JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 16, NO 2, DECEMBER 2024 P-ISSN 2086 – 4132 E-ISSN 2615

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika & Komputasi Statistik

"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.

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 Aplikasi Statistika & Komputasi Statistik Journal are cited by referring to the source material.

Editorial Foreword

.

“Jurnal Aplikasi Statistika & Komputasi Statistik (JASKS)” starting Volume 16 Number 1 June 2024 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.

Hopefully, the articles in this journal can increase readers' knowledge about the use of statistical methods and computational statistics on various types of data. The editorial eagerly awaits further scientific articles from fellow statisticians so that the resulting publication becomes one of the means to provide statistical socialization for the community.

Jakarta, December 2024

Editor in Chief,

Rani Nooraeni

JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 16, NO 2, DECEMBER 2024

Contents

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.652

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Patterns, Determinants, and Elasticity of Household Food Consumption in Indonesia (Period 2021-2022)

Wifa Darma Aulia[1*], Rita Yuliana[2]

[1]*BPS-Statistics Dharmasraya Regency, West Sumatera, Indonesia*
[2]*Politeknik Statistika STIS, Jakarta, Indonesia*
**Corresponding Author: E-mail address:* [1]wifa.darma@bps.go.id,[2]rita@stis.ac.id

ARTICLE INFO

Article history:
Received 27 October, 2023
Revised 2 September, 2024
Accepted 2 September, 2024
Published 31 December, 2024

Keywords:
Elasticity; food consumption patterns; LA-AIDS; strategic food commodities; volatile food.

Abstract

Introduction/Main Objectives: The increase in strategic food commodity prices contributed significantly to the inflation rate. In March 2022, the inflation rate for the food, beverage and tobacco category reached 3.59% (y-o-y). This increase in the price of strategic food commodities makes it difficult for households. The continued increase in food prices resulted in a decrease in household purchasing power, thereby reducing the level of household welfare in Indonesia **Background Problems:** Based on these problems, research is needed on the impact of rising prices of strategic food commodities on changes in household food consumption patterns in Indonesia. Therefore, what is the description of household food consumption patterns and the factors that influence them; and what is the elasticity of household food demand in Indonesia in the period March 2021 and March 2022? **Novelty:** The novelty of this research is to calculate the elasticity of food demand in Indonesia for the period March 2021 and March 2022 by fulfilling the assumptions of demand theory so that the value can be trusted, while several studies do not apply the assumptions of demand theory. **Research Methods:** This study used the Linear Approximated-Almost Ideal Demand System (LA-AIDS) model with the Seemingly Unrelated Regression (SUR) method. **Finding/Results:** The research results show that in March 2022 there was an increase in the price of strategic food commodities and a change in household food consumption patterns in Indonesia. The own price elasticity value shows a negative number. The cross price elasticity of some food groups is negative and some is positive. The elasticity of total expenditure shows that all food groups are normal goods.

1. Introduction

The problem of food security is still a challenge for all countries. The unstable world economic condition is one of the causes of this problem. In 2020, conditions of Food security are getting worse due to problems in food distribution. This happened because of the social restrictions implemented in various countries during the Corona virus Disease 2019 (COVID-19) pandemic which brought economic activity to a halt [1]. These social restrictions cause food supply chains that are interconnected with each other to be disrupted, starting from the production, distribution, and consumption of food for the population both globally and domestically. As a result, there was an increase in prices for several food commodities, especially strategic food commodities.

The increase in prices of strategic food commodities makes a large contribution to the inflation rate (volatile foods) [2]. In March 2022, the Consumer Price Index (CPI) for the food, beverage, and tobacco group reached 113.32 with an inflation rate of 3.59 percent compared to March 2021 [3]. In the last five years, of the 15 strategic food commodities that contributed to increasing inflation, there was a dominant food commodity, namely cooking oil [4]. The rural consumer price for cooking oil in March 2021 was IDR 15,095.00/Kg and increased in March 2022 to IDR 19,640.00/Kg [5]. The average increase in consumer prices for cooking oil in 2022 will reach 30.81% compared to the previous year. The high demand and decreasing supply of cooking oil have resulted in a shortage and increase in cooking oil in Indonesia [6], even though cooking oil is a food ingredient that is often used by households for cooking. Apart from cooking oil, average rural consumer prices for several other strategic food commodities have also increased. This can be seen in the following table:

Table 1. Average rural prices for several strategic food commodities in 2021 and 2022 (Rupiah)

