes 2021, including the proportion of villages/subdistricts classified as rural (X_1), the proportion of families living under high-voltage transmission lines (X_2), the proportion of families living along riverbanks (X_3), the proportion of families living in slum areas (X_4), the ratio of primary schools per 10,000 population (X_5), the ratio of junior high schools per 10,000 population (X_6), the ratio of senior high schools/vocational schools per 10,000 population (X_7), the ratio of special needs primary schools per 100 children with disabilities (X_8), the ratio of special needs junior high schools per 100 children with disabilities (X_9), the ratio of special needs senior high schools per 100 children with disabilities (X_{10}), the ratio of hospitals per 10,000 population (X_{11}), the ratio of community health centers per 10,000 population (X_{12}), the ratio of PKK per 10,000 population (X_{13}), the ratio of youth organizations per 10,000 population (X_{14}), the ratio of customary institutions per 10,000 population (X_{15}), the ratio of micro and small industries per 10,000 population (X_{16}), the ratio of financial institution facilities per 10,000 population (X_{17}), the ratio of cooperatives per 10,000 population (X_{18}), the ratio of economic facilities and infrastructure per 10,000 population (X_{19}), and the ratio of issued poverty certificates per 100 families (X_{20}).

3. Results and Discussion

3.1. Direct Estimate

The direct estimation of the number of out-of-school children with disabilities was obtained through a design-based direct estimation approach. The calculation process utilized sampling weights from the March 2023 Susenas and was conducted across 34 provinces in Indonesia. Figure 3 presents the visualization of the direct estimates of out-of-school children with disabilities at the provincial level in Indonesia for 2023.

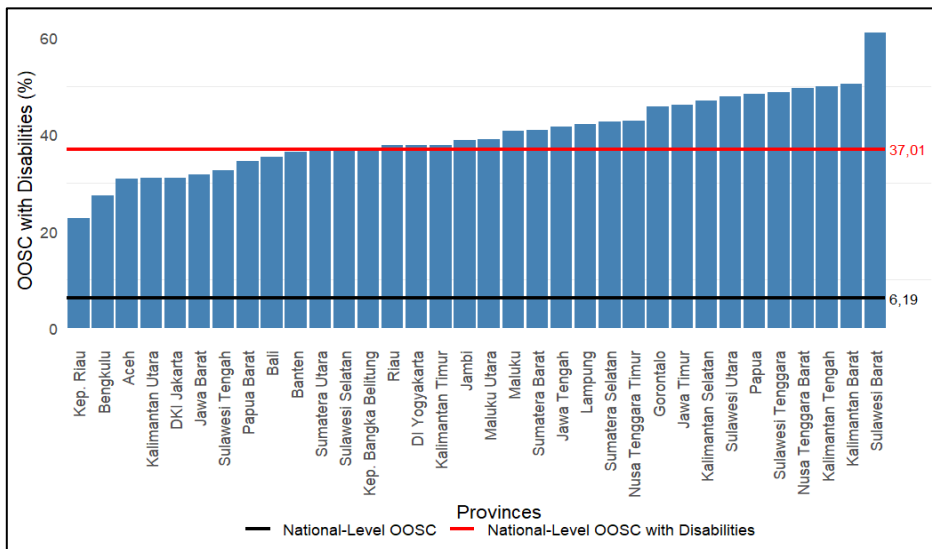


Figure 3. Direct estimates of out-of-school children with disabilities at the provincial level in Indonesia, 2023

Figure 3 shows the out-of-school children with disabilities across all provinces in Indonesia in 2023. The black horizontal line represents the national percentage of out-of-school for all children, regardless of disability status, which is 6.19%, while the red line indicates the national percentage of out-of-school for children with disabilities, which is 37.01%. It is evident that all provinces are above the national rate of 6.19%, indicating that children with disabilities face greater gaps and vulnerabilities in access to education compared to the general population. Moreover, more than half of the provinces have percentage of out-of-school for children with disabilities that exceed the national rate for this group. Only eleven provinces fall below the national percentage of out-of-school for children with disabilities. For a clearer overview of the direct estimates of out-of-school children with disabilities at the provincial level in Indonesia, descriptive statistics, along with their relative standard errors, are summarized in Table 1.

Table 1. Descriptive statistics of direct estimates of out-of-school children with disabilities at the provincial level in Indonesia, 2023

Descriptive Statistics	Direct Estimate (%)	RSE (%)
Minimum	22.64	13.51
First Quartile	35.68	24.29
Median	38.98	29.95
Mean	40.06	36.54
Third Quartile	46.05	43.65
Maximum	61.03	91.06

Based on Table 1, the lowest percentage of out-of-school children with disabilities is 22.64 percent, while the highest is 61.03 percent. The considerable gap between the highest and lowest values indicates significant disparities in education for children with disabilities. In addition to examining the direct estimates, it is also important to assess the accuracy of these estimates using the RSE values. According to Table 1, the highest and lowest RSE values for the direct estimates are 91.06 percent and 13.51 percent, respectively. The average RSE is 36.54 percent indicates that the estimates are not entirely accurate and, therefore, cannot yet serve as a strong basis for policy formulation.

The issue of out-of-school children with disabilities also needs to be analyzed at the district/city level. This aligns with the implementation of regional autonomy, whereby local governments are granted the rights and authority to manage their own affairs in accordance with the principles of decentralization. One of these affairs is education, which is primarily the responsibility of district/city governments. This means that policies related to education and interventions for out-of-school children are highly dependent on the capacity and commitment of the local government. By understanding the condition of out-of-school children with disabilities at a smaller level, namely districts/cities, local governments can formulate policies that are more responsive, data-driven, and tailored to local needs.

Direct estimates of out-of-school children with disabilities have also been conducted for 514 districts/cities in Indonesia. However, direct estimates at the district/city level are only available for 451 out of the 514 districts/cities. The remaining 63 districts/cities could not be directly estimated due to the absence of samples of children with disabilities in these areas. For a clearer understanding, the following is a visualization of districts/cities based on the availability of samples of children with disabilities.

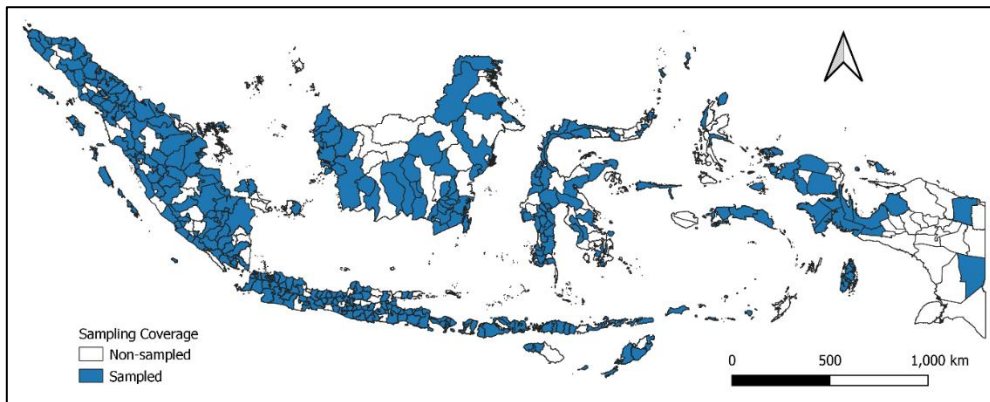


Figure 4. Map of districts/cities based on sample availability

As shown in Figure 4, there are several areas without samples of children with disabilities (non-sampled areas), indicated in white. A total of 63 districts/cities fall into the non-sampled category, and they are scattered across the country, though they tend to cluster in the northern part of Kalimantan and the western part of Papua. The existence of non-sampled areas reflects the limitations of surveys in reaching children with disabilities. This situation results in most regions lacking an adequate number of samples or even having no samples at all. Excluding these non-sampled areas, the following presents a summary of the descriptive statistics of the direct estimates of out-of-school children with disabilities at the district/city level in Indonesia.

Table 2. Descriptive statistics of direct estimates of out-of-school children with disabilities at the district/city level in Indonesia, 2023

Descriptive Statistics	Direct Estimate (%)	RSE (%)
Minimum	0	0
First Quartile	0	31.14
Median	18.77	60.58
Mean	30.36	58.47
Third Quartile	52.77	90.34
Maximum	100	137.99

Based on Table 2, the direct estimates of out-of-school children with disabilities range from a minimum of 0 percent to a maximum of 100 percent. There are 170 districts/cities with an estimated value of 0 percent, indicating that there were no recorded cases of children with disabilities being out of school. However, due to the limitations of the survey used, this value cannot be taken as conclusive evidence of the population condition of children with disabilities. In other words, it should not be interpreted as meaning that there are absolutely no out-of-school children with disabilities in those areas. Similarly, in regions with an estimated value of 100 percent, it does not imply that all children with disabilities are out of school. These extreme values rather reflect the limitations of the data in reaching small population groups.

Table 2 also shows that the minimum relative standard error (RSE) is 0 percent. An RSE value of zero does not indicate that the estimate is highly accurate or free from error; instead, it signals an anomaly in the data. Generally, an RSE of zero occurs when data variation cannot be calculated, usually due to very small and unrepresentative sample sizes. Moreover, the average RSE of 58.47 percent and the median RSE of 60.58 percent indicate that, overall, the quality of direct estimates at the district/city level is very inaccurate. An RSE exceeding 50 percent suggests that the estimates carry high uncertainty or are highly unreliable, making them unusable and uninterpretable. This is further reinforced by the

maximum RSE value of 137.99 percent, which indicates extremely high estimation uncertainty, far beyond statistically acceptable limits.

Table 3. Number of regions by RSE category

Category	District/City	Province
$RSE \leq 25\%$	62	11
$25\% < RSE \leq 50\%$	51	16
$RSE > 50\%$	159	7
NA	242	0

Based on the direct estimation results (Table 3), only 62 districts/cities demonstrate good accuracy. In fact, there are 242 districts/cities with RSE values reported as not available (NA). At the provincial level, most estimates are also relatively inaccurate. Only 11 provinces have accurate estimates, as indicated by RSE values less than or equal to 25 percent.

3.2. Auxiliary Variable Selection

Before entering the SAE modeling stage, auxiliary variables are selected by first examining the relationship between candidate auxiliary variables and the response variable. The selection of auxiliary variables is crucial because choosing the appropriate auxiliary variables is an important component for the success of indirect estimation using SAE [17]. One of the methods to statistically confirm the existence of relationships between variables is through Pearson correlation testing. In this study, the twofold HB beta SAE model is applied, so the candidate auxiliary variables are tested for their correlation with the logit-transformed direct estimates. Pearson correlation serves as an effective preliminary screening tool to identify variables that exhibit meaningful linear associations with the transformed response. It is considered sufficient at this stage because the purpose is only to filter out clearly irrelevant predictors before applying more comprehensive model-based evaluations [31]. Based on the correlation test results, six auxiliary variables were found to have a significant correlation with the logit of the direct estimator.

After conducting the correlation tests, it is necessary to ensure that there is no correlation among auxiliary variables to prevent multicollinearity in the model. Multicollinearity detection is carried out by examining Pearson correlation values. Collinearity occurs when the correlation between auxiliary variables exceeds 0.8 [32]. The results of multicollinearity detection among the auxiliary variables indicate that no variable has a correlation greater than 0.8 with another variable. Therefore, no auxiliary variables were eliminated, and all can proceed to the next stage.

The final step is variable selection using the backward elimination method. This approach is consistent with the study by Mellinda & Sumarni, which stated that backward elimination in SAE modeling can produce better results [33]. The variable selection process ultimately retained four variables: the ratio of junior high schools per 10,000 population (X_6), the ratio of public health centers per 10,000 population (X_{12}), the ratio of youth organizations per 10,000 population (X_{14}), and the ratio of issued poverty letters/SKTM per 100 families (X_{20}). These four variables are able to reflect various dimensions of life that are relevant in explaining the conditions of out-of-school children.

The ratio of junior high schools per 10,000 population (X_6) represents the educational dimension, while the ratio of public health centers per 10,000 population (X_{12}) represent the health dimension. Adequate availability of and access to schools and healthcare facilities can support the learning process of children optimally, both in terms of physical and mental readiness [34]. The ratio of youth organizations per 10,000 population (X_{14}) represent the social dimension. The organization is oriented toward achieving social welfare for the community, thereby serving as an inclusive and open space for all groups, including children with disabilities. Such social support can enhance the self-confidence and courage of children with disabilities to interact with the outside world, including in the context of attending school. The ratio of issued poverty letters per 100 families (X_{20}) represent the economic dimension. Economic limitations can increase the risk of children with disabilities being out of school [35].

3.3. Development SAE using Twofold HB Beta Model

In parameter estimation using the HB beta approach, calculations are based on the posterior distribution, which is derived from the prior distribution and the likelihood function. The resulting

posterior equation is highly complex, making direct integration extremely difficult. One method that can address this complexity is the Markov Chain Monte Carlo (MCMC) method. The MCMC method produces an accurate posterior distribution once it reaches equilibrium, that is, when the algorithm has converged [36]. In this study, several trials were conducted to achieve convergence, specifically using 60,000 iterations, 30 update iterations, a thinning interval of 25, and a burn-in of 15,000.

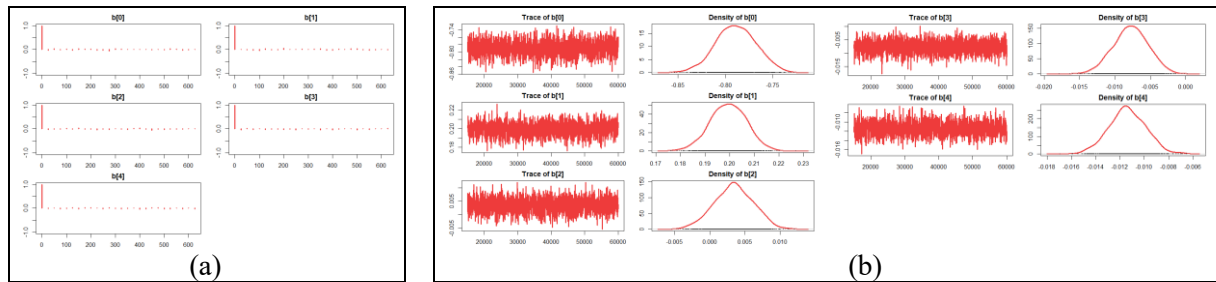


Figure 5. (a) Autocorrelation plot; (b) Trace plot and density plot

According to the Figure 5, convergence was confirmed by autocorrelation plots showing a clear cut-off, stationary trace plots without periodic patterns, and density plots for each parameter resembling a normal distribution. After obtaining convergence results for the four auxiliary variables included in the model, the following presents the estimated parameter coefficients from the twofold HB beta SAE model for out-of-school children with disabilities, with subareas defined as districts/cities and areas defined as provinces.

Table 4. Estimated parameter coefficients of the twofold HB beta SAE model

Estimated Parameter Coefficients	Mean	SD	2.5%	97.5%
b_0	-0.789	0.021	-0.835	-0.748
b_1	0.200	0.007	0.187	0.215
b_2	-0.003	0.002	-0.009	-0.001
b_3	-0.008	0.003	-0.013	-0.003
b_4	0.011	0.001	0.008	0.014

Based on Table 4, it can be seen that the parameter estimates associated with the variables used, namely X_6 , X_{12} , X_{14} , and X_{20} , are significant in the twofold HB beta model. Significance in the model is assessed using the credible interval ranging from 2.5 percent to 97.5 percent, where a variable is considered significant if the credible interval within this range does not include zero. An overview of the indirect estimation results of out-of-school children with disabilities using the twofold HB beta SAE model is presented in Table 5.

Table 5. Descriptive statistics of indirect estimate of out-of-school children with disabilities in Indonesia, 2023

Descriptive Statistics	District/City (%)	Province (%)
Minimum	12.43	28.07
First Quartile	36.70	38.12
Median	42.28	42.20
Mean	42.70	41.88
Third Quartile	48.58	46.06
Maximum	82.49	60.71

Based on Table 5, the indirect estimates of out-of-school children with disabilities in Indonesia in 2023 using the twofold HB beta SAE model show an average of 42.70 percent at the district/city level and 41.88 percent at the provincial level. At the district/city level, the lowest estimate is 12.43 percent in Banda Aceh City, while the highest is 82.49 percent in Lamandau Regency. At the provincial level, the lowest estimate is 28.07 percent in Aceh Province, and the highest is 60.71 percent in West Sulawesi Province.

3.4. Evaluation

The evaluation was carried out by comparing the estimates of out-of-school children with disabilities obtained from both direct estimation and indirect estimation using the twofold HB beta SAE model. The indirect estimates from the twofold HB beta SAE model are considered reliable and unbiased when their values are not significantly different from those of the direct estimates [37]. The following presents a comparative statistical summary of the estimates of out-of-school children with disabilities derived from direct and indirect estimation.

Table 6. Descriptive statistics of direct and indirect estimate of out-of-school children with disabilities in Indonesia, 2023

Descriptive Statistics	District/City	Province		
	Direct Est. (%)	Indirect Est. (%)	Direct Est. (%)	Indirect Est. (%)
Minimum	0	12.43	22.64	28.07
First Quartile	0	36.70	35.68	38.12
Median	18.77	42.28	38.98	42.20
Mean	30.36	42.70	40.06	41.88
Third Quartile	52.77	48.58	46.05	46.06
Maximum	100	82.49	61.03	60.71
NA	63	0	0	0

Table 6 shows a noticeable difference between the direct estimates and the indirect estimates using the twofold HB beta SAE model in estimating out-of-school children with disabilities, particularly at the district/city level. The most striking differences are observed in the minimum and maximum values. At the provincial level, no striking differences are observed; however, there is still a change in the average value, from 40.06 percent to 41.88 percent.

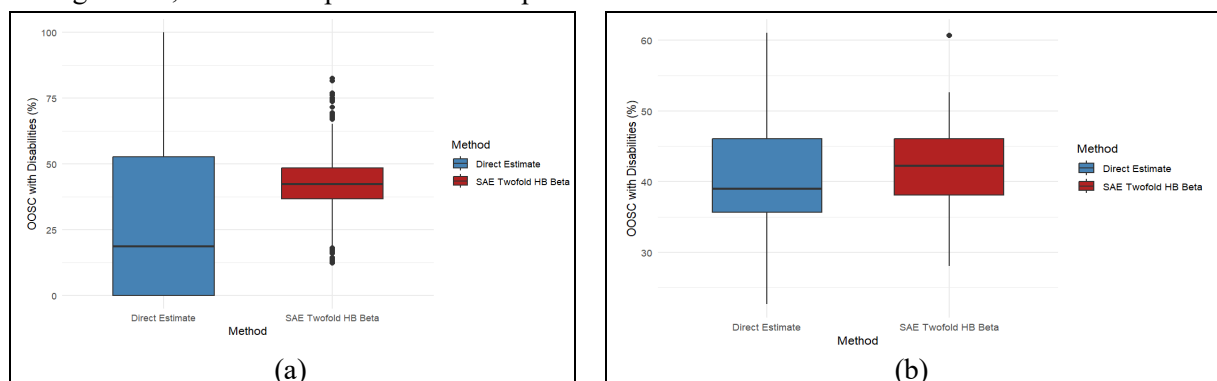


Figure 6. Boxplot of direct and indirect estimates of out-of-school children with disabilities at (a) the district/city level; (b) the provincial level in Indonesia, 2023

Further examination through the data distribution using the boxplot in Figure 6 reveals that the estimates from the twofold HB beta SAE model still fall within the main distribution range of the direct estimates, indicating that the differences do not exceed the extreme limits of data dispersion. The notable differences arise from the limitations of the direct estimation method, which is highly influenced by the number and representativeness of the sample in reflecting the actual population conditions. Direct estimates tend to be less accurate due to these sample constraints, whereas the SAE model is able to correct such issues. To determine the most appropriate estimation method for estimating out-of-school children with disabilities, an evaluation must be conducted based on the accuracy level of the estimates. In this study, the measure of accuracy used is the relative standard error (RSE). Presented below is the number of regions classified by RSE category.

From the comparison of RSE categories in Table 7, it can be observed that the use of the twofold HB beta SAE model is able to reduce RSE values and address the problem of non-sampled areas. Even at the provincial level, all 34 provinces have RSE values below 25 percent, indicating that the estimates are accurate. Most notably, the SAE method also succeeded in providing estimates for all areas without a single missing value. These findings reinforce that the twofold HB beta SAE model is not only superior in producing more accurate estimates but also capable of generating estimates for all areas, including

non-sampled ones. Its ability to lower RSE and produce comprehensive estimates makes this method the most appropriate choice for addressing the problem of limited sample data.

Table 7. Number of regions by RSE category

Category	District/City		Province	
	Direct Est. (%)	Indirect Est. (%)	Direct Est. (%)	Indirect Est. (%)
$RSE \leq 25\%$	62	70	11	34
$25\% < RSE \leq 50\%$	51	406	16	0
$RSE > 50\%$	159	38	7	0
NA	242	0	0	0

3.5. Mapping of the Estimated OOSC with Disabilities

To observe the distribution of out-of-school children with disabilities (OOSC) in Indonesia more clearly, a thematic map is used. Since there is no specific guideline for classifying OOSC with disabilities, the classification is carried out using three groups generated through the natural breaks classification (Jenks optimization method). The natural breaks classification method divides groups by minimizing within-group variation and maximizing between-group variation, making it suitable for data that are not normally distributed [38]. The use of the same classification for OOSC with disabilities at both the district/city and provincial levels aims to enable better comparison across aggregation levels. Based on the natural breaks classification with three groups, the resulting breakpoints are 35.36 percent and 45.34 percent. The first group consists of areas with OOSC with disabilities of less than 35.36 percent. The second group consists of areas with OOSC values ranging from a minimum of 35.36 percent to a maximum of 45.34 percent. Finally, the third group consists of areas with OOSC values above 45.34 percent.

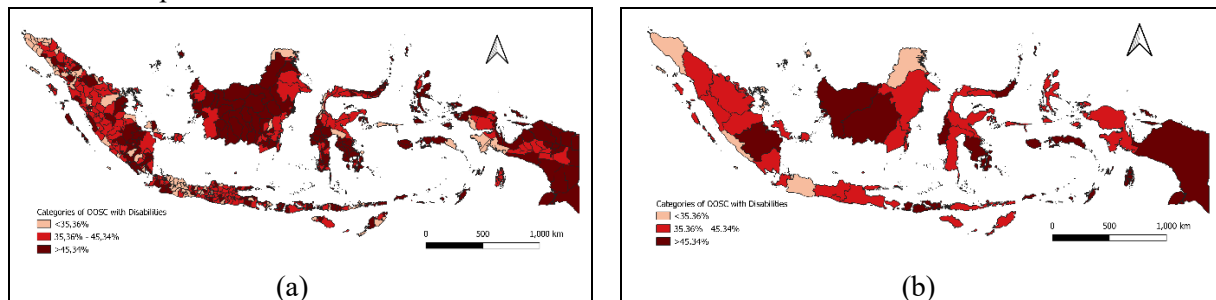


Figure 7. Thematic map based on estimates of out-of-school children with disabilities at (a) the district/city level; (b) the provincial level in Indonesia, 2023

Higher percentages of out-of-school children with disabilities are represented by darker colors, and vice versa. Based on Figure 7, most areas, both at the provincial and district/city levels, fall within the 35.36 to 45.34 percent category, while 195 districts/cities and 9 provinces have values exceeding 45.34 percent. This indicates that the issue of out-of-school children with disabilities is not confined to specific regions but is a widespread challenge across Indonesia. The dominance of areas in the 35.36 to 45.34 percent range reflects the systemic nature of limited educational access for children with disabilities, while the existence of regions in the highest category underscores more concerning conditions that demand greater attention and more serious interventions in policy planning and implementation.

As shown in the thematic map in Figure 7, the highest values of out-of-school children with disabilities tend to cluster in Kalimantan and Papua. This is closely related to geographical factors, which serve as a major barrier to educational access in these areas. Many regions in eastern Indonesia, such as Papua, Maluku, and several remote parts of Kalimantan, are characterized by challenging topography, including mountainous areas or small isolated islands. These conditions hinder access to educational services, particularly for children with disabilities who require mobility support and special accommodations [39]. Katheryn Bennet, representing UNICEF Indonesia, also stated that the high rate of out-of-school children in Indonesia is caused by various factors, such as economic pressures that push children into work, limited geographical access, and other barriers [40]. For children with disabilities, these barriers are often exacerbated by the family's low economic condition, which prevents them from affording additional needs such as special transportation, learning aids, or companion costs. As a result,

schooling is no longer seen as a main priority, and many of them instead choose to work to help support their family's economy [41].

The Government of Indonesia has also introduced various policies to address the issue of out-of-school children with disabilities. However, in practice, implementation on the ground remains uneven. For instance, in the provision of inclusive education, many regular schools still lack disability-friendly facilities such as wheelchair ramps, special toilets, or visual/auditory aids. In addition, the shortage of Special Education Support Teachers poses a serious challenge in supporting the learning process of children with disabilities [42]. Furthermore, data collection on children with disabilities remains suboptimal, hindering evidence-based and well-targeted policy planning.

The estimation results indicate that addressing out-of-school children with disabilities requires more than education policies alone. A multisectoral approach is necessary, involving education, social affairs, health, infrastructure, and other sectors. Key measures include improving disability-friendly infrastructure, providing social protection such as financial aid and learning support, and enhancing educational quality through special education teachers, training, and inclusive curricula. These efforts must be coordinated across national and regional agencies and supported by accurate, up-to-date data to monitor progress, evaluate policies, and design programs that reflect the real needs of the field.

4. Conclusion

Overall, the twofold HB beta SAE model is the most suitable method for estimating the number of out-of-school children with disabilities in Indonesia at both provincial and district/city levels, as it addresses sample limitations and improves estimation accuracy. The estimated proportion ranges from 35.36 to 45.34 percent, with the highest concentrations observed in Kalimantan and Papua. These results provide evidence for local governments to support inclusive education policy formulation and evaluation. More targeted regional analysis and improved inter-agency data integration are needed, especially in areas with complex socio-geographic conditions.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data from BPS Statistics Indonesia.

Credit Authorship

Aisha Maharani: Conceptualization, Methodology, Writing, Editing, Visualization. **Azka Ubaidillah:** Reviewing, Validation, Supervision. **Adhi Kurniawan:** Reviewing, Editing, Validation.

References

- [1] D. R. Taneo, S. E. P. R. A. Ratu, and D. R. . Lodang, "Persepsi peserta didik terhadap penggunaan teknologi pembelajaran pada mata pelajaran seni budaya di Sekolah Menengah Kejuruan Negeri 2 Kupang," *Metta J. Ilmu Multidisiplin*, vol. 4, no. 3, pp. 132–143, Oct. 2024, doi: <https://doi.org/10.37329/metta.v4i3.3607>.
- [2] N. Larasati and I. Rindaningsih, "The quality improve of education: implementation, development, and human resources," *Int. J. Multidiscip.*, vol. 1, no. 1, 2024, [Online]. Available: <https://journal.antispublisher.com/index.php/ijmi>
- [3] UNDP, *Human development report 1990*. New York: Oxford University Press, 1990.
- [4] Bappenas, "Sustainable development goals 4: pendidikan berkualitas," <https://sdgs.bappenas.go.id/17-goals/goal-4/>.
- [5] UNESCO, "UNESCO: 250 million children now out of school," Sep. 2023. Accessed: Jan. 20, 2025. [Online]. Available: <https://news.un.org/en/story/2023/09/1140882>
- [6] UNESCO, "250 million children out-of-school: what you need to know about UNESCO's latest education data," Sep. 2023. Accessed: Jun. 13, 2025. [Online]. Available: <https://www.unesco.org/en/articles/250-million-children-out-school-what-you-need-know-about-unescos-latest-education-data>
- [7] BPS, *Statistik pendidikan 2023*, vol. 12. Badan Pusat Statistik, 2023.
- [8] I. Rosmilawati and M. F. T. Mutaqin, "Eradicating out-of-school-children: strategies and best practices from alternative education context," *JPPI (Jurnal Penelit. Pendidik. Indones.)*, vol. 10, no. 2, p. 364, Jun. 2024, doi: 10.29210/020244055.
- [9] Y. Anggrayni, "Major influencing factors of out-of-school in West Java Province, Indonesia," *Asia-Japan Res. Acad. Bull.*, vol. 5, no. 12, Jul. 2024, doi: https://doi.org/10.34389/asiajapanbulletin.5.0_12.
- [10] I. R. Marfu'ah, D. N. Yasmara, and M. N. F. Al Amin, "Analisis program pendidikan inklusif bagi disabilitas di Kota Kediri dengan perspektif problem tree analysis," *J. Sos. Dan Hum.*, vol. 1, no. 4, pp. 178–189, May 2024, doi: 10.62017/arima.
- [11] United Nations, "Convention on the rights of persons with disabilities and optional protocol," 2006.
- [12] E. A. Lauer and A. J. Houtenville, "Estimates of prevalence, demographic characteristics and social factors among people with disabilities in the USA: a cross-survey comparison," Feb. 01, 2018, *BMJ Publishing Group*. doi: 10.1136/bmjopen-2017-017828.
- [13] United Nations, "Laporan PBB menyoroti kesenjangan data disabilitas di Indonesia," 2022. [Online]. Available: <https://indonesia.un.org/id/209911-laporan-pbb-menyoroti-kesenjangan-data-disabilitas-di-indonesia>
- [14] B. Berlian and A. Firdaus, "Good governance melalui kebijakan berbasis bukti reformulasi sistem peradilan pidana nasional," *J. Yuridis*, vol. 9, no. 1, pp. 27–36, Jun. 2022, doi: 10.35586/jjur.v9i1.3847.
- [15] A. H. Hakim and N. Hajarisman, "Pendugaan rata-rata pengeluaran per kapita menurut Kabupaten/Kota di Provinsi Jawa Barat melalui empirical best linear unbiased prediction dalam pendugaan area kecil," *Bandung Conf. Ser. Stat.*, vol. 2, no. 2, pp. 474–481, Aug. 2022, doi: 10.29313/bcss.v2i2.4747.
- [16] J. N. K. Rao and I. Molina, *Small area estimation*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2015.
- [17] A. L. Erciulescu, N. B. Cruze, and B. Nandram, "Model-based county level crop estimates incorporating auxiliary sources of information," *J. R. Stat. Soc. Ser. A Stat. Soc.*, vol. 182, no. 1, pp. 283–303, Jan. 2019, doi: 10.1111/rssa.12390.
- [18] R. Saadi and A. Ubaidillah, "Pendugaan area kecil angka partisipasi kasar perguruan tinggi di Indonesia menggunakan model subarea twofold dengan pendekatan hierarchical bayes," *Semin. Nas. Off. Stat.*, vol. 2022, no. 1, pp. 123–136, Nov. 2022, doi: <https://doi.org/10.34123/semnasoffstat.v2022i1.1217>.
- [19] R. F. Yudasena, "Penerapan SAE twofold HB beta dengan kecamatan sebagai subarea dan kabupaten/kota sebagai area (studi kasus estimasi persentase anak miskin tingkat Kecamatan dan Kabupaten/Kota di Jawa Tengah pada tahun 2022)," *Jakarta Politek. Stat. STIS*, 2024.
- [20] W. R. Eliezer *et al.*, "Small area estimation for percentage of out-of-school children aged 7-17 years in Sumatera Island, 2023," *J. Mat. Stat. dan Komputasi*, vol. 21, no. 1, pp. 215–233, Sep. 2024, doi: 10.20956/j.v21i1.36043.
- [21] L. F. Trihandika, A. Ubaidillah, A. Az-Zahra, A. Amirudin, M. Rusydiana, and Z. Maharani, "Small area estimation anak tidak sekolah di Pulau Kalimantan tahun 2023," *Limits J. Math. Its*

- Appl.*, vol. 21, no. 2, p. 273, Oct. 2024, doi: 10.12962/limits.v21i2.21497.
- [22] BPS, *Buku 4 Konsep dan definisi susenas maret 2023*. Jakarta, 2022.
 - [23] UNICEF, *Definition and Classification of Disability*. New York, 2014.
 - [24] H. Anttila, S. Tallqvist, M. Muñoz, S. Leppäjoki-Tiistola, O. Mäkitie, and S. Hiekkala, "Towards an ICF-based self-report questionnaire for people with skeletal dysplasia to study health, functioning, disability and accessibility," *Orphanet J. Rare Dis.*, vol. 16, no. 1, Dec. 2021, doi: 10.1186/s13023-021-01857-7.
 - [25] S. van der Veen, N. Evans, M. Huisman, P. Welch Saleeby, and G. Widdershoven, "Toward a paradigm shift in healthcare: using the International Classification of Functioning, Disability and Health (ICF) and the capability approach (CA) jointly in theory and practice," *Disabil. Rehabil.*, vol. 45, no. 14, pp. 2382–2389, 2023, doi: 10.1080/09638288.2022.2089737.
 - [26] S. Mitra, S. K. Mishra, and R. K. Abhay, "Out-of-school girls in India: a study of socioeconomic-spatial disparities," *GeoJournal*, vol. 88, no. 1, pp. 341–357, Feb. 2023, doi: 10.1007/s10708-022-10579-7.
 - [27] UNICEF and UNESCO Institute for Statistics, *Operational manual: global out-of-school children initiative*. UNICEF, 2015. [Online]. Available: www.unicef.org/education/
 - [28] A. Shanker, D. Marian, and C. Swimmer, "Effective interventions aimed at reaching out-of-school children a literature review," Kathmandu, 2015. Accessed: Apr. 23, 2025. [Online]. Available: <https://files.eric.ed.gov/fulltext/ED573790.pdf>
 - [29] M. Torabi and J. N. K. Rao, "On small area estimation under a sub-area level model," *J. Multivar. Anal.*, vol. 127, pp. 36–55, May 2014, doi: 10.1016/j.jmva.2014.02.001.
 - [30] W. Ren, J. Li, A. Erciulescu, T. Krenzke, and L. Mohadjer, "A variable selection method for small area estimation modeling of the proficiency of adult competency," pp. 689–713, 2022, [Online]. Available: <https://www.mdpi.com/2571-905X/5/3/41>
 - [31] D. Gujarati, D. Porter, and S. Gunasekar, *Basic Econometrics*. 2013.
 - [32] Mellinda and C. Sumarni, "Estimation of total fertility rate (TFR) using small area estimation (SAE) in Nusa Tenggara Timur (NTT) Province," *Proc. Int. Conf. Data Sci. Off. Stat.*, vol. 2021, no. 1, pp. 796–809, 2022, doi: <https://doi.org/10.34123/icdsos.v2021i1.107>.
 - [33] G. S. Becker, *Human capital: a theoretical and empirical analysis with special reference to education (third edition)*, 3rd ed. The University of Chicago Press, 1993.
 - [34] UNESCO Institute for Statistics and UNICEF, *Fixing the broken promise of education for all: findings from the global initiative an out-of-school children*. Montreal: UNESCO Institute for Statistics, 2015. doi: 10.15220/978-92-9189-161-0-en.
 - [35] A. Ubaidillah, "Small area estimation dengan pendekatan hierarchical bayesian neural network untuk pemetaan kemiskinan di Kota Jambi (Tesis S2)," *Inst. Teknol. Sepuluh Novemb.*, 2014, [Online]. Available: <https://www.researchgate.net/publication/340172483>
 - [36] Eurostat, *Guidelines on small area estimation for city statistics and other functional geographies*. European Union, 2019. doi: 10.2785/763467.
 - [37] A. Shafira, F. Kristiani, and B. Yong, "Penerapan metode klasifikasi perangkat lunak arcmap pada pemetaan penyebaran penyakit dengue di Bandung," *Limits J. Math. Its Appl.*, vol. 20, no. 1, p. 39, Mar. 2023, doi: 10.12962/limits.v20i1.9226.
 - [38] Bappenas, *Pedoman teknis penyusunan rencana aksi tujuan pembangunan berkelanjutan (TPB) / sustainable development goals (SDGs)*, II. Kedeputan Bidang Kemaritiman dan Sumber Daya Alam, Kementerian Perencanaan Pembangunan Nasional / Badan Perencanaan Pembangunan Nasional, 2020. Accessed: Apr. 23, 2025. [Online]. Available: <https://sdgs.bappenas.go.id/website/wp-content/uploads/2020/10/Buku-Pedoman-Rencana-Aksi-SDGs.pdf>
 - [39] R. Luthfia, "Mengatasi masalah anak tidak sekolah melalui pendidikan nonformal," PSKP Kemendikdasmen. [Online]. Available: <https://pskp.kemendikdasmen.go.id/kabar/detail/mengatasi-masalah-anak-tidak-sekolah-melalui-pendidikan-nonformal>
 - [40] M. I. Nasrallah, "Opportunities and challenges to support out- of-school children and youth through accelerated learning programmes: case study of Jordan," 2022, [Online]. Available: <https://inee.org/evidence/e-cubed>.
 - [41] P. Lydiasari, "Sekolah inklusi di Indonesia baru 14,8 persen yang memiliki guru pembimbing khusus," Jakarta, Apr. 2024. Accessed: Jun. 13, 2025. [Online]. Available: <https://jakarta.suaramerdeka.com/pendidikan/13412344129/sekolah-inklusi-di-indonesia-baru-148-persen-yang-memiliki-guru-pembimbing-khusus>



Spatial Heterogeneity of Food Security in Indonesia: Unpacking the Roles of Technology and Democracy Index

Ditto Satrio Wicaksono¹, Novi Hidayat Pusponegoro^{2*}, Arbi Setiyawan³

¹BPS-Statistics Bulukumba Regency, Bulukumba, Indonesia, ²Politeknik Statistika STIS, Jakarta, Indonesia,

³School of Management, Jiangsu University, Zhenjiang, China

*Corresponding Author: E-mail address: novie@stis.ac.id

ARTICLE INFO

Article history:

Received 11 June, 2025

Revised 10 December, 2025

Accepted 22 December, 2025

Published 31 December, 2025

Keywords:

Democracy; Food Security;
ICT; Indonesia; Spatial
Heterogeneity; Urbanization

Abstract

Introduction/Main Objectives: Food security is a key concern for all countries, especially Indonesia. Technological development and democratic quality are vital for sustainable food security. This study aims to determine the impact of technology and democracy on food security. **Background Problems:** The relationship between food security and these two factors remains uncertain. Moreover, the extant literature on the spatial impacts on food security yields results that are inconclusive. **Novelty:** This study offers a comprehensive depiction of the impact of spatial relationships between variables, with a particular focus on the quality of democracy and technology, on the multidimensionality of food security. **Research Methods:** A spatial lag model is applied to ascertain the impact of technological and democratic on multidimensional food security using data from 34 provinces in 2022. **Finding/Results:** The results reveal significant spatial dependence in Indonesia's food security. Technological development and democratic quality positively and significantly affect food security, while urbanization and food crop land expansion show negative and positive effects, respectively. Spatial spillover accounts for approximately 37%–38% of the total impact of each explanatory variable. These findings suggest that technology adoption, democratic strengthening, and interprovincial collaboration are crucial for improving food security.

1. Introduction

Food security represents a critical and multifaceted challenge with the potential to significantly hinder a nation's development and compromise its national security. At its core, food security is a region's ability to produce enough food to meet its own consumption needs [1] and to ensure that all people have consistent material and economic access to sufficient, nutritious, and safe food [2]. Alarmingly, a report from the Food and Agriculture Organization (FAO) indicates that approximately 783 million individuals globally faced hunger in 2022, with about one in ten experiencing severe food insecurity [3], [4]. The projected increase in the global population is expected to further amplify food demand and exacerbate existing food insecurity [5]. Consequently, food security is a crucial issue, explicitly recognized within the Sustainable Development Goals (SDGs), particularly for developing nations such as Indonesia [6].