Food Commodities	Average rural prices (Rupiah)		% Change
	2021	2022	
Rice	11,348	11,554	1.815
Beef	112,870	117,218	3.852
Purebred chicken meat	37,809	40,008	5.816
Chicken eggs	27,196	29,051	6.821
Cooking oil	15,701	20,538	30.807
Red chili pepper	42,194	51,104	21.117
Red onion	30,641	36,345	18.616
Garlic	30,271	30,713	1.460

Source: BPS (2021 and 2022b)

This increase in the price of strategic food commodities makes it difficult for households, especially for households with lower middle income. The short-term impact of increasing food prices is that it can reduce household purchasing power. When food prices increase, households will respond by reducing demand for that food or replacing it with cheaper food. This statement is in accordance with demand theory [8]. This phenomenon indicates that there is a decline in purchasing power and this will subsequently change household food consumption patterns. The long-term impact of increasing food prices is to reduce the level of household welfare [9]. This can happen because households cannot meet their basic needs.

Changes in consumption patterns will have an impact on changes in food demand. The magnitude of changes in food demand can be seen from the elasticity value of household food demand. The elasticity value of household food demand can be estimated using the Almost Ideal Demand System (AIDS). Estimation of food demand models using the AIDS model has been carried out by several researchers, namely [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], and [21]. Some of this research only examines the elasticity of food demand in one province and only focuses on one food commodity group, with the exception of research by [21]. In [21], Yuliana conducted research on food demand and changes in household welfare levels covering all foodstuffs in Indonesia in March 2016.

Based on these problems, research is needed on the impact of rising prices of strategic food commodities on changes in household food consumption patterns in Indonesia. Therefore, the aim of this research is to analyze the description of household food consumption patterns and the factors that influence them; and analyze the elasticity of household food demand in Indonesia in the period March 2021 and March 2022. This is because, in the period March 2021 to March 2022 there were several phenomena that resulted in an increase in food prices as previously explained. It is hoped that this research can be used as a reference for the government to create policies that have an impact on household food consumption patterns so that they can improve household welfare.

2. Material and Methods

2.1. Type of Research

This study uses quantitative research methods. This method is carried out by processing household sample data using statistical analysis and hypothesis testing has been carried out previously.

2.2. Location and Time Research

This research related to Patterns, Determinants, and Elasticity of Household Food Consumption in Indonesia was conducted during the lecture period at STIS Statistics Polytechnic as a requirement for the final thesis assignment. This research was conducted for approximately 9 months.

2.3. Data Collection Sources and Strategies

This research uses primary data from the National Socioeconomic Survey (SUSENAS) sourced from the Central Statistics Agency (BPS). The data collected is SUSENAS data for the period March 2021 and March 2022 in cross-sectional form by sampling household units in Indonesia. SUSENAS data for March 2021 and March 2022 covers 345,000 sample households with a response rate for each period of 99.75 percent or 344,148 households and 99.92 percent or 343,879 households.

All food commodities contained in SUSENAS data are grouped into 12 food groups. This grouping was formed based on the grouping of CPI calculations, policy targets (rice), and nutritional content [21]. Apart from that, the grouping of food commodities refers to research by [21] so that comparisons can be made to the results of this research. The following 12 food groups have been formed:

- Rice
- Non-rice (grains other than rice) and tubers
- Fresh fish
- Meat, eggs, and milk
- Vegetables
- Nuts
- Fruits
- Oil and coconut
- Drink ingredients
- Spices
- Other foods: other foods and preserved fish
- Food, beverages, and tobacco cigarettes

2.4. Analysis Method

Estimation of the demand function is carried out using the Linear Approximated – Almost Ideal Demand System (LA-AIDS) model. This model is a development of the Engel curve and the Marshallian equation which is derived from maximizing utility. The LA-AIDS equation in this study can be formulated as follows [9]:

$$w_i = \alpha_{0i} + \sum_{j=1}^{12} \gamma_{ij} lnp_j + \beta_i \, ln\_expfood\_defl + \alpha_{1i} ln\_nhm + \alpha_{2i} ln\_fac + \alpha_{3i} ln\_agehead + \alpha_{4i} schoolhead + \alpha_{5i} genderhead + \alpha_{6i} maritstatushead + \alpha_{7i} typearea + \alpha_{8i} poorstatus + \alpha_{9i} sourcehh + \alpha_{10i} healthhead + \alpha_{11i} internet + \alpha_{12i} asset + \alpha_{13i} businesshead + \alpha_{14i} accessfood + \alpha_{15i} aidfood + \alpha_{16i} IMR_i + \varepsilon_i \quad (1)$$