Indonesia, as an emerging nation, confronts a persistent challenge in achieving comprehensive food security. Despite its classification as an upper middle-income country, Indonesia grapples with

significant impediments to adequate food access, further compounded by the triple burden of malnutrition, encompassing the co-existence of undernutrition, overnutrition, and widespread micronutrient deficiencies. Statistical evidence from the *Badan Pusat Statistik* (BPS) reveals that approximately 4.85 percent of Indonesian households experienced moderate to severe food insecurity in 2022 [7]. This situation is underscored by a concurrent rise in undernourishment, escalating from 7.63 percent in 2019 to 10.21 percent in 2022, with a concerning prevalence across geographical regions, as 23 out of the nation's 34 provinces exhibited undernourishment rates exceeding the national average [8]. Furthermore, the most recent findings from the *Riset Kesehatan Dasar* (RISKESDAS) highlight the pervasive nature of micronutrient deficiencies, particularly affecting vulnerable demographic groups including children aged 0-59 months, adolescent girls, and pregnant and lactating women [9]. When viewed within a global context, Indonesia's 2022 rankings of 63rd in food security and 77th in hunger [10], [11]. Adding to the complexity of ensuring future food security for its populace, Indonesia faces formidable challenges posed by uncertain climate change patterns, ongoing environmental degradation, and the pressures associated with a large and rapidly growing population. Consequently, there is a compelling imperative for intensified and strategically focused interventions to effectively address the multifaceted and persistent challenges to food security across the Indonesian landscape.

A number of factors have been identified as potentially influential on food security, including technology and democracy. Technological development is a continuous and rapid process, with applications in all areas of life, including the food sector. [12] revealed that the application of technology (such as artificial intelligence (AI), blockchain technologies, mobile applications, Internet of Things (IoT), big data, and drones) in the food sector has been carried out in many ways, ranging from controlled environment agriculture, farming automation, and genetic editing. The utilization of such technologies has been demonstrated to enhance agricultural productivity [13], optimize supply chain efficiency [14], stabilize food prices [15], and improve the quality of food produced [16]. Hence, the advancement of technology is regarded as a pivotal solution for the establishment of sustainable food security [17], [18]. Furthermore, the capacity of technology to innovate, optimize, and revolutionize traditional practices offers a plethora of opportunities to mitigate the risks of food insecurity. However, in Indonesia, the rate of technology adoption in the agricultural sector remains low [19], [20]. This is further compounded by the disparate development of technology across different regions. Another pivotal factor that merits greater scrutiny in the pursuit of food security is the quality of democracy. Democracy is defined as institutional arrangements that enable individuals to engage in political processes freely [21]. In this sense, democracy is closely linked to governance and institutions. In the context of a flourishing democracy, the resultant governance and institutions are characterized by transparency and accountability, thereby fostering a foundation for political stability and policies that support food security [22], [23], [24]. Conversely, when democracy is poor, it engenders political tensions, government bureaucratic incompetence, and corruption, which hinder the achievement of food security [25]. As [26] have observed, a nation that respects democratic rules has never suffered from a famine. This perspective is predicated on the notion that the safeguarding of economic and social rights, including the right to food, is contingent on the promotion of political and civil liberties. While the academic literature acknowledges a correlation between the quality of democracy and the prevalence of food security, the precise nature and mechanisms underpinning this relationship remain to be systematically delineated. To address this gap in understanding, this study poses the following research questions: (1) Does technological development exert a significant influence on food security outcomes in Indonesia? (2) What specific role does democratic governance play in shaping food security within the Indonesian context?

A considerable number of studies have already been conducted on the subject of food security. However, the majority of these have adopted an approach that has been critiqued as one-sided [27] or have focused exclusively on agricultural performance [28], [29], which limits their ability to accurately depict food security conditions in their totality. Achieving sustainable food security necessitates consideration of all pertinent factors. The Food and Agriculture Organization (FAO) has identified four interrelated elements of food security: availability, affordability, utilization, and stability. Despite the existence of numerous studies that employ indicators encompassing these four dimensions, these studies have not incorporated spatial effects in their modelling [30], [31], [32]. Indeed, it is imperative to incorporate spatial effects into the modelling of the relationship between explanatory variables and response variables when location is designated as the unit of analysis [33]. This necessity arises from the recognition that distinct regions possess unique food security characteristics, which can exert

influence on one another. Moreover, with the advent of rapid developments in the social economy, logistics and transportation, information dissemination, and modernization of agricultural equipment, there has been a significant increase in the flow of food production, distribution, and consumption between regions [34]. As a result, spatial dependence on food security in various regions has become increasingly prominent. Consequently, the achievement of food security is inextricably linked to spatial considerations. However, despite the crucial concern of food security in Indonesia, scholarly attention to its nuanced dimensions, particularly spatial heterogeneity, remains limited. Moreover, the determinants of food security, such as the roles of technology and governance, are also underexplored within the Indonesian context.

In light of the aforementioned observations, the research objective of this study is to empirically estimate the relationships among food security, technology, and democratic governance, while also examining the spatial effects inherent within these relationships across the diverse landscape of Indonesia. This study contributes to the existing body of knowledge by employing food security indicators that comprehensively encompass all relevant dimensions, thereby providing a more nuanced understanding of the condition. To achieve these objectives, a spatial model that accounts for area-specific effects is applied. This methodological choice enables the determination of both direct and indirect relationships between the variables under investigation, thereby promising to yield more insightful research findings. Additionally, the significance of this study lies in its potential to offer crucial perspectives and evidence-based insights for planners and policymakers, facilitating the formulation of effective strategies aimed at achieving sustainable food security throughout Indonesia.

The remainder of the article is divided into three sections. Section 2 provides a detailed discussion of the research method, including the data that was employed and the analytical tools that were applied. Section 3 contains the results and further discussion. Finally, Section 4 offers a conclusion in the form of a synopsis of the article's findings and a set of policy recommendations.

2. Material and Methods

2.1. Data

This study was constructed using secondary data encompassing 34 Indonesian provinces in 2022. That year was chosen because Indonesia's food security situation is poor globally. Furthermore, the year 2022 provides a clearer picture of post-pandemic recovery and the challenges that remain. The dependent variable applied is a food security index created from three dimensions: availability, affordability, and utilization [35]. The employment of these three dimensions is predicated on their consideration of the availability of data at the smallest geographical level. The study incorporates the Information and Communication Technology Development Index and the Indonesian Democracy Index as the primary variables to represent technological development and the quality of democracy, respectively. Concomitantly, the study includes urbanization and agricultural land utilization as additional explanatory variables. These two variables were selected on the grounds that they are capable of describing the conditions from upstream to downstream with regard to food security, particularly in light of the challenges currently faced by Indonesia. The rapid process of urbanization, particularly in developing countries such as Indonesia, exerts a significant influence on the demand for food, consequently affecting food consumption and distribution patterns. Conversely, the density of food crop land emerges as a pivotal factor in determining agricultural productivity. Moreover, in Indonesia, with its substantial population and rapid growth, pressure on agricultural land is increasing. A comprehensive description and the data sources utilized are presented in Table 1.

Table 1. Detail of research variables

Variable	Definition	Unit	Source
Food Security Index (FSI)	An index that measures the achievement of food security development in a region on a scale of 0 to 100	Point	Ministry of Agriculture
Information, technology and communication development index (ICT)	An index that measures the standardised level of ICT development in a region on a scale of 0 to 10	Point	BPS-Statistics of Indonesia
Indonesia democracy index (IDI)	A composite indicator that shows the level of democratic development in Indonesia on a scale of 0 to 100	Point	BPS-Statistics of Indonesia
Urbanization (URB)	Ratio of urban population to total population	Percent	BPS-Statistics of Indonesia
Food crop land density (FCLD)	Ratio of harvested area of food crops (rice, corn, soybean, sweet potato, and cassava) to the total area of the province	Percent	Ministry of Agriculture and BPS-Statistics of Indonesia

2.2. Spatial Autocorrelation Presence Test

The analytical apparatus employed in this study is spatial area modelling, a methodology that utilizes the unit of analysis as a location, thereby enabling the consideration of spatial effects within the model. The initial step involves ascertaining the existence of spatial correlation. Spatial autocorrelation emerges when there is a correlation among the values of a single variable that is exclusively attributable to its relatively proximate location. The Global Moran's I and Geary's C indices are generally applied to this assessment, with the calculation formula presented below [36], [37].

$$\text{Moran's I} = \frac{n \sum_{i=1}^n \sum_{j=1}^n W_{ij} (y_i - \bar{y}) (y_j - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n \sum_{j=1}^n W_{ij}} \quad (1)$$

$$C^* = \frac{(n-1) \sum_{i=1}^n \sum_{j=1}^n W_{ij} (y_i - y_j)^2}{2 \sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n \sum_{j=1}^n W_{ij}} \quad (2)$$

where n represents total regions, w_{ij} is the spatial weight matrix between the i^{th} and j^{th} location, y and $\bar{y}$ are the dependent variable and the average value of the dependent variable, respectively. After the Moran's I index is normalised by variance, the value will be between -1 and 1. $I > 0$ indicates positive spatial correlation, the larger the value, the more obvious the spatial correlation; $I < 0$ indicates negative spatial correlation, the smaller the value, the greater the spatial difference; $I = 0$, spatial randomness. The value of Geary's C index is generally between 0 and 2. When $C > 1$, it means negative correlation, and when $C < 1$, it means positive correlation.

The significance of Moran's I and Geary's C is determined by computing a p-value and a z-score based on the null hypothesis of no spatial dependence (i.e., complete spatial randomness). If the p-value is greater than 0.05, we fail to reject the null hypothesis, thereby suggesting that the spatial distribution of the values could be attributable to random chance. Conversely, if the p-value is less than 0.05, the null hypothesis is rejected, indicating the presence of spatial dependence.

2.3. Spatial Area Modelling

The spatial economics literature provides two main approaches to modelling spatial correlation [37]. Spatial dependence is theorized as a long-run equilibrium of underlying spatial-temporal processes. Spatial dependence based on time-lag relationships has been demonstrated to describe diffusion

processes in space. Consequently, spatial autoregressive (SAR) models encompass spatial lags of the dependent. Conversely, when spatial dependence emanates from omitted variables that demonstrate spatial dependence, the resultant model incorporates the spatial lags of the explanatory variables and the dependent variable, hereby establishing the corresponding model as a spatial error model (SEM). Meanwhile, the integration of models with spatially autoregressive models (Spatial lag) and a spatial error model, designated as the "general spatial autocorrelation model"[38]. The incorporation of a weight matrix is pivotal in introducing spatial dependence into the analysis, thereby defining the spatial component structure. Specifically, the spatial matrix W in the horizontal model assumes the form $n \times n$, where the elements are non-negative and sparsity is observed. The elements of this matrix are defined as $w_{ij} = 1$ for neighborhoods and $w_{ij} = 0$ to avoid observations being identified as neighbors. The spatial weighting matrix applied in this research is the k-Nearest Neighbor (KNN) distance weighting. The KNN was selected due to its accessibility and its capacity to establish connections between disconnected neighbors, a functionality that is particularly advantageous in sparsely populated regions, such as islands [39].

The general form of the spatial area model that includes all types of interaction effects can be written as [38]:

$$\mathbf{y} = \lambda \mathbf{W}\mathbf{y} + \delta \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u} \quad (3)$$

$$\mathbf{u} = \gamma \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \quad (4)$$

where $\mathbf{y}$ represents a vector of the dependent variable for each unit in the sample with the size of $n \times 1$, $\mathbf{X}$ is a matrix of independent variables with the size of $n \times k$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are a vector of parameter estimation coefficients with the dimension of $k \times 1$, and $\mathbf{1}_n$ is a vector of ones corresponding to the constant term, with parameter δ to be estimated. γ denotes the intensity of interdependence among regression residuals, while λ is the spatial dependence coefficient that captures the spatially lagged dependent variable and describes the intensity of spatial interaction. In addition, $\mathbf{W}\mathbf{y}$ represents the interaction effects on the dependent variable, $\mathbf{W}\mathbf{X}$ reflects the interaction effects between the independent variables, and $\mathbf{W}\mathbf{u}$ denotes the interaction effects between the error terms of the different units. Both $\mathbf{u}$ and $\boldsymbol{\varepsilon}$ represent the total error and the error term, respectively.

The SAR model is predicated on the principle that the spatial matrix, otherwise known as spatial dependence, manifests exclusively in relation to the endogenous variable. This enables the examination of whether the formation of a distinct middle or class influences the formation of a different class in neighboring areas. The model is obtained by restricting the coefficients to the spatial autocorrelation coefficients of the residuals, designated as ε ($\gamma = 0$), and the absence of interaction between the independent variables ($\theta = 0$). Consequently, the SAR model equation is as follows:

$$\mathbf{y} = \lambda \mathbf{W}\mathbf{y} + \delta \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (5)$$

Meanwhile, when spatial autocorrelation is observed in the disturbances ($\lambda = 0$ and $\theta = 0$), the SEM model becomes an appropriate approach. In such instances, the formulation of the modeling equation is as follows:

$$\mathbf{y} = \delta \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (6)$$

The selection of an appropriate spatial dependency model is a critical initial step. The diagnostic tool used is called the Lagrange Multiplier (LM) test. Two types of Lagrange multiplier (LM) tests were conducted: the classical LM test proposed by [40] and the robust LM test proposed by [41]. Each of these tests consists of LM Error and LM Lag tests. In instances where the outcomes of the classical LM test demonstrate significant LM error and LM lag, the implementation of robust LM testing is imperative. In the event that either the LM error or the LM lag is found to be significant, the model will be selected. Conversely, if both are found to be significant, the spatial model is selected based on the largest test statistic. The description of the LM test is listed in Table 2. In the event that the LM error or robust LM error test results are found to be significant, this serves as an indication that the SEM model is a more appropriate one to utilize. Conversely, when the LM Lag or robust LM Lag test results are significant, it indicates that the SAR model is more representative.

Table 2. Lagrange Multiplier (LM) test

LM Test	Detects	Hypothesis
LM Error	The effect of spatial dependency in errors	$H_0: \gamma = 0$ $H_1: \gamma \neq 0$
LM Lag	The effect of spatial dependency in lag	$H_0: \lambda = 0$ $H_1: \lambda \neq 0$
Robust LM Error	The effect of spatial dependency in errors	$H_0: \gamma = 0$ $H_1: \gamma \neq 0$
Robust LM Lag	The effect of spatial dependency in lag	$H_0: \lambda = 0$ $H_1: \lambda \neq 0$

Subsequent to the acquisition of the requisite specifications, a meticulous interpretation of the results is imperative. In this context, [38] propose a methodology that facilitates the calculation of direct and indirect effects. The direct effect is defined as the result of the impact of a change in the explanatory variable a . Meanwhile, the indirect effect is defined as the interaction where a change in the condition of the independent variable of region j has an effect on the social class of region i . Furthermore, the summation of these two effects is denoted as the total effect. Meanwhile, the modeling in this study uses R as its statistical software.

2.4. Model Evaluation

Classical assumptions

The validity and reliability of the statistical model are contingent upon the fulfillment of certain classical assumptions, as posited by [42]. There are several assumptions that must be met for this method to be employed. These include the assumptions of normality, homoscedasticity, and non-multicollinearity. In order to ascertain the fulfillment of the aforementioned assumptions, classical assumption testing is conducted. The results of this testing are documented in Table 3.

Table 3. Classical assumptions test

Classical Assumptions Test	Detects	Hypothesis/Value
Jarque-Berra Test	Non-Normality	H_0 : the errors are normally distributed H_1 : the errors are not normally distributed
Breusch-Pagan Test	Heteroscedasticity	H_0 : constant variance of errors H_1 : non-constant variance of errors
Variance Inflation Factor (VIF)	Multicollinearity	$VIF < 5$ (No evidence of collinearity) $5 \leq VIF < 10$ (Weak collinearity) $VIF > 10$ (Strong collinearity)

Goodness of fit

The goodness of fit of the model is determined using two measures, the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), to determine the most appropriate model, whether it is a non-spatial model or a spatial regression model. The calculation of AIC and BIC is outlined as follows [43], [44].

$$AIC = -2 \ln L + 2k \quad (7)$$

$$BIC = -2 \ln L + 2k \ln N \quad (8)$$

In the above equation, L denotes the value of the likelihood, N signifies the number of recorded measurements, and k represents the number of estimated parameters. The superiority of a model is regarded when its AIC and BIC are lower.

3. Results and Discussion

3.1. Characteristics of food security in Indonesia

The characteristics of Indonesia's food security for 2022 are obtainable in Table 4, which provides a summary of the statistical data for each variable included in this study. The mean value of the food security index is 71.971, with a high range and standard deviation. This suggests that discrepancies exist with respect to the levels of food security observed among Indonesian provinces. A further examination of the remaining four variables revealed that the ICT variable exhibited the lowest standard deviation, followed by the quality of democracy. In contrast, the variable with the greatest standard deviation is the density of harvested land area for food crops. This indicates that there is a relatively consistent level of technological advancement and quality of democracy across different locations in Indonesia, whereas the harvested land area of each individual province varies considerably.

Table 4. Summary of statistics on research variables

Variable	Obs	Min	Max	Mean	Med	Std
FSI	34	37.800	85.190	71.971	73.910	9.916
ICT	34	3.220	7.640	5.881	5.825	0.698
IDI	34	62.930	85.620	77.949	78.780	5.433
URB	34	21.393	100	48.850	44.052	18.459
FCLD	34	0.076	67.634	13.745	4.780	19.587

Figure 1 illustrates the spatial distribution of the food security index across Indonesia's provinces. Overall, the index values exhibit notable variation nationwide, with the western and central regions generally displaying higher levels of food security compared to the eastern regions. However, an exception within the western region is the Riau Islands province, which demonstrates low food security levels. An examination of the distribution pattern reveals a tendency for areas with high food security to cluster geographically. For instance, the provinces of Java and Bali exemplify this, exhibiting high food security indices. This situation is likely reinforced by factors such as ease of access, the presence of robust infrastructure, a high quality of human capital, and the advancement of the regional economy. This observed clustering pattern suggests a spatial dependency on the surrounding areas.



Figure 1. Food security index distribution by provinces, 2022

3.2. Spatial dependence of Indonesia's Food Security

Table 5 presents the results of spatial correlation testing conducted using Moran's I and Geary's C tests. The Moran's I index value for provincial food security is significantly greater than zero, while the Geary's C index is significantly less than one across all significance levels. These findings provide robust statistical evidence confirming the presence of spatial autocorrelation in provincial food security across Indonesia during the 2022 period. In other words, the geographical distribution of provincial food security is not random but exhibits clear spatial dependencies. Regions with similar food security profiles tend to be located in close proximity, forming spatial clusters or agglomerations. This spatial dependence can be attributed to the spillover effects of shared needs, such as resource flows, between neighboring regions. Furthermore, it suggests an indirect association with factors like economic interdependencies and the diffusion of policies aimed at achieving food security at the regional level.

Table 5. Global spatial correlation test

Variable	Morans'I	Geary's C
FSI	0.621*** (0.000)	0.265*** (0.000)

Note: () represents the *p-value*

*, **, *** denote 10%, 5%, and 1% significance respectively

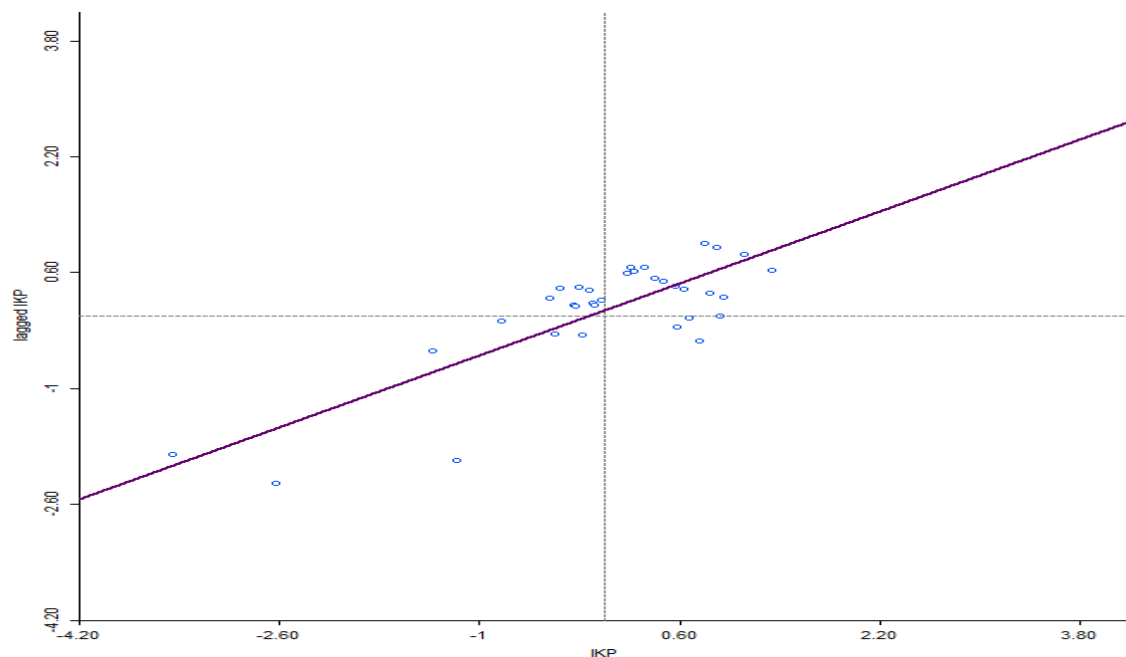


Figure 2. Moran scatter plot for Indonesia provincial food security index

The occurrence of local spatial correlation can be identified through the use of the Moran scatter plot presented in Figure 2. The figure illustrates the phenomenon of spatial clustering through four quadrants with most provinces located in regions showing positive correlation in the first and third quadrants. These results provide evidence that food security has significant positive spatial spillover effects.

3.3. Spatial econometric modelling of Indonesia's food security

As outlined by [42], the process of spatial model construction commences with the estimation of a non-spatial model (see Table 6). This is followed by the utilization of the LM test, which serves to ascertain the most appropriate spatial area model. The outcomes of LM tests are presented in Table 8. The classical LM test yielded significant results for both the LM for the spatial error model and the LM for the spatial lag model. Therefore, the model selection is determined by examining the outcomes of

the robust LM test, wherein the LM results for the spatial lag model are found to be statistically significant.

Table 6. Regression estimates

Variable	OLS	Spatial Area Model	
		SEM	SAR
Intercept	-38.549** (16.542)	-10.081 (14.143)	-39.857*** (12.956)
ICT	7.896*** (2.362)	6.269*** (1.743)	6.449*** (1.860)
IDI	0.936*** (0.247)	0.625*** (0.186)	0.631*** (0.202)
URB	-0.217** (0.083)	-0.129* (0.076)	-0.166** (0.065)
FCLD	0.124** (0.059)	0.187*** (0.060)	0.087* (0.051)
W X FSI (λ)	-	-	0.421*** (0.115)
W X ϵ (γ)	-	0.517** (0.125)	-
Indicators of model goodness			
R ²	0.715	0.794	0.799
AIC	220.784	216.292	213.446
BIC	229.942	226.976	224.131
Log-Likelihood	-104.392	-101.1459	-99.72315
Diagnostic Test			
Jarque-Berra test	0.227	0.347	2.404
Breusch-Pagan Test	5.994	7.736	2.096

Note: () represents the standard error

*, **, *** denote 10%, 5%, and 1% significance respectively

Moreover, an examination of the model goodness indicators (see Table 6) reveals that the spatial lag model exhibits the lowest AIC, BIC, and Log-Likelihood values and the highest R² among the alternative models [37]. Thus, the spatial lag model is selected as the model to be utilized for subsequent analysis.

Table 7. Variance Inflation Factor

Independent Variable	Value
ICT	2.818
IDI	1.866
URB	2.424
FCLD	1.361

Prior to any interpretation of the selected model, it is first necessary to ascertain whether the model assumptions have been met (see Table 6 and Table 7). As can be seen from the table, all of the assumptions have been validated. This is attributable to the fact that the test results indicate insignificant

values and a VIF value that is less than 5. It can thus be concluded that the estimation results of the selected model are accurate and reliable.

Table 8. Testing spatial effects

Spatial Dependence Test	Statistics	Prob
LM _p for spatial lag model	7.462	0.006***
LM _λ for spatial error model	3.182	0.074*
Robust LM _p for spatial lag model	4.534	0.033**
Robust LM _λ for spatial error model	0.253	0.615

Note: *, **, *** denote 10%, 5%, and 1% significance respectively

Table 6 displays the parameter estimates for each variable within the spatial lag model. In general, the explanatory variables are found to be significant in relation to provincial food security in Indonesia. The ICT variable, which is the primary focus of this study, has been demonstrated to exert a positive influence on provincial food security. This suggests that the progression of ICT can assist in attaining food security within each province. This outcome aligns with research conducted by [13] and [43]. The quality of democracy, an important variable in this study, has also been found to encourage the realization of food security in Indonesia. This finding is consistent with the studies of [28] and [44].

The following variables that were found to have a significant effect on provincial food security were urbanization and food crop land density. Urbanization has been shown to exert a negative influence on food security, indicating that an increase in urbanization will result in a decline in food security. This relationship is further validated by empirical evidence derived from studies [48] and [49]. In contrast, the density of agricultural land exhibits a positive correlation. Consequently, increasing the area dedicated to food crop agriculture is expected to enhance regional food security in Indonesia. This direction of relationship is also found in studies [50] and [51].

The modeling results also demonstrate that there are interrelated food security conditions between provinces, as indicated by the direction of the estimates and the significance of the parameter λ . This indicates that food security in Indonesia has a positive externality. In other words, the food security conditions in neighboring provinces exert a beneficial effect on food security in the focal province. Such a phenomenon may result from the implementation of policies or efforts to strengthen food security that are aligned with the needs of geographically proximate regions. The aforementioned favorable impact may also be attributed to the stimulation of a desire to enhance food security in the region, prompted by the observation of more robust food security in neighboring regions.

3.4. Average marginal effect of independent variables on food security

Subsequent analyses were conducted to determine how much influence the changes in each explanatory variable had on the food security index. Since the estimated SAR coefficients are biased and cannot be interpreted as marginal effects [38], the regression results provide only a preliminary assessment of the direction of action of each factor. Consequently, a more detailed decomposition of the overall effect (as illustrated in Table 6) is necessary. The decomposition of the effect is divided into three effects, namely direct, indirect, and total, which are presented in Table 9. Approximately 37%–38% of the effect of each independent variable on food security is attributed to spatial spillover effects.

As the primary variable, ICT exerts a direct influence on food security, with a coefficient of 6.966. Thus, a one-unit increase in the technology development index will result in a 6.966-point increase in the food security index. Concurrently, the spillover effect from the ICT variable is 4.170, whereby an increase in the technology development index in neighboring provinces will have an impact on food security in the main province to the extent of 4.170 points. Therefore, the overall effect of ICT is 11.136. All three effect magnitudes are statistically significant. The relationship between ICTs and food security can be examined through a number of channels [52]. Firstly, ICT can function as a potential medium for disseminating information to farmers regarding novel technologies, superior input management techniques, and refined farming methodologies, capable of engendering an escalation in productivity levels. Secondly, farmers can be made aware of superior sales opportunities for their produce. Finally, ICTs have the capacity to optimize the efficiency of agricultural markets, reduce price volatility, and

increase food availability. Furthermore, ICT has the potential to enhance the flow of information between farmers, food producers, traders, and consumers, thereby reducing food wastage and augmenting food reliability in the supply chain [5].

Table 9. Direct, indirect, and total effects of each independent variable

Variable	Direct	Indirect	Total
ICT	6.966*** (1.989)	4.170* (2.355)	11.136*** (3.861)
IDI	0.682*** (0.213)	0.408* (0.220)	1.090*** (0.379)
URB	-0.180** (0.070)	-0.107* (0.068)	-0.287** (0.126)
FCLD	0.094* (0.055)	0.056 (0.037)	0.150* (0.086)

Note: *, **, *** denote 10%, 5%, and 1% significance respectively

Democratic development, which represents the core variable, has a discernible impact on the three types of effects. The total effect is 1.090, which is the sum of the direct and indirect effects. The direct effect of the variable is 0.682, indicating that a one-point increase in the democracy index is associated with a 0.682-point increase in the food security index. The spillover effect of the democratic development variable is 0.408, suggesting that a one-point rise in the democracy index is linked to a 0.408-point rise in the food security index. As articulated by [44], a democratic system empowers the public to engage in the decision-making processes pertaining to food policy, thereby ensuring that the policies implemented are more attuned to the interests of the people and more responsive to the food needs of the community. Furthermore, the enhancement of democratic quality will precipitate the formation of institutions and governance that prioritize food security. The strengthening of institutions and the refinement of governance can serve to mitigate the factors that impede food security, including corruption and political instability. In this context, the promotion of better democracy engenders the formation of policies that are more pro-community, more transparent, and more responsive to food security issues. Consequently, this contributes to sustainable food security and addresses the challenges that Indonesia has been facing, such as climate change, natural disasters, and food distribution inequality. On the other hand, the promotion of mature democracy is conducive to the development of improved infrastructure, greater access to information, and more equitable social and economic policies, which contribute to augmented food security. As a nation that endorses democratic principles, the evolution of this system fosters the development of enhanced infrastructure, improved access to information, and more equitable social and economic policies, which contribute to food security.

Temporarily, the overall impact of urbanization is -0.287, with direct impacts contributing -0.180 and indirect impacts contributing -0.107. This magnitude signifies that an incremental one percentage point increase in the urban population results in a decrement of 0.180 points in the food security index. A similar phenomenon is observed when considering the spillover effect of an incremental one percentage point increase in the urban population of a neighboring province, which exerts a negative influence on the food security index of the focal province, reducing it by 0.107 points. The urbanization effect channel has been statistically substantiated. [53] explained that urbanization has a negative effect on food security from two perspectives, namely food production and consumption. The relocation of workers from rural to urban areas will reduce the amount of labor dedicated to food production. Concurrently, the expanding urban population necessitates the acquisition of additional land for urban expansion, resulting in encroachment on adjacent agricultural land and the limitation of agricultural investment. This phenomenon is further compounded by the fact that while food production is set to decline, demand is set to rise, resulting in a state of food instability and food accessibility. Another channel through which urbanization exerts its influence is through extreme weather events. This is because urbanization encourages extreme weather events, such as rising temperatures, rainfall, and flooding, which will affect production yields, distort food supply, and disrupt the distribution of agricultural logistics [50].

The direct, indirect, and total effects of changes in harvested land density have been calculated to be 0.094, 0.056, and 0.150, respectively. Although the effects are positive, the spillover effects of this variable are found to be insignificant. This is due to the fact that land density does not invariably translate into sufficient food diversity to ensure robust food security. In the context of agricultural production, a region's exclusive focus on a singular crop, such as rice or corn, may not yield the optimal spillover effect for neighboring areas. Consequently, the sole effect can be attributed to the direct effect, whereby a one-percent point increase in food crop harvesting density is associated with a 0.094-point increase in the food security index. [4] explained that land utilization for agriculture has an impact on increasing food availability, and with the fulfilment of the amount of food production, the demand for food can also be met, thus realizing stability.

4. Conclusion

The degree of food security within Indonesian provinces in 2022 demonstrates variability and agglomeration patterns, with the eastern region of the nation exhibiting consistently suboptimal food security levels. The observed phenomenon is further substantiated by the presence of a spatial correlation in the food security capabilities of provinces within Indonesia. In other words, provinces with similar food security characteristics are proximate to each other or can exert influence on each other. Consequently, spatial modelling emerges as a suitable approach to analyze the interplay between technological advancement and the quality of democracy on food security. The efficacy of both variables in supporting Indonesia's regional food security has been substantiated through empirical evidence. The impact of technological development and democratic quality as core variables is not only direct, but also indirect. These findings elucidate the relationship between the two variables and food security, which has been ambiguous until now. Moreover, evidence has been demonstrated that urbanization and agricultural land have a role in ensuring food security, albeit in contradictory ways. The expansion of land for food production is identified as a key factor in improving food security. However, urbanization is hypothesized to hinder the realization of this expansion in Indonesia. Furthermore, approximately 37%–38% of the impact of each variable on food security is attributable to spatial spillover effects. This figure underscores the necessity of incorporating considerations of interregional interactions and cross-provincial influences in the pursuit of food security.

The results of this study suggest several policy recommendations for policymakers and government authorities in achieving sustainable food security at both the national and regional levels. Firstly, the development and implementation of sophisticated technological frameworks is imperative within all domains pertinent to the production, distribution, and consumption of foodstuffs. Secondly, the quality of democracy must be strengthened to create policies and institutions that support food security by involving public participation and decentralizing food decision-making. Thirdly, it is imperative to leverage the full potential of resources present in both rural and urban regions to ensure food security in all provinces. Finally, the enhancement of collaboration and mutual learning is imperative to cultivate coherent and sustainable interregional food policies.

This study may serve to address the gaps in previous research, yet there is still room for improvement in similar studies in the future. The use of regional data necessitates the consideration of omitted variable bias (OVB) at the provincial level, a component that has not been addressed in this study. The incorporation of robust and sensitivity analyses could offer a solution to bolster the rigor of similar studies in the future.

Ethics approval

Approval for ethics was not required for this study.

Acknowledgments

We would like to express our gratitude to Politeknik Statistika STIS, who provided the necessary resources and facilities to conduct this research.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

Derived data supporting the findings of this study are available from the corresponding author on request.

Credit Authorship

Ditto Satrio Wicaksono: Conceptualization, Methodology, Formal Analysis, Writing—Original Draft, **Novi Hidayat Pusponegoro:** Conceptualization, Supervision, Resources, Final Draft. **Arbi Setiyawan:** Supervision, Final Draft.

References

- [1] Z. Zhang, K. Shi, L. Tang, K. Su, Z. Zhu, and Q. Yang, “Exploring the spatiotemporal evolution and coordination of agricultural green efficiency and food security in China using ESTDA and CCD models,” *J Clean Prod*, vol. 374, Nov. 2022, doi: 10.1016/j.jclepro.2022.133967.
- [2] FAO, “Rome Declaration on World Food Security,” Rome, 1996.
- [3] FAO, IFAD, UNICEF, WFP, and WHO, *The state of food security and nutrition in the world 2023*. Rome: FAO, 2023. doi: 10.4060/cc3017en.
- [4] C. M. Viana, D. Freire, P. Abrantes, J. Rocha, and P. Pereira, “Agricultural land systems importance for supporting food security and sustainable development goals: a systematic review,” Feb. 01, 2022, *Elsevier B.V.* doi: 10.1016/j.scitotenv.2021.150718.
- [5] M. K. Anser *et al.*, “Impact of ict adoption and governance interaction on food security in West Africa,” *Sustainability*, vol. 13, no. 10, May 2021, doi: 10.3390/su13105570.
- [6] United Nations, “Work of the statistical commission pertaining to the 2030 agenda for sustainable development,” 2017.
- [7] BPS-Statistics Indonesia, “Prevalence of moderate or severe food insecurity in the population, based on the food insecurity experience scale (FIES.” Accessed: Dec. 13, 2024. [Online]. Available: <https://www.bps.go.id/en/statistics-table/2/MTQ3NCMy/prevalence-of-moderate-or-severe-food-insecurity-in-the-population--based-on-the-food-insecurity-experience-scale--fies--percent-.html>
- [8] BPS-Statistics Indonesia, “Prevalence of undernourishment,” Accessed: Dec. 13, 2024. [Online]. Available: <https://www.bps.go.id/en/statistics-table/2/MTQ3MyMy/prevalence-of-undernourishment.html>
- [9] World Food Programme, “Indonesia annual country report 2022,” 2022.
- [10] Economist Impact, “Global food security index 2022,” 2022. [Online]. Available: <http://economistimpact>.
- [11] K. Von Grebmer *et al.*, “Global hunger index food systems transformation and local governance,” 2022.
- [12] A. Schnack, F. Bartsch, V. S. Osburg, and A. Errmann, “Sustainable agricultural technologies of the future: determination of adoption readiness for different consumer groups,” *Technol Forecast Soc Change*, vol. 208, Nov. 2024, doi: 10.1016/j.techfore.2024.123697.
- [13] R. Gouvea, D. Kapelianis, S. Li, and B. Terra, “Innovation, ICT & food security,” *Glob Food Sec*, vol. 35, Dec. 2022, doi: 10.1016/j.gfs.2022.100653.

- [14] I. H. Su, L. Wu, and K. H. Tan, "The future of the food supply chain: a systematic literature review and research directions towards sustainability, resilience, and technology adoption," *Journal of Digital Economy*, vol. 2, pp. 303–316, Dec. 2023, doi: 10.1016/j.jdec.2024.03.001.
- [15] A. M. Heshmati, "Impacts of improved agricultural technologies on food security and child nutrition in rural Ethiopia," *Cogent Food Agric*, vol. 9, no. 2, 2023, doi: 10.1080/23311932.2023.2276565.
- [16] G. Muha-Ud-Din *et al.*, "CRISPR/Cas9-based genome editing: A revolutionary approach for crop improvement and global food security," Jan. 01, 2024, *Academic Press*. doi: 10.1016/j.pmpp.2023.102191.
- [17] B. T. W. Hertel, U. L. C. Baldos, and K. O. Fuglie, "Trade in technology: a potential solution to the food security challenges of the 21st century," *Eur Econ Rev*, vol. 127, Aug. 2020, doi: 10.1016/j.eurocorev.2020.103479.
- [18] M. B. Zegeye, A. H. Fikire, and A. B. Assefa, "Impact of agricultural technology adoption on food consumption expenditure: evidence from Rural Amhara Region, Ethiopia," *Cogent Economics and Finance*, vol. 10, no. 1, 2022, doi: 10.1080/23322039.2021.2012988.
- [19] S.- Suprehatin, "Determinants of agricultural technology adoption by smallholder farmers in developing countries: perspective and prospect for Indonesia," *Jurnal Penelitian dan Pengembangan Pertanian*, vol. 40, no. 1, p. 21, May 2021, doi: 10.21082/jp3.v40n1.2021.p21-30.
- [20] B. Hidayati, Ermitayusida, V. A. Qurrata, Nasikh, and W. Huang, "The digital agriculture model for sustainable food system: an analysis of agricultural technology adoption in East Java, Indonesia," *J Sustain Sci Manag*, vol. 18, no. 4, pp. 173–191, Apr. 2023, doi: 10.46754/jssm.2023.04.013.
- [21] Md. A. Rahaman, "Governance and democracy," in *Global Encyclopedia of Public Administration, Public Policy, and Governance*, Springer International Publishing, 2017, pp. 1–6. doi: 10.1007/978-3-319-31816-5_1819-1.
- [22] D. Cassimon, O. Fadare, and G. Mavrotas, "The combined effect of institutional quality and capital flows on food and nutrition security and undernourishment in Sub-Saharan Africa," *PLoS One*, vol. 17, no. 10 October, Oct. 2022, doi: 10.1371/journal.pone.0275345.
- [23] N. N. Soko, S. Kaitibie, and N. N. Ratna, "Does institutional quality affect the impact of public agricultural spending on food security in Sub-Saharan Africa and Asia?," *Glob Food Sec*, vol. 36, Mar. 2023, doi: 10.1016/j.gfs.2022.100668.
- [24] Y. Khan, Š. Bojnec, U. Daraz, and F. Zulpiqar, "Exploring the nexus between poor governance and household food security," *Economic Change and Restructuring*, vol. 57, no. 2, Apr. 2024, doi: 10.1007/s10644-024-09679-w.
- [25] Abdullah, W. Qingshi, M. A. Awan, and J. Ashraf, "The impact of political risk and institutions on food security," *Current Research in Nutrition and Food Science*, vol. 8, no. 3, pp. 924–941, Dec. 2020, doi: 10.12944/CRNFSJ.8.3.21.
- [26] A. Sen, *Development as freedom*. New York: Oxford University Press, 1999.
- [27] A. Awad, "The determinants of food insecurity among developing countries: are there any differences?," *Sci Afr*, vol. 19, Mar. 2023, doi: 10.1016/j.sciaf.2022.e01512.
- [28] G. Mohamed, F. Chiad, M. Abdesslam, B. Omar, and M. S. M. AL-Absy, "Identifying determinants of food security using panel data analysis: evidence from maghreb countries," *Economies*, vol. 12, no. 4, Apr. 2024, doi: 10.3390/economies12040091.
- [29] A. I. Ogunniyi, G. Mavrotas, K. O. Olagunju, O. Fadare, and R. Adedoyin, "Governance quality, remittances and their implications for food and nutrition security in Sub-Saharan Africa," *World Dev*, vol. 127, Mar. 2020, doi: 10.1016/j.worlddev.2019.104752.
- [30] S. Kaur and H. Kaur, "Determinants of food security in Sub-Saharan Africa, South Asia and Latin America," in *Global Economic Cooperation*, New Delhi: Springer, 2016. doi: https://doi.org/10.1007/978-81-322-2698-7_6.
- [31] M. K. Anser *et al.*, "Social inclusion, innovation and food security in West Africa," *Sustainability (Switzerland)*, vol. 13, no. 5, pp. 1–12, Mar. 2021, doi: 10.3390/su13052619.
- [32] E. B. Yiadom, R. K. Dziwornu, J. K. M. Mawutor, and R. F. Amankwah, "Exploring the relationship between extreme weather events, urbanization, and food insecurity: Institutional quality perspective," *Environmental Challenges*, vol. 13, Dec. 2023, doi: 10.1016/j.envc.2023.100775.
- [33] D. S. Wicaksono *et al.*, "Modeling factors affecting educated unemployment on java island using geographically weighted poisson regression model," *Barekeng: Jurnal Ilmu Matematika dan Terapan*, vol. 18, no. 1, pp. 0615–0626, Mar. 2024, doi: 10.30598/barekengvol18iss1pp0615-0626.