Information:

i, j	= 1, 2, ..., 12 (i/jth food group)
w_i	= proportion of expenditure on food group i
lnp_j	= natural logarithm of the estimated price of the jth food group
$ln\_expfood\_defl_h$	= natural logarithm of a household's total monthly food expenditure which has been deflated by the stone price index (P), namely$\ln P = \sum w_i \, lnp_i$
ln_nhm	= natural logarithm of the number of household members (people)
ln_fac	= natural logarithm of floor area per capita (m^2)
ln_agehead	= natural logarithm of the age of head of household (years)
schoolhead	= length of school of the age of head of household (years)
genderhead	= dummy head of household gender (1 = male, 0 = female)
maritstatushead	= dummy household head's marital status (1 = Married, 0 = Not married/divorced)
typearea	= dummy type of area where the household lives (1 = urban, 0 = rural)

poorstatus = dummy household poor status (1 = Poor, 0 = Not poor)
sourcehh = dummy largest source of household financing (1 = working household member, 0 = others)
healthhead = dummy physical health status of head of household (1 = Having difficulty taking care of themselves, 0 = Having no difficulty taking care of themselves)
internet = dummy household internet use (1 = Yes, 0 = No)
asset = dummy household asset ownership (1 = Has at least 1 asset, 0 = Has no assets)
businesshead = dummy head of household business field (1 = Agricultural sector, 0 = Non-agricultural sector)
accessfood = dummy difficulty accessing healthy and nutritious household food (1 = Yes, 0 = No/don't know/refuse to answer)
aidfood = dummy food aid for February 2021 and February 2022 (1 = Yes, 0 = No)
IMR_i = Inverse Mill's Ratio, correction variable for selectivity bias for the i[th] food group
$\alpha_0, \dots, \alpha_{16}, \gamma_{ij}$ = parameter
ε_i = error

The price variable for each food group is proxied by the amount of food expenditure divided by the total quantity also called the unit value of that food group. The following is the formula for calculating the unit value of the i-th food group (uv_i) [22] :

$$uv_i = \sum_{j=1}^{J_i} \left[uv_j \frac{e_j}{\sum_{j=1}^{J_i} e_j} \right] \quad (2)$$

with is the unit value of the j[th] commodity paid by the household which is formulated as follows :

$$uv_j = \frac{e_j}{q_j} \quad (3)$$

e_j is the expenditure value for the j[th] commodity and q_j is the amount of the j[th] commodity consumed by the household.

The use of this unit value can cause several problems [22]. First, it produces biased estimates because the unit value is influenced by the quality and quantity purchased. This problem can be overcome by using instrument variables. Second, there is a contemporaneous correlation problem or a correlation between errors in different equations. This can be overcome by using the Seemingly Unrelated Regression (SUR) estimation method. The final problem is selectivity bias. This bias arises because within a week there are households that do not consume certain food groups. If this value is ignored it will cause bias in the estimation results. The way to overcome this problem is to group all food commodities and add the Inverse Mill's Ratio (IMR) variable to the LA-AIDS model [22].

This model has several advantages compared to other demand functions, namely that it makes the estimation process easier because it involves many parameters without having to use non-linear methods. Apart from that, this model produces a good estimator because it is able to overcome basic assumption problems in Ordinary Least Squares (OLS) such as heteroscedasticity problems [11]. To overcome the problem of the basic assumptions of OLS, the LA-AIDS model is estimated using the SUR method with the Three-Stage Least Squares (3SLS) procedure. The next advantage is that there are several restrictions placed on the LA-AIDS model, resulting in estimates that are in accordance with demand theory. The following restrictions were applied in this study:

- Additivity : $\sum_{i=1}^{n} \alpha_i = 1; \sum_{i=1}^{n} \gamma_{ij} = 0; \sum_{i=1}^{n} \beta_i = 0$
- Homogeneity : $\sum_j \gamma_{ij} = 0$
- Slutsky Symmetry : $\gamma_{ij} = \gamma_{ji}$

The application of restrictions in this model also makes elasticity calculations simple and consistent with demand theory. The following is the formula for calculating demand elasticity [23]:

1. Own price elasticity (ε_{ii})

$$\varepsilon_{ii} = -(1 + \beta_i) + \frac{\gamma_{ii}}{w_i} \quad (4)$$

2. Cross price elasticity (ε_{ij})

$$\varepsilon_{ij} = \frac{\gamma_{ij}}{w_i} - \beta_i \left(\frac{w_j}{w_i}\right) \quad (5)$$