- [34] C. Zhao, C. Li, J. Liu, H. Lian, and W. Yamaka, "Analysis of factors affecting the spatial association network of food security level in China," *Agriculture (Switzerland)*, vol. 14, no. 11, Nov. 2024, doi: 10.3390/agriculture14111898.
- [35] National Food Agency, "Food security index 2022," Jakarta, 2022.
- [36] R. C. Geary, "The contiguity ratio and statistical mapping," 1954.
- [37] L. Anselin, *Spatial Econometrics: Methods and models*. Kluwer Academic, 1988.
- [38] J. LeSage and R. K. Pace, *Introduction to spatial econometrics*. Taylor & Francis Group, LLC, 2009. [Online]. Available: <https://www.researchgate.net/publication/339938974>
- [39] M. Mawarni and I. Machdi, "Dynamic nearest neighbours for generating spatial weight matrix," in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Malang: IEEE, 2016, pp. 257–262. doi: 10.1109/ICACSIS.2016.7872771.
- [40] L. Anselin, A. K. Bera B', R. Florax, and M. J. Yoon, "Simple diagnostic tests for spatial dependence," *Reg Sci Urban Econ*, vol. 26, pp. 77–104, 1996.
- [41] J. Elhorst, "Spatial panel data analysis," in *Handbook of Applied Spatial Analysis*, Springer Berlin Heidelberg, 2010. doi: 10.1007/978-3-642-03647-7.
- [42] D. N. Gujarati and D. C. Porter, *Basic econometrics*, 5th ed. McGraw-Hill, 2009.
- [43] H. Akaike, "A new look at the statistical model identification," *IEEE*, vol. 19, no. 6, 1974.
- [44] M. Stone, "Comments on model selection criteria of akaike and," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 276–278, 1979.
- [45] J. P. Elhorst, *Spatial econometrics from cross-sectional data to spatial panels*. New York, 2014. [Online]. Available: <http://www.springer.com/series/10096>
- [46] M. Masha, E. Bojago, Y. Abraham, D. Leja, and M. W. Delango, "Determinants of food insecurity and coping mechanisms in Offa District, Southern Ethiopia," *J Agric Food Res*, vol. 14, Dec. 2023, doi: 10.1016/j.jafr.2023.100782.
- [47] D. Rossignoli and S. Balestri, "Food security and democracy: do inclusive institutions matter?," *Canadian Journal of Development Studies*, vol. 39, no. 2, pp. 215–233, Apr. 2018, doi: 10.1080/02255189.2017.1382335.
- [48] S. Szabo, "Urbanisation and food insecurity risks: assessing the role of human development," *Oxford Development Studies*, vol. 44, no. 1, pp. 28–48, Jan. 2016, doi: 10.1080/13600818.2015.1067292.
- [49] M. U. Farrukh, M. K. Bashir, M. F. Rola-Rubzen, and A. Ahmad, "Dynamic effects of urbanization, governance, and worker's remittance on multidimensional food security: an application of a broad-spectrum approach," *Socioecon Plann Sci*, vol. 84, Dec. 2022, doi: 10.1016/j.seps.2022.101400.
- [50] E. B. Yiadom, R. K. Dziwornu, J. K. M. Mawutor, and R. F. Amankwah, "Exploring the relationship between extreme weather events, urbanization, and food insecurity: institutional quality perspective," *Environmental Challenges*, vol. 13, Dec. 2023, doi: 10.1016/j.envc.2023.100775.
- [51] X. Cao, J. Han, and X. Li, "Analysis of the impact of land use change on grain production in Jiangsu Province, China," *Land (Basel)*, vol. 13, no. 1, Jan. 2024, doi: 10.3390/land13010020.
- [52] E. Nakasone and M. Torero, "A text message away: ICTs as a tool to improve food security," *Agricultural Economics (United Kingdom)*, vol. 47, pp. 49–59, Nov. 2016, doi: 10.1111/agec.12314.
- [53] C. C. Lee, M. Zeng, and K. Luo, "The impact of urbanization on food security in China," *International Review of Economics and Finance*, vol. 93, pp. 1159–1175, Jun. 2024, doi: 10.1016/j.iref.2024.05.037.



Forecasting Farmer Exchange Rate (FER) in Southeast Sulawesi Province Using Cheng's Fuzzy Time Series Method

Rastina¹, Lilis Laome^{2*}, Bahriddin Abapihi³, Gusti Ngurah Adhi Wibawa⁴, Mukhsar⁵, Makkulau⁶, Gama Putra Danu Sohibien⁷, Sukim⁸, Fathurrahman Yahyasatrio⁹

^{1,2,3,4,5,6}Faculty of Mathematics and Natural Sciences, Universitas Halu Oleo, Kendari, Indonesia, ^{7,8}Politeknik Statistika STIS, Jakarta, Indonesia, ⁹School of Computing, University of Portsmouth, Portsmouth, United Kingdom

*Corresponding Author: E-mail address: lilis.la_ome@uho.ac.id

ARTICLE INFO

Article history:

Received 4 September, 2024

Revised 1 December, 2025

Accepted 23 December, 2025

Published 31 December, 2025

Keywords:

FER; Fuzzy Time Series; Fuzzy Time Series Cheng; MAPE; Forecasting

Abstract

Introduction/Main Objectives: This study aims to forecast the Farmer Exchange Rate (FER) in Southeast Sulawesi Province for 2024 as a basis for short-term economic assessment and policy-related analysis. **Background Problems:** FER is a key indicator of farmers' purchasing power and agricultural welfare; however, its monthly dynamics are characterized by fluctuations and uncertainty, making conventional forecasting methods less effective in capturing its behavior. **Novelty:** This study contributes by implementing the the Fuzzy Time Series (FTS) Cheng approach for FER forecasting in Southeast Sulawesi, emphasizing its suitability for handling vagueness and nonlinear patterns inherent in agricultural economic indicators. **Research Methods:** The analysis utilizes monthly secondary FER data obtained from BPS-Statistics of Southeast Sulawesi Province, covering the period from January 2014 to December 2023. Forecast accuracy is evaluated using the Mean Absolute Percentage Error (MAPE). **Finding/Results:** The forecasting results indicate that the FER values for January, February, and March 2024 are each estimated at 105.93. The model achieved a MAPE of 0.3027%, corresponding to an accuracy level of 99.6973%, which places the forecasting performance in the "excellent" category.

1. Introduction

Forecasting is an approach used to estimate something that will happen in the next few periods based on historical data [1]. There are two types of forecasting methods, namely qualitative and quantitative. Qualitative methods are predictions based on the opinions of experts as a basis for decision making, and the data cannot be expressed clearly in numerical form. Meanwhile, quantitative methods are predictions that can be arranged in the form of numbers based on historical data commonly referred to as time series data [2].

Time series data is a type of data that is collected according to a time series within a certain period of time. The time used can be days, weeks, months, years, and so on. Time series data is very useful for decision makers to estimate or predict future events. This is because the pattern of changes in time series data in a period in the past is likely to be repeated in the future. Time series forecasting can predict future values and decide on certain policies that are influenced by the results of forecasting.

Southeast Sulawesi is one of the provinces in Indonesia where most of the population lives in rural areas and generally still depends on the agricultural sector for their livelihood. The agricultural sector plays an important role in the national economy. This is evident in terms of Production, Agricultural, Forestry, and Fisheries Business Fields which make the most dominant contribution to the GDP of Southeast Sulawesi Province, which is 21.99% in the economic growth of Southeast Sulawesi Province in the second quarter of 2023. Therefore, it can be said that the agricultural sector is one of the sectors to increase economic growth in Southeast Sulawesi Province.

Based on what has been explained, the development of the agricultural sector must be the main thing. Farmers are often disadvantaged with relatively small incomes, while production costs are higher than the costs received [3]. As a form of agricultural development efforts for the welfare of farmers in an area, a benchmark is needed. The benchmark in question is the Farmer Exchange Rate (FER). FER is a comparison between the price index received by farmers (*It*) and the price index paid by farmers (*Ib*) expressed as a percentage, namely the price of consumption and production needs [4]. The meaning of FER is: $FER < 100$ means the farmer is experiencing a surplus. $FER = 100$ means the farmer is breaking even. $FER > 100$ means the farmer is experiencing a deficit.

As an indicator of welfare, FER plays an important role in supporting the evaluation of agricultural sector development because it can illustrate the balance between farmers' income and expenditure. Recent research indicates that FER remains relevant for assessing farmer welfare across various commodities and regions in Indonesia, and serves as a basis for formulating agricultural policies aimed at improving farmer well-being [5]. However, the value of FER is fluctuating and influenced by changes in agricultural product prices and production costs, necessitating forecasting efforts to gain insight into farmers' conditions in the future [6]. Forecasting FER is crucial to enable the government and stakeholders to anticipate changes in farmer welfare and develop more targeted and data-driven policies [7].

Data obtained from BPS shows that the growth of FER in Southeast Sulawesi Province has fluctuated from time to time, where the state of FER in Southeast Sulawesi Province from January 2014 to December 2023 has increased or decreased. FER that tends to be low can affect the level of welfare and productivity of the agricultural sector to decrease, so that in the long run it results in a decrease in production rate and an increase in consumption rate [8]. In fact, it is expected that FER from time to time should always increase. Therefore, the right method is needed to forecast FER in Southeast Sulawesi Province.

FER data is available on a monthly basis so it includes the time series data type. One of the methods that can be used to forecast time series data is the Fuzzy Time Series. Fuzzy Time Series (FTS) is a data forecasting method that uses fuzzy sets as the basis for forecasting modeling. FTS is a method introduced by Song and Chissom (1993) which is a concept used to predict problems where actual data is formed in linguistic values [9]. This method is capable of handling data fluctuations, uncertainty and subjectivity in data, compared to classical statistics. FTS has the advantage that it does not require a large amount of historical data and does not require assumptions in forecasting [10]. There are several FTS methods that have been developed, one of which is the Fuzzy Time Series Cheng method [11]. FTS Cheng is a development and improvement of the method proposed by Chen in 1996. This method was proposed by Cheng, Cheng has a different way of determining intervals, namely seen from the frequency obtained. Then in the formation of the fuzzy set for each relationship is inserted and given weights based on the same sequence and loops [12].

Research on the Fuzzy Time Series method has previously been conducted by [13], namely comparing Cheng's Fuzzy Time Series method with the Box-Jenkins method in predicting the Composite Stock Price Index (JCI), then the forecasting accuracy value is calculated using MAE, MSE, MAPE. The result is Cheng's Fuzzy Time Series method is more accurate [13]. Furthermore, research conducted by [2] is to compare the Fuzzy Time Series Cheng and Fuzzy Time Series Markov Chain methods in forecasting the Farmer Exchange Rate (FER) in Indonesia, then the forecasting accuracy value is calculated using MAE, MSE, MAPE. The result is that Cheng's Fuzzy Time Series method is more accurate. [14] compared the Fuzzy Time Series Cheng and Lee methods for forecasting clean water supply and reported that the FTS Cheng method achieved a MAPE value below 3%, indicating very high forecasting accuracy. Additionally, Tursina [15] also used FTS Cheng to predict the Consumer Price Index with a MAPE of 0.23%. Based on these backgrounds and problems, the researcher is interested in conducting research on Farmer Exchange Rate Forecasting (FER) in Southeast Sulawesi Province using the Fuzzy Time Series Cheng Method, then to determine the level of accuracy using the Mean Absolute Percentage Error (MAPE) value. The purpose of this study is to find out the results of

forecasting the Farmer Exchange Rate (FER) in Southeast Sulawesi Province using the Fuzzy Time Series Cheng method.

2. Material and Methods

2.1. Type of Research

The type of research used is applied research. Applied research was chosen because the results of this study are not only intended to increase knowledge, but can also be directly used as a basis for formulating policies related to farmer management and welfare improvement. Thus, the Farmer Exchange Rate (FER) forecasting conducted is expected to assist local governments in making more accurate decisions to improve the welfare of farmers in Southeast Sulawesi Province.

2.2. Data Source

The data used in this study is secondary data obtained from [4] The data is monthly data on the Farmer Exchange Rate (FER) in Southeast Sulawesi Province from January 2014 to December 2023. The data used amounted to 120. Then the data was divided into two, namely 80% of the training data (January 2014 – December 2021) and 20% of the test data (January 2022 – December 2023). This data split ratio was chosen because empirical studies show that optimal results are achieved when using 20-30% of the data for testing and the remaining 70-80% for training. Prediction using data splitting is done by dividing the historical data used according to the ratio of training data and testing data, which is then processed sequentially following the steps in the FTS Cheng method to obtain prediction results and MAPE as a comparison to evaluate how well the model that has been created performs [16].

2.3. Data Analysis

The selection of Cheng's Fuzzy Time Series method in this study is supported by recent empirical research showing that this method often produces lower forecasting error rates compared to classical time series models when applied to economic and agricultural data. Therefore, Cheng's FTS method is considered suitable for use in forecasting FER, which has fluctuating and uncertain characteristics [17]. The steps of data analysis in this study are as follows:

- 1) Describe FER data through line charts to find out data patterns.
- 2) Model Fuzzy Time Series Cheng using FER data, with the following steps:
 - a. Universe Set

$$U = [D_{min} - D_1, D_{max} + D_2], \quad (1)$$

where D_{min} is the lowest data, while D_{max} is the highest data with D_1 and D_2 are arbitrary positive real numbers determined by the researcher. D_1 is a real number determined by the researcher to widen the lower limit of the interval and D_2 is a real number determined by the researcher to widen the upper limit of the interval [18].

- b. The length of the interval.
 - Determine the number of intervals using the Sturges equation,

$$n = 1 + 3,322 \times \log(N), \quad (2)$$

where n is the number of intervals, and N is the amount of data used. From these results, a number of linguistic values will be formed to represent a fuzzy set at intervals formed from the universe set (U).

$$U = \{u_1, u_2, \dots, u_n\} \quad (3)$$

where,

U : Universe set

u_i : the number of intervals in U , for $i = 1, 2, \dots, n$.

- The value of the range or range,

$$R = (D_{max} + D_2) - (D_{min} - D_1) \quad (4)$$

- The length of the interval,

$$l = \frac{R}{n}, \quad (5)$$

- The middle value

$$m_i = \frac{(\text{lower limit} + \text{upper limit})}{2} \quad (6)$$

Where i is a fuzzy set.

So each interval can be calculated by,

$$u_1 = [D_{min} - D_1; D_{min} - D_1 + l]$$

$$u_2 = [D_{min} - D_1 + l; D_{min} - D_1 + 2l]$$

$$u_3 = [D_{min} - D_1 + 2l; D_{min} - D_1 + 3l]$$

⋮

$$u_n = [D_{min} - D_1 + (n - 1)l; D_{min} - D_1 + nl] \quad (7)$$

- c. The Fuzzy Set and fuzzification.

Roughly speaking, a fuzzy set can be defined as a class of numbers with vague boundaries. If the universe of discourse (U) is the set of universes $U = \{u_1, u_2, \dots, u_n\}$, then a fuzzy set A_i of U with membership degree is generally expressed as follows.

$$A_i = \frac{\mu_{A_i}(u_1)}{u_1} + \frac{\mu_{A_i}(u_2)}{u_2} + \frac{\mu_{A_i}(u_3)}{u_3} + \dots + \frac{\mu_{A_i}(u_n)}{u_n} \quad (8)$$

Where $\mu_{A_i}(u_j)$ is the degree of membership from u_j to A_i where $\mu_{A_i}(u_j) \in [0,1]$ and $1 \leq j \leq n$. The membership degree value of $\mu_{A_i}(u_j)$ is defined as follows.

$$\mu_{A_i}(u_j) = \begin{cases} 1 & ; i = j \\ 0,5 & ; j = i - 1 \text{ or } j = i + 1 \\ 0 & ; i \neq j \end{cases} \quad (9)$$

This can be illustrated with the following rules:

Rule 1: If the actual data X_t belongs to u_j , then the membership degree for u_j is 1, and u_{j+1} is 0.5 and if not u_j and u_{j+1} , then it is zero.

Rule 2: If the actual data X_t belongs to u_i , $1 \leq i \leq n$ then the membership degree for u_j is 1, for u_{j-1} and u_{j+1} is 0.5 and if not u_j , u_{j-1} and u_{j+1} then it is zero.

Rule 3: If the actual data X_t belongs to u_j , then the membership degree for u_j , and for u_{j-1} is 0.5 and if it is not u_j and u_{j-1} then it is zero.

Suppose $A_1, A_2, \dots, A_n$ is a fuzzy set that has a linguistic value of a linguistic variable, the fuzzy set an $A_1, A_2, \dots, A_n$ in the set universe U is defined as follows:

$$A_1 = \left\{ \frac{1}{u_1} + \frac{0,5}{u_2} + \frac{0}{u_3} + \dots + \frac{0}{u_n} \right\}$$

$$A_2 = \left\{ \frac{0,5}{u_1} + \frac{1}{u_2} + \frac{0,5}{u_3} + \dots + \frac{0}{u_n} \right\}$$

$$\begin{aligned}
 A_3 &= \left\{ \frac{0}{u_1} + \frac{0,5}{u_2} + \frac{1}{u_3} + \dots + \frac{0}{u_n} \right\} \\
 &\vdots \\
 A_n &= \left\{ \frac{0}{u_1} + \frac{0}{u_2} + \dots + \frac{0,5}{u_{n-1}} + \frac{1}{u_n} \right\}
 \end{aligned} \tag{10}$$

Where u_j ($i = 1, 2, \dots, n$) is an element of the universe set (U) and the number given the symbol “-” expresses the degree of membership $\mu_{A_i}(u_j)$ to A_i ($i = 1, 2, \dots, n$) whose value is 0, 0.5 or 1 [8].

d. Fuzzy Logical Relations (FLR) and Fuzzy Logical Relations Group (FLRG)

Establish fuzzy logical relationships based on historical data. In fuzzified data, two consecutive fuzzy sets $A_i(t-1)$ and $A_j(t)$ can be expressed as FLR $A_i \rightarrow A_j$. The relationship is identified based on the results of fuzzification of time series data. If the time series variable $F(t-1)$ is fuzzified as A_i dan $F(t)$ is fuzzified as A_j , then A_i and A_j can be denoted as $A_i \rightarrow A_j$, where A_i is the current state and A_j is the next state. Suppose if FLR $A_1 \rightarrow A_1$, $A_1 \rightarrow A_1$, $A_1 \rightarrow A_3$ then the FLRG formed is $A_1 \rightarrow A_1, A_2, A_3$.

e. Weighting matrix from the FLRG

Determine the weights of FLR relations into FLRG by including all relations and assigning weights based on the same order and recurrence. FLRs that have the same current state (A_i) are combined into a group in the form of a weighting matrix. Then transfer the weights into a weighting matrix whose equation is written as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & w_{ij} & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{bmatrix} \tag{11}$$

where W is the weight matrix and W_{ij} is the weight matrix at the i -th row and j -th column with $i = 1, 2, \dots, n$; $j = 1, 2, \dots, n$.

f. Standardized weighting matrix

Next, transfer the FLRG weights into the form of a standardized weight matrix (W^*) which has the following equation.

$$W^* = \begin{bmatrix} w_{11}^* & w_{12}^* & \dots & w_{1n}^* \\ w_{21}^* & w_{22}^* & \dots & w_{2n}^* \\ \vdots & \vdots & w_{ij}^* & \vdots \\ w_{n1}^* & w_{n2}^* & \dots & w_{nn}^* \end{bmatrix} \tag{12}$$

Where W^* is a standardized weight matrix with the following formula.

$$W_{ij}^* = \frac{w_{ij}}{\sum_{i=1}^n w_{ij}} \tag{13}$$

g. Defuzzification

To generate forecasting values, the standardized weight matrix W^* is multiplied by m_i . So the forecasting calculation becomes as follows.

$$F_i = w_{i1}^*(m_1) + w_{i2}^*(m_2) + \dots + w_{in}^*(m_n) \tag{14}$$

Where F_i is the forecasting result, with W_{in}^* is equation 13. If the fuzzification result of the i -th period is A_i , and A_i has no FLR in FLRG or can be written with the condition $A_i \rightarrow \emptyset$, where the maximum value of the membership degree is at u_i , then the forecasting value (F_i) is the middle value of u_i , or defined by m_i [19].

h. MAPE

MAPE is an error measurement that calculates the size of the percentage deviation between actual data and forecasting data. If the lower the error rate on the forecast data, the more accurate the level of the forecast results [20]. The MAPE calculation is shown in the following equation:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \frac{|X_t - F_t|}{X_t} \times 100\% \quad , t = 1, 2, \dots, n \quad (15)$$

Where X_t is the actual data in the t -th period, F_t is the value of the forecasting results on the data, and n is the amount of actual data. The accuracy criteria using the MAPE value can be seen in Table 1 [21].

Table 1. MAPE accuracy criteria

MAPE Value	Accuracy Criteria
$MAPE \leq 10\%$	Very good
$10\% < MAPE \leq 20\%$	Good
$20\% < MAPE \leq 50\%$	Fair
$MAPE > 50\%$	Poor

3) Making conclusions

To clarify the research procedure, Figure 1 outlines the forecasting process using Cheng's Fuzzy Time Series method in a structured manner. The flowchart illustrates each step, starting from data input to model evaluation and forecasting results. This representation provides a clear overview of the methodological framework applied to forecast the Farmer Exchange Rate (FER).

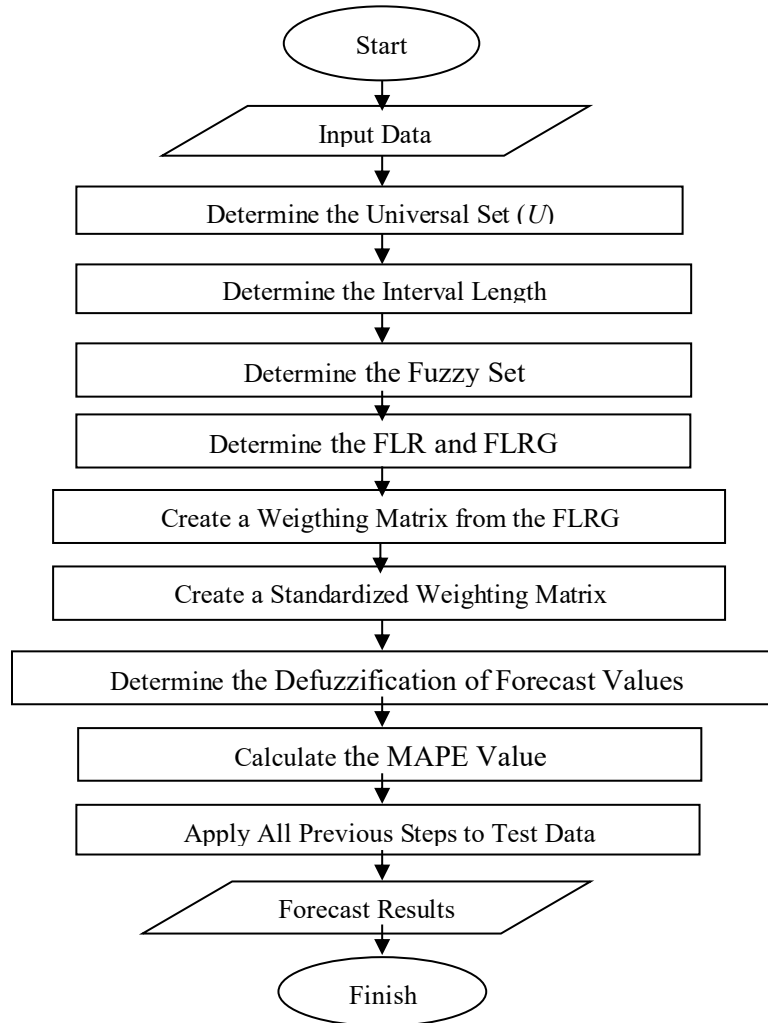


Figure 1. Flowchart of the Cheng's Fuzzy Time Series Method for Forecasting the FER

3. Results and Discussion

3.1. Characteristics of the Farmer Exchange Rate (FER) in Southeast Sulawesi Province

The data used in this study is the monthly time series data of the Farmer Exchange Rate (FER) in Southeast Sulawesi Province from January 2014 to December 2023. The graph of the FER can be seen in Figure 2. The FER pattern shows fluctuations over time, as evidenced by the increase and decrease in value during each period. For example, from November to December 2014, there was a decrease in value from 101.23 to 99.63, while in the same period in 2015, there was an increase from 100.64 to 101.01, followed by another decrease in 2016.

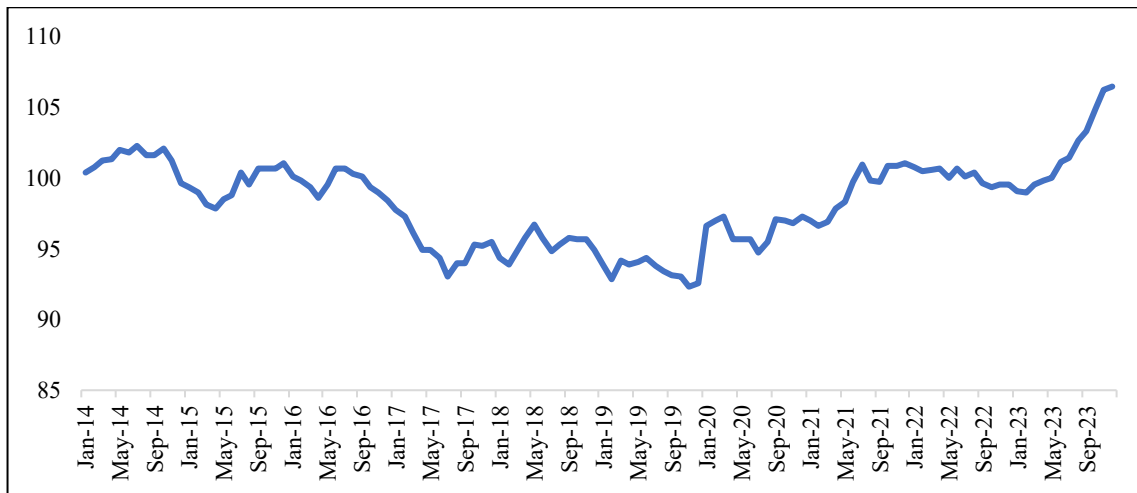


Figure 2. FER chart in Southeast Sulawesi January 2014 - December 2023

This fluctuation occurs because FER is influenced by changes in agricultural commodity prices, which determine farmers' income, as well as changes in production costs and household consumption, which determine the level of expenditure. When production and consumption costs increase faster than the selling price of agricultural products, FER will decrease, weakening farmers' purchasing power. Conversely, when income from harvests is greater than expenditure, FER increases and farmers gain an economic surplus. This fluctuating condition indicates that the welfare of farmers in Southeast Sulawesi Province is highly vulnerable to price and production cost dynamics, necessitating a forecasting method capable of representing this uncertainty to support informed decision-making.

3.2. Cheng's Fuzzy Time Series Prediction

In forecasting using the fuzzy time series cheng, the first step that must be done is to determine the values of D_1 and D_2 , where the values of D_1 and D_2 are positive numbers determined by the researcher himself. So that to obtain optimal D_1 and D_2 values, experiments were carried out on the training data on several D_1 and D_2 values, where in this study the D_1 and D_2 values were equalized, namely 0, 0.005, 0.05, and 0.1, respectively. So that the MAPE value for four variations of D_1 and D_2 value can be seen in Table 2.

Table 2. MAPE values for a number of variations of D_1 and D_2

D_1	D_2	MAPE
0	0	0.6619%
0.005	0.005	0.6618%
0.05	0.05	0.6494%
0.1	0.1	0.6463%

From the table, it can be seen that the D_1 and D_2 values that produce the smallest MAPE value are 0.1, which is 0.6463%. So, the next step is to do FER forecasting in Southeast Sulawesi Province using the D_1 and D_2 values of 0.1 each in the test data. The steps of applying Cheng's Fuzzy Time Series method using test data in this study are as follows. So, the next step is to carry out FER forecasting in

Southeast Sulawesi Province using D_1 and D_2 values of 0.1 each on the test data. The steps to apply the Fuzzy Time Series Cheng method using test data in this study are as follows:

1) Formation of the U Universe

The first step in the fuzzy time series method is to form a universe of discourse. The following is the calculation process in the set of universes.

$$\begin{aligned} U &= [D_{min} - D_1; D_{max} + D_2] \\ &= [98.97 - 0.1; 106.47 + 0.1] \\ &= [98.87; 106.57] \end{aligned}$$

Where D_{min} is the lowest data of 98.87 and D_{max} is the highest data of 106.57. In this study, a D_1 value of 0.1 and a D_2 of 0.1 were taken.

2) Forming Interval Length

In this step, the formation of the length of the interval goes through several stages, first by calculating the number of intervals, determining the value of the range or distance, then determining the size of the interval, and calculating the midpoint value. The following is the process of calculating the interval length.

a. Calculating the number of intervals

In determining the number of intervals, the method used by the researcher uses the Sturges formula, as follows.

$$\begin{aligned} n &= 1 + 3.322 \log N \\ &= 1 + 3.322 \log 24 \\ &= 5.5851 \approx 6 \end{aligned}$$

So the number of intervals used is 6 intervals.

b. Specifying a range value

$$\begin{aligned} R &= (D_{maks} + D_2) - (D_{min} + D_1) \\ &= 106.57 - 98.87 \\ &= 7.7 \end{aligned}$$

c. Determining the length of the interval

$$\begin{aligned} l &= \frac{R}{n} \\ &= \frac{7.7}{6} \\ &= 1.2833 \end{aligned}$$

From these results, the partitions of the set of the universe are obtained according to the length of the interval.

$$u_1 = [98.87; 100.15]$$

$$u_2 = [100.15; 101.44]$$

$$u_3 = [101.4; 102.72]$$

$$u_4 = [102.72; 104.00]$$

$$u_5 = [104.00; 105.29]$$

$$u_6 = [105.29; 106.57]$$

Next, a table of interval classes, middle or midpoint values obtained from the upper and lower limit values, and the data frequency can be seen in Table 3.

Table 3. Frequency at each interval of the set of universes

u_i	Lower Limit	Upper Limit	Middle Value	Frequency
u_1	98.87	100.15	99.51	11
u_2	100.15	101.44	100.80	8
u_3	101.44	102.72	102.08	1
u_4	102.72	104.00	103.36	1
u_5	104.00	105.29	104.65	1
u_6	105.29	106.57	105.93	2

Then the next step is to calculate the average frequency value for each interval. The average frequency for each interval is the result of division between the number of frequencies and the number of intervals, namely 4, so the interval value must be smaller than the average frequency. Based on Table 3, it can be seen that there are two intervals that have a frequency value greater than the average frequency, namely the u_1 and u_2 intervals. So it is necessary to divide the interval again on the two intervals whose division results can be seen in Table 4.

Table 4. Frequency at each interval of the universal set after division

u_i	Lower Limit	Upper Limit	Middle Value	Frequency
u_1	98.87	99.19	99.03	2
u_2	99.19	99.51	99.35	3
u_3	99.51	99.83	99.67	2
u_4	99.83	100.15	99.99	4
u_5	100.15	100.58	100.37	3
u_6	100.58	101.01	100.80	3
u_7	101.01	101.44	101.22	2
u_8	101.44	102.72	102.08	1
u_9	102.72	104.00	103.36	1
u_{10}	104.00	105.29	104.65	1
u_{11}	105.29	106.57	105.93	2

d. Fuzzification Process

The fuzzification process or converting numerical data into linguistic data. The process assumes $A_1, A_2, \dots, A_n$ or a fuzzy collection of linguistic values of linguistic variables. The number of A_n is as many as the number of interval classes that have been obtained, namely 11 class intervals formed. From each interval class, fuzzy set A_i will be defined, with $1 \leq i \leq 11$. Then the linguistic variables will be formed as follows.

$$A_1 = \{1/u_1 + 0.5/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7 + 0/u_8 + 0/u_9 + 0/u_{10} + 0/u_{11}\}$$

$$A_2 = \{0.5/u_1 + 1/u_2 + 0.5/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7 + 0/u_8 + 0/u_9 + 0/u_{10} + 0/u_{11}\}$$

$$A_3 = \{0/u_1 + 0.5/u_2 + 1/u_3 + 0.5/u_4 + 0/u_5 + 0/u_6 + 0/u_7 + 0/u_8 + 0/u_9 + 0/u_{10} + 0/u_{11}\}$$

:

$$A_{11} = \{0/u_1 + 0/u_2 + 0/u_3 + 0/u_4 + 0/u_5 + 0/u_6 + 0/u_7 + 0/u_8 + 0/u_9 + 0.5/u_{10} + 1/u_{11}\}$$

Furthermore, this section will explain the formation of fuzzification based on intervals and obtain linguistic values according to the intervals formed. Determination of the fuzzification value by defining the data into the appropriate interval. For example, the first data in this study, January 2022, is 100.79, so the data is included in the u_6 class interval. In the fuzzification process, the data will be converted into a linguistic value of A_6 because in the fuzzification process a data will enter into a linguistic value that has a membership degree value equal to 1 which indicates a true value. The fuzzification results can be seen in Table 5.

Table 5. Fuzzification

t	Month	FER	Fuzzification
1	Jan-22	100.79	A_6
2	Feb-22	100.5	A_5
3	Mar-22	100.58	A_5
:	:	:	:
22	Okt-23	104.92	A_{10}
23	Nov-23	106.23	A_{11}
24	Des-23	106.47	A_{11}

In Table 5, the results of the fuzzification process on the FER value can be explained that the January 2022 period is 100.79, can be defined into the interval [100.58; 101.01] which is included in the range of linguistic values A_6 . Furthermore, in the February 2022 period worth 100.5, can be defined into the interval [100.15; 100.58] which is included in the range of linguistic values A_5 , and so on until the December 2023 period worth 106.47 with the interval [105.29; 106.57] which is included in the range of linguistic values A_{11} .

e. Formation of Fuzzy Logic Relationship and FLR Group

From Table 5, it is known that in January 2022 and February 2022, the FER values were fuzzified into sets A_6 and A_5 , respectively, so this relationship can be notated as $A_6 \rightarrow A_5$. Economically, this relationship indicates a tendency for a decline in farmers' purchasing power in the subsequent period, which could be caused by rising production costs or falling prices for agricultural commodities.

Table 6. FLR

Data Sequence	FLR	Data Sequence	FLR
1	-	12→13	$A_3 \rightarrow A_1$
1→2	$A_6 \rightarrow A_5$	13→14	$A_1 \rightarrow A_1$
2→3	$A_5 \rightarrow A_5$	14→15	$A_1 \rightarrow A_2$
3→4	$A_5 \rightarrow A_6$	15→16	$A_2 \rightarrow A_4$
4→5	$A_6 \rightarrow A_4$	16→17	$A_4 \rightarrow A_4$
5→6	$A_4 \rightarrow A_6$	17→18	$A_4 \rightarrow A_7$
6→7	$A_6 \rightarrow A_4$	18→19	$A_7 \rightarrow A_7$
7→8	$A_4 \rightarrow A_5$	19→20	$A_7 \rightarrow A_8$
8→9	$A_5 \rightarrow A_3$	20→21	$A_8 \rightarrow A_9$
9→10	$A_3 \rightarrow A_2$	21→22	$A_9 \rightarrow A_{10}$
10→11	$A_2 \rightarrow A_2$	22→23	$A_{10} \rightarrow A_{11}$
11→12	$A_2 \rightarrow A_3$	23→24	$A_{11} \rightarrow A_{11}$

The results of the fuzzy logic relationship (FLR) formation presented in Table 6 show that dominant relationships such as $A_6 \rightarrow A_4$ indicate that when FER is in the medium range, the probability of FER in the next period will tend to decrease to a lower value group. Meanwhile, the relationship $A_{11} \rightarrow A_{11}$

shows stability at high FER levels, where the value tends to remain in the same category in the next period. Thus, the dominant FLRG provides an overview of the most frequent patterns of change in FER. If the dominant FLRG shows a declining transition, this indicates economic pressure on farmers. Conversely, a stable or increasing FLRG reflects relatively maintained or improved farmer welfare conditions.

Furthermore, to form the FLRG obtained from the FLR results. If there is a fuzzy set that has a relationship or can predict with more than one fuzzy set, it can be combined. Based on Table 6, it can be seen that fuzzy A_1 has a relationship that can predict A_1, A_2 , so for FLRG can be formed with the notation $A_1 \rightarrow A_1 A_2$. Complete results for the formation of FLRG can be seen in Table 7.

Tabel 7. FLRG

<i>Current State</i>		<i>Next State</i>
A_1	$\rightarrow$	$A_1 A_2$
A_2	$\rightarrow$	$A_2 A_3 A_4$
A_3	$\rightarrow$	$A_1 A_2$
A_4	$\rightarrow$	$A_4 A_5 A_6 A_7$
A_5	$\rightarrow$	$A_3 A_5 A_6$
A_6	$\rightarrow$	$A_4 A_5$
A_7	$\rightarrow$	$A_7 A_8$
A_8	$\rightarrow$	A_9
A_9	$\rightarrow$	A_{10}
A_{10}	$\rightarrow$	A_{11}
A_{11}	$\rightarrow$	A_{11}

f. Weighting

After the FLRG construction is completed, the weighting process is carried out by identifying the frequency of identical relationships within the FLRG. This process, which involves assigning weights to each fuzzy relation, is presented in Table 8. Suppose there is a fuzzy set $A_1 \rightarrow A_1, A_2$ then it can be known for the fuzzy relation $A_1 \rightarrow A_1$ as much as one and $A_1 \rightarrow A_2$ as much as one. Then based on the FLRG, the weighted $w_{11} = 1$ (from A_1) and $w_{12} = 1$ (from A_2) are obtained. So that a weight matrix is formed which can be written as $W_t = [w_1 w_2] = [1 1]$. If written in the form of a weighted fuzzy relation, the result is $A_1 \rightarrow A_1, A_2$. Where the weighting matrix can be seen in Table 9.

Table 8. FLRG weighting

<i>Current State</i>		<i>Next State</i>
A_1	$\rightarrow$	$A_1 A_3$
A_2	$\rightarrow$	A_3
A_3	$\rightarrow$	$A_1 A_2 A_3 A_4$
A_4	$\rightarrow$	$A_4 A_5 A_6 A_7$
A_5	$\rightarrow$	$A_3 A_5 A_6$
A_6	$\rightarrow$	$2(A_4) A_5$
A_7	$\rightarrow$	$A_7 A_8$
A_8	$\rightarrow$	A_9
A_9	$\rightarrow$	A_{10}
A_{10}	$\rightarrow$	A_{11}
A_{11}	$\rightarrow$	A_{11}

Table 9. Weighting matrix

FLRG	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈	A ₉	A ₁₀	A ₁₁
A ₁	1	1	0	0	0	0	0	0	0	0	0
A ₂	0	1	1	1	0	0	0	0	0	0	0
A ₃	1	1	0	0	0	0	0	0	0	0	0
A ₄	0	0	0	1	1	1	1	0	0	0	0
A ₅	0	0	1	0	1	1	0	0	0	0	0
A ₆	0	0	0	2	1	0	0	0	0	0	0
A ₇	0	0	0	0	0	0	1	1	0	0	0
A ₈	0	0	0	0	0	0	0	0	1	0	0
A ₉	0	0	0	0	0	0	0	0	0	1	0
A ₁₀	0	0	0	0	0	0	0	0	0	0	1
A ₁₁	0	0	0	0	0	0	0	0	0	0	1

The next step is to transfer the FLRG weights into the form of a standardised weight matrix (W^*). Suppose there is a fuzzy set $A_1 \rightarrow A_1, A_2$. which has weighted $w_{11} = 1$ (from A_1) and $w_{12} = 1$ (from A_2) with matrix $W_t = [w_{11} \ w_{12}] = [1 \ 1]$, so that the standardised weight matrix becomes.

$$W_{ij}^* = \frac{w_{ij}}{\sum_{j=1}^n w_{ij}}$$

$$W_{11}^* = \frac{1}{1+1} = 0.5$$

$$W_{12}^* = \frac{1}{1+1} = 0.5$$

So for the fuzzy set in $A_1 \rightarrow A_1, A_2$ after transferring the FLRG weights into a standardised weighting matrix form, the standardised weighting matrix value is obtained, namely $W_{ij}^* = [W_{11}^* \ W_{12}^*] = [0.5 \ 0.5]$. The step applies onwards for other fuzzy sets to transfer the standardised weighting matrix values which can be seen in Table 10.

Table 10. Standardisation weighting matrix

FLRG	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇	A ₈	A ₉	A ₁₀	A ₁₁
A ₁	0.5	0.5	0	0	0	0	0	0	0	0	0
A ₂	0	0.3333	0.3333	0.3333	0	0	0	0	0	0	0
A ₃	0.5	0.5	0	0	0	0	0	0	0	0	0
A ₄	0	0	0	0.25	0.25	0.25	0.25	0	0	0	0
A ₅	0	0	0.3333	0	0.3333	0.3333	0	0	0	0	0
A ₆	0	0	0	0.6667	0.3333	0	0	0	0	0	0
A ₇	0	0	0	0	0	0	0.5	0.5	0	0	0
A ₈	0	0	0	0	0	0	0	0	1	0	0
A ₉	0	0	0	0	0	0	0	0	0	1	0
A ₁₀	0	0	0	0	0	0	0	0	0	0	1
A ₁₁	0	0	0	0	0	0	0	0	0	0	1

g. Forecasting Value Defuzzification Process

In the Cheng model forecasting method, after the FLRG generation process is formed, weights are added to calculate defuzzification. In the calculation process, it will use a standardised weight matrix

(W^*) and the middle value in the interval or midpoint (m_i) for each relation in fuzzy $A_i = A_1, A_2, \dots, A_n$.

Suppose the researcher will calculate the defuzzification result of the forecasting value in the fuzzy set A_1 . Then according to the calculation of the Cheng model formula for the F_i forecasting value defuzzification process as follows.

$$\begin{aligned} F_1 &= w_{11}^*(m_1) + w_{12}^*(m_2) \\ &= 0.5(99.03) + 0.5(99.35) \\ &= 99.19 \end{aligned}$$

From the results of the forecasting calculation above, the defuzzification result of the forecasting value in the fuzzy set A_1 is 99.19. This step applies to each fuzzy set relation A_i to get the results or future forecasting values. The complete results of the forecasting value defuzzification process on all fuzzy set relations, namely 11 groups, can be seen in Table 11.

Table 11. Defuzzification result forecasting value

Current State		Next State	Defuzzification
A_1	→	$A_1 \ A_3$	99.19
A_2	→	$A_2 \ A_3 \ A_4$	99.67
A_3	→	$A_1 \ A_2$	99.19
A_4	→	$A_4 \ A_5 \ A_6 \ A_7$	100.59
A_5	→	$A_3 \ A_5 \ A_6$	100.28
A_6	→	$2(A_4) \ A_5$	100.12
A_7	→	$A_7 \ A_8$	101.65
A_8	→	A_9	103.36
A_9	→	A_{10}	104.65
A_{10}	→	A_{11}	105.93
A_{11}	→	A_{11}	105.93

After the defuzzification process was carried out, the predicted FER value for January 2024 was obtained using the FTS Cheng method, based on the FER data from January 2022 to December 2023 presented in Table 12. However, the Fuzzy Time Series Cheng method has limitations, namely that the prediction results tend to stabilize after several periods, making it less than optimal for long-term forecasting. Additionally, this method has not yet considered external factors such as planting seasons, climate change, and food price policies, which can significantly affect the FER value.

Table 12. Data estimation

t	Month	Actual Data	Estimation
1	Jan-22	100.79	
2	Feb-22	100.5	100.12
3	Mar-22	100.58	100.28
⋮	⋮	⋮	⋮
22	Okt-23	104.92	104.65
23	Nov-23	106.23	105.93
24	Des-23	106.47	105.93

h. Calculating the MAPE Value

The purpose of calculating the Mean Absolute Percentage Error (MAPE) value is to determine the level of accuracy in forecasting by measuring the accuracy of the forecasting results. In a forecasting situation, it contains a degree of uncertainty or a level of error (Error). So, what must be done is to find the level of error or error obtained from the difference between actual data and forecasting data. The MAPE value of 0.3027% and the accuracy of the forecasting results is 99.6973%. From these results, it can be concluded that forecasting using the fuzzy time series method in forecasting FER in Southeast Sulawesi can be said to be 'Very Good' because it has a MAPE value of less than 10%.

i. Forecasting Farmer Exchange Rate in Southeast Sulawesi Province in 2024

FER data for future periods can be forecasted by considering each forecasting output produced as input data to forecast the next period. For example, forecasting in January 2024 ($t=15$). It is known that the forecast result in December 2023 was 105.93. This result will be used as actual data in January 2024 as well as in the February and March 2024 periods. Because 105.93 is in the u_{11} interval and by looking at the previous period, December 2023 ($t=14$), it is fuzzified to A_{11} . So the FLR obtained is $A_{11} \rightarrow A_{11}$ and the FLRG obtained is $A_{11} \rightarrow A_{11}$. so the following forecasting results are obtained.

$$\begin{aligned} F_{11} &= w_{111}^*(m_{11}) \\ &= 1(105.93) \\ &= 105.93 \end{aligned}$$

In this study, FER forecasting will be carried out in Southeast Sulawesi Province for the next three months, namely January, February, March 2024. The forecast results can be seen in Table 13.

Table 13. Forecast results for the next three months

t	Month	Fuzzification	Forecast
25	Jan-24	A_{11}	105.93
26	Feb-24	A_{11}	105.93
27	Mar-24	A_{11}	105.93

The plot of actual data and FER forecast results for the next three months can be seen in Figure 3.

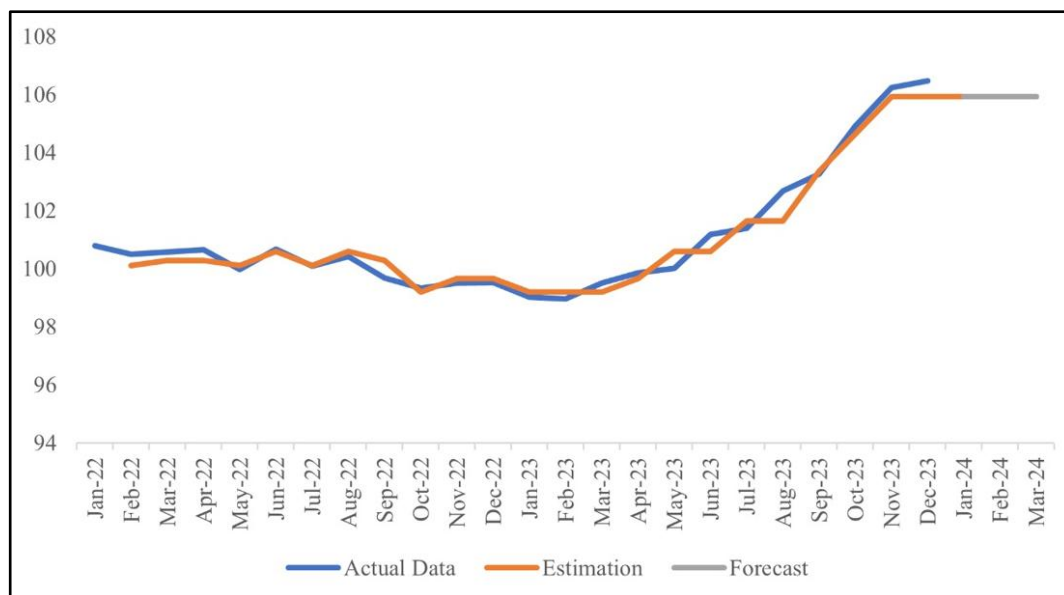


Figure 3. FER forecast result

In Figure 3, it can be seen that the actual data plot is shown with a blue line, the estimated value is shown with an orange line, and the forecast results are shown with a grey line. The forecast results produce the same fuzzification value for the next three periods, namely A_{11} . The forecasting results graph shows that the FER value for the next three periods tends to be stable. This condition indicates that farmers' purchasing power is expected to remain at a moderate surplus level as long as there are no significant changes in commodity prices and consumption costs. Fluctuations in FER in the previous period were primarily influenced by changes in income from commodity sales and the level of production costs and household needs. When the selling price of agricultural products decreased or production costs increased, FER tended to fall, weakening farmers' purchasing power. Conversely, when income increased or costs could be kept down, FER rose and farmers' welfare improved. Thus, the pattern of FER fluctuations reflects the level of vulnerability of farmers' welfare in Southeast Sulawesi

Province to price and cost of living dynamics. Although the forecasting results show a stable trend. periodic monitoring is still necessary to anticipate future changes in economic conditions.

4. Conclusion

The conclusion that can be drawn in this study is that forecasting with the fuzzy time series cheng method can be used to forecast the Farmer Exchange Rate (FER) in Southeast Sulawesi Province in 2024. The FER forecast results in January 2024 were obtained at 105.93 and the FER was predicted to be the same as in February and March 2024. The forecast results of $FER > 100$. which means that farmers experience a surplus. the price of production rises greater than the increase in consumption prices. Farmers' income increased more than their expenditure. Accuracy by looking at the accuracy of the mean absolute percentage error (MAPE). The MAPE obtained is 0.3027%. which means that the accuracy of the model obtained is 99.6973%. from these results it can be concluded that the criteria for forecasting results using the Cheng fuzzy time series method are included in the 'very good' criteria. because they have a MAPE value of less than 10%.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data from BPS-Statistics Indonesia.

Credit Authorship

Rastina: Conceptualization, Methodology, Writing, Editing, Visualization. **Lilis Laome:** Conceptualization, Methodology, Editing, Validation. **Bahriddin Abapihi:** Methodology, Reviewing. **Gusti Ngurah Adhi Wibawa:** Methodology. **Mukhsar:** Reviewing. **Makkulau:** Reviewing. **Gama Putra Danu Sohieben:** Editing, Reviewing, Supervision. **Sukim:** Supervision. **Fathurrahman Yahyasatrio:** Supervision.

References

- [1] M. Arumsari. A. Tri. and R. Dani. "Peramalan data runtun waktu menggunakan model hybrid time series regression – Autoregressive integrated moving average." vol. 02. no. 01. pp. 1–12. 2021.
- [2] Noviani. "Perbandingan metode fuzzy time series markov chain dan fuzzy time series cheng (Studi kasus: Rata-rata harga beras)." Universitas Islam Negeri Maulana Malik Ibrahim.. 2020.
- [3] C. Keumala and Z. Zainuddin. "Indikator kesejahteraan petani melalui nilai tukar petani (NTP) dan pembiayaan syariah sebagai solusi Cut Muftia Keumala Zamzami Zainuddin pendahuluan salah satu sumber kebutuhan utama manusia berasal dari sektor." vol. 9. pp. 129–149. 2018.
- [4] BPS. "Statistik Nilai Tukar Petani Provinsi Sulawesi Tenggara Triwulan III." 2023.
- [5] Purba. S. F., Yulianti. A., Raphael. Y., & Khotimah. H. (2024). Determinan kesejahteraan petani tanaman pangan di Provinsi Jawa Barat. *Jurnal Ilmu Pertanian Indonesia*. 29(1). 59–67.
- [6] Rahmah. P. A., Hayati. M. N., Cahyaningsih. A. (2025). Peramalan nilai tukar petani Kalimantan Timur menggunakan metode neural network. *Indonesian Journal of Applied Statistics and Data Science*. 2(1).
- [7] Kuddi. B. F and A. D. Saputro (2024). Peramalan nilai tukar petani di Provinsi Papua menggunakan model hybrid ARIMA-NN. SAINTIFIK: *Jurnal Matematika. Sains. dan Pembelajarannya*. 10(1). 135-140
- [8] M. I. Riyadh. "Analisis nilai tukar petani komoditas tanaman pangan di Sumatera Utara (Analysis of farmers term of trade of crops commodities in North Sumatra)." no. 191. pp. 17–32. 2015.
- [9] L. N. Mustika. "Perbandingan metode fuzzy time series cheng dan markov chain pada peramalan nilai tukar petani di Indonesia." Universitas Islam Negeri Maulana Malik Ibrahim.. 2021.
- [10] M. Muhammad *et al.*. "Peramalan nilai tukar petani subsektor peternakan menggunakan fuzzy time series lee." vol. 3. no. 1. pp. 1–15. 2021.
- [11] Sumartini. M. N. Hayati. and S. Wahyuningsih. "Peramalan menggunakan metode fuzzy time series cheng forecasting using fuzzy time series cheng method." vol. 8. pp. 51–56. 2017.
- [12] A. R. Prayogi. "Demand forecasting penggunaan energi listrik (Kwh) menggunakan fuzzy time series cheng." Universitas Islam Indonesia. 2018.
- [13] M. L. Tauryawati and M. I. Irawan. "Perbandingan metode fuzzy time series cheng dan metode box-jenkins untuk memprediksi." vol. 3. no. 2. 2014.
- [14] Prayetno. R. S., F. Rakhmawati. (2025). Comparison of the cheng and lee fuzzy time series methods for predicting clean water supply in Asahan Regency. *Jatilima: Jurnal Multimedia dan Teknologi Informasi* 7(3). 507-515.
- [15] Tursina. R. Septiriana. & I. Varian. (2023). Prediksi indeks harga konsumen menggunakan metode fuzzy time series cheng. *Jurnal Locus: Penelitian & Pengabdian* 2(1). 51-59.
- [16] Putri. I. A., N. E. Maidah., & M. A. Furqon. (2024). Penerapan metode fuzzy time series cheng pada peramalan inflasi di Indonesia. *Komputika: Jurnal Sistem Komputer*. 13(2). 183-191.
- [17] Amalani. M. Z. B., Santoso. N. A., & Syefudin. (2025). Perbandingan metode ARIMA dan fuzzy time series dalam peramalan harga eceran daging sapi di Indonesia. *RIGGS: Journal of Artificial Intelligence and Digital Business*. 4(3). 2232-2242.
- [18] Ella. A.P., & Deni. A. (2020). Penerapan fuzzy time series dalam peramalan harga minyak serai pada CV Agam Jaya Atsiri. *Jember: Fakultas Teknik Universitas Muhammadiyah Jember*.
- [19] T. Fahmi. Sudarno. and Y. Wilandari. "Perbandingan metode eksponensial tunggal dan fuzzy time series untuk memprediksi indeks harga saham gabungan." *Gaussian*. vol. 2. no. 2. pp. 137–146. 2013.
- [20] A. Fadhillah. B. Martalelli. and N. Rita. "Perbandingan model chen dan model cheng pada algoritma fuzzy time series untuk prediksi harga bahan pokok." Universitas Maritim Raja Ali Haji. 2017.
- [21] Bidin. J., dkk. (2022). Cheng fuzzy time series model to forecast the price of crude oil in Malaysia. *Journal of Computing Research and Innovation (JCRINN)*. 7(2). 196-210.



Satellite Imagery for Classification Analysis of Abrasion Areas on Panaitan, Banten

Ghaffar Ismail^{1*}, Robert Kurniawan², Silvia Ni'matul Maula³

^{1,2}Politeknik Statistika STIS, Jakarta, Indonesia, ³School of Public Policy and Management, Tsinghua University, Beijing, China

*Corresponding Author: E-mail address: 222112071@stis.ac.id

ARTICLE INFO

Article history:

Received 30 September, 2025

Revised 6 December, 2025

Accepted 23 December, 2025

Published 31 December, 2025

Keywords:

Abrasion; Classification;
Machine learning; Panaitan;
Satellite imagery

Abstract

Introduction/Main Objectives: Abrasion causes severe environmental degradation and socio- economic losses, Waton and Karang Gundul Islands have already subsided due to erosion, posing risks to Panaitan Island, a national park that also faces deforestation, infrastructure development, and vegetation loss which may intensify abrasion. **Background Problems:** Limited spatial data on coastal abrasion in Panaitan Island hampers effective monitoring and management, highlighting the need for spatially explicit analysis. **Novelty:** This study identified and classified abrasion-prone areas on Panaitan Island (a rarely exposed island) with rarely variables which have impactful indices such as MVI, TCI, and LSWI. **Research Methods:** Landsat 8 and Sentinel-2 imagery from 2018 and 2023 were analyzed to assess changes in vegetation, mangroves, surface temperature, and soil moisture. Random Forest, Support Vector Machine, and Logistic Regression were employed to classify abrasion-prone areas. **Finding/Results:** The analysis revealed signs of abrasion covering 2.04 km², with Random Forest achieving the highest accuracy (82.23%) and NDVI as the most influential variable; abrasion was mainly associated with declining forest and mangrove cover, soil moisture showed weak correlation, while moderate surface temperature had a positive effect. Preventive measures such as reforestation and mangrove rehabilitation are recommended to mitigate risks and ensure long-term environmental sustainability.

1. Introduction

Shorelines are threatened by coastal abrasion, which also damages infrastructure and communities, jeopardizes human safety, and degrades the environment [1]. Land erosion in Europe reached about 4,500 km² between 1975 and 2006, and by 2050, it is expected to reach 5,800 km² [2]. Over one-third of China's coastline has been eroded, and in certain places, shoreline retreat can approach 300 meters annually [3]. It is also anticipated that Indonesia, which has more than 17,000 islands and 80,000 km of coastline, will be badly impacted. With 115 islands expected to be at risk of submersion by 2100 and 28 islands reported to have sunk in 2011, abrasion poses a concern to Indonesia's small islands [4]. Banten Province is one of the several coastal regions in Indonesia where abrasion is now occurring.

Karang Gundul Island and Waton Island were lost to coastal abrasion in 2015, according to the Banten Province Government [5]. Panaitan Island and other nearby islands may also be under danger due to this circumstance. According to the Republic of Indonesia's Decree of the Minister of Forestry

No. SK.3658/Menhut-VII/KUH/2014, issued May 5, 2014, regarding the Designation of the Ujung Kulon National Park Forest Area, Panaitan Island is a protected forest area that is home to a variety of plants and animals and is a well-liked destination for ecotourism and surfing. But there are now serious risks: the number of unlawful operations in Ujung Kulon National Park has increased dramatically, from 216 in 2019 to 680 in 2023. Abrasion can be accelerated by a variety of activities, including logging, illegal road construction, and the removal of vegetation [6].

Because their root systems prevent soil and sand from eroding, green vegetation, especially mangroves, is crucial in reducing abrasion [7]. Paradoxically, despite having the largest mangrove environment, Pandeglang Regency also saw the highest mangrove loss from abrasion, with 130.73 hectares lost in 2023 [8]. Rising sea levels and temperatures brought on by climate change make abrasion much worse [1]. These circumstances give rise to worries over possible abrasion on Panaitan Island, which may eventually cause the island to sink. Monitoring signs of abrasion on Panaitan Island is further hampered by the absence of island-level statistics and information as opposed to district/city-level data.

This study examines and categorizes regions that are prone to abrasions while examining important contributing elements. A number of indices are used, such as the Normalized Difference Vegetation Index (NDVI) for vegetation density [9], the Modified Normalized Difference Water Index (MNDWI) for shoreline changes using an overlay approach [10, 11], the Mangrove Vegetation Index (MVI) for detecting mangrove distribution [12], the Land Surface Temperature (LST) for tracking changes in surface temperature related to climate change [13], and the Land Surface Water Index (LSWI) for soil moisture conditions [14]. For classification modeling, machine learning techniques like Random Forest, Support Vector Machine (SVM), and Logistic Regression are used [15].

Various techniques for abrasion analysis and satellite image classification have been widely applied in related studies. One study assessed the influence of vegetation density on shoreline change using NDVI and DSAS and found that its effect was relatively small, accounting for only about 6.5% [16]. Another study comparing machine learning algorithms such as Random Forest, SVM, and stacking reported that SVM achieved the best performance, with accuracy reaching 0.983 [17]. Abrasion modeling using overlay and SWIR-1 techniques indicated a potential abrasion area of 118.50 hectares over a ten-year period [11]. Shoreline change analysis has also been conducted using tidal variables and linear regression methods [18]. In addition, the application of a Decision Tree algorithm using MNDWI, NDVI, EVI2, and GRVI variables resulted in an accuracy of 82.25% in coastal abrasion studies [19]. An integrated approach combining statistical techniques (EPR, AOR, and LR), remote sensing, and GIS achieved an RMSE of ± 10 m [20]. Furthermore, in multispectral image classification, Random Forest produced the highest accuracy (81.2%) compared to Decision Tree and Rule-Based approaches [21].

This study is unusual because it combines remote sensing technology with a special set of indices like MVI, LSWI, and TCI—that are rarely used in earlier abrasion classification models. Imagery for each variable in this study was obtained from Landsat 8 and Sentinel-2. However, the study is limited by the omission of other variables that could affect abrasion, such as wave height. Additionally, it thoroughly examines the interactions between variables. Additionally, research concentrating on the risk of abrasion on a small island have not yet been conducted on Panaitan Island. It is anticipated that this study will provide light on how to stop small islands from being lost to abrasion. This work intends to create a machine learning-based categorization model of abrasion sites by identifying probable abrasion risks on Panaitan Island, taking into account factors including vegetation density, soil conditions, surface temperature, and mangrove distribution. In order to classify abrasion-prone areas of Panaitan Island using optimal machine learning techniques (Support Vector Machine, Random Forest, and Logistic Regression) with satellite imagery, this study aims to identify key variables for classifying abrasion areas and estimating impacted zones, while providing remote-sensing-based insights to support abrasion monitoring and mitigation on Panaitan Island.

2. Material and Methods

2.1. Data Source

The data were obtained from Sentinel-2 and Landsat imagery (Landsat-5, Landsat-7, and Landsat-8) accessed through the Google Earth Engine (GEE) Data Catalog to analyze shoreline abrasion on Panaitan Island. Shoreline abrasion was defined as the dependent variable and classified into abrasion and non-abrasion categories based on coastline overlay analysis, while the independent variables consisted of changes in the Mangrove Vegetation Index (MVI), Normalized Difference Vegetation Index (NDVI), Land Surface Water Index (LSWI), and Temperature Condition Index (TCI) observed from January 1 to December 31 in 2018 and 2023. A five-year observation interval was selected because

it is sufficient to capture abrasion dynamics compared to shorter periods [11,16]. The imagery was processed using annual median pixel values, which were averaged to generate a $20 \text{ m} \times 20 \text{ m}$ spatial grid in QGIS. This spatial resolution was chosen to preserve detailed spatial variation while effectively covering coastal areas and is consistent with previous studies employing grid sizes of $50 \times 50 \text{ m}$ [22] and $10 \times 10 \text{ m}$ [23] for vegetation and erosion analyses. The characteristics of the variables and data sources used in this study are summarized in Table 1.

Table 1. Data sources

No.	Variable	Description	Sources	References
1.	X1	Change in Mangrove Vegetation Index (ΔMVI) 2018 and 2023	Sentinel-2	[12, 24, 25]
2.	X2	Change in Normalized Difference Vegetation Index (ΔNDVI) 2018 and 2023	Sentinel-2	[26]
3.	X3	Change in Land Surface Water Index (ΔLSWI) 2018 and 2023	Sentinel-2	[14, 27]
4.	X4	Change in Temperature Condition Index (ΔTCI) 2018 and 2023	Landsat-8	[28]
5.	Y	Change in Abrasion (1) and Non-Abrasion (0) Area Categories 2018 and 2023	Landsat-5,7,8	[10]

2.2. Analysis Method

The analytical procedure began with the preparation of satellite image data, including extraction, scaling, and categorization, followed by class balancing using the Synthetic Minority Over-sampling Technique (SMOTE) to address the lower number of abrasion samples compared to non-abrasion samples. The dataset was then divided into training and testing subsets. Exploratory data analysis was conducted using descriptive statistics, while variable distributions and relationships were examined through boxplots and correlation analysis, complemented by multicollinearity testing using correlation matrices and the Variance Inflation Factor (VIF), where VIF values greater than 10 indicate multicollinearity [29]. This step supports attribute selection and improves the stability and performance of machine learning models [30]. Abrasion classification was subsequently performed using Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression, with optimal model parameters determined through Bayesian optimization and five-fold cross-validation. Model performance was evaluated using accuracy, precision, recall, and F1-score, supported by confusion matrix validation, georeferencing, and spatial visualization of the classification results. In addition, Feature Importance Ranking (FIR) was applied to assess the contribution of each variable, with RF and SVM selected based on their documented superior performance in satellite image classification [17,21], while Logistic Regression was included due to the binary nature of the abrasion classification.

2.2.1. Random Forest

Random Forest is an ensemble machine learning technique including many decision trees utilized for categorization purposes [31]. The method employed for identifying optimal parameters is Grid Search. This study's Random Forest encompasses $n_estimators$ from 1 to 100, max_depth from 1 to 10, $min_samples_split$ from 2 to 10, and $min_samples_leaf$ from 1 to 10. Furthermore, there exists a bootstrap parameter with options of True or False, along with $max_features$ varying from 0.1 to 1.

2.2.2. Support Vector Machine

The Support Vector Machine (SVM) seeks to identify the optimal separating hyperplane for precise data classification. The objective is to reduce classification errors in both training and unseen test datasets [32]. This study's SVM parameters consist of C values of 0.1, 1, and 10; gamma values of 1, 0, 0.1, and 0.01; and kernel options of 'rbf' and 'linear'.

2.2.3. Logistic Regression

Logistic regression is frequently selected for binary response models. Its reliance on chances facilitates interpretation, rendering the outcomes comprehensible to the general populace. Conversely, the probit and complementary log-log models are exclusively applicable to prospective data due to their dependence on probability [33]. This study's Logistic Regression settings comprise C values of 0.01, 0.1, 1, and 10; max_iter values of 100, 200, and 300; solver options of 'lbfgs', 'liblinear', and 'saga'; and class_weight options of None and 'balanced'.

2.2.4. Model Evaluation

The optimal machine learning model was selected to achieve the most effective categorization results for abrasion-prone areas. This study employed four primary metrics to evaluate model performance: accuracy, precision, recall, and F1-score [34]. The equations for each assessment metric are as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \quad (4)$$

where TP (True Positive) denotes the number of abrasion-prone areas correctly classified as prone, TN (True Negative) represents the number of non-abrasion-prone areas correctly classified as non-prone, FP (False Positive) refers to the number of non-abrasion-prone areas incorrectly classified as prone, and FN (False Negative) indicates the number of abrasion-prone areas incorrectly classified as non-prone.

3. Results and Discussion

This chapter aims to identify and analyze each change in variables in this study, after which a classification of abrasion and non-abrasion areas is carried out so that all existing information can be elaborated and analyzed. This study was performed on Panaitan Island, Ujung Kulon National Park, utilizing satellite imagery with a resolution of 20m x 20m. The abrasion on Panaitan Island was examined by overlay analysis. A total of 11,545 data records were collected, comprising 4,867 records in the abrasion category (1) and 6,678 records in the non-abrasion category (0).

3.1. Exploration of variables for changes in environmental conditions

Prior to modeling the classification of abrasion and non-abrasion zones, researchers examined the environmental conditions in Panaitan, focusing on changes that transpired between 2018 and 2023 as indicators or variables for the model.

3.1.1. Coastline

Shoreline abrasion in this study was identified through an overlay analysis of coastline maps derived from satellite imagery in 2018 and 2023, enabling the detection of coastline retreat as an indicator of abrasion. Grid cells with a spatial resolution of 20 m × 20 m were classified as abrasion (1) when they intersected areas showing coastline retreat—depicted as red zones between the two observation years—and as non-abrasion (0) otherwise, as illustrated in Figure 1. The spatial distribution reveals that multiple sections of Panaitan Island exhibit clear signs of abrasion, with the most pronounced changes occurring along the eastern coastline. This pattern suggests that the eastern coast is more exposed to wave action and hydrodynamic forces, making it particularly vulnerable to shoreline retreat over time.

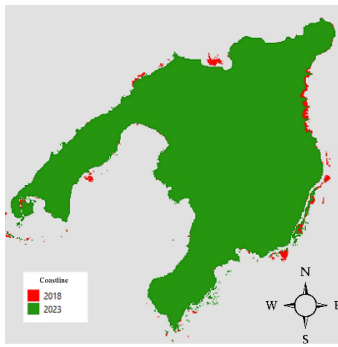


Figure 1. Map of indication of abrasion area of Panaitan Island

3.1.2. Mangrove Vegetation Index

The Mangrove Vegetation Index (MVI) assesses the health and spread of mangrove vegetation through the comparison of NIR, Green, and SWIR values. Mangroves serve as a significant signal closely associated with abrasion overall. The Mangrove Vegetation Index (MVI) assesses mangrove health, distribution, and density using Sentinel-2 data, calculated from the green, near-infrared (NIR), and shortwave infrared (SWIR) bands as: $MVI = (NIR - Green) / (SWIR - Green)$ [12, 24, 25]. Figures 2(a) and 2(b) indicate that certain coastal regions of Panaitan Island possess mangrove vegetation. Nevertheless, in comparing the two years in their entirety, numerous regions exhibited a decline in mangrove vegetation. Moreover, an increased index value correlates with greater mangrove density. Mangroves function as wave breakers and enhance soil stability during significant wave events.



Figure 2. (a) MVI distribution map 2018 Panaitan; (b) MVI distribution map 2023 Panaitan

3.1.3. Normalized Difference Vegetation Index

Mangroves serve as a significant signal closely associated with abrasion overall. The Normalized Difference Vegetation Index (NDVI) monitors vegetation density with an index value range from -1 to +1. The Normalized Difference Vegetation Index (NDVI) assesses vegetation health using the red and near-infrared (NIR) bands and is calculated as $(NIR - Red) / (NIR + Red)$ [35]. According to Figures 3(a) and 3(b), from 2018 to 2023, certain regions of the coast of Panaitan Island, specifically the southwest and east, exhibited a diminishing green hue, signifying a reduction in vegetation density. In contrast, the central portion of the forest retained its density and verdancy due to the absence of residential developments.



Figure 3. (a) NDVI distribution map 2018 Panaitan; (b) NDVI distribution map 2023 Panaitan

3.1.4. Land Surface Water Index

The Land Surface Water Index (LSWI) is used to measure soil moisture with an index range of -1 to +1. The Land Surface Water Index (LSWI) evaluates current soil moisture using the near-infrared (NIR) and shortwave infrared (SWIR) bands and is calculated as $(NIR - SWIR) / (NIR + SWIR)$ [14]. Figures 4(a) and 4(b) demonstrate that negative vegetation index values signify aquatic regions, including beaches and the ocean. From 2018 to 2023, Panaitan Island's coastline witnessed a reduction in vegetation density, however the interior forest retained its density and verdancy due to the absence of settlements. In the figure, there is an indication that the 2018 image appears darker. This may be because throughout that year, the dry season was more dominant, resulting in lower LSWI values, which appear darker. Conversely, during years with a longer rainy season, LSWI values are generally higher and therefore appear brighter. The imagery in GEE has been processed using a cloud-mask filter to remove clouds as effectively as possible, and the images shown represent the best results obtained in this study for the LSWI imagery.



Figure 4. (a) LSWI distribution map 2018 Panaitan; (b) LSWI distribution map 2023 Panaitan

3.1.5. Temperature Condition Index

The Temperature Condition Index (TCI) assesses coastal thermal pressure. The TCI index spans from 0 to 100, indicating that a value closer to zero corresponds to a higher surface temperature pressure. The Temperature Condition Index (TCI) evaluates drought conditions using land surface temperature (LST) and is calculated as $((LST_{max} - LST_i) / (LST_{max} - LST_{min})) \times 100$ [13]. Figures 5(a) and 5(b) show a striking difference in the surface temperature index between sea and land in 2018 and 2023. Higher TCI values indicate cooler and more stable temperatures, with islands having higher values due to the presence of forests. Overall, 2018 had more areas with cooler temperatures than 2023, although the average coastal TCI showed the opposite. There are several limitations when using satellite imagery, one of which is the presence of clouds. However, in this study, a cloud-mask filter in GEE was applied and tested multiple times, and the images shown above represent the best results obtained for the TCI imagery.



Figure 5. (a) TCI distribution map 2018 Panaitan; (b) TCI distribution map 2023 Panaitan

3.1.6. Relationship Between Variables

All independent variables have a correlation of less than 0.8 with one another. Abrasion exhibits a negative correlation with alterations in NDVI (-0.49) and MVI (-0.24), suggesting that a decline in vegetation and mangrove density heightens the risk of abrasion. LSWI exhibits a correlation of 0.00,

signifying an absence of linear association with abrasion. TCI exhibits a positive correlation of 0.23, suggesting that regions with declining surface temperatures are more prone to abrasion; however, this value may be affected by external variables such as climate change or pandemics, as illustrated by the correlation coefficients in Figure 6.



Figure 6. Heatmap of correlation coefficient between variables

3.1.7. Characteristics of Abrasion and Non-Abrasion Area Categories

The index delta is calculated by subtracting the 2018 index from the 2023 index; thus, a negative value indicates that the 2023 index is lower than the 2018 index, and vice versa. According to Table 2, the abrasion zone on Panaitan Island exhibits an average decline in NDVI and MVI, signifying a reduction in the density of green vegetation and mangroves that mitigate abrasion, this could be caused by illegal activities that occur in the Ujung Kulon area, one of which is Panaitan Island such as logging, land clearing, and resource extraction. The variation in LSWI is minimal between the abrasion and non-abrasion zones; however, the abrasion zone appears drier in 2023 owing to diminished water-retaining vegetation. The TCI in the abrasion zone is positive, signifying that the surface temperature in 2023 was lower than in 2018, potentially affected by external forces such as the pandemic that diminished human activity [36].

Table 2. Average value of index change by category

Index	Non-Abrasion (0)	Abrasion (1)
ΔMVI	0.206443	-0.28326
ΔTCI	-0.319361	0.438194
$\Delta NDVI$	0.420855	-0.577454
$\Delta LSWI$	-0.000786	0.001078

3.2. Classification of Abrasion Areas

Data extracted from satellite photos will be utilized to predict the classification of abrasion zones through machine learning techniques. The modeling step encompasses multicollinearity assessment, machine learning model creation, evaluation of model performance, essential feature selection, and classification analysis of abrasion-susceptible regions on Panaitan Island.

3.2.1. Multicollinearity Test

This investigation necessitates a multicollinearity test due to the utilization of logistic regression. The test findings demonstrate that all VIF values for the variables in this study are below 10 ($VIF < 10$).

The VIF test findings indicate no multicollinearity in the data. By adhering to the criteria for multicollinearity, all independent variables may be utilized in the training and validation processes to identify optimal parameters and evaluate estimations.

3.2.2. Classification Models

This study analyzes machine learning algorithms based on supervised learning, specifically Random Forest, Support Vector Machine, and Logistic Regression. The optimal parameter search was conducted via Bayesian optimization with k-fold cross-validation [37]. The findings indicate that the ideal parameters for Random Forest are a max_depth of 8, max_features of 0.9395, min_samples_leaf of 4, min_samples_split of 7, n_estimators of 42, and random_state of 42. The ideal settings for the Support Vector Machine are $C = 10$, $\gamma = 1$, and a linear kernel. Logistic Regression attains optimal performance with $C = 0.1$, max_iter = 100, solver = liblinear, and a balanced class weight.

3.2.3. Machine Learning Model Performance

Model evaluation is an essential process for identifying the optimal model. The evaluation of the classification model in this study was performed utilizing accuracy, precision, recall, and F1-score [34]. The subsequent evaluation measures pertain to each machine learning algorithm. However, the analysis is limited to the period between 2018 and 2023. The spatial resolution of the data, derived from Landsat 8 and Sentinel-2 at 20×20 m per record in coastline area Panaitan's island, generates a large dataset that is effectively suitable for machine learning applications.

According to Table 3, the Random Forest model demonstrated superior performance compared to the other machine learning algorithms evaluated in this study. This model achieved an accuracy of 82.23%, with a precision of 0.82, a recall of 0.82, and an F1-score of 0.82, indicating a balanced and robust classification capability for both abrasion and non-abrasion classes. The strong performance of Random Forest can be attributed to its ability to handle nonlinear relationships and complex interactions among environmental variables without requiring strict distributional assumptions. Moreover, Random Forest is less sensitive to noise and multicollinearity than parametric models, making it particularly suitable for large, high-dimensional datasets derived from multi-source satellite imagery. Consequently, this model was selected to further analyze feature importance and assess the relative contribution of each environmental variable to shoreline abrasion classification.

Table 3. Machine learning model performance

Parameter	Random forest	Support vector machine	Logistic regression
Accuracy	0.8223	0.7434	0.7422
Precision	0.82	0.76	0.75
Recall	0.82	0.74	0.74
F1-score	0.82	0.74	0.74

3.2.4. Feature Importance

Feature importance for identifying factors that significantly influence the constructed model, as illustrated in Figure 7. NDVI is the most significant variable in the model. The values of each variable in the abrasion categorization are as follows: NDVI is 0.5304, TCI is 0.2131, LSWI is 0.154, and MVI is 0.1025. While NDVI exerts the most significant influence, all variables contribute quite uniformly to the model. This contrasts with another research, who determined that NDVI exerts only a 6.5% influence on coastline alterations, which is statistically insignificant [16]. This research is in line with other researches indicating that vegetation mitigates wave action, preserves soil moisture, endures wind, and stabilizes sediment [7, 38, 39]. Climate change has widespread impacts on maritime and coastal sectors, leading to rising temperatures, sea-level rise, and extreme weather events, which contribute to coastal abrasion that threatens ecosystems and coastal communities [1].