3. Income elasticity (η_i)

$$\eta_i = 1 + \frac{\beta_i}{w_i} \quad (6)$$

To get the elasticity of demand for food groups towards total household expenditure (food and non-food) this can be done by multiplying the equation by the elasticity of total food expenditure towards total household expenditure η_i [24]. The following is a linear equation to obtain the elasticity of total food expenditure on total household expenditure:

$$ln\, y_{food} = a + b\, lny_{expendituretotal} + \varepsilon \quad (7)$$

$$e_{tf} = \frac{\partial lny_{food}}{\partial lny_{expendituretotal}} = b \quad (8)$$

Information :

e_{tf} = elasticity of total food expenditure on total household expenditure
y_{food} = total monthly household food expenditure
$y_{expendituretotal}$ = total monthly household expenditure

Thus, we obtain the elasticity of demand for food groups towards total household expenditure (φ_i) which can be formulated as follows:

$$\varphi_i = \eta_i \,.\, e_{tf} \quad (9)$$

3. Results and Discussion

3.1. Analysis Overview of Household Food Consumption Patterns in Indonesia and the Factors that Influence Them in the Period March 2021 and March 2022

The amount of household food demand is related to price changes. When food prices increase, household demand tends to decrease. Conversely, if food prices decrease, the quantity demanded will increase.

Table 2. Average food prices (unit value) and changes according to food groups in Indonesia in the period March 2021 and March 2022 (Rupiah)

Food Group	Average Food Prices (Rupiah) (Unit Value)		
	March 2021	March 2022	% Change
(1)	(2)	(3)	(4)
Rice	10,359	10,791	4.165
Non rice and tubers	8,492	9,401	10.695
Fresh fish	29,029	30,874	6.356
Meat, eggs, and milk	27,200	28,496	4.763
Vegetables	19,264	16,343	-15.163
Nuts	11,687	12,598	7.794
Fruits	12,562	13,603	8.288
Oil and coconut	12,824	19,491	51.990
Drink ingredients	2,250	2,423	7.693
Spices	849	861	1.407
Other foods	3,993	4,272	6.972

Food, beverages, and tobacco cigarettes	4,773	5,170	8.317

Source: Susenas Primary Data March 2021 and March 2022 (processed)

Table 2 is the result of price calculations for each food group using the approach unit value. In Table 2 column 4, it can be seen that the prices of almost all food groups have increased except for the vegetables group. In March 2022, several commodities in the vegetable group experienced price declines, such as cabbage, green beans, long beans, carrots, red chilies, and cayenne peppers [5]. Then, the oil and coconut group experienced the highest increase when compared to other food groups. This happened because, at the beginning of 2022, Indonesia experienced a shortage of cooking oil which caused the price of cooking oil to increase [25]. Apart from that, it can also be seen that strategic food commodity groups experienced price increases, such as rice; meat, eggs, and milk. Therefore, Table 2 proves that in March 2022, overall food prices will increase compared to March 2021.

Table 3. Average total monthly food expenditure, the proportion of household food expenditure, and changes according to food groups in Indonesia in the period March 2021 and March 2022

Food Group	Average total monthly food expenditure (Rupiah)			Proportion of food expenditure		
	March 2021	March 2022	% Change	March 2021	March 2022	% Change
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Rice	265,118	267,540	0.914	0.119	0.113	-5.162
Non rice and tubers	60,069	63,566	5.822	0.027	0.027	-0.549
Fresh fish	187,795	203,292	8.252	0.084	0.086	1.735
Meat, eggs, and milk	227,001	239,458	5.488	0.102	0.101	-0.863
Vegetables	203,616	212,732	4.477	0.091	0.090	-1.813
Nuts	40,850	42,939	5.114	0.018	0.018	-1.214
Fruits	88,721	109,747	23.699	0.040	0.046	16.252
Oil and coconut	61,618	83,016	34.727	0.028	0.035	26.616
Drink ingredients	74,455	76,069	2.168	0.033	0.032	-3.983
Spices	49,382	54,630	10.627	0.022	0.023	3.967
Other foods	73,098	78,008	6.717	0.033	0.033	0.292
Food, beverages, and tobacco cigarettes	894,923	938,289	4.846	0.402	0.396	-1.466
Total	2,226,646	2,369,285	6.406	1.000	1.000	

Source: Susenas Primary Data March 2021 and March 2022 (processed)

Rising food prices will further influence household food consumption patterns. This can be seen in Table 3 which shows that the average total monthly food expenditure of households in Indonesia in March 2022 has increased compared to March 2021. The food group that has the highest average total monthly food expenditure is the food group, finished drinks, cigarettes, and tobacco. both in March 2021 and March 2022. Then followed by the rice group; meat, eggs, and milk groups. In column 4 it can be seen that in March 2022, the average total monthly household food expenditure for all food groups has increased compared to March 2021. However, column 7 shows that in March 2022 households decreased the proportion of their food expenditure compared to March 2021 for some food groups, namely the rice group; non rice and tubers; meat, eggs, and milk; vegetables; nuts; beverage ingredients; and food and drink become tobacco cigarettes. This shows that the increase in food prices makes households reduce the proportion of food expenditure for the six food groups and divert it to other food groups. The decrease in the proportion of food expenditure due to the increase in food prices shows that there has been a change in household food consumption patterns due to the increase in food prices in March 2022.

Apart from the price of goods, the diversity of household food expenditure is also influenced by several other factors. In this study, these factors are the socio-demographic characteristics of each household which can be seen in Table 4 as follows:

Table 4. Socio-demographic characteristics of households in Indonesia for the period March 2021 and March 2022

Sociodemographic Characteristics of Households	Mar-21	Mar-22
(1)	(2)	(3)
Average number of household members (people)	3.757	3.645
Average age of head of household (years)	48.157	48.744
Average length of school for head of household (years)	8.216	8.201
Average floor area per capita (m^2)	24.233	25.948
Area type: Urban (%)	42.090	41.848
Head of household gender: Male (%)	85.133	84.874
Household Marital Status: Married (%)	80.671	80.352
Physical health of head of household: having difficulty taking care of themselves (%)	3.635	1.596
Head of household business field: Agricultural sector (%)	41.435	41.682
Poor status: poor (%)	10.006	8.697
Main source of income: working household members (%)	91.992	92.356
Internet usage: Yes (%)	45.317	51.541
Asset ownership: own at least one asset (%)	95.377	95.927
Difficulty accessing healthy and nutritious food: Yes (%)	10.374	10.019
Receiving food aid: Yes (%)	15.887	14.394

Source: Susenas Primary Data March 2021 and March 2022 (processed)

Table 4 shows that all household characteristics in Indonesia in the two periods are almost the same except for several characteristics that have changed. In March 2022, households with the head of the household having difficulty taking care of themselves decreased by 2.04 percentage points when compared to March 2022. Likewise, poor households in March 2022 also experienced a decrease of 1.31 percentage points. Then, due to technological developments, internet use in households experienced a large increase, namely by 6.224 percentage points. Using the internet can make it easier for households to consume food by shopping online. Households receiving food assistance experienced a decrease of 1.493 percentage points. In conditions of increasing food prices, the government's role is needed to make it easier for households to meet their food needs. However, recipients of food aid actually experienced a decline.

3.2. Analysis of the Elasticity of Household Food Demand in Indonesia in the Period March 2021 and March 2022

The estimation results of the LA-AIDS model (equation 1) show that simultaneously obtained a p-value for all small food groups of 0.01, which means that overall the independent variables have a significant influence on the dependent variable with a significance level of 1%. The coefficient of determination (R-Square) value was obtained in the range of 6%-49% for March 2021 and 6%-44% for March 2022. This means that the variation in the proportion of food group expenditure that can be explained by the independent variable is 6% to 49% for March 2021 and 6% to 44% for March 2022, while the rest is influenced by factors outside the model. If a partial test is carried out, it is found that there are several variables that have no effect on the expenditure proportion variable for certain food groups.

The coefficient value of the LA-AIDS model cannot yet describe household sensitivity or household response to price changes. Therefore, to be able to see this, the LA-AIDS estimation results are used to calculate the elasticity value of food demand according to equations 4, 5, 6, and 9.

Consumer behavior is a theory that explains how consumers allocate their resources to consume various kinds of goods and services in order to maximize consumer satisfaction. Consumer decisions to make purchases are influenced by income and price [8]. The relationship between the quantity of goods consumed at a certain price level and time can be shown from the demand function. The law of demand in economics explains that when the price of a good increases, consumers will reduce the quantity of that good. Consumer responses to price changes can be analyzed using the elasticity of demand value.