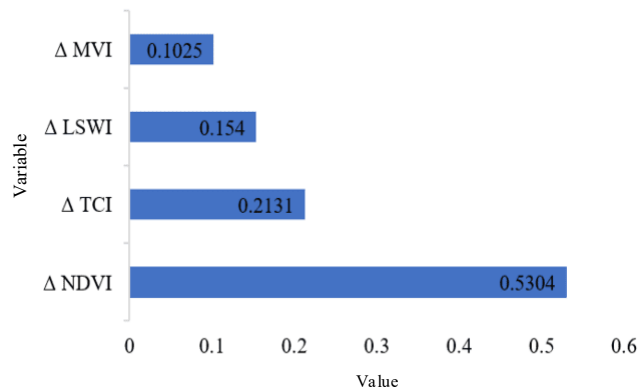


Figure 7. Feature importance graph random forest model

3.3. *Area Estimation and Classification Analysis of the Abrasion Area of Panaitan Island*

Spatial analysis conducted using QGIS indicates that Panaitan Island experienced shoreline abrasion covering an estimated area of 2.04 km² during the period 2018–2023, out of a total land area of 125.77 km² in 2023. Areas affected by abrasion are generally characterized by noticeable reductions in the density of green vegetation, particularly coastal forests and mangroves that play a critical role in stabilizing shorelines and dissipating wave energy. Correlation analysis shows that changes in soil conditions and moisture, as represented by LSWI, do not exhibit a strong linear relationship with abrasion occurrence. However, the machine learning results demonstrate that these variables still contribute meaningfully to abrasion classification, suggesting that their influence may be nonlinear or interact with other environmental factors. This finding underscores that shoreline abrasion on Panaitan Island is driven by a combination of vegetation degradation and environmental dynamics rather than by a single dominant factor.

4. Conclusion

This study employs remote sensing data from Panaitan Island, Banten, comprising 11,545 coastal records, to classify abrasion-prone areas based on variations in vegetation density (NDVI), mangrove density (MVI), soil moisture (LSWI), and surface temperature (TCI) using machine learning, identifying Random Forest as the best-performing model (82.23% accuracy) with NDVI as the most influential factor, estimating approximately 2.04 km² of abraded area from 2018 to 2023 linked to human activities such as illegal road construction and logging, while surface temperature, climatic changes, and vegetation conditions also serve as key indicators. The classification could achieve better performance if additional marine environmental variables, such as wave height and length, were included, although the study remains limited by its temporal span, spatial resolution, and unobserved environmental factors.

Ethics approval

Not required.

Acknowledgments

The author wishes to convey appreciation to the Ujung Kulon National Park Unit for its website and pertinent information concerning the natural conditions of Panaitan Island. The author extends profound gratitude to Mr. Robert for his helpful feedback and ongoing support during the study's development.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data from the Google Earth Engine (GEE) Data Catalog.

Credit Authorship

Ghaffar Ismail: Conceptualization, Methodology, Writing, Editing, Visualization. **Robert Kurniawan:** Reviewing, Methodology, Validation, Editing, Supervision. **Silvia Ni'matul Maula:** Data Curation, Editing, Validation.

References

- [1] N. S. Gravitationi, E., Rahardjo, M., & Pratiwi, "Apakah turis bersedia membayar untuk mitigasi abrasi? Sebuah studi di daerah pesisir di Yogyakarta, Indonesia," *Glob. J. Bus. Soc. Sci. Rev.*, pp. 5(2), 46–51, 2017, [Online]. Available: <https://ssrn.com/abstract=3002423>
- [2] D. Paprotny, P. Terefenko, A. Giza, P. Czaplinski, and M. I. Voutsoukas, "Future losses of ecosystem services due to coastal erosion in Europe," *Sci. Total Environ.*, vol. 760, p. 144310, 2021, doi: 10.1016/j.scitotenv.2020.144310.
- [3] F. Cai, X. Su, J. Liu, B. Li, and G. Lei, "Coastal erosion in China under the condition of global climate change and measures for its prevention," *Prog. Nat. Sci.*, vol. 19, no. 4, pp. 415–426, 2009, doi: 10.1016/j.pnsc.2008.05.034.
- [4] M. A. Marfai, *Catatan kecil tentang informasi geospasial (Edisi ke-1)*. Badan Informasi Geospasial, 2023.
- [5] Y. Deslatama, "Tidak dirawat, 2 pulau di Banten hilang," *Liputan6.com*. Accessed: Feb. 20, 2025. [Online]. Available: <https://www.liputan6.com/news/read/2253209/tidak-dirawat-2-pulau-di-banten-hilang>
- [6] Balai Taman Nasional Ujung Kulon, "Taman Nasional Ujung Kulon," Kementerian Lingkungan Hidup dan Kehutanan Republik Indonesia. Accessed: Mar. 28, 2025. [Online]. Available: <https://tnujungkulon.menlhk.go.id/>
- [7] A. Hakam, B. Istijono, and T. Ophiandri, "Vegetations as beach protection: Simulations of root protection mechanism against abrasion," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 6, pp. 2316–2321, 2017, doi: 10.18517/ijaseit.7.6.3445.
- [8] Dinas Lingkungan Hidup dan Kehutanan Prov. Banten, "Updating data luas kawasan mangrove di Provinsi Banten," 2023.
- [9] G. Kaplan and U. Avdan, "Mapping and monitoring wetlands using sentinel-2 satellite imagery," *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. 4, no. 4W4, pp. 271–277, 2017, doi: 10.5194/isprs-annals-IV-4-W4-271-2017.
- [10] R. Bishop-Taylor, R. Nanson, S. Sagar, and L. Lymburner, "Mapping Australia's dynamic coastline at mean sea level using three decades of Landsat imagery," *Remote Sens. Environ.*, vol. 267, no. March, p. 112734, 2021, doi: 10.1016/j.rse.2021.112734.
- [11] M. N. Nandaniko, Supriatna, and M. D. M. Manessa, "Modeling of coastline change by viewing trends from the abrasion and accretion rate in karawang regency," *IEEE*, pp. 0–4, 2019, doi: 10.1109/ICST47872.2019.9166274.
- [12] A. B. Baloloy, A. C. Blanco, R. R. C. Raymund Rhommel, and K. Nadaoka, "Development and application of a new mangrove vegetation index (MVI) for rapid and accurate mangrove mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, no. January, pp. 95–117, 2020, doi:

- 10.1016/j.isprsjprs.2020.06.001.
- [13] F. N. Kogan, "Global drought watch from space," *Bull. Am. Meteorol. Soc.*, vol. 78, no. 4, pp. 621–636, 1997, doi: 10.1175/1520-0477(1997)078<0621:GDWFS>2.0.CO;2.
- [14] Y. Zhou *et al.*, "Open surface water mapping algorithms: A comparison of water-related spectral indices and sensors," *Water (Switzerland)*, vol. 9, no. 4, 2017, doi: 10.3390/w9040256.
- [15] A. Ghorbanian, S. Zaghian, R. M. Asiyabi, M. Amani, A. Mohammadzadeh, and S. Jamali, "Mangrove ecosystem mapping using sentinel-1 and sentinel-2 satellite images and random forest algorithm in google earth engine," *Remote Sens.*, vol. 13, no. 13, pp. 1–18, 2021, doi: 10.3390/rs13132565.
- [16] E. D. Kusumaningrum, A. Saputra, and A. Nurwijayanti, "Analysis of vegetation density changes on coastline changes in the Glagah Coastal Area, Kulon Progo Regency 2018 and 2023," *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1357, no. 1, 2024, doi: 10.1088/1755-1315/1357/1/012009.
- [17] A. Rahman *et al.*, "Remote sensing applications: Society and environment performance of different machine learning algorithms on satellite image classification in rural and urban setup," *Remote Sens. Appl. Soc. Environ.*, vol. 20, no. July, p. 100410, 2020, doi: 10.1016/j.rsase.2020.100410.
- [18] M. Mahapatra, R. Ratheesh, and A. S. Rajawat, "Shoreline change analysis along the coast of South Gujarat, India, using digital shoreline analysis system," *J. Indian Soc. Remote Sens.*, vol. 42, no. 4, pp. 869–876, 2014, doi: 10.1007/s12524-013-0334-8.
- [19] D. N. D. Pratama, N. Khakhim, A. Wicaksono, A. Musthofa, and W. Lazuardi, "Spatio-temporal analysis of abrasion susceptibility effect on land cover in the coastal area of Bantul regency, Yogyakarta, Indonesia," *Int. J. Geoinformatics*, vol. 17, no. 4, pp. 109–126, 2021, doi: 10.52939/ijg.v17i4.1961.
- [20] B. Deepika, K. Avinash, and K. S. Jayappa, "Shoreline change rate estimation and its forecast: Remote sensing, geographical information system and statistics-based approach," *Int. J. Environ. Sci. Technol.*, vol. 11, no. 2, pp. 395–416, 2014, doi: 10.1007/s13762-013-0196-1.
- [21] T. M. Berhane *et al.*, "Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory," *Remote Sens.*, vol. 10, no. 4, 2018, doi: 10.3390/rs10040580.
- [22] I. N. Hidayati, R. Suharyadi, and P. Danoedoro, "Kombinasi indeks citra untuk analisis lahan terbangun dan vegetasi perkotaan," *Majalah Geografi Indonesia*, vol. 32, no. 1, p. 24, 2018. [Online]. Available: <https://doi.org/10.22146/mgi.31899>
- [23] R. H. J. Bagus Pamungkas, "Pemanfaatan citra penginderaan jarak jauh dan sistem informasi geografis untuk pemetaan erosi di DAS Serang Kabupaten Kulon Progo," *Jurnal Bumi Indonesia*, vol. 15, no. 1, pp. 165–175, 2016.
- [24] M. Conopio, A. B. Baloloy, J. Medina, and A. C. Blanco, "Spatio-Temporal mapping and analysis of mangrove extents around manila bay using landsat satellite imagery and mangrove vegetation index (MVI)," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 46, no. 4/W6-2021, pp. 103–108, 2021, doi: 10.5194/isprs-Archives-XLVI-4-W6-2021-103-2021.
- [25] M. P. Neri, A. B. Baloloy, and A. C. Blanco, "Limitation assessment and workflow refinement of the mangrove vegetation index (mvi)-based mapping methodology using sentinel-2 imagery," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch.*, vol. 46, no. 4/W6-2021, pp. 235–242, 2021, doi: 10.5194/isprs-Archives-XLVI-4-W6-2021-235-2021.
- [26] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao, "A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing," *J. For. Res.*, vol. 32, no. 1, pp. 1–6, 2021, doi: 10.1007/s11676-020-01155-1.
- [27] X. Zhao, H. Xia, L. Pan, H. Song, W. Niu, R. Wang, R. Li, X. Bian, Y. Guo, and Y. Qin, "Drought monitoring over yellow river basin from 2003–2019 using reconstructed MODIS land surface temperature in google earth engine," *Remote Sens.*, vol. 13, no. 18, 2021, doi: 10.3390/rs13183748.
- [28] J. Brema, T. S. Rahul, and J. J. Julius, *Study on Drought Monitoring Based on Spectral Indices in Noyyal River Sub-watershed Using Landsat-8 Imageries*. Springer International Publishing, 2019. doi: 10.1007/978-3-319-77276-9_42.
- [29] M. W. Browne, "Cross-validation methods. Journal of Mathematical Psychology," *J. Math. Psychol.*, vol. 44, pp. 108–132, 2000.
- [30] N. Shrestha, "Detecting multicollinearity in regression analysis," *Am. J. Appl. Math. Stat.*, vol. 8, no. 2, pp. 39–42, 2020, doi: 10.12691/ajams-8-2-1.

- [31] L. Breiman, "Support vector machines for classification and regression. Analytical Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 6, pp. 135(2), 230–240, 2001, doi: 10.1039/b918972f.
- [32] A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018, doi: 10.1613/jair.1.11192.
- [33] J. R. Wilson and K. A. Lorenz, "Short history of the logistic regression model," pp. 17–23, 2015, doi: 10.1007/978-3-319-23805-0_2.
- [34] W. P. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, pp. 16, 321–357, 2002, doi: 10.1613/jair.953.
- [35] K. Ichii, A. Kawabata, and Y. Yamaguchi, "Global correlation analysis for NDVI and climatic variables and NDVI trends: 1982–1990," *International Journal of Remote Sensing*, vol. 23, no.
- [36] Z. Feng, X. Wang, J. Yuan, Y. Zhang, and M. Yu, "Changes in air pollution, land surface temperature, and urban heat islands during the COVID-19 lockdown in three Chinese urban agglomerations," *Science of the Total Environment*, vol. 892, art. no. 164496, 2023.
- [37] P. I. Frazier, "A tutorial on bayesian optimization," *CoRR*, vol. abs/1807.02811, Jul. 2018.18, pp. 3873–3878, 2002.
- [38] M. A. Wojtas and K. Chen, "Feature importance ranking for deep learning," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, no. NeurIPS, 2020.
- [39] D. C. Donato, J. B. Kauffman, D. Murdiyarso, S. Kurnianto, M. Stidham, and M. Kanninen, "Mangroves among the most carbon-rich forests in the tropics," *Nat. Geosci.*, vol. 4, no. 5, pp. 293–297, 2011, doi: 10.1038/ngeo1123.



Estimating Economic Activity Using Geospatial Big Data in East Java, Indonesia: Relative Spatial GDP Index Approach

Rifqi Ramadhan^{1*}, I Made Satria Ambara Putra², Taufiq Agung Kurniawan³, Fitri Kartiasih⁴, Raden Muaz Munim⁵, Somethea Buoy⁶

¹BPS-Statistics Indonesia, Jakarta, Indonesia, ²BPS-Statistics Kotawaringin Barat Regency, Pangkalan Bun, Indonesia, ³BPS-Statistics Dharmasraya Regency, Pulau Punjung, Indonesia, ⁴Politeknik Statistika STIS, Jakarta, Indonesia, ⁵Department of Computer Science, Hamad Bin Khalifa University, Qatar, ⁶Department of the Statistical Information Services, National Institute of Statistics, Ministry of Planning, Cambodia

*Corresponding Author: E-mail address: rifqi.ramadhan@bps.go.id

ARTICLE INFO

Article history:

Received 17 October, 2025

Revised 21 November, 2025

Accepted 28 November, 2025

Published 31 December, 2025

Keywords:

East Java; Geospatial Big Data; GRDP; Relative Spatial GDP Index; Remote Sensing; Robust Regression

Abstract

Introduction/Main Objectives: GRDP serves as a fundamental indicator for assessing regional economic performance in Indonesia and plays a critical role in development planning. **Background Problems:** Conventional GRDP measurement in Indonesia relies on survey-based approaches, which are time-consuming, costly, and provide limited spatial detail. **Novelty:** This study introduces a Relative Spatial GDP Index (RSGI) constructed from geospatial big data such as remote sensing and point of interest (POI) to estimate GRDP more granular in East Java. This approach represents the first geospatial data driven GRDP index developed at such fine spatial resolution in Indonesia. **Research Methods:** Four weighting schemes were applied to generate RSGI variations, which were then evaluated through regression modeling against official GRDP. They are equal weight, pearson correlation, spearman correlation, and principal component analysis (PCA). **Finding/Results:** The RSGI PCA produced the best performance (RMSE = 0.73047; MAE = 0.48185; MAPE = 7.00%; $R^2 = 0.7618$). PCA weight outperformed other weight by capturing shared variance and generating objective weights that better represent spatial economic intensity. The RSGI PCA demonstrates a strong and significant correlation with GRDP at the sub-district level and provides a robust tool for fine-scale economic estimation.

1. Introduction

Economic growth is one of the international development agendas included in the eighth point of the 2030 Sustainable Development Goals (SDGs). This agenda was also adopted by the Indonesian government in the National Medium-Term Development Plan (RPJMN) for the 2020–2024 period [1]. The expected economic growth target for the 2020–2024 period will increase by an average of 5.7–6.0 percent per year [1]. National economic growth, according to data from the Central Bureau of Statistics (BPS) of the Republic of Indonesia as of February 2023, grew by 5.31% [2]. This achievement is still below the economic growth target to be achieved in the prepared RPJMN. From this description, the Gross Domestic Product (GDP) has a special place, which is the focus of the SDGs development. According to [2], GDP is an important indicator to determine economic conditions in a country over a certain period.



Economic development is a series of efforts to reduce unemployment, improve people's welfare, and minimize inequality in people's incomes. In measuring the progress or health of a country's economy, the Gross Domestic Product (GDP) is usually used as an important indicator both of current prices and based on constant prices (fixed prices). Meanwhile, to measure the economic progress or health of a province, district, or city, the Gross Regional Domestic Product (GRDP) is used as one of the indicators. GRDP is one indicator that can measure economic activity in a region. GRDP is prepared using three approaches, namely production, income, and expenditure, and is presented based on current prices and constant (real) prices. Although the components of presenting GRDP data are different for each calculation approach, conceptually, these three approaches will produce the same figure [3]. GRDP at current prices (nominal GRDP) is prepared based on prices prevailing during the calculation period and aims to look at the structure of the economy. Meanwhile, GRDP at constant (real) prices is prepared based on prices in the base year [2]. GRDP at current prices shows the ability of economic resources produced by a region. GRDP at constant prices can be used to show the rate of economic growth as a whole or for each category from year to year. GRDP per capita at current prices shows the value of GRDP per population, while GRDP per capita at constant prices is used to determine the real economic growth per capita population of a region [3].

GRDP using the expenditure approach is calculated based on all components of final demand, such as Final Household Consumption Expenditure, Final Consumption Expenditures for Non-Profit Institutions serving Households, Government Final Consumption Expenditure, Gross Domestic Fixed Capital Formation, Changes in Inventories, and Net Exports. In the final household consumption expenditure (PKRT) component, there are various data sources used to compile GRDP data, such as the National Socio-Economic Survey (SUSENAS). Within one year, SUSENAS data collection was carried out twice, namely in March and in September [4]. However, to carry out periodic monitoring in a shorter time and in more detail in terms of area or granularity, SUSENAS data cannot accommodate this, considering that SUSENAS is conducted every six months and the level of presentation is only at districts or cities.

On the other hand, the calculation of GRDP uses a production approach using various data sources collected by BPS as well as compilation results from various other data sources. For example, in calculating GRDP for the agricultural, forestry, and fisheries components, one of the data sources used comes from the agricultural census, which is carried out once every ten years. Conventional survey data collection conducted in Indonesia for the calculation of GRDP has various weaknesses, such as SUSENAS, which is conducted every six months with a method that is limited in scope, costs a lot, and takes a long time. If monitoring in the short term, periodically, and in more detail in terms of SUSENAS data areas cannot provide it, a supporting data source with a faster and more efficient method is needed. In addition, similar problems can also be found in other GRDP-forming data sources, such as the SKLNP, which is carried out every year on a quarterly basis, as well as the agricultural census, which is carried out once every 10 years.

Today, many studies have turned to alternatives data sources such as satellite imagery, which is the result of remote sensing, and other big data sources, such as Point of Interest (POI) from OpenStreetMap (OSM). The use of other big data sources has been carried out by [5], and it can be concluded that POI can effectively differentiate industrial and commercial areas and therefore has the potential to improve GRDP mapping. Previous research has shown that Visible Infrared Imaging Radiometer Suite (VIIRS) remote sensing data that captures Nighttime Light Intensity (NTL) has a positive correlation to GDP in a region [6]–[10]. In addition, Sentinel-2 Multispectral Instrument Level-2A, which captures the Normalized Difference Vegetation Index (NDVI), influences GDP and has a significant correlation [11]–[16]. Other composite indices, such as the Normalized Difference Water Index (NDWI), detect a reduction in water sources associated with GDP and population growth [17]. Then, the Normalized Difference Built-Up Index (NDBI) relates to the area of built-up land and the expansion of urban areas [18]. Sentinel-5P satellite imagery is designed to detect pollution emissions in an area such as CO, which is closely related to urban areas and economic growth [15], [19], and NO₂ shows results that have a strong correlation with the value of GDP [20]. and SO₂ emissions, along with NO₂, are indicated to have a relationship to the intensity of energy use and economic growth [21]. Therefore, pollution emissions from CO, SO₂, and NO₂ can be used to determine the GDP value of an area because they have an association with economic growth. Images derived from MODIS can detect the magnitude of the ground surface temperature in an area during the day and night where the ground surface temperature will have a significant impact due to an increase in population, industrial activity, infrastructure development, and socio-economic activities [22]. The accurate use of elevation data, namely the Digital Elevation Model (DEM) derived from ASTER satellite imagery, can provide a scientific and socio-economic point of view, and provide important things in geomorphology [23]. Another data source

originating from OpenStreetMap (OSM) can be utilized as a POI data source, which is a geospatial data source with geographic locations and descriptions based on categories and can effectively differentiate industrial and commercial areas, so that it has the potential to improve the precise GDP mapping of secondary and tertiary sectors [5]. Thus, the source of big data from satellite imagery and POI can be used as an alternative data source in data collection to estimate the GDP value of a region using relatively low costs, a short time frame, and continuous updating as needed.

Furthermore, recent literature shows a growing shift toward leveraging diverse geospatial big data sources to strengthen the ability of models to map economic indicators at the sub-national scale. The previously dominant linear regression approaches are now considered less adequate due to their limitations in capturing non-linear relationships and inter-variable spatial interactions. This limitation has encouraged researchers to adopt more flexible Machine Learning (ML) methods, including Random Forest which have demonstrated notable improvements in GDP estimation, particularly when combining multisource remote sensing data with POI data for regional economic activity mapping in China [24]. Empirical evidence shows that the Random Forest model demonstrated strong predictive performance, achieving R^2 values of 0.95 for non-agricultural GDP. Alongside these developments, Deep Learning (DL) approaches, especially Convolutional Neural Networks (CNN) applied to high-resolution daytime satellite imagery and nightlights data have shown promise in identifying subtle spatial patterns linked to economic activity in China. Using an attention-enhanced VGG-16 architecture to extract image features, this approach produced a strong predictive result in predicting localized economic activity, reaching an R^2 value of 0.71 [25]. Incorporating remote sensing data has also been shown to substantially improve the ability for GRDP mapping, particularly within dense urban environments such as NTL from NPP-VIIRS and built-up area extraction from Sentinel-2 images, where the predicting of GRDP shows the R^2 values to 0.82 in Shandong Province, China [26]. In addition, downscaling mapping techniques are increasingly adopted to redistribute official aggregate statistics into finer spatial units. Recent advances that combine NTL, POI, and multispectral features within multi-scale deep learning frameworks have further demonstrated the potential to generate GDP maps that are more stable, detailed, and representative of real economic conditions [27]. However, despite the rapid growth of geospatial and machine learning based approaches, most existing studies remain centered on improving prediction accuracy rather than constructing a spatially explicit economic index that can capture relative economic intensity at finer administrative levels, which is essential for evidence-based local policy and development planning. Therefore, this study develops a spatial GDP index that is crucial for capturing economic activity as the alternatives data to provides a more robust evidence base for development-oriented policymaking.

For the record, the alternatives data can only be used as support for SUSENAS data and cannot be used as the essence or core of data from the field obtained from SUSENAS. Provinces in Java dominate the performance and structure of the Indonesian economy spatially, with a contribution of 56.48% and a growth of 5.31% in 2022 [2]. East Java Province was chosen as the case study in this research because East Java Province is in second place after DKI Jakarta in terms of achieving the GRDP level in Java Island (BPS, 2022). To anticipate the weakness of GRDP data collection, which still uses conventional surveys, the contribution of this research is to implement the utilization of integrating remote sensing data and POI data as geospatial big data sources a combination that has not yet been explored in the Indonesian context to estimate GRDP at a smaller granular level down to the sub-district level and map it so that it can be easily interpreted. This granularity data is crucial because economic activities and development patterns often exhibit significant heterogeneity within a single district or city. Relying solely on aggregate district/city level data masks critical localized disparities in wealth, infrastructure needs, and the effectiveness of local policies, thereby limiting the government's ability to conduct targeted intervention, policy formulation, and balanced regional planning. Then, this research also offers the first empirical contribution of its kind in Indonesia, demonstrating the potential of geospatial big data to generate spatially detailed economic indicators beyond the limitations of conventional survey-based GRDP statistics.

2. Material and Methods

2.1. Study Area

This study was conducted in East Java Province, Indonesia, with 2022 as the reference year. East Java consists of 38 regencies and municipalities and 666 sub-districts, covering an area of approximately 48,000 km², making it the largest province on Java Island. During the period 2020–2022, East Java recorded the second-highest Gross Regional Domestic Product (GRDP) at constant 2010 prices, after

DKI Jakarta, which motivated its selection as the study area. Moreover, the province exhibits substantial economic diversity across manufacturing, agro-industry, agriculture, services, and trade sectors, reflecting significant geographic and socio-cultural heterogeneity. Figure 1 presents the study area along with the official GRDP distribution of East Java Province in 2022.

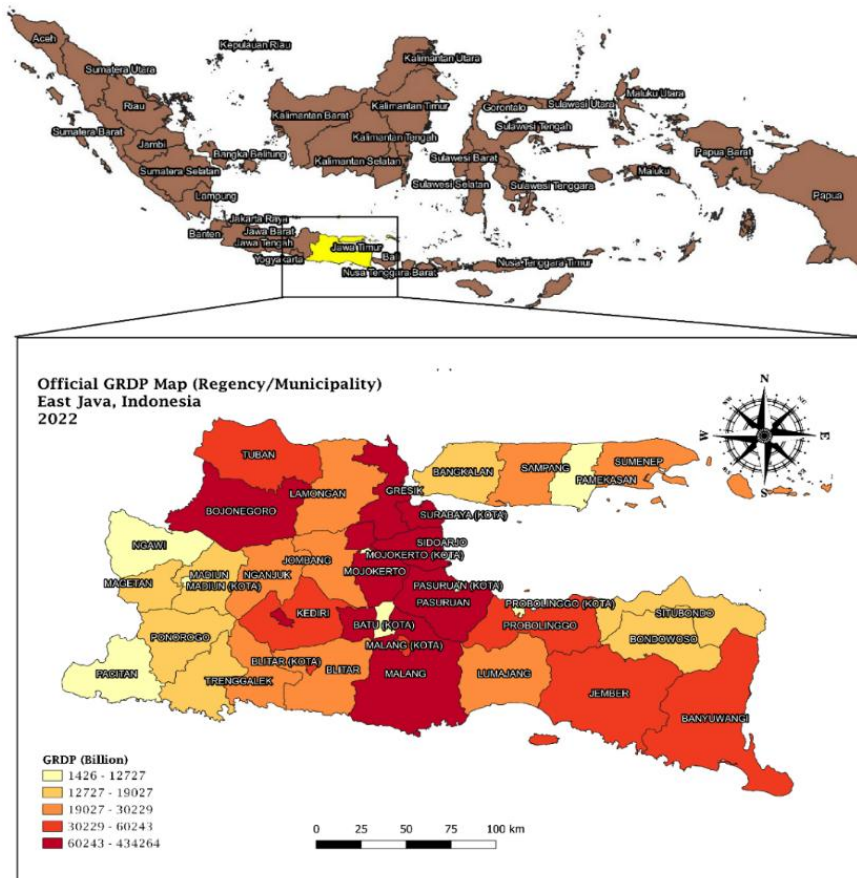


Figure 1. The official GRDP at constant prices mapping in East Java, Indonesia in 2022

2.2. Big Data

Big data has become a crucial concept in both industry and academia, encompassing various meanings. Big data as an information asset characterized by volume, velocity, and variety (the 3Vs), requiring specific technologies and analytical methods to derive value [29]. Its application is expanding rapidly across sectors such as industry, environment, and economic development. A specialized form, geospatial big data, refers to spatially referenced data, including land characteristics like land use [30]. This data generally appears in two formats: raster data, comprising images from sources like drones or satellites (e.g., Google Earth), and vector data, consisting of points, lines, and polygons obtained through platforms such as the OpenStreetMap API [30].

2.3. Satellite Imagery

Satellite imagery refers to images captured by sensors mounted on satellites orbiting more than 400 km above Earth's surface [31]. These images result from remote sensing, a process in which sensors record energy emitted or reflected by objects on Earth [31]. Remote sensing involves collecting data from aircraft or satellites to measure and observe terrestrial features [32] and includes all techniques for capturing and processing electromagnetic recordings of the Earth's surface [33]. Its ability to provide consistent, repetitive observations enables the monitoring of both short- and long-term environmental changes caused by natural or human activities [32].

Remote sensing serves as a data source, a science, and a tool simultaneously. As a data source, it records physical properties of distant objects through emitted or reflected energy. As a science, it applies systematic processes such as measurement, data analysis, and interpretation. As a tool, it supports applications ranging from resource inventory to ecological assessment [34].

Remote sensing data are characterized by four key resolutions:

1. Spatial Resolution, the smallest detectable object in imagery, represented by pixel size [31]. Higher resolution (smaller pixels) yields more detailed information [35]. Examples include very high-resolution imagery (Piades, 0.5 m), high (Iconos, 4 m), medium (Landsat, 15–30 m), and low (MODIS, 250–1000 m) [31].
2. Temporal Resolution, the frequency with which a satellite revisits the same location [31]. A shorter revisit time means higher temporal resolution; this depends on orbit type, sensor characteristics, and image swath width [35], [36].
3. Spectral Resolution, the sensor's ability to distinguish between different wavelengths of the electromagnetic spectrum. Higher spectral resolution means more and narrower spectral bands [31].
4. Radiometric Resolution, the number of bits used to represent energy levels per pixel, determining brightness detail. Greater bit depth (e.g., 8-bit, 10-bit, 16-bit) indicates higher radiometric resolution [31], [37].

Satellite imagery is available from various sources such as VIIRS, Sentinel-2, MODIS, Sentinel-5P, and ASTER. Subsequent analyses often utilize imagery from these satellites to explore the relationship between environmental observations and socio-economic indicators, including GRDP formation.

2.4. Visible Infrared Imaging Radiometer Suite (VIIRS)

The Suomi National Polar Partnership (SNPP) is a satellite operated jointly by NASA and NOAA, equipped with the Visible Infrared Imaging Radiometer Suite (VIIRS) as its primary imaging sensor. VIIRS is distinct among meteorological satellite sensors for its ability to capture short-wave infrared, near-infrared, and nighttime visible light emissions [38]. Research utilizing nighttime light data expanded following the release of annual stable light products from the DMPS-OLS by NOAA's Earth Observation Group (EOG). Since the launch of VIIRS in 2011, EOG has provided improved nighttime light datasets featuring monthly updates, higher spatial resolution, and enhanced data quality [39]. These advancements make VIIRS data highly valuable for studies analyzing social and economic activities, particularly those related to economic growth [39][40].

2.5. Sentinel-2 Multispectral Instrument Level-2A

Sentinel-2 is a European satellite equipped with the Multispectral Instrument (MSI) Level-2A, launched in 2015, designed for high-resolution optical imaging of the Earth's surface. Unlike radar satellites that utilize Synthetic Aperture Radar Ground Range Detected (SAR-GRD) technology to capture images through cloud cover [41], Sentinel-2 focuses on multispectral observation. Compared to other medium-resolution satellites, Sentinel-2 offers superior spatial resolution, reaching 10 meters in the B4 (Red), B3 (Green), B2 (Blue), and B8 (Near-Infrared/NIR) bands. The satellite's temporal resolution, its revisit frequency is 10 days for a single satellite and 5 days when two satellites operate together.

Sentinel-2 is a multispectral imaging satellite mission that supports the Copernicus Land Monitoring program, focusing on observing vegetation, land cover, water bodies, inland waterways, and coastal areas. Its combination of sensing bands enables the development of various composite indices, including the Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Normalized Difference Built-Up Index (NDBI). Sentinel-2 offers high spatial resolution of up to 10 meters in the B4 (Red), B3 (Green), B2 (Blue), and B8 (Near Infrared/NIR) bands, with a temporal resolution of 10 days, or 5 days when two satellites are operational.

The NDVI quantifies vegetation greenness by analyzing reflectance differences between the NIR and Red bands, as formulated:

$$NDVI = \frac{NIR_{band\ 8} - RED_{band\ 4}}{NIR_{band\ 8} + RED_{band\ 4}} \quad (1)$$

NDVI reflects vegetation health and ecological changes [42] and has shown a positive correlation with economic growth, particularly with GDP and population [16].

The NDWI distinguishes between water and land surfaces, as it is sensitive to liquid water content in vegetation [43]. Its values range from below 0 (non-water areas) to above 0 (water areas) [44], with the following formula:

$$NDWI = \frac{Green_{band\ 3} - NIR_{band\ 8}}{Green_{band\ 3} + NIR_{band\ 8}} \quad (2)$$

The NDBI identifies built-up or urban areas using the difference between Short-Wave Infrared (SWIR) and NIR reflectance [45]. According to [46], it is formulated as:

$$NDBI = \frac{SWIR_{band\ 11} - NIR_{band\ 8}}{SWIR_{band\ 11} + NIR_{band\ 8}} \quad (3)$$

Higher NDBI values indicate a greater likelihood of urban development [47].

2.6. MODIS

The Moderate Resolution Imaging Spectroradiometer (MODIS) is a sensor onboard the Terra satellite, operating under NASA's Earth Observing System [48]. MODIS–Terra captures Land Surface Temperature (LST) data both day and night, providing valuable insights into Earth's thermal conditions. The satellite orbits polarly at an altitude of 705 km, crossing the equator around 10:30 a.m. local time [48].

LST data are particularly useful for analyzing the Urban Heat Island (UHI) phenomenon [22]. Increases in population density, industrial activity, and infrastructure development significantly influence LST variations. Consequently, areas exhibiting higher LST values typically correspond to regions with intense socio-economic activities and denser populations [22].

2.7. Sentinel-5P

The Sentinel-5 Precursor (Sentinel-5P) is an atmospheric monitoring satellite launched by the European Space Agency (ESA) on October 13, 2017, designed to observe and measure various air pollutants [49]. Sentinel-5P captures data on gases such as carbon monoxide (CO), nitrogen dioxide (NO₂), and sulfur dioxide (SO₂), which are key indicators of air quality and human environmental impact.

Carbon Monoxide (CO)

CO is a crucial atmospheric gas for assessing tropospheric chemistry and serves as a major pollutant in urban environments (Copernicus Sentinel-5P, processed by ESA, 2021). It is primarily generated from biomass burning, fossil fuel combustion, and atmospheric oxidation of methane and hydrocarbons, with significant sources including motor vehicles, forest fires, power plants, and incinerators [50].

Nitrogen Dioxide (NO₂)

NO₂ occurs in both the troposphere and stratosphere, produced by anthropogenic activities (e.g., fossil fuel and biomass combustion) and natural processes such as soil microbial reactions, lightning, and wildfires (Copernicus Sentinel-5P, processed by ESA, 2021). It is one of the major air pollutants in industrial and urban regions worldwide [51], [52]. Areas with high economic and industrial activity generally exhibit elevated NO₂ concentrations.

Sulfur Dioxide (SO₂)

SO₂ emissions are closely linked to population density and energy consumption [53]. High population areas tend to produce more SO₂ pollutants. Key contributors include electricity generation, manufacturing, and mining activities. As economic activity increases, greater energy demand for both residential and industrial purposes results in higher SO₂ emissions into the atmosphere [54].

2.8. The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER)

The Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) is a satellite jointly operated by the Japanese Ministry of Economy, Trade, and Industry (METI) and the U.S. National Aeronautics and Space Administration (NASA). Launched in December 1999, ASTER carries 14 imaging sensors capable of capturing thermal infrared, shortwave infrared, and visible-near infrared (VNIR) spectral data [55].

ASTER is notable for its ability to generate Digital Elevation Models (DEMs) three-dimensional representations of the Earth's surface that exclude above-ground objects such as vegetation and buildings [56]. DEMs are widely used in geographic information systems (GIS) and serve as the foundation for creating digital maps. Variations in elevation influence economic activity patterns: higher elevations are typically dominated by agricultural activities, while lower-lying areas tend to support more industrial development.

2.9. Point of Interest (POI)

Point of Interest (POI) refers to specific locations that attract frequent human activity, such as restaurants, convenience stores, transportation hubs, parks, cafés, and tourist attractions, places that provide value or utility to visitors [57]. With the widespread use of mobile devices, service providers and technology companies increasingly analyze user mobility data with high geospatial-temporal resolution to identify and map nearby POIs [57]. Common POI data sources include OpenStreetMap (OSM), Google Maps, HERE Maps, and OneMap. In this study, POI data are obtained from OpenStreetMap (OSM).

2.10. Robust Linear Regression

The modeling used is robust regression modeling that can address the problem of heteroskedasty, and the image (oulier) on the data. Robust regression has many kinds of methods in terms of dealing with the problems of various classical regression assumptions such as M-Regression, S-Regresion, and MM- Regression. In this study, the MM-Regression method will be used because this method combines between M- Regression and S-Regresion so that it is resistant and has a higher efficiency [58]. Robust regression will not make the modeling error following the normal distribution because it has a different estimate method based on the weighing of the data so that the outlier will give a small weight to the parameter, but produce a level of accuracy generated better than the Ordinary Least Square method. (OLS) [59]. The MM-Estimation procedure is a procedure by combining S-Estimation and M-Estimation with the initial step of searching for the Estimator S so that a breakdown point is obtained to minimize the residual in the estimator M which is used to set parameters on its regression so that its efficiency is high and usually uses Tukey Bisquare. Based on [60], the equation form of the estimate method Method of Moment (MM) is as follows.

$$\sum_{i=1}^n p_i'(u_i)X_{ij} = 0 \quad (4)$$

$$\sum_{i=1}^n p_i'\left(\frac{Y_i - \sum_{j=0}^k X_{ij}\hat{\beta}_j}{S_{MM}}\right)X_{ij} = 0 \quad (5)$$

where s_{MM} is the standard deviation derived from S-estimation and ρ is Tukey's biweight function formulated as follows:

$$\rho(u_i) = \frac{u_i^2}{2} - \frac{u_i^4}{2c^2} + \frac{u_i^6}{6c^2}, -c \leq u_i \leq c \text{ dan} \quad (6)$$

$$\rho(u_i) = \frac{c^2}{6}, u_i < -c \text{ atau } u_i > c$$

calculating the weighted value derived from Tukey Bisquare in the following equation.

$$w_i = [1 - (\frac{u_i}{c})^2]^2, u_i \leq c \quad (7)$$

$$w_i = 0, u_i > c$$

where $u_i = \frac{e_i}{s_{MM}}$ and $c = 4.865$

The MM estimate method uses WLS (Weighted Least Square) to estimate the regression parameters in the form of the following equation.

$$\hat{\beta}_i^{(n)} = (X^T W^{(n)} X)^{-1} (X^T W^{(n)} Y) \quad (8)$$

where n is the value of the iteration performed and this iterative process is performed so that the weighing value becomes convergent or $\hat{\beta}_i^{(n+1)} - \hat{\beta}_i^{(n)} \approx 0$.

2.11. Data Transformation

The transformation used in this study is the Yeo-Johnson power transformation, which is a family of Box-Cox transformations and can transform positive and negative data. The Yeo-Johnson power transformation is well used in handling variables that do not have the same units in all ranges by making the distribution of these variables like a Gaussian distribution or more normally distributed [61]. Data transformation in this study is defined as follows:

$$y_\lambda(x) = \begin{cases} \frac{(1+x)^\lambda - 1}{\lambda}, \lambda \neq 0 \text{ dan } x \geq 0 \\ \log(1+x), \lambda = 0 \text{ dan } x \geq 0 \\ -\frac{(1-x)^{2-\lambda}}{2-\lambda}, \lambda \neq 2 \text{ dan } x < 0 \\ -\log(1-x), \lambda = 2 \text{ dan } x < 0 \end{cases} \quad (9)$$

where x is the value of the variable or data input and λ is the parameter value estimated using the Maximum Likelihood technique with the assumption that these variables follow a normal distribution. A linear relationship will be generated by families with parameter $\lambda = 1$, then the transformation will make the right tail of the distribution thicker or denser at 1 while widening the left tail of the distribution to make the distribution right-skewed (skewed) towards a symmetrical distribution. In $\lambda > 1$, the distribution that is left skewed becomes a more symmetrical distribution as well.