Demand elasticity is divided into uncompensated demand elasticity (Marshallian) and compensated demand elasticity (Hicksian). While elasticity is based on causal factors, elasticity is divided into three, namely:

a. Price elasticity

Price elasticity shows the percentage change in demand for goods due to a change in the price of the good itself by 1 percent [8]. The price elasticity value itself is usually negative. When the absolute value of elasticity is greater than 1, then the goods are goods that are elastic or responsive to changes in the price of the goods themselves. On the other hand, if the absolute value of price elasticity is less than 1, then it is an inelastic good. If the price elasticity value is equal to 1, then it is a unitary item, or the demand for the item is not influenced by changes in the price of the item itself.

b. Income elasticity

Income elasticity shows the percentage change in quantity demanded due to an increase in income of 1 percent [8]. If the income elasticity is greater than 1, then it is a luxury or superior product because it is more responsive to changes in income, for examples, luxury cars, jewelry, and so on. Meanwhile, if the income elasticity value is between 0 and 1, then it is a normal item which is a basic necessity item.

c. Cross price elasticity

Cross-price elasticity is the percentage change in demand for a good due to a 1 percent increase in the price of another good [20]. If the cross price elasticity value has a positive number then the relationship between the goods is substitution. If the cross-price elasticity value is negative, then the two goods are complementary.

Table 5. Own price elasticity and total household expenditure elasticity according to food groups for the period March 2021 and March 2022

Food Group	Own Price Elasticity		Elasticity of Total Expenditures	
	March 2021	March 2022	March 2021	March 2022
(1)	(2)	(3)	(4)	(5)
Rice	-0.430	-0.411	0.311	0.337
Non rice and tubers	-0.994	-1.140	0.525	0.596
Fresh fish	-0.780	-0.767	0.732	0.735
Meat, eggs, and milk	-0.695	-0.697	0.915	0.936
Vegetables	-0.774	-0.784	0.514	0.537
Nuts	-0.903	-0.904	0.476	0.530
Fruits	-0.648	-0.610	1.011	1.032
Oil and coconut	-1.036	-0.784	0.445	0.534
Drink ingredients	-0.801	-0.810	0.509	0.534
Spices	-0.852	-0.856	0.600	0.662
Other foods	-0.496	-0.512	0.833	0.853
Food, beverages, and tobacco cigarettes	-1.083	-1.073	1.180	1.148

Source: Susenas Primary Data March 2021 and March 2022 (processed)

Table 5 shows that the price elasticity value for all food groups has a negative number, which means that when the price of that food group increases, the quantity demanded of that food group will decrease, and vice versa. This shows that there is conformity with the demand theory.

In March 2021, all food groups had absolute price elasticity values smaller than 1 except for the oil and coconut groups; food, beverages, cigarettes, and tobacco. This means that the two groups are elastic with their price elasticity values of -1.036; and -1.083, which means that if the price of the two food groups increases by 10%, then the quantity demanded will decrease respectively by 10.360% and 10.830%. The results of this research are in line with research conducted by [21]. Meanwhile, food groups with an absolute value of their price elasticity that is smaller than 1, are inelastic goods, which means that if there is a 10% increase in the price of a food group, then the quantity demanded of that group will decrease by less than 10%. In March 2022, the food groups that are elastic are the non-rice and tuber groups; food groups, ready-made drinks, cigarettes, and tobacco. The results of the study [21]

also show that in March 2016 the food groups that were elastic were the non-rice and tuber groups; oil and coconut group; ready-made beverages, cigarettes, and tobacco food groups with price elasticity values each of -1.134; -1.148; and -1.064.

The rice price elasticity value of rice from the study [21] was -0.549. In [20], the price elasticity value for rice was also -0.411, which is almost the same as the results of this study. The inelastic rice price elasticity value indicates that households have low sensitivity to changes in rice prices. When the price of rice increases by 10%, the demand for rice will decrease by less than 10%. This is because the majority of households in Indonesia still have a high dependence on consuming rice as a source of carbohydrates. This means that rice has an important role in household consumption in Indonesia.

The total expenditure elasticity in Table 5 is an approximation of household income elasticity. The values of all total expenditure elasticity have positive numbers, which means that all food groups are included in normal goods. If total household expenditure increases, then the demand for that food group will also increase. In March 2021 and March 2022, the fruit group; The food, beverage, cigarette, and tobacco groups have a total expenditure elasticity value of more than one, which means that both groups include luxury goods or superior goods. The total expenditure elasticity value for the two groups is 1.011; 1.179 in March 2021 and 1.031; 1.148 in March 2022, which means that if total household expenditure increases by 10%, then the total demand for these two food groups will increase respectively by 10.110% and 11.790% in March 2021 and 10.310% and 11.480% in March 2022. In contrast to the results of [21], in March 2016 the food groups included in luxury goods were meat, eggs, milk; fruits; ready-made food and beverage groups, and cigarettes. Meanwhile, for food groups that have a total expenditure elasticity value of less than 1, it can be said that this food group is less responsive to changes in total household expenditure (income) or includes normal goods.

The rice group has the smallest food expenditure elasticity and total expenditure elasticity values compared to other food groups. Meanwhile, the food and beverages, cigarettes, and tobacco groups have the greatest value. This high sensitivity of households in consuming food, drink, cigarettes, and tobacco means that changes in household lifestyles occur when households become richer. The richer a household is, the more likely the household is to consume ready-made food, drinks, cigarettes, and tobacco which are more expensive [21].

Apart from the food drinks, cigarettes, and tobacco group, the fruit group is also a luxury good. Several fruit commodities experienced price increases during the COVID-19 pandemic. This is also proven by research [26] which found that there was an increase in the average price of fruit before and during the pandemic in traditional and modern markets with changes of 12.44% and 7.52% respectively in Jember City/Regency in 2021. High fruit prices can reduce the purchasing power of households, especially households with lower middle income. In fact, consuming fruit will help improve people's nutrition, thereby increasing a country's food security. Basic Health Research in 2018 stated that 95.5% of the Indonesian population consumed less than the recommended number of vegetables and fruit [27]. Therefore, the government's role is very important in monitoring food prices, especially strategic food groups which are basic needs and nutritional fulfillment for households in Indonesia.

Appendix 1 shows the cross-price elasticity for 12 food groups. A positive cross-price elasticity indicates that the two food groups have a substitution relationship. For example in March 2022, if the price of the rice group increases, then demand for the non-rice and tuber group and the beverage ingredients group will increase, while the other nine groups will experience a decrease in demand (complementary relationship). Then, symmetrically it can also be interpreted that in March 2022, if the prices of the non-rice and tuber groups and the beverage ingredients group increase, then demand for rice will increase, whereas if the prices of the other nine food groups increase, then demand for rice will decrease.

4. Conclusion

In March 2022, saw an increase in prices for strategic food commodities compared to March 2021. This resulted in a change in household food consumption patterns in Indonesia in March 2022, which was shown by a decrease in the proportion of food expenditure due to an increase in food prices. Apart from being influenced by food prices, the diversity of household food expenditure is also influenced by the socio-demographic characteristics of the household.

The own price elasticity value has a negative number which indicates that there is conformity with demand theory. Cross-price elasticities for some food groups have negative values (mutual complementarity), while some have positive values (mutual substitutes). The elasticity of total household expenditure shows a positive value, which means that all food groups are included in normal

goods. Fruit group; the food, ready-made drinks, cigarettes, and tobacco group are normal goods that are considered luxury goods in both March 2021 and March 2022.

The government is expected to be able to overcome and monitor rising food prices, especially for food groups which are important commodities for households in Indonesia. Suggestions for the next study could be to separate the ready-to-drink food group and the cigarette and tobacco group to find out how sensitive households are to price changes in these two food groups.

Ethics approval

This study was conducted in accordance with the ethical standards. 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 for providing support and resources that have been provided. We also thank Badan Pusat Statistik (BPS) for providing data to support this study so that it can be useful for readers. The support and cooperation of all parties are very valuable for the success of this study.

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

Wifa Darma Aulia: Conceptualization, Data Collection, Formal Analysis, Writing – Original Draft, Visualization. **Rita Yuliana**: Methodology, Writing – Review & Editing, Supervision.

References

[1]. Gloria, Pandemi Covid-19 Munculkan Kompleksitas Masalah Pangan [The COVID-19 Pandemic Creates Complexities in Food Issues], Mei 8. 2020, Berita Universitas Gadjah Mada: https://ugm.ac.id/id/berita/19397-pandemi-covid-19-munculkan-kompleksitas-masalah-pangan/. [Accessed Mei 14, 2024].

[2]. Firdaus. M, Disparitas Harga Pangan Strategis Sebelum dan Saat Pandemi COVID-19 [Price Disparity of Strategic Food Items Before and During the COVID-19 Pandemic], Jurnal Ekonomi Inodnesia. 107-120, 2021.

[3]. BPS, Indikator Ekonomi Maret 2022 [Economic Indicators March 2022], Jakarta: Badan Pusat Statistik, 2022a

[4]. BPS, Pengeluaran untuk Konsumsi Penduduk Indonesia [Consumption Expenditure of Population of Indonesia], Jakarta: Badan Pusat Statistik, 2017.

[5]. BPS, Statistik Harga Konsumen Pedesaan Kelompok Makanan 2022 [Rural Consumer Price Statistics Food Groups 2022], Jakarta: Badan Pusat Statistik, 2022b.