2.12. Corelation Analysis and Selection of Important Variabels

The weighted sum model with the linear model is a weighting calculation to build the RSGI index used in this study. Correlation analysis and selection of important variables need to be carried out to ensure that the variables used can linearly represent the GRDP mapping in East Java Province. The numerical relationship of integrated multi-source satellite imagery and POI data with regional GRDP data at the sub-district level is carried out using correlation analysis. Correlation analysis was carried out by using the correlation coefficient of Pearson and Spearman Rank with a p-value that showed significant results at a significance level of 5% ($\alpha = 0.05$). The following equation is a formula for obtaining the value of the Pearson correlation coefficient.

$$r_{xy} = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sqrt{(n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2) (n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2)}} \quad (10)$$

The value of r_{xy} is the correlation value between x_i and y_i , and n is the number of observations. Meanwhile, the Spearman correlation value is calculated by changing the observed value to the ranking value as in the following equation.

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n^3 - n} \quad (11)$$

The value of the correlation coefficient is between 0 and 1, with the direction of the relationship indicated by a positive or negative sign. Table 1 below shows the guidelines for interpreting the results of the correlation coefficient used in this study [62].

Table 1. Correlation coffecient interpretation

Correlation Coefficient	Interpretation
$0.00 \leq r \leq 0.199$	Very Weak
$0.20 \leq r \leq 0.399$	Weak
$0.40 \leq r \leq 0.599$	Moderate
$0.60 \leq r \leq 0.799$	Strong
$0.80 \leq r \leq 1.000$	Very Strong

Furthermore, a correlation significance test was performed to determine whether the correlation coefficient obtained was statistically significant with a significance level of 5% ($\alpha = 0.05$). The null hypothesis used in this study is that there is no correlation or relationship between the two variables x_i and y_i , compared with the alternative hypothesis, which assumes that there is a relationship between the variables x_i and y_i . Based on the results of the correlation analysis, a statistically significant correlation with the GRDP variable in East Java will be selected at the sub-district level and will then be used for the construction of the RSGI. Thus, RSGI is made based on variables originating from satellite imagery big data and other geospatial data, namely POI, which linearly has a significant relationship or correlation to estimate GRDP in East Java Province down to the sub-district granular level.

2.13. Weighted Sum Model

The relative spatial GDP index (RSGI) is calculated by integrating satellite imagery and geospatial variables that represent the economic activity in East Java Province by calculating the weighted sum model. Several previous studies have used the index building method in their research, and some of them have used geospatial data [40], [63]–[66]. The variables used in the index construction include NTL Total, POI density, POI distance, NDVI, NDBI, NDWI, LST Daytime, LST Nighttime DEM, and atmospheric pollutants (NO_2 , SO_2 , CO). The following formula is used to build RSGI with a weighted sum model:

$$\begin{aligned} RSGI = & w_1 \cdot NTL + w_2 \cdot POI_{density} + w_3 \cdot POI_{distance} + w_4 \cdot NDVI + w_5 \cdot NDWI \\ & + w_6 \cdot NDBI + w_7 \cdot LST_{daytime} + w_8 \cdot LST_{nighttime} + w_9 \cdot DEM + w_{10} \cdot CO \\ & + w_{11} \cdot NO_2 + w_{12} \cdot SO_2 \end{aligned} \quad (12)$$

where p is the number of variables used in building the index, w is the weight used, and x_i is the value of each observation.

In this study, the RSGI is not a single model instead, it is constructed in four distinct variations based on different weighting approach. They are equal weight sum (RSGI EWS), pearson correlation (RSGI Pearson Weight), spearman correlation (RSGI Spearman Weight), and principal component analysis loadings (RSGI PCA). First, an equal-weighted correlation method was applied, where all variables were assigned identical weights, influenced only by the direction of their correlation with GRDP. Second, weights were determined based on the Pearson correlation coefficient, using only variables that showed significant correlations with GRDP in East Java. Third, the Spearman Rank correlation coefficient was used, assigning higher weights to variables with stronger correlations, representing GRDP more accurately. Fourth, weights were derived from Principal Component Analysis (PCA), utilizing components that contributed cumulatively significant variance above 70%. Several previous studies have also employed PCA for index construction in socio-economic research, who developed the Regional Digital Development Index (RDDI) for Indonesia [67]; who created a Spatial

Deprivation Index [68]; and who mapped the Relative Spatial Poverty Index (RSPI) in East Java, showing a strong correlation with poverty levels [65].

2.14. Evaluation

To find out how far the RSGI has come in estimating the GRDP of an area in East Java, an evaluation measure is used. In this study, the evaluation used is a numerical and descriptive evaluation. A descriptive evaluation measure is carried out by visually identifying the RSGI results obtained. Then, numerical evaluation is used to numerically measure the accuracy of the results between the results obtained and the ground truth results, where in this case the Pearson, Spearman Rank, RMSE, MAE, MAPE, and R^2 correlation measures are used in the robust linear regression model at the sub-district level. RMSE, MAE, MAPE, and R^2 values are calculated using the following formula:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (13)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (14)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100\% \quad (15)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (16)$$

2.15. Analysis Method

This study utilizes multi-source satellite imagery and Point of Interest (POI) data from OpenStreetMap (OSM), collected between January 1 and December 31, 2022. The satellite data were collected and processed using the Google Earth Engine (GEE) platform. The variables used are summarized in Table 1. The target variable, Gross Regional Domestic Product (GRDP) at constant prices (ADHK), was chosen because it reflects the real economic growth rate. Since official GRDP data are only available at the regency or city level, GRDP for the 666 sub-districts was estimated using the Broad Area Ratio Estimation (BARE) method, a Small Area Estimation (SAE) technique that employs population data as a proportional auxiliary variable.

The remote sensing data used in this study consist of multi-source satellite imagery and geospatial big data. The satellite imagery includes Nighttime Light (NTL) intensity from NOAA-VIIRS; Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Normalized Difference Built-Up Index (NDBI) from Sentinel Multispectral Level 2A; Daytime and Nighttime Land Surface Temperature (LST) from MODIS; Nitrogen Dioxide (NO₂), Carbon Monoxide (CO), and Sulfure Dioxide (SO₂) from Sentinel-5P; as well as Digital Elevation Model (DEM) data from ASTER. These variables serve as GRDP estimation indicators based on relevant literature, previous studies, and image characteristics identified for East Java Province. The data were collected and processed through Google Earth Engine (GEE), a cloud-based platform for managing and analyzing Earth observation data, offering high-resolution, freely accessible imagery suitable for advanced processing such as classification [69]. Meanwhile, POI data represent another form of geospatial big data collected from OSM to complement the satellite-based analysis.

POI data describe the accessibility and activity level of an area. In this study, more than 17,000 POI points were identified across East Java Province, representing public facilities categorized into sectors such as education, health, economy, and tourism. The highest satellite spatial resolution is provided by Sentinel Multispectral Level 2A (10 m), while the lowest is Sentinel-5P (1113.2 m). To address varying resolutions, GRDP mapping was constructed through the Relative Spatial GDP Index (RSGI) with a resolution of 1 km × 1 km, in which each grid cell aggregates extracted features from all variables at the sub-district level. All geospatial data sources used in this study, including POI data from OSM and multisource remote sensing data underwent a validation and quality-checking procedure prior to modeling. The POI dataset was refined through classification cleaning, removal of duplicated or irrelevant entries, and spatial consistency checks to ensure its representativeness for economic activity. In addition, the completeness and distribution of POI categories were examined to confirm that

economically relevant facilities such as commercial, industrial, and service-related points were adequately captured within each sub-district. Then, the satellite-derived variables were obtained from products that include sensor calibration and atmospheric correction. Further validation was implicitly ensured through the standardized pre-processing workflow applied to each satellite dataset, which included cloud masking, median compositing, band selection, and pollution-product selection as appropriate for each sensor. These procedures reduce atmospheric noise, remove anomalous observations, and enhance the reliability of the extracted indicators for representing underlying economic conditions. These steps ensured that each data source met the standards of methodological appropriateness and was fit for use in constructing the RSGI. Additionally, official population data for 2022 from Statistics Indonesia (BPS) at the regency, municipality, and sub-district levels were used to develop a GRDP estimation model using the SAE approach with the BARE method, an approach widely applied in previous studies [70]. The GRDP at constant prices (ADHK) was employed as it reflects the real rate of economic growth across sectors and over time [28].

This research proposes a method to overcome the limitations of conventional surveys for GRDP calculation, which are often costly, slow, and lack granularity. The proposed solution utilizes multi-source satellite imagery and OpenStreetMap (OSM) Point of Interest (POI) data to estimate GRDP at the sub-district and district/city levels using a robust linear regression model. The overall research framework is depicted in Figure 2.

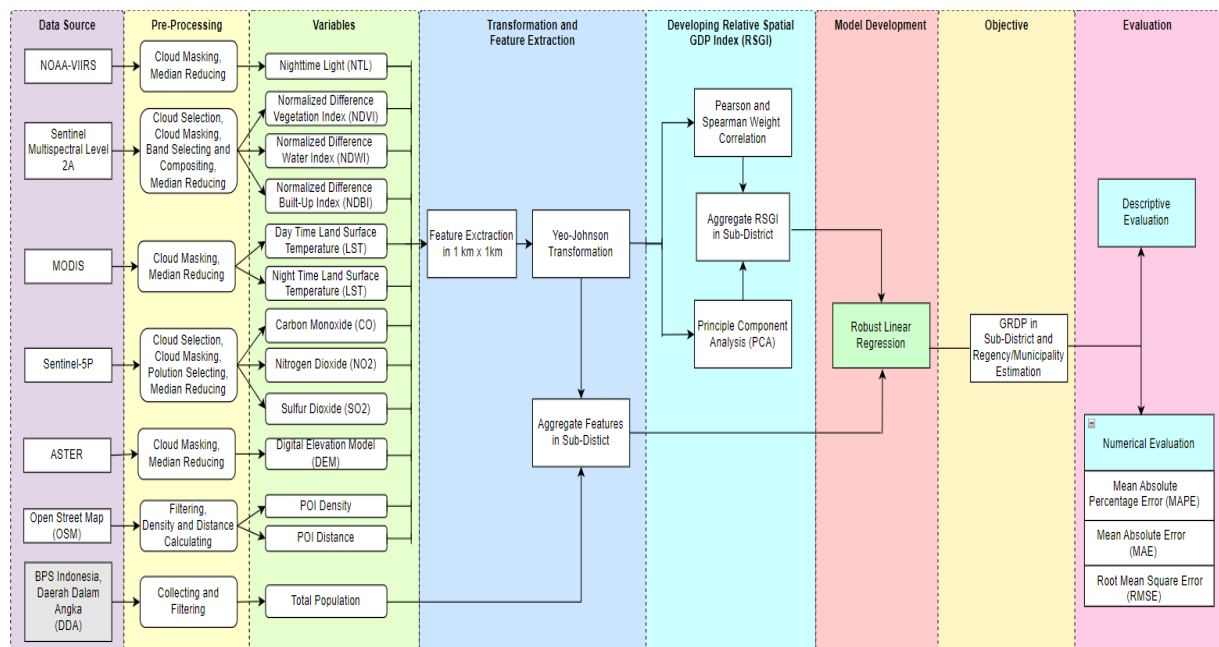


Figure 2. The research framework

The methodology is centered on the creation of a Relative Spatial GDP Index (RSGI). This process begins with preprocessing and feature extraction from a diverse set of variables: NTL, NDVI, NDWI, NDBI, NO_2 , CO, SO_2 , daytime and nighttime LST, POI Density, POI Distance, and DEM. These features are aggregated into a 1 km x 1 km grid, transformed using the Yeo-Johnson method, and then combined to construct the RSGI.

This index serves as the primary predictor in a robust linear regression model to estimate GRDP. The performance and accuracy of the model are evaluated using several metrics: the correlation coefficient, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the coefficient of determination (R^2). Finally, the GRDP estimation results are presented as visualized maps to facilitate interpretation and support stakeholder decision-making.

3. Results and Discussion

3.1. Correlation Analysis and Principal Component Analysis (PCA)

In this study, correlation analysis was applied to determine the direction and strength of the relationship between geospatial variables and GRDP in East Java. Due to the limited availability of official GRDP data at the $1 \text{ km} \times 1 \text{ km}$ grid level, the analysis was conducted at the district or city level (38 observations) and at the sub-district level (666 observations) using GRDP estimates derived from the BARE method. To obtain more accurate insights, both Pearson and Spearman Rank correlation tests were employed, with p-values used to assess statistical significance at $\alpha = 5\%$ and correlation coefficients indicating the direction and closeness of relationships. Following the interpretation approach of [62], the Pearson test results revealed that all variables, except Nighttime LST, were significantly correlated ($p\text{-value} < 0.05$) with GRDP at the sub-district level, while the Spearman Rank test showed significant correlations ($p\text{-value} < 0.05$) for all variables. The results of the correlation analysis are explained in Table 2.

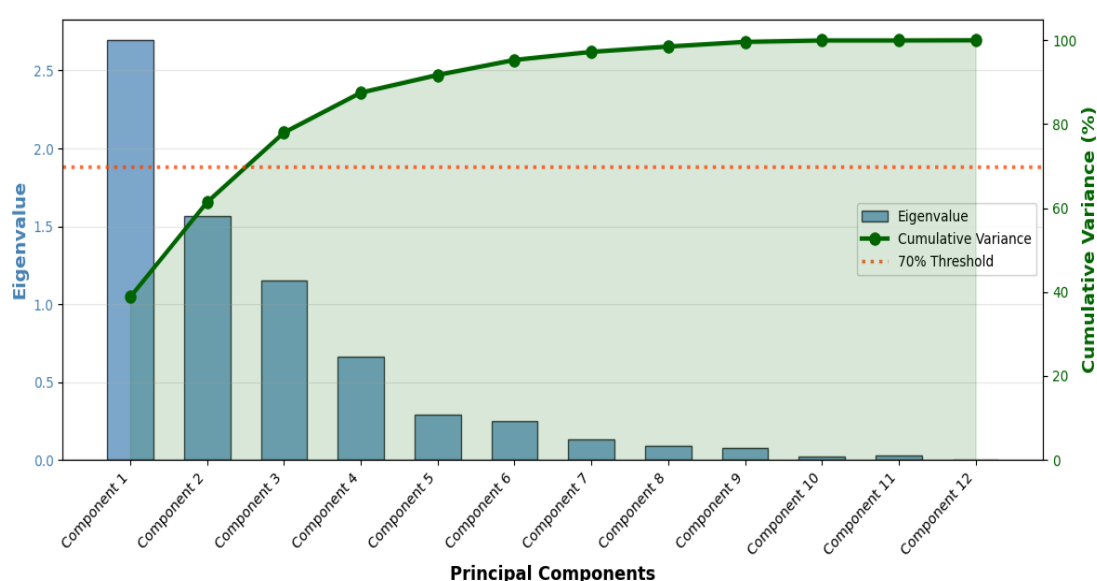
Table 2. Results of correlation analysis of geospatial data variabls on GRDP in East Java

Variable	Pearson Correlation Test		Spearman Rank Test		Direction	Closeness	Statistically Significant
	Correlation Coefficient	p-value	Correlation Coefficient	p-value			
NTL	0.6031	2.2×10^{-16}	0.5677	2.2×10^{-16}	Positive	Strong & Moderate	Significant
NDVI	-0.3645	2.2×10^{-16}	-0.3019	1.8×10^{-15}	Negative	Weak	Significant
NDWI	0.3506	2.2×10^{-16}	0.3014	2.1×10^{-15}	Positive	Weak	Significant
NDBI	0.3840	2.2×10^{-16}	0.326	2.2×10^{-16}	Positive	Weak	Significant
Day LST	0.2395	4.1×10^{-10}	0.2699	1.5×10^{-12}	Positive	Weak	Significant
Night LST	0.0627	1.1×10^{-1}	0.1116	4.0×10^{-3}	Positive	Very Weak	Not Significant
CO	0.2099	4.8×10^{-8}	0.2553	2.5×10^{-11}	Positive	Weak	Significant
NO ₂	0.3753	2.2×10^{-16}	0.3389	2.2×10^{-16}	Positive	Weak	Significant
SO ₂	0.2845	7.8×10^{-14}	0.2947	9.04×10^{-15}	Positive	Weak	Significant
POI Density	0.4974	2.2×10^{-16}	0.4288	2.2×10^{-16}	Positive	Moderate	Significant
POI Distance	-0.4017	2.2×10^{-16}	-0.3853	2.2×10^{-16}	Negative	Moderate	Significant
DEM	0.2845	7.8×10^{-14}	0.2947	9.04×10^{-15}	Positive	Weak	Significant

The PCA method was applied to reduce data dimensions and extract principal components that capture the highest variance proportions from the original data. Based on [71], components contributing a cumulative variance above 70% were selected. As shown in Table 3 and Figure 3, three components were chosen, accounting for 78.003% of the total variance. Component 1 explains 38.821% of the variance, dominated by high loadings from NTL, NDBI, NO₂, and CO variables. Component 2 contributes 22.6% (cumulative 61.4252%), with SO₂ and DEM showing the highest loadings. Component 3 adds 16.58%, bringing the cumulative variance to 78.003%, with dominant loadings from NTL, NDVI, and NDWI. These selected components serve as weighted coefficients in constructing the RSGI using the weighted sum method based on the obtained loading values.

Table 3. Loadings value (eigen vector) of each main component of PCA

Variable	Component											
	1	2	3	4	5	6	7	8	9	10	11	12
NTL	0.308	0.100	0.708	-0.164	0.023	0.026	-0.114	0.800	0.454	-0.016	-0.027	0.0000
NDVI	-0.171	-0.117	-0.555	-0.142	-0.222	0.065	0.143	0.155	-0.009	-0.015	0.727	1.8×10^{-15}
NDWI	0.137	0.130	0.580	0.172	0.257	0.039	-0.238	-0.094	-0.014	-0.050	0.678	1.6×10^{-15}
NDBI	0.434	-0.049	0.084	-0.047	-0.045	0.502	0.677	-0.146	0.157	0.178	0.084	-3.4×10^{-16}
NO2	0.486	0.004	-0.136	0.553	-0.103	0.624	0.184	-0.062	0.016	-0.006	0.008	-1.2×10^{-16}
CO	0.395	-0.244	-0.286	0.308	-0.150	0.506	-0.508	0.053	-0.245	-0.073	-0.020	4.3×10^{-16}
SO2	-0.017	0.621	-0.242	0.128	0.133	0.134	-0.009	-0.022	-0.002	0.041	-0.002	-7.1×10^{-16}
Day LST	0.291	-0.061	-0.229	-0.339	0.481	0.069	0.021	-0.209	0.050	-0.677	-0.015	3.1×10^{-15}
Night LST	0.197	-0.163	-0.225	-0.237	0.503	0.146	-0.205	-0.128	0.020	0.701	0.025	-3.6×10^{-15}
POI Density	0.288	0.208	0.138	-0.393	-0.143	0.142	0.097	0.222	-0.772	0.044	-0.002	2.3×10^{-16}
POI Distance	-0.260	-0.224	-0.035	0.400	0.553	0.109	0.324	0.431	-0.327	-0.045	-0.029	5.4×10^{-16}
DEM	-0.017	0.621	-0.232	0.128	0.133	0.134	-0.009	-0.022	-0.002	0.041	-0.002	-1.2×10^{-16}

**Figure 3.** Principal component selection and cumulative variance contribution

3.2. Calculation of the RSGI (Relative Spatial GDP Index)

To provide an initial understanding of the linear relationships among the key geospatial indicators and the dependent variable, bivariate analyses were conducted, focusing on the three parameters exhibiting the strongest association with the logarithm of GRDP (Ln GRPD), they are NTL, POI Density, and POI Distance. The visual representation of these relationships is presented through scatter plots in Figure 4. The first plot illustrates the moderately strong positive relationship between NTL Total and Ln GRDP, indicate that areas with higher aggregate night-time light intensity consistently exhibit higher economic output, affirming the known strong link between energy consumption, urbanization, and economic activity. The second plot examines the association between POI Density and Ln GRDP, also demonstrates a positive relationship and this finding reflects the principle that regions with a denser concentration of economic activities, services, and infrastructure (as captured by the POI count per unit area) tend to be associated with significantly higher GRDP values. Conversely, the third plot focuses on POI Distance, which shows a clear negative association with Ln GRDP, and this pattern confirms the expectation that greater average distance from economic centers is typically linked to a decrease in overall economic activity. Overall, these initial visual patterns and correlation values, strongly reinforce the relevance and potential of these geospatial indicators in effectively capturing and explaining the observed variations in regional economic intensity across the study area.

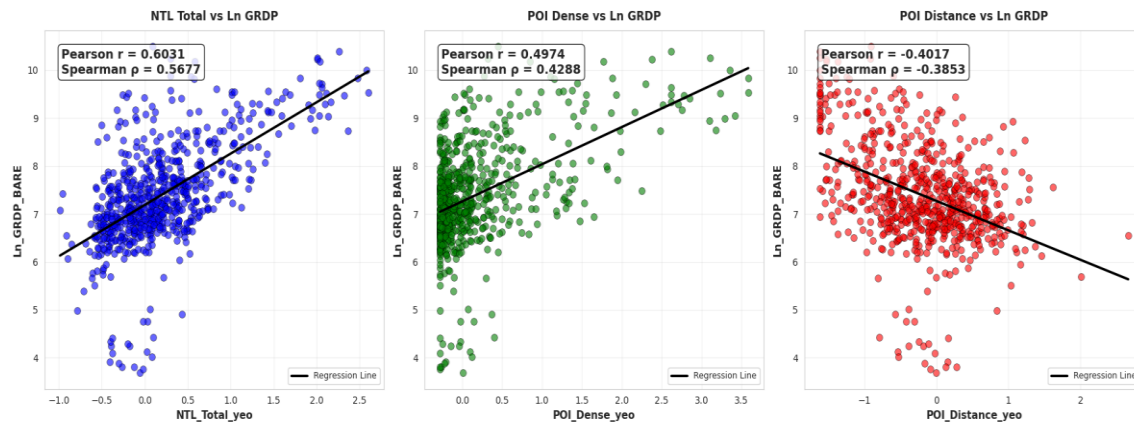


Figure 4. Correlation analysis between socioeconomic indicators (NTL, POI density, and POI distance) and regional economic output in East Java districts

In building a linear regression model, selection is carried out using the Pearson correlation test with a moderate to strong relationship. The results of the calculations are shown in Table 4, sorted from the highest value. The Relative Spatial GDP Index (RSGI) was calculated using the weighted sum model method, incorporating geospatial variables that showed significant correlations with GRDP in East Java Province, namely NTL, NDVI, NDWI, NDBI, NO₂, CO, SO₂, Daytime LST, Nighttime LST, POI Density, POI Distance, and DEM. Two weighting strategies were applied: an equal weighted sum, which assigns identical weights to all variables based solely on the direction of their significant correlations, and four specific weighting approaches, equal weights based on correlation direction (W_1), Pearson correlation coefficients (W_2), Spearman Rank correlation coefficients (W_3), and PCA-based weights derived from the first three components (W_4), which together account for 78% of cumulative variance. The results of these weight calculations are presented in the following table.

Table 4. RSGI weight calculation results

Variable	Equal Weight	Pearson Weight	Spearman Weight	PCA Weight
NTL	+1	0.6031	0.5677	1.116
NDVI	-1	-0.3645	-0.3019	-0.843
NDWI	+1	0.3506	0.3014	0.847
NDBI	+1	0.3840	0.3260	0.469
NO ₂	+1	0.3753	0.3389	0.354
CO	+1	0.2099	0.2553	-0.135
SO ₂	+1	0.2845	0.2947	0.309
Daytime LST	+1	0.2395	0.2699	0.001
Nighttime LST	+1	Not Significant	0.1116	-0.191
POI Density	+1	0.4974	0.4288	0.634
POI Distance	-1	-0.4017	-0.3853	-0.519
DEM	+1	0.2845	0.2947	0.372

Using four different weighting approaches, the weighting was applied to each 1 km × 1 km grid cell to generate a relative spatial GDP (RSGI) map, as shown in Figure 5. The mapping process employed min–max scaling to normalize RSGI values within a 0–1 range for easier interpretation. The RSGI values were classified into six categories using the equal interval method, ensuring each class spans an identical range, and the color gradient signifies an increasing index value. Very Low RSGI areas, ranging from 0 to 0.17, are depicted in black, representing the lowest combined socio-economic, demographic, and physical-geographic activity. This is followed by Low RSGI (0.17–0.33) in dark purple, Moderate Low RSGI (0.33–0.5) in purple, and Moderate High RSGI (0.5–0.67) in dark orange. The higher end of the spectrum includes High RSGI (0.67–0.83), shown in orange, and finally, Very High RSGI (0.83–1), depicted in bright cream, which indicates the highest concentration of the combined index variables.

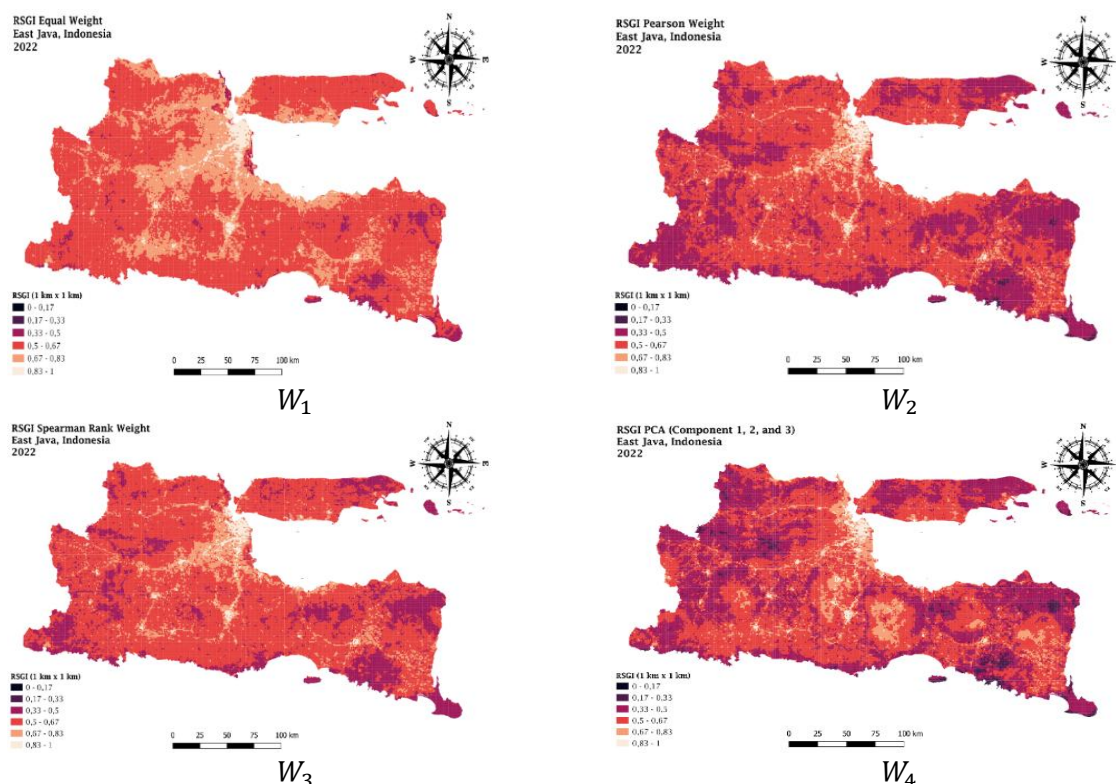


Figure 5. RSGI with equal weight sum method visualization

The resulting of four different weights for RSGI maps consistently indicate a stark spatial dichotomy. Higher index values, predominantly characterized by orange and bright cream colors are primarily concentrated in major urban centers and their expanding peri-urban hinterlands. The most significant concentration of Very High RSGI forms a dense, continuous cluster centered around the Gerbangkertosusila metropolitan region, encompassing the core area of Kota Surabaya and its neighboring highly industrialized municipalities and Kota Malang. This expansive, high-value stretch directly reflects the region's status as the uncontested economic and administrative engine of East Java, fueled by high population density, intensive industrial activities, and commercial infrastructure development. Furthermore, other key regional urban clusters, including Jember, Blitar, and Kota Madiun, also exhibit prominent, localized pockets of high RSGI, radiating outward from their municipal boundaries and serving as vital secondary nodes for regional socio-economic growth.

In sharp contrast, lower RSGI values, represented by dark purple and black colors, dominate the less urbanized and peripheral areas, indicating significant disparities. The lowest RSGI values are widespread across the central and eastern parts of Madura Island and are strongly correlated with structural factors such as lower levels of industrial investment, limited access to advanced education and healthcare, and a predominant reliance on traditional economic sectors like agriculture and fisheries. The low values in the south are primarily linked to the difficult topography of the Southern Mountains, which limits agricultural productivity, hinders transportation network development, and restricts high-density economic and social infrastructure. Similarly, low RSGI values in several eastern mountainous areas are associated with forest dominance, conservation land use, and lower population carrying capacity. This spatial dichotomy underscores the geographically driven relationship between urbanization, infrastructure development, and socio-economic conditions in East Java Province.

3.3. Robust Linear Regression Modelling

The study employed robust regression modeling to address issues of heteroscedasticity and outliers. Two models were developed using GRDP (natural logarithm) from the SAE BARE method as the dependent variable (Table 5). Model 1 included only the RSGI as the independent variable, while Model 2 added the population variable (natural logarithm), denoted as P .

In Model 1, four robust regression equations were estimated, each yielding significant results for both the intercept and RSGI (p -value < 0.05). The model using RSGI with Pearson weighting achieved an R^2 of 0.4755, indicating that RSGI explains approximately 47.55% of the GRDP variance. The Breusch–Pagan test confirmed no heteroscedasticity (p -value > 0.05).

In Model 2, both RSGI and population variables were significant (p -value < 0.05). The highest Adjusted R^2 value of 0.7658 was obtained using the RSGI Pearson Weight combined with population, meaning 76.58% of the GRDP variance was explained, while 23.42% remained unexplained. The Breusch–Pagan test again indicated no heteroscedasticity (p -value > 0.05), and the Variance Inflation Factor (VIF) values were all below 10, confirming the absence of multicollinearity [72].

Table 5. Robust Regression modelling results

Information	RSGI Equal Weighted	RSGI Pearson Weighted	RSGI Spearman Weighted	RSGI PCA Weighted
Model 1	7.212 + 0.116 <i>RSGI</i>	7.209 + 0.334 <i>RSGI</i>	7.207 + 0.349 <i>RSGI</i>	7.282 + 0.263 <i>RSGI</i>
Student-t Test (p -value)				
Intercept	2×10^{-16}	2×10^{-16}	2×10^{-16}	2×10^{-16}
RSGI	2×10^{-16}	2×10^{-16}	2×10^{-16}	2×10^{-16}
R^2	0.4262	0.4755	0.4611	0.4392
Breusch Pagan Test (p -value)	0.2904	0.1587	0.1885	0.1819
Model 2	−4.071 + 0.074 <i>RSGI</i> 1.035 <i>Ln (P)</i>	−3.589 + 0.221 <i>RSGI</i> 0.991 <i>Ln (P)</i>	−3.071 + 0.229 <i>RSGI</i> 0.999 <i>Ln (P)</i>	−4.271 + 0.1719 <i>RSGI</i> 1.056 <i>Ln (P)</i>
Student-t Test (p -value)				
Intercept	2×10^{-16}	2.44×10^{-15}	2.44×10^{-15}	2×10^{-16}
RSGI	2×10^{-16}	2×10^{-16}	2×10^{-16}	2×10^{-16}
<i>Ln (Population)</i>	2×10^{-16}	2×10^{-16}	2×10^{-16}	2×10^{-16}
Adjusted R^2	0.7445	0.7658	0.7565	0.7618
Breusch Pagan Test (p -value)	0.2237	0.1726	0.193	0.1032
VIF				
RSGI	1.3236	1.3706	1.3599	1.3362
<i>Ln (Population)</i>	1.3236	1.3706	1.3599	1.3362

3.4. RSGI Numerical and Descriptive Evaluation Results

The first numerical evaluation aimed to assess the closeness between RSGI values and GRDP. Since official GRDP data are only available at the district or city level, numerical evaluation at the 1 km $\times$ 1 km grid level was not feasible. To achieve a more accurate assessment at a finer scale, RSGI values were aggregated by sub-district, resulting in 666 observations. The relationship and degree of association between RSGI and GRDP were then evaluated using Pearson and Spearman Rank correlation analyses. The results of these analyses are presented in Table 6.

Table 6. RSGI correlation analysis with GRDP (BARE) in sub-district

RSGI	Pearson Correlation Test		Spearman Rank Test		Direction	Closeness	Statistically Significant
	Correlation Coefficient	p -value	Correlation Coefficient	p -value			
<i>RSGI Equal Weighted</i>	0.58429	2.2×10^{-16}	0.54944	2.2×10^{-16}	Positive	Moderate	Significant
<i>RSGI Pearson Weighted</i>	0.61250	2.2×10^{-16}	0.58458	2.2×10^{-16}	Positive	Strong, Moderate	Significant
<i>RSGI Spearman Weighted</i>	0.60601	2.2×10^{-16}	0.57782	2.2×10^{-16}	Positive	Strong, Moderate	Significant
<i>RSGI PCA Weighted</i>	0.59622	2.2×10^{-16}	0.54167	2.2×10^{-16}	Positive	Moderate	Significant

Based on Table 6, all RSGI values show significant correlations in both Pearson and Spearman Rank tests with sub-district level GRDP data. All correlations are positive, indicating that higher RSGI values tend to align with higher GRDP values. According to the interpretation by [62], the Pearson correlation results reveal that RSGI weighted by correlation coefficients exhibit strong relationships with GRDP, while other RSGI variants show moderate correlations. Meanwhile, the Spearman Rank test indicates moderate correlations across all RSGI values. The second numerical evaluation assessed

the predictive accuracy of Model 1 and Model 2 using robust regression (Table 7). Model 1 achieved the lowest RMSE, MAE, and MAPE values when using the RSGI Pearson Weighted variable as the independent variable. In contrast, Model 2, which included population and RSGI PCA Weighted as independent variables, yielded the smallest RMSE, MAE, and MAPE values 0.730, 0.482, and 7.000%, respectively. Overall, Model 2 demonstrated better predictive performance than Model 1, making it the most suitable model for estimating and predicting sub-district level GRDP in East Java Province.

Table 7. Robust Regression evaluation

Information	RMSE	MAE	MAPE
Model 1			
RSGI Equal Weighted	0.87611	0.63364	9.30421%
RSGI Pearson Weighted	0.85345	0.61129	9.00263%
RSGI Spearman Weighted	0.85883	0.61742	9.08007%
RSGI PCA Weighted	0.86731	0.62695	9.24068%
Model 2			
RSGI Equal Weighted	0.74442	0.49498	7.19380%
RSGI Pearson Weighted	0.73358	0.47902	6.98911%
RSGI Spearman Weighted	0.73831	0.48622	7.07949%
RSGI PCA Weighted	0.73047	0.48185	7.00055%

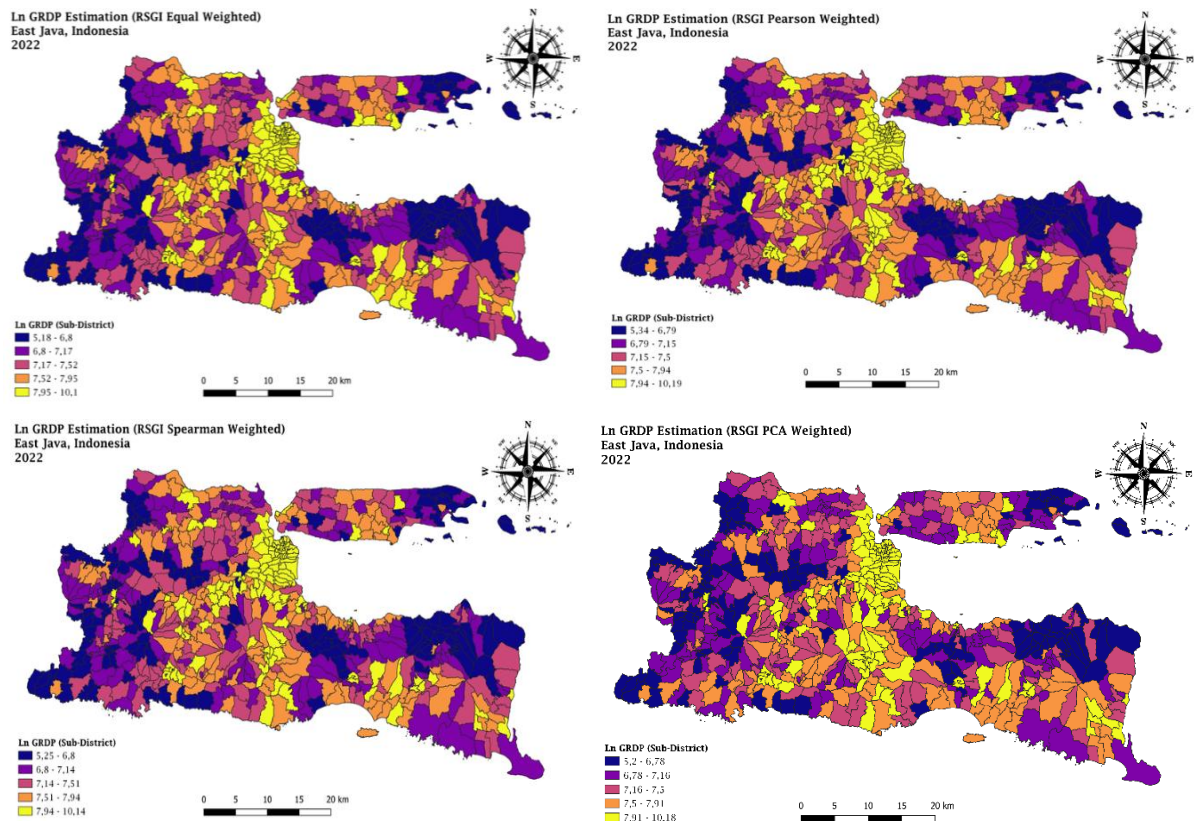


Figure 6. Map of GRDP estimated results in sub-district, East Java 2022 with Robust Regression.

The GRDP estimation results for each model using the robust regression method in Model 2 were visualized to facilitate interpretation, as shown in Figure 6. The sub-district-level GRDP estimates were mapped based on the best-performing model from the numerical evaluations. As illustrated in Figure 7, the model using RSGI PCA and population as independent variables produced the most accurate results, with RMSE, MAE, and MAPE values of 0.73047, 0.48185, and 7.00055%, respectively. Mapping was performed by classifying the estimated GRDP values into five equal intervals at both sub-district and

district/city levels. The sub-district-level results indicate that high GRDP estimates are concentrated in urban centers such as Surabaya City, Malang City, Madiun City, and Kediri City, as well as in several sub-districts of Malang and Banyuwangi Regencies. When aggregated to the district or city level, the highest estimated GRDP values were found in Surabaya City, Sidoarjo Regency, Batu City, Malang City. Overall, the results reveal that high GRDP estimates are concentrated in the northern and southern regions of East Java Province, while lower estimates are predominantly found in the western areas.

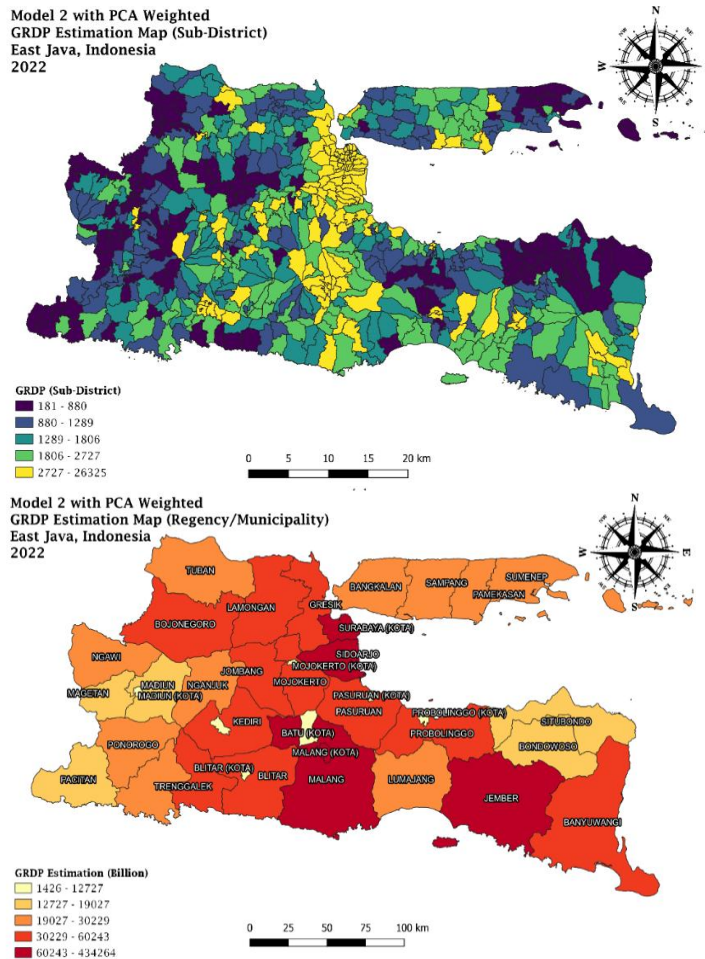


Figure 7. The map of GRDP estimation in regency/municipality level

4. Conclusion

This study introduces a new approach to GRDP mapping in East Java with higher spatial granularity, extending coverage to the sub-district level through the construction of a Relative Spatial GDP Index (RSGI) at a $1 \text{ km} \times 1 \text{ km}$ resolution. The method offers a cost-effective and timely update mechanism to support SDG 8, promoting sustained economic growth and monitoring regional development. The research integrates multi-source satellite imagery and Points of Interest (POI) data to build and model the RSGI, estimate GRDP using robust regression, and visualize the resulting spatial distribution.

The developed RSGI demonstrates a strong and significant correlation with sub-district-level GRDP in East Java. Visual analysis indicates that areas with high RSGI values correspond to urban regions with high accessibility, while low RSGI values are concentrated in spatially deprived or less accessible areas. This finding confirms that higher RSGI values are associated with higher GRDP levels.

Two robust regression models were tested, Model 1 using only RSGI and Model 2 incorporating both RSGI and population (log-transformed). Model 2 outperformed Model 1, yielding the best accuracy metrics with $\text{RMSE} = 0.73047$, $\text{MAE} = 0.48185$, and $\text{MAPE} = 7.00055\%$, and explaining 76.18% of GRDP variance using the RSGI PCA Weighted variable and population as predictors. The results further reveal that areas with high GRDP estimates are densely populated and well-connected, whereas low GRDP estimates are observed in sparsely populated or undeveloped regions.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This research uses secondary data obtained from Badan Pusat Statistik, Satellite Imagery, and OpenStreetMap.

Credit Authorship

Rifqi Ramadhan: Writing – review & editing, Resources, Methodology, Data Collection, Investigation, Formal analysis, Conceptualization. **I Made Satria Ambara Putra:** Writing – original draft, Investigation, Formal analysis. **Taufiq Agung Kurniawan:** Writing – original draft, Investigation, Formal analysis. **Fitri Kartiasih:** Validation, Supervision, Writing-Reviewing and Editing. **Raden Muaz Munim:** Writing-Reviewing, Validation. **Somethea Buoy:** Writing-Reviewing, Final Draft.

References

- [1] Bappenas, “Rancangan teknokratik rencana pembangunan jangka menengah nasional (RPJMN) 2020-2024,” *Kementrian Perenc. Pambang. Nas.*, pp. 2015–2019, 2019.
- [2] Badan Pusat Statistik, “Pertumbuhan ekonomi indonesia triwulan IV-2022,” *Badan Pusat Statistik*, 2023. .
- [3] BPS Jawa Timur, “Produk domestik regional bruto provinsi Jawa Timur menurut lapangan usaha 2018-2022,” *Badan Pusat Statistik Provinsi Jawa Timur*, 2023. .
- [4] Badan Pusat Statistik, “Indikator perumahan dan kesehatan lingkungan,” *Badan Pusat Statistik*, 2022. .
- [5] Q. Chen *et al.*, “Mapping China’s regional economic activity by integrating points-of-interest and remote sensing data with random forest,” *Environ. Plan. B Urban Anal. City Sci.*, vol. 48, no. 7, pp. 1876–1894, 2021, doi: 10.1177/2399808320951580.
- [6] Y. Zhou, T. Ma, C. Zhou, and T. Xu, “Nighttime light derived assessment of regional Inequality of socioeconomic development in China,” pp. 1242–1262, 2015, doi: 10.3390/rs70201242.
- [7] Y. Lu and N. C. Coops, “Bright lights, big city: Causal effects of population and GDP on urban brightness,” *PLoS One*, vol. 13, no. 7, pp. 1–15, 2018, doi: 10.1371/journal.pone.0199545.
- [8] J. A. Pagaduan, “Do higher-quality nighttime lights and net primary productivity predict subnational GDP in developing countries? Evidence from the Philippines,” *Asian Econ. J.*, vol. 36, no. 3, pp. 288–317, 2022, doi: 10.1111/asej.12278.
- [9] G. C. McCord and M. Rodriguez-Heredia, “Nightlights and subnational economic activity:

- estimating departmental GDP in Paraguay,” *Remote Sens.*, vol. 14, no. 5, pp. 1–16, 2022, doi: 10.3390/rs14051150.
- [10] Y. Gu, Z. Shao, X. Huang, and B. Cai, “GDP forecasting model for China’s provinces using nighttime light remote sensing data,” *Remote Sens.*, vol. 14, no. 15, 2022, doi: 10.3390/rs14153671.
- [11] X. M. Jin, L. Wan, Y. K. Zhang, and M. Schaepman, “Impact of economic growth on vegetation health in China based on GIMMS NDVI,” *Int. J. Remote Sens.*, vol. 29, no. 13, pp. 3715–3726, 2008, doi: 10.1080/01431160701772542.
- [12] W. Li, J. D. M. Saphores, and T. W. Gillespie, “A comparison of the economic benefits of urban green spaces estimated with NDVI and with high-resolution land cover data,” *Landsc. Urban Plan.*, vol. 133, pp. 105–117, 2015, doi: 10.1016/j.landurbplan.2014.09.013.
- [13] M. Hu and B. Xia, “A significant increase in the normalized difference vegetation index during the rapid economic development in the Pearl River Delta of China,” *L. Degrad. Dev.*, vol. 30, no. 4, pp. 359–370, 2019, doi: 10.1002/ldr.3221.
- [14] Y. Guo *et al.*, “Spatial and temporal changes in vegetation in the Ruoergai Region, China,” *Forests*, vol. 12, no. 1, pp. 1–17, 2021, doi: 10.3390/f12010076.
- [15] G. Wang and W. Peng, “Detecting influences of factors on GDP density differentiation of rural poverty changes,” *Struct. Chang. Econ. Dyn.*, vol. 56, pp. 141–151, 2021, doi: 10.1016/j.strueco.2020.10.004.
- [16] Z. M. Fan, X. Y. Bai, and N. Zhao, “Explicating the responses of NDVI and GDP to the poverty alleviation policy in poverty areas of China in the 21st century,” *PLoS One*, vol. 17, no. 8 August, pp. 1–16, 2022, doi: 10.1371/journal.pone.0271983.
- [17] H. W. Duolaiti, X., Kasim, A., Rehemani, R., & Laing, “Water body extraction of Ebinur Lake based on four water indexes and analysis of spatial-temporal changes,” *J. Yangtze River Sci. Res. Inst.*, 2022.
- [18] S. Ahmed, “Assessment of urban heat islands and impact of climate change on socioeconomic over Suez Governorate using remote sensing and GIS techniques,” *Egypt. J. Remote Sens. Sp. Sci.*, vol. 21, no. 1, pp. 15–25, 2018, doi: 10.1016/j.ejrs.2017.08.001.
- [19] A. D. Sakti *et al.*, “School location analysis by integrating the accessibility, natural and biological hazards to support equal access to education,” *ISPRS Int. J. Geo-Information*, vol. 11, no. 1, 2022, doi: 10.3390/ijgi11010012.
- [20] X. Zhang *et al.*, “Long-term trends in NO₂ columns related to economic developments and air quality policies from 1997 to 2016 in China,” *Sci. Total Environ.*, vol. 639, no. 2, pp. 146–155, 2018, doi: 10.1016/j.scitotenv.2018.04.435.
- [21] A. Sinha, “Trilateral association between SO₂/NO₂ emission, inequality in energy intensity, and economic growth: A case of Indian cities,” *Atmos. Pollut. Res.*, vol. 7, no. 4, pp. 647–658, 2016, doi: 10.1016/j.apr.2016.02.010.
- [22] D. Dissanayake, T. Morimoto, Y. Murayama, M. Ranagalage, and H. H. Handayani, “Impact of urban surface characteristics and socio-economic variables on the spatial variation of land surface temperature in Lagos City, Nigeria,” *Sustain.*, vol. 11, no. 1, pp. 1–23, 2019, doi: 10.3390/su11010025.
- [23] M. Khal, A. Algouti, A. Algouti, N. Akdim, S. A. Stankevich, and M. Menenti, “Evaluation of open digital elevation models: estimation of topographic indices relevant to erosion risk in the Wadi M’Goun watershed, Morocco,” *AIMS Geosci.*, vol. 6, no. 2, pp. 231–257, 2020, doi: 10.3934/geosci.2020014.
- [24] Q. Chen *et al.*, “Mapping China ’ s regional economic activity by integrating points-of-interest and remote sensing data with random forest,” vol. 0, no. 0, pp. 1–19, 2020, doi: 10.1177/2399808320951580.
- [25] H. Liu *et al.*, “Erratum to: Nightlight as a proxy of economic indicators: Fine-grained GDP inference around Chinese mainland via attention-augmented CNN from daytime satellite imagery (Remote SENS. 2021, 13, 2067),” *Remote Sens.*, vol. 13, no. 17, 2021, doi: 10.3390/rs13173360.
- [26] Z. Chen, W. Wang, H. Zong, and X. Yu, “Combining the NPP-VIIRS-like dataset and sentinel-2 images,” 2024.
- [27] N. Wu, J. Yan, D. Liang, Z. Sun, R. Ranjan, and J. Li, “High-resolution mapping of GDP using multi-scale feature fusion by integrating remote sensing and POI data,” *Int. J. Appl. Earth Obs. Geoinf.*, vol. 129, no. April, p. 103812, 2024, doi: 10.1016/j.jag.2024.103812.
- [28] Badan Pusat Statistik Provinsi Jawa Timur, “Produk domestik regional bruto Provinsi Jawa Timur menurut lapangan usaha 2018-2022,” *Badan Pusat Statistik Jawa Timur*, 2023. .

- [29] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," *AIP Conf. Proc.*, vol. 1644, pp. 97–104, 2015, doi: 10.1063/1.4907823.
- [30] J. G. Lee and M. Kang, "Geospatial big data: Challenges and opportunities," *Big Data Res.*, vol. 2, no. 2, pp. 74–81, 2015, doi: 10.1016/j.bdr.2015.01.003.
- [31] BPS RI, *Teknik pengumpulan data dan preprocessing citra satelit*. Badan Pusat Statistik, 2022.
- [32] R. A. Schowengerdt, *Remote sensing*. Elsevier, 2007.
- [33] J. B. Campbell, *Introduction to remote sensing*. Guilford Press, 2011.
- [34] G. K. Moore, "What is a picture worth? a history of remote sensing," *Hydrol. Sci. Bull.*, vol. 24, no. 4, pp. 477–485, 1979, doi: 10.1080/02626667909491887.
- [35] N. Suwargana, "Temporal dan spektral pada citra satelit landsat, spot Dan ikonos," *J. Ilm. Widy.*, vol. 1, no. 2, pp. 167–174, 2013.
- [36] A. F. Syah, "Penginderaan jauh dan aplikasinya di wilayah pesisir dan lautan," *J. Kelaut.*, vol. 3, no. 1, pp. 18–28, 2010.
- [37] A. Oktaviani and Y. Johan, "Perbandingan resolusi spasial, temporal dan radiometrik serta kendalanya," *J. Enggano*, vol. 1, no. 2, pp. 74–79, 2016, doi: 10.31186/jenggano.1.2.74-79.
- [38] C. D. Elvidge, M. Zhizhin, F.-C. Hsu, and K. Baugh, "What is so great about nighttime VIIRS data for the detection and characterization of combustion sources?," in *Proceedings of the Asia-Pacific Advanced Network*, 2013, vol. 35, no. 0, p. 33, doi: 10.7125/apan.35.5.
- [39] X. Chen and W. D. Nordhaus, "VIIRS nighttime lights in the estimation of cross-sectional and time-series GDP," *Remote Sens.*, vol. 11, no. 9, pp. 1–11, 2019, doi: 10.3390/rs11091057.
- [40] R. Ramadhan, A. W. Wijayanto, and S. Pramana, "Geospatial big data approaches to estimate granular level poverty distribution in East Java, Indonesia using machine learning and deep learning regressions," in *Proceedings of The International Conference on Data Science and Official Statistics*, 2023(1), 2023, pp. 186–200, doi: <https://doi.org/10.34123/icdsos.v2023i1.359>.
- [41] A. W. Wijayanto, D. W. Triscowati, and A. H. Marsuhandi, "Maize field area detection in East Java, Indonesia: An integrated multispectral remote sensing and machine learning approach," *ICITEE 2020 - Proc. 12th Int. Conf. Inf. Technol. Electr. Eng.*, pp. 168–173, 2020, doi: 10.1109/ICITEE49829.2020.9271683.
- [42] Y. Li, Z. Cao, H. Long, Y. Liu, and W. Li, "Dynamic analysis of ecological environment combined with land cover and NDVI changes and implications for sustainable urban–rural development: The case of Mu Us Sandy Land, China," *J. Clean. Prod.*, vol. 142, pp. 697–715, Jan. 2017, doi: 10.1016/j.jclepro.2016.09.011.
- [43] Bo-Cai Gao, "NDWI A normalized difference water index for remote sensing of vegetation liquid water from space," *Elsevier*, vol. 266, no. April, Feb. 1996.
- [44] S. K. McFeeters, "The use of the normalized difference water index (NDWI) in the delineation of open water features," *Int. J. Remote Sens.*, vol. 17, no. 7, pp. 1425–1432, 1996, doi: 10.1080/01431169608948714.
- [45] H. Krishna Karanam, "Study of normalized difference built-up (Ndbi) index in automatically mapping urban areas from landsat TM imagery," *Int. J. Sci. Res. Rev.*, vol. 7, no. 1, pp. 1–8, 2018.
- [46] Y. Zha, J. Gao, and S. Ni, "Use of normalized difference built-up index in automatically mapping urban areas from TM imagery," *Int. J. Remote Sens.*, vol. 24, no. 3, pp. 583–594, 2003, doi: 10.1080/01431160304987.
- [47] C. He, P. Shi, D. Xie, and Y. Zhao, "Improving the normalized difference built-up index to map urban built-up areas using a semiautomatic segmentation approach," *Remote Sens. Lett.*, vol. 1, no. 4, pp. 213–221, 2010, doi: 10.1080/01431161.2010.481681.
- [48] E. D. Meutia, M. Fikri, R. Munadi, and Yunida, "Analisis data citra satelit terra modis sebagai indikasi potensi ikan di wilayah kreung raya," *KITEKRO J. Komputer, Teknol. Informasi, dan Elektro*, vol. 7, no. 1, pp. 9–14, 2022.
- [49] J. P. Veefkind *et al.*, "TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications," *Remote Sens. Environ.*, vol. 120, no. 2012, pp. 70–83, 2012, doi: 10.1016/j.rse.2011.09.027.
- [50] R. D. Stephens and S. H. Cadle, "Remote sensing measurements of carbon monoxide emissions from on-road vehicles," *J. Air Waste Manag. Assoc.*, vol. 41, no. 1, pp. 39–46, 1991, doi: 10.1080/10473289.1991.10466823.
- [51] C. Nistor, M. Virghileanu, and B. Mihai, "Monitoreo de la contaminación por dióxido de nitrógeno (NO₂) con imágenes satelitales Sentinel-5P en Europa durante el brote pandémico de

- coronavirus,” no. 2, 2020.
- [52] R. Kurniawan, R. Hasabi, S. K. Wongsonadi, P. U. Gio, A. Purwanto, and B. Sumargo, “Examining the influence of congestion, industry, and green open space on air quality vulnerability in towards green development in Jakarta,” *Innov. Green Dev.*, vol. 4, no. 3, 2025, doi: 10.1016/j.igd.2025.100247.
 - [53] K. Bakhsh, T. Akmal, T. Ahmad, and Q. Abbas, “Investigating the nexus among sulfur dioxide emission, energy consumption, and economic growth: empirical evidence from Pakistan,” *Environ. Sci. Pollut. Res.*, vol. 29, no. 5, pp. 7214–7224, 2022, doi: 10.1007/s11356-021-15898-9.
 - [54] L. Shikwambana, P. Mhangara, and M. Kganyago, “Assessing the relationship between economic growth and emissions levels in south africa between 1994 and 2019,” *Sustain.*, vol. 13, no. 5, pp. 1–15, 2021, doi: 10.3390/su13052645.
 - [55] M. Abrams, R. Crippen, and H. Fujisada, “ASTER global digital elevation model (GDEM) and ASTER global water body dataset (ASTWBD),” *Remote Sens.*, vol. 12, no. 7, pp. 1–12, 2020, doi: 10.3390/rs12071156.
 - [56] I. Elkhrachy, “Vertical accuracy assessment for SRTM and ASTER digital elevation models: A case study of Najran city, Saudi Arabia,” *Ain Shams Eng. J.*, vol. 9, no. 4, pp. 1807–1817, 2018, doi: 10.1016/j.asej.2017.01.007.
 - [57] L. W. Yeow, R. Low, Y. X. Tan, and L. Cheah, “Point-of-interest (Poi) data validation methods: An urban case study,” *ISPRS Int. J. Geo-Information*, vol. 10, no. 11, 2021, doi: 10.3390/ijgi10110735.
 - [58] A. P. Utomo *et al.*, “Regresi robust untuk memodelkan pendapatan usaha industri makanan non-makloon berskala mikro dan kecil di Jawa Barat tahun 2013,” pp. 63–74, 2013.
 - [59] Soemartini, *Pencilan (Outliers)*. Jatinangor: Universitas Padjajaran, 2007.
 - [60] J. Yohai, D. B. Aires, and C. E. Ma, “High breakdown-point and high efficiency robust estimates for regression,” vol. 15, no. 2, pp. 642–656, 2014.
 - [61] J. Raymaekers and P. J. Rousseeuw, “Transforming variables to central normality: Machine learning,” *Mach. Learn.*, vol. 113, no. 8, pp. 4953–4975, 2024, doi: 10.1007/s10994-021-05960-5.
 - [62] Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif dan R&D*. 2011.
 - [63] W. Lin, “Temporal and spatial analysis of integrated energy and environment efficiency in China based on a green GDP index,” pp. 1376–1390, 2011, doi: 10.3390/en4091376.
 - [64] Y. Wang and Y. Chen, “Using VPI to measure poverty-stricken villages in China,” *Soc. Indic. Res.*, no. June, 2016, doi: 10.1007/s11205-016-1391-5.
 - [65] S. R. Putri and A. W. Wijayanto, “Developing relative spatial poverty index using integrated remote sensing and geospatial big data approach : A case study of East Java , Indonesia,” 2024.
 - [66] R. Ramadhan and A. W. Wijayanto, “Integrating satellite imageries and multiple geospatial big data for granular mapping of spatial distribution of the human development index in East Java , Indonesia,” in *Proceedings of The International Conference on Data Science and Official Statistics*, 2023, vol. 2023(1), pp. 274–295, doi: <https://doi.org/10.34123/icdsos.v2023i1.369>.
 - [67] F. Kartiasih, N. D. Nachrowi, I. D. Gede, and K. Wisana, “Information technology for development inequalities of Indonesia ’ s regional digital development and its association with socioeconomic characteristics : a spatial and multivariate analysis,” *Inf. Technol. Dev.*, vol. 0, no. 0, pp. 1–30, 2022, doi: 10.1080/02681102.2022.2110556.
 - [68] A. Cartone and P. Postiglione, “Principal component analysis for geographical data : the role of spatial effects in the definition of composite indicators Principal component analysis for geographical data : the role of spatial effects in the de fi nition of composite,” *Spat. Econ. Anal.*, vol. 0, no. 0, pp. 1–22, 2020, doi: 10.1080/17421772.2020.1775876.
 - [69] Y. Nurmasari and A. W. Wijayanto, “Oil palm plantation detection in Indonesia using sentinel-2 and landsat-8 optical satellite imagery (Case Study: Rokan Hulu Regency, Riau Province),” *Int. J. Remote Sens. Earth Sci.*, vol. 18, no. 1, p. 1, 2021, doi: 10.30536/j.ijreses.2021.v18.a3537.
 - [70] D. Coondoo, I. Researcher, and A. Majumder, “District-level poverty estimation : A proposed method District-level poverty estimation : a proposed method,” no. October 2014, 2011, doi: 10.1080/02664763.2010.547568.
 - [71] L. Mládková, V. Pení, and R. Va, “Forest soil acidification assessment using principal component analysis and geostatistics,” vol. 140, pp. 374–382, 2007, doi: 10.1016/j.geoderma.2007.04.018.
 - [72] F. J. Riggins, “The digital divide : Current and future research directions,” vol. 6, no. 12, 2005, doi: 10.17705/1jais.00074.



Integrating Multi-Criteria Decision Analysis and Machine Learning for Fine-Scale Mapping of Safe Drinking Water Access in Bengkulu Province, Indonesia

Andrew Maruli Tua Tampubolon^{1*}, Bony Parulian Josaphat², Asriadi Sakka³,
Yohanes Wahyu Trio Pramono⁴

^{1,2}Politeknik Statistika STIS, Jakarta, Indonesia, ³Korea National University of Science and Technology (UST), Daejeon, Republic of Korea, ⁴School of Information Management, Victoria University of Wellington, New Zealand

*Corresponding Author: E-mail address: 222111894@stis.ac.id

ARTICLE INFO

Article history:

Received 12 December, 2025

Revised 21 December, 2025

Accepted 23 December, 2025

Published 31 December, 2025

Keywords:

Big Data; Machine Learning;
Multi-Criteria Decision
Analysis; Safe Drinking Water
Access; Satellite Imagery

Abstract

Introduction/Main Objectives: This study aims to develop a 1 km × 1 km level estimation model of safe drinking water access using multisource satellite imagery, point of interest (POI), and aquifer productivity maps. **Background Problems:** There is a lack of alternative data sources for estimating safe drinking water access that are cost-, time-, and labor-efficient while maintaining high accuracy and frequent updates. **Novelty:** This study integrates Multi-Criteria Decision Analysis (MCDA) and machine learning methods to estimate and map safe drinking water access at a 1 km × 1 km resolution. **Research Methods:** Multisource geospatial data were used to construct the model. Within the MCDA approach, the Weighted Product Model (WPM) was employed to develop the Safe Drinking Water Access Index (SDWAI). Meanwhile, the machine learning regression algorithms Adaptive Boosting Regression (ABR) and Gradient Boosting Regression (GBR) were applied to estimate safe drinking water access at a fine spatial scale. The study was conducted in Bengkulu Province, Indonesia. **Finding/Results:** WPM yielded the best MCDA performance ($R^2 = 0.3699$, RMSE = 10.6566, MAE = 9.5427, MAPE = 0.1405), while ABR showed the best machine learning performance ($R^2 = 0.4361$, RMSE = 10.0813, MAE = 8.3750, MAPE = 0.1333).

1. Introduction

Water constitutes a vital element for human well-being and the advancement of development [1], [2]. Clean water refers to water of sufficient quality suitable for consumption and daily use, including sanitation [3]. The growing population has increased demand for clean water, prompting the Indonesian government to enhance access to safe water across various regions [4]. Adequate drinking water availability is vital for supporting healthy and sustainable living [5]. However, approximately 25% of the global population still lacks access to safe water [6]. In Indonesia, drinking water quality standards are regulated by Minister of Health Regulation No. 492 of 2010, mandating that water must be physically, chemically, and microbiologically safe [6] – [8], with providers responsible for ensuring safety throughout distribution [7].

According to WHO/UNICEF, access to safe drinking water includes sources capable of providing safe water such as piped supplies, bore wells, protected springs, and rainwater harvesting [9] – [11]. The Sustainable Development Goals (SDGs) emphasize universal access to safe and affordable drinking water [12]. Despite this, challenges to equitable access in Indonesia include limited funding, insufficient infrastructure, capacity constraints, and natural disaster risks [13]. The Ministry of Public Works and Housing (PUPR) targets increased household connections to piped water networks, requiring accurate and detailed spatial data at the household level to prioritize interventions [14], [15].

Detailed data availability is also essential for village governments to manage safe drinking water and plan village fund allocations per recent regulations [16], [17]. Currently, access data primarily derives from the National Socio-Economic Survey (Susenas), which is costly and limited to district/city levels, complicating priority area identification and program evaluation [18], [19]. As an alternative, satellite imagery, points of interest (POI), and other spatial data offer advantages in speed, cost-efficiency, ease of collection, and spatial detail [20]. Aquifer productivity maps further support assessment of regional groundwater capacity for clean water supply [21]. To support comprehensive spatial decision-making, Multi-Criteria Decision Analysis (MCDA) integrates relevant indicators into composite indices for policy guidance [22]. However, MCDA faces limitations with high-dimensional and multicollinear data [23]; thus, machine learning techniques complement MCDA by effectively extracting patterns from complex datasets [21], [24].

Recent studies have identified several geospatial indicators as proxies for understanding safe drinking water access. Elevation serves as a proxy for piped infrastructure accessibility, as high-altitude areas often face pipeline construction challenges [25]. Slope indicates groundwater potential [26], while the Topographic Wetness Index (TWI) relates to surface water utilization, a major water source for household piped systems in Indonesia [27]. Night-time Light (NTL) proxies electricity consumption, indicating potential access to electric water pumps, with proven efficacy in developing countries [28]. The Human Settlement Index (HSI) reflects urban areas with high socio-economic activities [29]. Carbon monoxide (CO) pollution correlates with economic growth and poverty, with poorer areas often lacking private water infrastructure such as pipes, pumps, or protected wells [25], [30]. Land Surface Temperature (LST), linked to urban heat island effects, distinguishes urban and rural areas, correlating with disparities in safe water access—urban residents commonly use piped water, while rural populations rely on groundwater or unprotected sources [31].

Additionally, points of interest (POI) related to economic facilities serve as proxies because communities without piped water often depend on accessible economic outlets such as bottled water vendors [32]. POI distance water bodies and POI distance IPA are used in modeling access to safe drinking water [25], [33], while aquifer productivity maps remain critical for assessing groundwater availability [21]. To complement official data on safe drinking water access, this study develops an estimation mapping approach using multisource remote sensing, POI, and aquifer productivity maps. This method aggregates data at a fine 1 x 1 km scale to produce detailed maps that can be updated more efficiently. Consequently, policy decisions aimed at improving equitable access to safe drinking water are expected to be better informed and more effective.

2. Material and Methods

2.1. Study Area

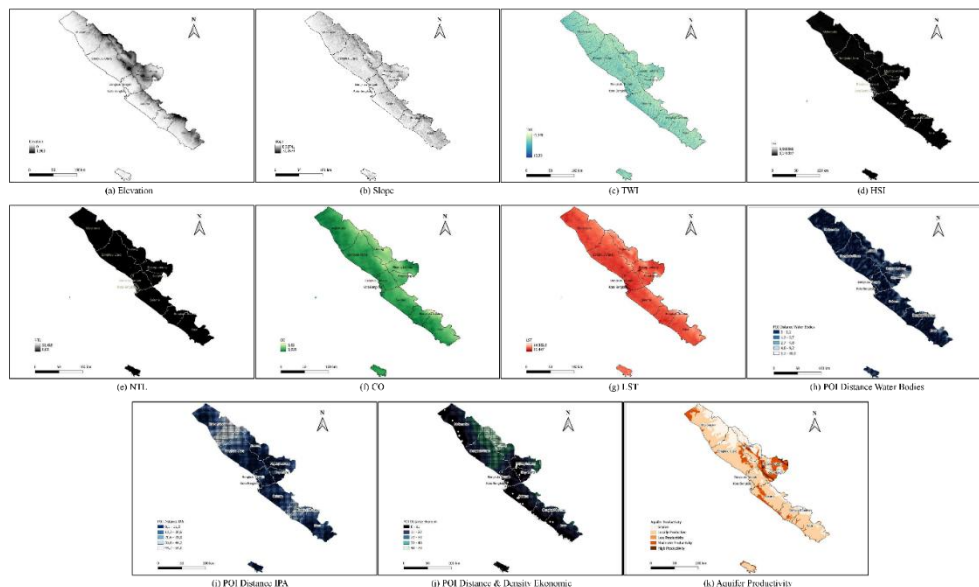
Bengkulu Province is located to the west of the Bukit Barisan mountain range. It comprises 10 regencies/municipalities and 129 districts. In Bengkulu, protected wells are the most widely used primary source of drinking water for households, accounting for 29.72 percent. According to data from BPS-Statistics Indonesia in 2022, Bengkulu ranks as the second lowest province in Indonesia in terms of the percentage of access to safe drinking water, following Papua.

2.2. Data Source

In this study, satellite imagery, points of interest (POI), and other geospatial data were used to develop an index and estimation model for access to safe drinking water. The datasets are described in detail in Table 1.

Table 1. Data summary

Variable	Spatial Resolution	Temporal Resolution	Data Source	References
Elevation	30 m	-	NASA SRTM	[25], [33], [34]
Slope	30 m	-	NASA SRTM	[25], [33], [34]
Topographic Wetness Index (TWI)	30 m	-	NASA SRTM	[35]
Human Settlement Index (HSI)	463,83 m	monthly	Sentinel-2, Suomi-NPP VIIRS	[34]
Night-time Light (NTL)	463,83 m	monthly	Suomi-NPP VIIRS	[15], [33]
Carbon Monoxide (CO)	1113,2 m	< 1 day	Sentinel-5P	[30]
Land Surface Temperature (LST)	1000 m	1 day	MODIS	[36], [37]
POI Distance Water Bodies	Point	dynamically	Sentinel-2, OpenStreetMap	[25], [33], [38]
POI Distance & Density Economic	Point	dynamically	OpenStreetMap	[33]
POI Distance IPA	Point	dynamically	Ministry of Public Works and Housing	[25]
Aquifer Productivity	100 m	-	Ministry of Energy and Mineral Resources	[21]
Population Grid	100 m	-	Worldpop	[22], [25], [33]
Percentage of households with access to safe drinking water	Regency, district	2022	BPS-Statistics Indonesia	

**Figure 1.** Data visualization

This research utilizes remote sensing data, specifically multisource satellite imagery, combined with spatial POI data obtained from OpenStreetMap (OSM). The satellite imagery sources used in this study comprise Elevation, Slope, and the Topographic Wetness Index (TWI) derived from NASA's SRTM; Nighttime Light (NTL) and Human Settlement Index (HSI) obtained from Suomi-NPP; Carbon Monoxide (CO) data from Sentinel-5P; and Land Surface Temperature (LST) data from MODIS. POI data were collected from both OpenStreetMap (OSM) and the Ministry of Public Works. Additional geospatial datasets included the Aquifer Productivity Map sourced from the Ministry of Energy and Mineral Resources and population distribution data from WorldPop. The data visualizations used in this study are presented in Figure 1.

2.3. Research Framework

This research generated a map of safe drinking water access at a 1 km × 1 km spatial resolution to offer a more efficient and update-ready alternative to household survey data, while maintaining accuracy and detailed spatial representation. The methodological steps involved data acquisition, preprocessing, integration and transformation, correlation analysis and variable selection, model construction, and subsequent validation and interpretation. Analysis and visualization were carried out using QGIS 3.26.2 and Python 3.12. The expected outcome is a spatially detailed 1 km × 1 km map of access to safe drinking water, accompanied by its validation results, as illustrated in Figure 2.

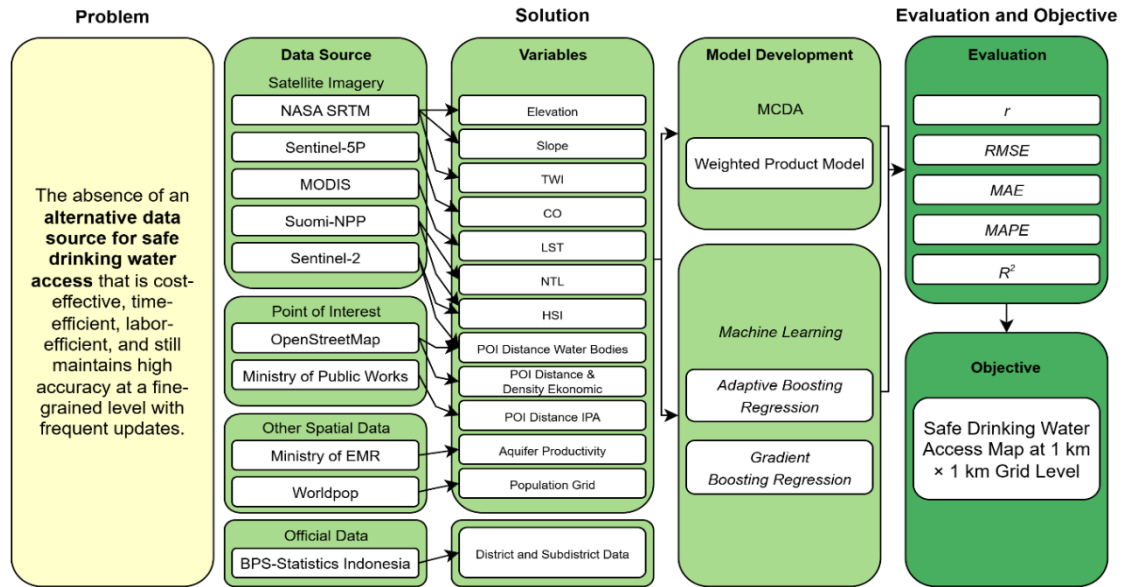


Figure 2. The research framework

2.4. Data Transformation

Feature extraction from the dataset produced median values for each 1 km × 1 km geographic zone across various variables. Subsequently, the values were normalized using the Yeo–Johnson power transformation, a method capable of processing variables with either positive or negative ranges [39]. This transformation aims to approximate a normal (Gaussian) distribution, thereby enhancing analytical and modeling performance [40], as formulated in Equation (1) [39].

$$y_{\lambda}(x) = \begin{cases} \frac{(1+x)^{\lambda-1}}{\lambda}, \lambda \neq 0, x \geq 0 \\ \log(1+x), \lambda = 0, x \geq 0 \\ -\frac{(1+x)^{2-\lambda}}{2-\lambda}, \lambda \neq 2, x < 0 \\ -\log(1-x), \lambda = 2, x < 0 \end{cases} \quad (1)$$

In this transformation, x denotes the variable value or input data, and λ is a parameter estimated using the Maximum Likelihood method under the assumption of normally distributed variables. When $\lambda = 1$, the transformation is linear; for $\lambda < 1$, it compresses the right tail and stretches the left tail, making a right-skewed distribution more symmetric; and for $\lambda > 1$, it adjusts a left-skewed distribution toward symmetry.

2.5. Pearson Correlation

The Pearson correlation coefficient (r) has a value range from 0 to 1, with the sign indicating whether the relationship is positive or negative. The interpretation of the strength of the Pearson correlation is shown in Table 2 [41].

Table 2. Interpretation of correlation coefficient

Range of Values	Level of Intensity
$0.00 \leq r \leq 0.199$	Very Weak
$0.20 \leq r \leq 0.399$	Weak
$0.40 \leq r \leq 0.599$	Strong Enough
$0.60 \leq r \leq 0.799$	Strong
$0.80 \leq r \leq 1.000$	Very Strong

2.6. Multi-Criteria Decision Analysis (MCDA)

Safe Drinking Water Access Index (SDWAI) was constructed using a Multi-Criteria Decision Analysis (MCDA) approach with the Weighted Product Model (WPM). In WPM, the calculation involves a multiplicative process, where each alternative's value is raised to the power of its corresponding criterion weight and then multiplied across all criteria to obtain the final score for each alternative [42]. The weight of each indicator was determined based on its Pearson correlation coefficient. The model is formulated as follows:

$$WPM_j = \prod_{i=1}^n |X_{ij}|^{w_i} \quad (2)$$

where WPM_j represents the WPM index value for the j -th grid, X_{ij} denotes the i -th indicator at the j -th grid level, and w_i refers to the weight of the i -th indicator.

2.7. Machine Learning

The estimation model was developed using a machine learning approach, which is widely applied in estimation tasks due to its focus on achieving high accuracy—an essential requirement in predictive modeling [43]. In this approach, the algorithms employed include Adaptive Boosting Regression (ABR) and Gradient Boosting Regression (GBR). Optimal parameters and hyperparameters for each model were determined through a grid search combined with 5-fold cross validation.

Model development was conducted at the sub-district level using data from two different time periods. Data from 2021 were used for model training, while data from 2022 were reserved exclusively for model testing. This temporal holdout (time-based split) strategy was adopted to prevent data leakage and to provide a more realistic assessment of model performance when applied to future data [35]. Such a separation between training and testing datasets is commonly recommended in predictive modeling to ensure the generalizability of the developed models [44].

2.8. Model Evaluation

The evaluation metrics used include RMSE, MAE, MAPE, Pearson's correlation coefficient, and the coefficient of determination (R^2). RMSE, MAE, and MAPE are used to quantify the magnitude of error between predicted and observed values, where lower values indicate higher prediction accuracy [45]. Pearson's correlation coefficient is employed to assess the strength of the linear relationship between the estimated values and official statistics, while R^2 measures the proportion of variance in the official data explained by the model. Together, these metrics provide a comprehensive evaluation of model performance by capturing both prediction accuracy and the consistency of statistical and spatial patterns relative to official data. The formulations of RMSE, MAE, MAPE, and R^2 are presented in Equations (3) to (6).

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3)$$

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{\hat{y}_i} \right| \quad (4)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6)$$

3. Results and Discussion

3.1. Variable Selection

Before constructing the Safe Drinking Water Access Index (SDWAI), a Pearson correlation analysis was conducted to examine the relationship between spatially derived variables and the percentage of access to safe drinking water. All variables were aggregated to the regency/municipality level and transformed using the Yeo–Johnson method to improve their distributional properties prior to analysis. Pearson’s correlation coefficient was employed to quantify the strength and direction of linear relationships between the transformed variables and official safe drinking water access data. Variables exhibiting statistically significant linear correlations with the percentage of access to safe drinking water were considered suitable for inclusion in the SDWAI. The analysis was conducted using official statistics available for only 10 regencies/municipalities in Bengkulu Province.

Statistical significance was evaluated using a significance level of $\alpha = 0.15$ to reduce the risk of excluding potentially relevant variables, given the limited number of regency/municipality-level observations. Variables with p-values less than 0.15 were considered statistically significant. The significance test was used to identify variables exhibiting sufficiently strong empirical relationships for inclusion in the SDWAI. As shown in Table 3, most variables exhibited statistically significant correlations at the 15% level. However, the Topographic Wetness Index (TWI), POI Distance to Economic Facilities, and POI Distance to IPA did not meet this criterion and were therefore excluded from the SDWAI construction. The correlation coefficients of the retained variables were subsequently used as weights in the SDWAI formulation.

Table 3. Results of the Pearson correlation

Variable	Correlation	p-value	Significant
Elevation	-0.570	0.085	Yes
Slope	-0.508	0.134	Yes
Topographic Wetness Index (TWI)	0.486	0.154	No
Human Settlement Index (HSI)	0.514	0.129	Yes
Night-time light (NTL)	0.497	0.144	Yes
Carbon Monoxide (CO)	0.580	0.079	Yes
Land Surface Temperature (LST)	0.503	0.139	Yes
POI Distance Water Bodies	0.825	0.003	Yes
POI Distance Ekonomik	-0.189	0.601	No
POI Density Ekonomik	0.620	0.056	Yes
POI Distance IPA	0.284	0.427	No
Aquifer Productivity	-0.512	0.131	Yes

3.2. The SDWAI Evaluation and Development

The evaluation to determine the best SDWAI was conducted using correlation (r), R^2 , RMSE, MAE, and MAPE, with the SDWAI aggregated at the regencies/municipalities level. Table 4 presents the evaluation results at this level. The correlation coefficient (r) of 0.6491 suggests that the estimated values and the official data are strongly positively correlated. The coefficient of determination (R^2) of 0.3699 suggests that approximately 36.99% of the variation in official safe drinking water access data can be explained by the WPM-based estimates.

Table 4. Results of the SDWAI evaluation at the regency/municipality level

Method	r	R^2	RMSE	MAE	MAPE
Weight Product Model (WPM)	0.6491	0.3699	10.6566	9.5427	0.1405

Access to safe drinking water is closely related to population distribution. The Safe Drinking Water Access Index (SDWAI) was generated by overlaying it with grid-level population data obtained from WorldPop. This approach allows the index to better reflect spatial variations in demand for drinking water services and the concentration of populations potentially affected by limited access. The following section displays the SDWAI map generated at a $1 \text{ km} \times 1 \text{ km}$ spatial level, as shown in Figure 3.

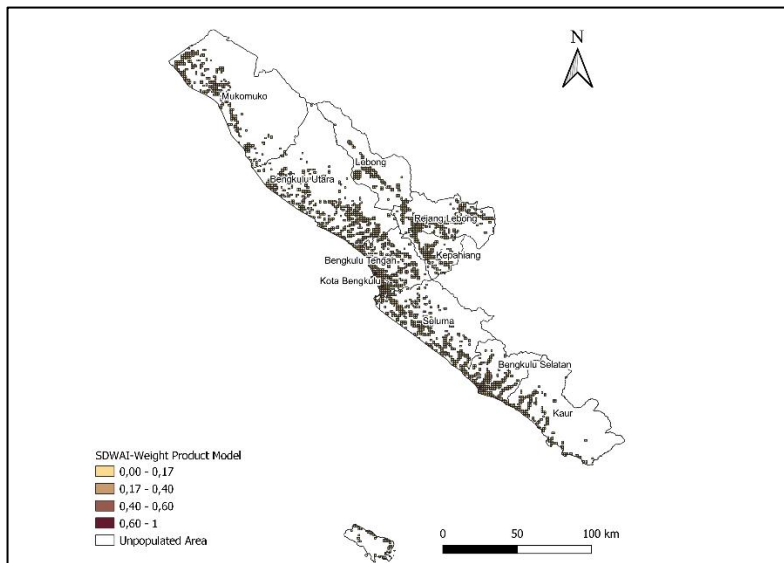


Figure 3. The obtained scaled SDWAI-WPM

The WPM-based mapping results show that areas with index values of 0.00–0.17 are generally located far from major activity centers, such as most of Kaur Regency and the western part of North Bengkulu Regency. Values of 0.17–0.40 are more widely distributed, covering large parts of Mukomuko, North Bengkulu, South Bengkulu, and Rejang Lebong Regencies. Values of 0.40–0.60 are concentrated around the center of South Bengkulu Regency and several locations in Bengkulu Municipality. Meanwhile, values of 0.60–1 are found only in limited areas in the city center of Bengkulu Municipality.

3.3. The Best Parameter/Hyperparameter Selection

The machine learning model was developed by selecting the optimal parameters and hyperparameters through a grid search process. Table 5 presents the best parameter and hyperparameter configurations derived from the 5-fold cross-validation performed with a random state of 42. This procedure ensures that the selected models achieve a balance between predictive accuracy and model generalizability while minimizing the risk of overfitting. The use of cross-validation further enhances the robustness of the parameter selection by evaluating model performance across multiple data partitions.

Each model was trained using grid search and 5-fold cross validation to determine the optimal hyperparameters. The model performance was evaluated through two validation approaches, which included 5-fold cross-validation and testing on a hold-out dataset, using RMSE, MAE, and MAPE as the evaluation metrics. The evaluation results are presented in Table 6.

Table 5. Results of hyperparameter selection

Model	Parameter/ Hyperparameter	Options	Selected
Adaptive Boosting Regression (ABR)	n_estimators	100, 500, 1000	100
	learning_rate	0.01, 0.1, 1.0	0.01
Gradient Boosting Regression (GBR)	n_estimators	500, 1000	500
	learning_rate	0.1, 0.01, 0.001	0.001
	max_depth	1, 3, 9, 12	1
	min_samples_split	8, 10, 12	8

Table 6. Results of the evaluation with 5-fold cross-validation and data testing

Model	RMSE		MAE		MAPE	
	<i>5-fold CV</i>	<i>Test</i>	<i>5-fold CV</i>	<i>Test</i>	<i>5-fold CV</i>	<i>Test</i>
ABR	0.9279	0.9709	0.7412	0.7896	0.0189	0.0103
GBR	0.9699	1.0228	0.8036	0.8574	0.0123	0.0104

Based on the RMSE, MAE, and MAPE values from the test data, the ABR model achieved the best performance, with an RMSE of 0.9709, MAE of 0.7896, and MAPE of 0.0103, the lowest among all models. This indicates that ABR was able to produce predictions that were closest to the actual values. The superior performance of ABR suggests its effectiveness in capturing complex, non-linear relationships between multisource geospatial predictors and safe drinking water access. Moreover, the ensemble-based nature of ABR enhances model stability by iteratively reducing prediction errors from weak learners.

Table 7. Results of the ABR evaluation at the regency/municipality level

Method	<i>r</i>	<i>R</i> ²	RMSE	MAE	MAPE
Adaptive Boosting Regression (ABR)	0.7706	0.4361	10.0813	8.3750	0.1333

As shown in Table 7, the correlation coefficient (*r*) of 0.7706 indicates a strong positive relationship between the estimated values and the official data. The coefficient of determination (*R*²) of 0.4361 suggests that approximately 43.61% of the variation in official safe drinking water access data can be explained by the ABR model estimates. This level of explanatory power is notable given the inherent spatial heterogeneity and the limited availability of fine-scale ground truth data. The results indicate that the ABR model provides a reliable approximation of official statistics while enabling spatial disaggregation at a much finer resolution.

3.4. The Safe Drinking Water Access Mapping

The machine learning models were trained using data on the percentage of households with access to safe drinking water, enabling estimation at the 1 km × 1 km spatial grid. Since SDWAI values range from 0 to 1, the first step to enable comparison with the mapped estimates was to align their data ranges. The evaluation employed simple linear regression, assigning SDWAI as the independent variable and the official district-level estimates as the dependent variable. The resulting regression model was then applied to generate SDWAI estimates at the spatial grid. Figure 4 presents the estimated distribution of safe drinking water access derived from the SDWAI-WPM model, while Figure 5 displays the estimation outcomes generated using the ABR model.

SDWAI and machine learning maps show similar patterns. Areas with poorer access to safe drinking water are located in the northeast, which consists of hilly and mountainous terrain, while areas with better access are found in urban regions. The convergence of patterns across both approaches indicates that the spatial distribution of safe drinking water access is consistently captured by the index-based and data-driven models. Topographic constraints and limited infrastructure development in mountainous areas likely contribute to lower accessibility, whereas urban regions benefit from more developed piped water networks and supporting facilities. This consistency reinforces the reliability of the proposed framework in identifying spatial disparities in safe drinking water access at a fine resolution.

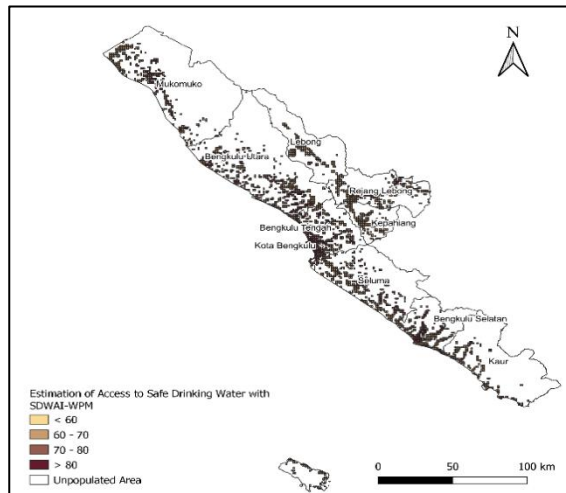


Figure 4. Results of the SDWAI-WPM estimation

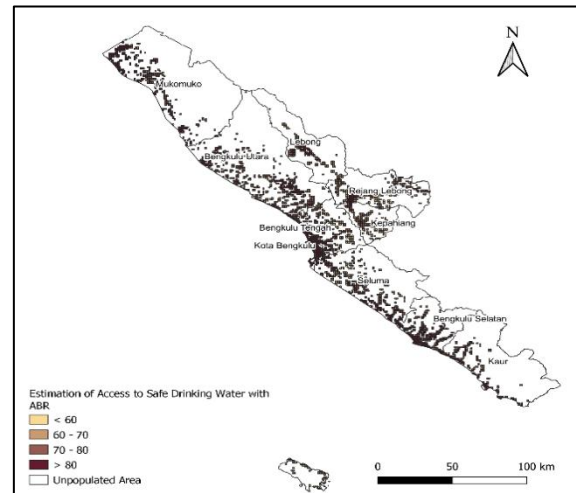


Figure 5. Results of the ABR estimation

3.5. SDWAI Ground Truth Analysis

Numerous previous studies have shown that ground-truth identification based on high-resolution imagery offers an evaluation perspective that cannot be fully represented by purely numerical, pixel-level validation. In this study, four randomly selected 1 km × 1 km grid cells were examined through visual interpretation using Google Earth imagery to characterize their geographic conditions.

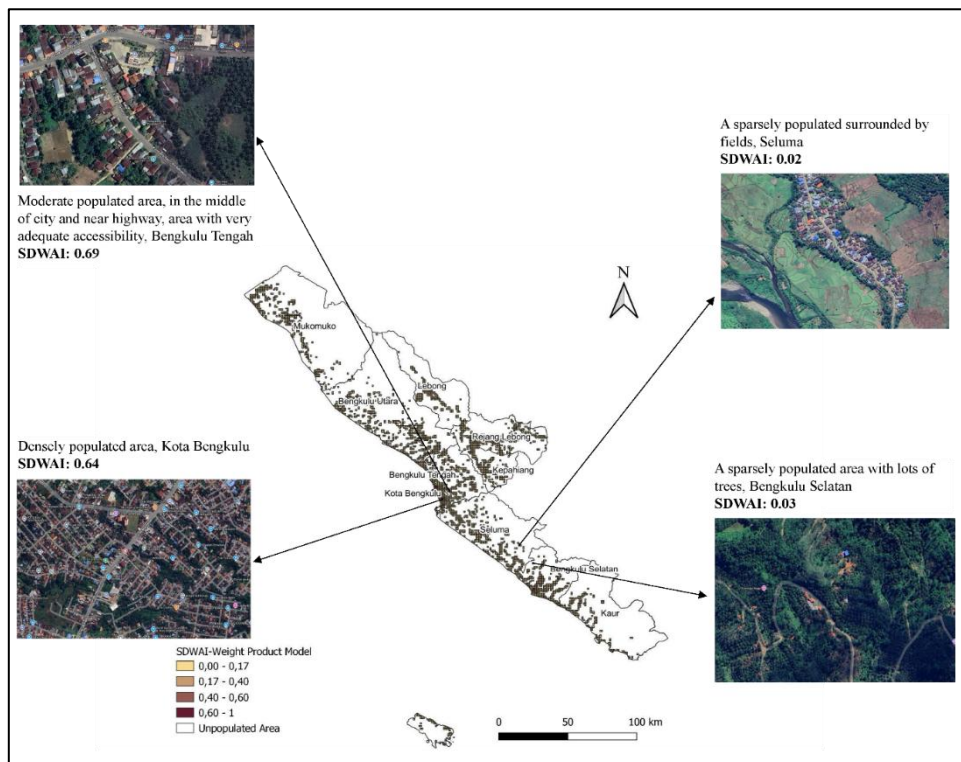


Figure 6. SDWAI (scaled index) ground-truth check

The results indicate that areas with high SDWAI scores tend to correspond to moderately to densely populated areas with more adequate accessibility (Figure 6). Urban areas generally exhibit higher SDWAI values, reflecting better availability of infrastructure and services. In contrast, areas with low SDWAI scores are more commonly associated with sparsely populated regions characterized by inadequate infrastructure and restricted access to basic services.

4. Conclusion

This study successfully achieved its objective of developing a 1 km × 1 km resolution estimation model of safe drinking water access by integrating multisource satellite imagery, point of interest (POI) data, and aquifer productivity maps. By combining Multi-Criteria Decision Analysis (MCDA) and machine learning approaches, the proposed framework enables the generation of fine-scale spatial estimates in regions where official statistics are available only at the regency/municipality level.

The Safe Drinking Water Access Index (SDWAI), developed using the Weighted Product Model (WPM), demonstrated a strong positive relationship with official data ($r = 0.6491$; $R^2 = 0.3699$), indicating that the MCDA-based approach provides a consistent and interpretable baseline for spatial disaggregation. The incorporation of machine learning further improved estimation performance, with Adaptive Boosting Regression (ABR) achieving the highest accuracy ($r = 0.7706$; $R^2 = 0.4361$). These results show that the complementary use of MCDA and machine learning enhances the robustness of the proposed estimation framework.

The proposed framework has strong potential for practical application, particularly in regions where safe drinking water statistics are available only at the regency/municipality level or are updated infrequently. The resulting 1 km × 1 km spatial estimates can support evidence-based planning by identifying priority areas for the development or expansion of drinking water facilities, enabling spatially targeted policy interventions, and informing spatial planning based on local clean water service needs. Moreover, this approach offers a cost-effective and scalable alternative to conventional survey-based methods, complementing official statistics for monitoring progress toward safe drinking water access targets in data-scarce regions.

Despite these contributions, this study has limitations. Validation is constrained by the availability of official data only at the regency/municipality level, resulting in an indirect assessment of the 1 km × 1 km estimates. In addition, field-based ground truth validation, which would ideally support fine-scale accuracy assessment, could not be conducted.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

All the authors declare that there are no conflicts of interest.

Funding

This study received no external funding.

Underlying data

This study uses secondary data obtained from BPS-Statistics Indonesia, Ministry of Public Works, Ministry of Energy and Mineral Resources, Satellite Imagery, and OpenStreetMap

Credit Authorship

Andrew Maruli Tua Tampubolon: Methodology, Software, Data Collection, Data Curation, Writing-Original Draft Preparation, Visualization. **Bony Parulian Josaphat:** Conceptualization, Supervision, Reviewing and Editing. **Asriadi Sakka:** Reviewing, Validation. **Yohanes Wahyu Trio Pramono:** Reviewing, Final Draft.

References

- [1] V. Sheel, A. Kotwal, N. Dumka, V. Sharma, R. Kumar, and V. Tyagi, "Water as a social determinant of health: bringing policies into action," *J Glob Health Rep*, vol. 8, 2024, doi: 10.29392/001c.92160.
- [2] G. Mujtaba, M. U. H. Shah, A. Hai, M. Daud, and M. Hayat, "A holistic approach to embracing the United Nation's Sustainable Development Goal (SDG-6) towards water security in Pakistan," Jan. 01, 2024, *Elsevier Ltd*. doi: 10.1016/j.jwpe.2023.104691.
- [3] J. Abellán and J. A. Alonso, "Promoting global access to water and sanitation: A supply and demand perspective," *Water Resour Econ*, vol. 38, Apr. 2022, doi: 10.1016/j.wre.2022.100194.
- [4] A. Umami, H. Sukmana, E. A. Wikurendra, and E. Paulik, "A review on water management issues: potential and challenges in Indonesia," *Sustain Water Resour Manag*, vol. 8, no. 3, Jun. 2022, doi: 10.1007/s40899-022-00648-7.
- [5] M. L. Crouch, H. E. Jacobs, and V. L. Speight, "Defining domestic water consumption based on personal water use activities," *Aqua Water Infrastructure, Ecosystems and Society*, vol. 70, no. 7, pp. 1002–1011, Nov. 2021, doi: 10.2166/aqua.2021.056.
- [6] G. E. Adjovu, H. Stephen, D. James, and S. Ahmad, "Overview of the application of remote sensing in effective monitoring of water quality parameters," Apr. 01, 2023, *MDPI*. doi: 10.3390/rs15071938.
- [7] Kementerian Kesehatan, "Peraturan menteri kesehatan Republik Indonesia Nomor 492/Menkes/PER/IV/2010 tentang persyaratan kualitas air minum," Jakarta, 2010.
- [8] Kementerian Kesehatan, "Peraturan menteri kesehatan Republik Indonesia Nomor 2 Tahun 2023 tentang peraturan pelaksanaan peraturan pemerintah Nomor 66 Tahun 2014 tentang kesehatan lingkungan," Jakarta, 2023. [Online]. Available: www.peraturan.go.id
- [9] WHO and UNICEF, *Progress on drinking water, sanitation and hygiene: 2017 update and SDG baselines*. Geneva: World Health Organization and the United Nations Children's Fund, 2017.
- [10] WHO and UNICEF, *Safely managed drinking water*. Geneva: World Health Organization and the United Nations Children's Fund, 2017.
- [11] J. D. Tetteh *et al.*, "Spatial heterogeneity in drinking water sources in the Greater Accra Metropolitan Area (GAMA), Ghana," *Popul Environ*, vol. 44, no. 1–2, pp. 46–76, Sep. 2022, doi: 10.1007/s11111-022-00407-y.
- [12] L. Ho, A. Alonso, M. A. Eurie Forio, M. Vanclooster, and P. L. M. Goethals, "Water research in support of the Sustainable Development Goal 6: A case study in Belgium," Dec. 20, 2020, *Elsevier Ltd*. doi: 10.1016/j.jclepro.2020.124082.
- [13] UNICEF, "WASH Acts 2023," 2024.
- [14] Kementerian PUPR, Kementerian PPN/Bappenas, USAID IUWASH PLUS, and SECO, "Peta jalan peningkatan kapasitas sumber daya manusia BUMD air minum," Jakarta, 2021.
- [15] B. Ajisegiri *et al.*, "Geo-spatial modeling of access to water and sanitation in Nigeria," *Journal of Water Sanitation and Hygiene for Development*, vol. 9, no. 2, pp. 258–280, Jun. 2019, doi: 10.2166/washdev.2019.089.
- [16] P. D. T. dan T. Kementerian Desa, "Peraturan menteri desa, pembangunan daerah tertinggal, dan transmigrasi Nomor 8 Tahun 2022 tentang prioritas penggunaan dana desa Tahun 2023," Jakarta, 2022.
- [17] Hasbudin, W. O. Aswati, and Nirwana, "Evaluasi pengelolaan dana desa di Desa Bone Kancintala Kecamatan Bone Kabupaten Muna," *Jurnal Akuntansi dan Keuangan (JAK)*, vol. 6, no. 1, 2021, [Online]. Available: <http://ojs.uho.ac.id/index.php/jak-uho/issue/archive>
- [18] Kementerian PPN/Bappenas, Pokja PPAS, and USAID IUWASH PLUS, "Meta data target indikator air minum: : Kupas Tuntas SDGs 6.1 Air Minum," Jakarta, 2021.
- [19] BPS Kabupaten Tanjung Jabung Timur, "Statistik kesejahteraan rakyat Kabupaten Tanjung Jabung Timur 2022," Muara Sabak, 2022. [Online]. Available: www.freepik.com

- [20] S. D. Kurniawan *et al.*, *Big data: Mengenal big data & implementasinya di berbagai bidang*. Jambi: PT. Sonpedia Publishing Indonesia, 2024. [Online]. Available: www.buku.sonpedia.com
- [21] C. Abesser and M. Lewis, "A semi-quantitative technique for mapping potential aquifer productivity on the national scale: example of England and Wales (UK)," *Hydrogeol J*, vol. 23, no. 8, pp. 1677–1694, Dec. 2015, doi: 10.1007/s10040-015-1295-5.
- [22] M. C. de O. Silva, R. S. Vasconcelos, and J. A. Cirilo, "Risk mapping of water supply and sanitary sewage systems in a City in the Brazilian semi-arid region using GIS-MCDA," *Water (Switzerland)*, vol. 14, no. 20, Oct. 2022, doi: 10.3390/w14203251.
- [23] M. Guo, Q. Zhang, X. Liao, F. Y. Chen, and D. D. Zeng, "A hybrid machine learning framework for analyzing human decision-making through learning preferences," *Omega (Westport)*, vol. 101, Jun. 2021, doi: 10.1016/j.omega.2020.102263.
- [24] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.
- [25] M. J. de S. Cordão, I. A. A. Rufino, P. Barros Ramalho Alves, and M. N. M. Barros Filho, "Water shortage risk mapping: a GIS-MCDA approach for a medium-sized city in the Brazilian semi-arid region," *Urban Water J*, vol. 17, no. 7, pp. 642–655, Aug. 2020, doi: 10.1080/1573062X.2020.1804596.
- [26] M. Radulović *et al.*, "Assessment of groundwater potential zones using GIS and fuzzy AHP techniques—A case study of the titel municipality (Northern Serbia)," *ISPRS Int J Geoinf*, vol. 11, no. 4, Apr. 2022, doi: 10.3390/ijgi11040257.
- [27] A. S. Suryani, "Persepsi masyarakat dalam pemanfaatan air bersih (Studi kasus masyarakat pinggir sungai di Palembang)," *Aspirasi*, vol. 7, no. 1, pp. 33–48, 2016, [Online]. Available: www.koran-sindo.
- [28] B. Guo *et al.*, "Estimating socio-economic parameters via machine learning methods using luojial-01 nighttime light remotely sensed images at multiple scales of China in 2018," *IEEE Access*, vol. 9, pp. 34352–34365, 2021, doi: 10.1109/ACCESS.2021.3059865.
- [29] Y. Zheng, Q. Zhou, Y. He, C. Wang, X. Wang, and H. Wang, "An optimized approach for extracting urban land based on log-transformed dmsp-ols nighttime light, ndvi, and ndwi," *Remote Sens (Basel)*, vol. 13, no. 4, pp. 1–22, Feb. 2021, doi: 10.3390/rs13040766.
- [30] Y. Wang *et al.*, "The impact of carbon monoxide on years of life lost and modified effect by individual- and city-level characteristics: Evidence from a nationwide time-series study in China," *Ecotoxicol Environ Saf*, vol. 210, Mar. 2021, doi: 10.1016/j.ecoenv.2020.111884.
- [31] E. A. Adams and S. L. Smiley, "Urban-rural water access inequalities in Malawi: implications for monitoring the Sustainable Development Goals," *Nat Resour Forum*, vol. 42, no. 4, pp. 217–226, Nov. 2018, doi: 10.1111/1477-8947.12150.
- [32] T. Prasetyawan, A. Nastiti, and B. S. Muntalif, "'Bad' piped water and other perceptual drivers of bottled water consumption in Indonesia," *Wiley Interdisciplinary Reviews: Water*, vol. 4, no. 4, Jul. 2017, doi: 10.1002/WAT2.1219.
- [33] N. Dejito *et al.*, "Mapping access to water and sanitation in Colombia using publicly accessible satellite imagery, crowd-sourced geospatial information and random forests," in *35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, Nov. 2021.
- [34] K. Shi, Z. Chang, Z. Chen, J. Wu, and B. Yu, "Identifying and evaluating poverty using multisource remote sensing and point of interest (POI) data: A case study of Chongqing, China," *J Clean Prod*, vol. 255, May 2020, doi: 10.1016/j.jclepro.2020.120245.
- [35] H. E. Winzeler, P. R. Owens, Q. D. Read, Z. Libohova, A. Ashworth, and T. Sauer, "Topographic wetness index as a proxy for soil moisture in a Hillslope Catena: Flow algorithms and map generalization," *Land (Basel)*, vol. 11, no. 11, 2022, doi: 10.3390/land11112018.
- [36] P. Rao, P. Tassinari, and D. Torreggiani, "Exploring the land-use urban heat island nexus under climate change conditions using machine learning approach: A spatio-temporal analysis of remotely sensed data," *Heliyon*, vol. 9, no. 8, Aug. 2023, doi: 10.1016/j.heliyon.2023.e18423.
- [37] Y. Bin Cai, Z. J. Wu, Y. H. Chen, L. Wu, and W. Bin Pan, "Investigate the difference of cooling effect between water bodies and green spaces: The study of fuzhou, China," *Water (Switzerland)*, vol. 14, no. 9, May 2022, doi: 10.3390/w14091471.
- [38] S. Mishra and D. R. Mishra, "Normalized difference chlorophyll index: A novel model for remote estimation of chlorophyll-a concentration in turbid productive waters," *Remote Sens Environ*, vol. 117, pp. 394–406, Feb. 2012, doi: 10.1016/j.rse.2011.10.016.
- [39] I. Yeo and R. A. Johnson, "A new family of power transformations to improve normality or symmetry," *Biometrika*, vol. 87, no. 4, pp. 954–959, 2000, [Online]. Available: <http://biomet.oxfordjournals.org/>

- [40] J. Raymaekers and P. J. Rousseeuw, "Transforming variables to central normality," *Mach Learn*, vol. 113, no. 8, pp. 4953–4975, Aug. 2021, doi: 10.1007/s10994-021-05960-5.
- [41] Sugiyono, *Metode penelitian pendidikan pendekatan kuantitatif, kualitatif dan R&D*. Alfabeta, 2010.
- [42] Ch. M. Rao and K. Venkatasubbaiah, "Application of WSM, WPM and TOPSIS methods for the optimization of multiple responses," *International Journal of Hybrid Information Technology*, vol. 9, no. 10, pp. 59–72, Oct. 2016, doi: 10.14257/ijhit.2016.9.10.07.
- [43] N. U. Moroff, E. Kurt, and J. Kamphues, "Machine Learning and Statistics: A study for assessing innovative demand forecasting models," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 40–49. doi: 10.1016/j.procs.2021.01.127.
- [44] R. D. Stephens and S. H. Cadle, "Remote sensing measurements of carbon monoxide emissions from on-road vehicles," *J Air Waste Manage Assoc*, vol. 41, no. 1, pp. 39–46, 1991, doi: 10.1080/10473289.1991.10466823.
- [45] A. Botchkarev, "Evaluating performance of regression machine learning models using multiple error metrics in Azure Machine Learning Studio," *SSRN Electronic Journal*, 2018, [Online]. Available: <https://ssrn.com/abstract=3177507>



TEMPLATE

Click Here, Type the Title of Your Paper, Capitalize First Letter of Each Word (Times New Romans (TNR), Size 17 pt, exactly spacing at 20 pt, 12 pt spacing for next heading, justify))

First Author Name^{1*}, Second Author Name², Third Author Name³ (TNR font, size 13 pt, exactly spacing at at 15 pt and 8 pt spacing for the next heading.)

¹First Affiliation, City, Country, ^{2,3}Author Affiliation, City, Country ²Second Affiliation, City, Country, ^{2,3}Author Affiliation, City, Country

*Corresponding Author: E-mail address: author@institute.xxx

(TNR font, size 10 pt, with single spacing and 0 pt spacing for the next heading. And for the corresponding author use 10 pt Times New Roman font with single spacing and 8 pt spacing for the next heading.)

ARTICLE INFO

Article history: (TNR, 10pt)

Received dd month, yyyy

Revised dd month, yyyy

Accepted dd month, yyyy

Published dd month, yyyy

Keywords: (TNR, 10 pt)

Type your keywords here, separated by semicolon (;) Capitalize first letter of each word, times new roman, use 10 pt and write alphabetically in 5-10 words

Abstract (Times New Roman, size font 12)

Introduction/Main Objectives: Describe the topic your paper examines. Provide a background to your paper and why is this topic interesting. Avoid unnecessary content. **Background Problems:** State the problem or statistical applied/statistic computing phenomena studied in this paper and specify the research question(s) in one sentence. **Novelty:** Summarize the novelty of this paper. Briefly explain why no one else has adequately researched the question yet. **Research Methods:** Provide an outline of the research method(s) and data used in this paper. Explain how did you go about doing this research. Again, avoid unnecessary content and do not make any speculation(s). **Finding/Results:** List the empirical finding(s) and write a discussion in one or two sentences. Abstract written in English, with a length of 150 - 200 words. Use 10 pt Times New Roman font with justified alignment, single spacing, and 1 pt spacing for the next heading.

1. Main Text (bold, TNR, 14 pt, spacing before- after 12 pt, line spacing 12 pt)

These instructions give you guidelines for preparing papers for Jurnal Aplikasi Statistika & Komputasi Statistik which is published by Politeknik Statistika STIS, effective from the June 2024 edition. Starting from June 2024 Volume 16 No. 1, please use the template available at the following link <https://s.stis.ac.id/TemplateJurnalASKS>. The paragraphs continue from here and are only separated by headings, subheadings, images and formulae. The section headings are arranged by numbers, bold and 14 pt. Here follow further instructions for authors.

The manuscript was created using Microsoft Office Word only and should be formatted for direct printing. As indicated in the template, manuscript should be prepared in single column format that suitable for direct printing onto paper with A4 paper size (21 x 29.7 cm). All parts of the manuscript are typed in Times New Roman font, size 11, line spacing exactly at 12 pt, with 0.2 line spacing for the next heading and margins of 3 cm of left and 2 cm for top, bottom, and right, the length of header from the top is 1.5 cm and the length of footer from the bottom is 1 cm. For the main text, use justify alignment and special indent for the first line in 0.76 cm. For the purpose of editing the manuscript, all parts of the manuscript (including tables, figures and mathematical equations) are made in a format that can be edited by the editor [1, 2].

The writing style for the Jurnal Aplikasi Statistika & Komputasi Statistik is written in English with a narrative style. Tracing is kept simple and as far as possible avoiding multilevel chronology.

1.1. Structure

Please make sure that you use as much as possible normal fonts in your documents. Special fonts, such as fonts used in the Far East (Japanese, Chinese, Korean, etc.) may cause problems during processing. To avoid unnecessary errors, you are strongly advised to use the ‘spellchecker’ function of MS Word. Follow this order when typing manuscripts: **Title, Authors, Affiliations, Abstract, Keywords, Main Text (Introduction, Material and Methods, Result and Discussion, Conclusion, including figures and tables), Acknowledgements, and References.**

Introduction coverage What is the purpose of the study? Why are you conducting the study? The main section of the article should start with an introductory section, which provides more details about the paper’s purpose, motivation, research methods and findings. The introduction should be relatively nontechnical, yet clear enough for an informed reader to understand the manuscript’s contribution. The Introduction is not an extended version of the abstract; never use the same sentences in both sections

The “introduction” in the manuscript is important to demonstrate the motives of the research. It analyzes the empirical, theoretical and methodological issues in order to contribute to the extant literature. This introduction will be linked with the following parts, most noticeably the literature review. Explaining the problem’s formulation should cover the following points: (1) Problem recognition and its significance; (2) clear identification of the problem and the appropriate research questions; (3) coverage of problem’s complexity; and (4) well-defined objectives.

The second part of the manuscript, “Method, Data, and Analysis” is designed to describe the nature of the data. The method should be well elaborated and enhance the model, the approach to the analysis and the step taken. Equations should be numbered as we illustrate.

This section typically has the following sub-sections: Sampling (a description of the target population, the research context, and units of analysis; the sample; and respondents’ profiles); data collection; and measures (or alternatively, measurements).

The research methodology should cover the following points: Concise explanation of the research’s methodology is prevalent; reasons for choosing the particular methods are well described; the research’s design is accurate; the sample’s design is appropriate; the data collection processes are properly conducted; the data analysis methods are relevant and state-of-the-art.

The second part of manuscript, “Result and Discussion” The author needs to report the results in sufficient detail so that the reader can see which statistical analysis was conducted and why, and later to justify their conclusions.

The “Discussion and Analysis” part, highlights the rationale behind the result answering the question “why the result is so?” It shows the theories and the evidence from the results. The part does not just explain the figures but also deals with this deep analysis to cope with the gap that it is trying to solve.

The “Conclusion and Suggestion”, in this section, the author presents brief conclusions from the results of the research with suggestions for advanced researchers or general readers. A conclusion may cover the main points of the paper, but do not replicate the abstract in the conclusion. Authors should explain the empirical and theoretical benefits, and the existence of any new findings. The author may present any major flaws and limitations of the study, which could reduce the validity of the writing, thus raising questions from the readers (whether, or in what way), the limits in the study may have affected the results and conclusions. Limitations require a critical judgment and interpretation of the impact of their research. The author should provide the answer to the question: Is this a problem caused by an error, or in the method selected, or the validity, or something else?

The manuscript including the graphic contents and tables should be around 15-20 pages. The manuscript is written in English. The Standard English grammar must be observed. The title of the article should be brief and informative and it is recommended not to exceed 12 words. When writing numbers, use a period to separate decimal points and a comma to separate thousands.

The use of abbreviation is permitted, but the abbreviation must be written in full and complete when it is mentioned for the first time and it should be written between parentheses. Terms/foreign words or regional words should be written in italics. Notations should be brief and clear and written according to the standardized writing style. Symbols/signs should be clear and distinguishable, such as the use of number 1 and letter l (also number 0 and letter O).

Bulleted lists may be included and should look like this:

- First point
- Second point
- And so on

Ensure that you return to the ‘body-text’ style, the style that you will mainly be using for large blocks of text, when you have completed your bulleted list.

Please do not alter the formatting and style layouts which have been set up in this template document.

1.2. Tables

All tables should be numbered with Arabic numerals. Every table should have a caption. Headings should be placed above tables with left justified alignment. Only horizontal lines should be used within a table, to distinguish the column headings from the body of the table, and immediately above and below the table. Tables must be embedded into the text and not supplied separately. Below is an example which the authors may find useful.

Table 1. Rice coefficient for various climatic conditions

Humidity	Wind Speed		
	Low	Medium	High
Dry	1.10	1.15	1.20
Medium	1.05	1.10	1.15
High	1.00	1.05	1.10

1.3. Construction of references

References must be listed at the end of the paper. Do not begin them on a new page unless this is absolutely necessary. Authors should ensure that every reference in the text appears in the list of references and vice versa. Indicate references by [1] or [2] or [3] in the text.

Some examples of how your references should be listed are given at the end of this template in the ‘References’ section, which will allow you to assemble your reference list according to the correct format and font size. The paper must include a reference list containing only the quoted work and using the Mendeley tool. Each entry should contain all the data needed for unambiguous identification. With the author-date system, use the following format recommended by IEEE Citation Style. The first line of each citation is left adjusted. Every subsequent line is indented 5-7 spaces. The references are arranged in alphabetical order, written in 11pt Times New Roman font with 0 pt spacing for the next heading.

The references shall contain at least 20 (twenty) references. For whole references, at least 16 references or 80% of them must be refer to primary sources (scientific journals, conference proceedings, research reference books) which are published within 5 (five) year. The IEEE citation guide can be access here:

<https://iee-dataport.org/sites/default/files/analysis/27/IEEE%20Citation%20Guidelines.pdf>

1.4. Section headings

Section headings should be left justified, bold, with the first letter capitalized and numbered consecutively, starting with the Introduction. Section headings use 14 pt Times New Roman and exactly spacing at 12 pt with before and after spacing in 12 pt, left alignment and special hanging indentation at 0.63 cm. Sub-section headings should be in capital and lower-case italic letters, numbered 1.1, 1.2, etc, exactly spacing at 12 pt with before and after spacing in 12 pt, left alignment with 0.12 cm left indentation and special hanging indentation at 0.63 cm, with second and subsequent lines indented. All headings should have a minimum of three text lines after them before a page or column break. Ensure the text area is not blank except for the last page. Both section heading and sub-section headings are in dark blue color with the code #034F84 (R: 3 G: 79 B: 132).

1.5. General guidelines for the preparation of your text

Avoid hyphenation at the end of a line. Symbols denoting vectors and matrices should be indicated in bold type. Scalar variable names should normally be expressed using italics. Weights and measures should be expressed in SI units. All non-standard abbreviations or symbols must be defined when first mentioned, or a glossary provided.

1.6. Footnotes

Footnotes should be avoided if possible.

2. Illustrations

All figures should be numbered with Arabic numerals (1,2,3,...). Every figure should have a caption. All photographs, schemas, graphs and diagrams are to be referred to as figures. Line drawings should be good quality scans or true electronic output. Low-quality scans are not acceptable. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be properly coded. Preferred format of figures are PNG, JPEG, GIF etc. Lettering and symbols should be clearly defined either in the caption or in a legend provided as part of the figure. Figures should be placed at the top or bottom of a page wherever possible, as close as possible to the first reference to them in the paper. Please ensure that all the figures are of 300 DPI resolutions as this will facilitate good output. Figures should be embedded and not supplied separately.

The figure number and caption should be typed below the illustration in 11 pt and left justified [**Note:** one-line captions of length less than column width (or full typesetting width or oblong) centered].

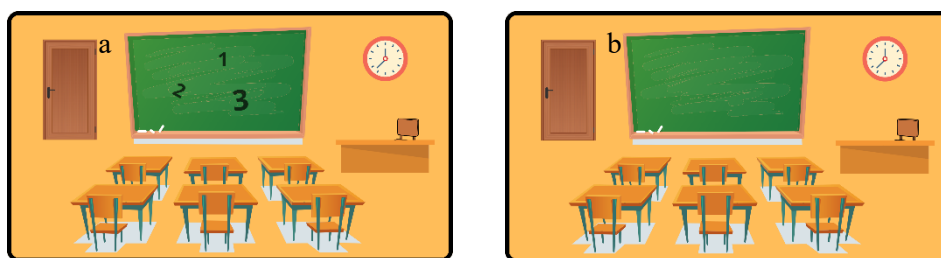


Figure 1. (a) first picture; (b) second picture

3. Equations

Equations and formulae should be typed in MathType or Microsoft Equation, and numbered consecutively with Arabic numerals in parentheses on the right hand side of the page (if referred to explicitly in the text). They should also be separated from the surrounding text by one space.

$$\rho = \frac{\vec{E}}{J_c(T = \text{const.}) \cdot \left(P \cdot \left(\frac{\vec{E}}{E_c} \right)^m + (1 - P) \right)} \quad (1)$$

Ethics approval

The Ethical approval statement should be provided including the consent. If not appropriate, authors should state: “Not required.”

Acknowledgments

This section contains a form of thanks to individuals or institutions who have provided assistance in carrying out research, preparing the article, providing language help, writing assistance or proof reading the article and others.

Competing interests

A competing interest statement should be provided, even if the authors have no competing interests to declare. If no conflict exists, authors should state: “All the authors declare that there are no conflicts of interest.”

Funding

List funding sources in this standard way to facilitate compliance to funder's requirements. It is not necessary to include detailed descriptions on the program or type of grants and awards. When funding is from a block grant or other resources available to a university, college, or other research institution, submit the name of the institute or organization that provided the funding. If no funding has been provided for the research, please include the following sentence: “This study received no external funding.”

Underlying data

This can be written as: “Derived data supporting the findings of this study are available from the corresponding author on request.”

Credit Authorship

Please outline the contributions of each co-author, using the following categories: conceptualization, methodology, software, validation, formal analysis, investigation, data collection, data curation, writing—original draft, writing—review & editing, visualization, supervision, and funding acquisition.

References

- [1] W.K. Chen, *Linear Networks and Systems*. Belmont, CA: Wadsworth Press, 2003.
- [2] R. Hayes, G. Pisano, and S. Wheelwright, *Operations, Strategy, and Technical Knowledge*. Hoboken, NJ: Wiley, 2007.
- [3] K. A. Nelson, R. J. Davis, D. R. Lutz, and W. Smith, “Optical generation of tunable ultrasonic waves,” *J Appl Phys*, vol. 53, no. 2, pp. 1144–1149, Feb. 2002.