[6]. Andriessa. R, Pusat Studi Perdagangan Dunia, Universitas Gadjah Mada, Maret 5. 2022, Minyak goreng langka ? Ternyata inilah penyebabnya! [Is Cooking Oil Scarce? Here is the Cause!]: https://cwts.ugm.ac.id/2022/03/05/minyak-goreng-langka-ternyata-inilah-penyebabnya/. [Accessed June 6, 2023].
[7]. BPS, Statistik Harga Konsumen Pedesaan Kelompok Makanan 2021 [Rural Consumer Price Statistics Food Groups 2021], Jakarta: Badan Pusat Statistik, 2021
[8]. Pindyck. R. S, Rubinfeld, D. L, and Rabasco. E, Microeconomics (8th ed.), Jersey (USA): Pearson Education, 2013.
[9]. Deaton. A and Muellbauer. J, An Almost Ideal Demand System, The American Economic Review. 312-326, 1980.
[10]. Asikin. M, Risyanto, and Said. A, Manfaat Program Bantuan Langsung Tunai Terhadap Ketahanan Pangan Rumah Tangga Miskin : Studi Kasus di Kabupaten Karangasem dan Buleleng [The Benefits of Direct Cash Assistance Program for Household Food Security: A Case Study in Karangasem and Buleleng Districts, Bali Province], Jakarta: Direktorat Analisis dan Pengembangan Statistik Badan Pusat Statistik, 2009.
[11]. Farras. M. F, Anindita. R, and Asmara. R, Pola Konsumsi dan Permintaan Protein Hewani di Kota Malang Model Almost Ideal Demand System (AIDS) [Consumption Patterns and Demand for Animal Protein in Malang City Using the Almost Ideal Demand System (AIDS) Model], Jurnal Ekonomi Pertaniandan Agribisnis (JEPA). 286-297, 2021.
[12]. Heriyanto, Perilaku Konsumsi Pangan Sumber Karbohidrat Rumah Tangga Petani Kelapa Sawit di Kecamatan Kandis Kabupaten Siak [Consumption Behavior of Carbohydrate Food Sources in Oil Palm Farmers' Households in Kandis Subdistrict, Siak Regency], Jurnal Ilmiah Pertanian. 15-38, 2016.
[13]. Miranti. A, Syaukat. Y, and Harianto, Pola Konsumsi Pangan Rumah Tangga Di Provinsi Jawa Barat [Household Food Consumption Patterns in Jawa Barat Province], Jurnal Agro Ekonomi. 34, 67-80, Mei 26. 2016.
[14]. Nasution. A, Krisnamurthi. B, and Rachmina. D, Analisis Permintaan Pangan Rumah Tangga Penerima Manfaat Bantuan Pangan Non Tunai (BPNT) di Kota Bogor [Analysis of Food Demand Among Households Receiving Non-Cash Food Assistance (BNPT) in Municipality of Bogor], Forum Agribisnis, 1-10, 2020.
[15]. Novarista. N, Syahni. R, and Jafrinur, Faktor-Faktor yang Mempengaruhi Konsumsi Pangan Hewani Pada Konsumen Rumahtangga di Kota Padang [Factors Affecting Animal-Based Food Consumption Among Household Consumers in Manucipality of Padang], Jurnal Agribisnis Kerakyatan. 64-74, 2013.
[16]. Nursamsi, Nurmalina. R, and Rifin. A, Kajian Sistem Permintaan Komoditas Sumber Protein di Enam Propinsi di Indoneisa [Study of Demand Systems for Protein Source Commodities in Six Province of Indonesia], Jurnal Agribisnis Indonesia. 142-156, 2019.
[17]. Ritonga. H, The impact of household characteristics on household consumption behavior : A demand system analysis on the consumption behavior of urban households in the province of Central Java. Indonesia, IOWA STATE UNIVERSITY, 1994.
[18]. Sari. N. A, Analisis Pola Konsumsi Pangan Daerah Perkotaan dan Pedesaan Serta Keterkaitannya Dengan Karakteristik Sosial Ekonomi di Provinsi Kalimantan Timur [Analysis of Food Consumption Patterns in Urban and Rural Areas and Their Relationship with Socio-Economic Characteristics in Kalimantan Timur Province], JEMI. 69-81, 2016.
[19]. Suryanty. M and Reswita. R, Analisis Konsumsi Pangan Berbasis Protein Hewani di Kabupaten Lebong : Pendekatan Model AIDS (Almost Ideal Demand System) [Animal Protein Based Food Consumption Analysis in District of Lebong: AIDS Approach], AGRISEP. 101-109, 2016.
[20]. Wijayanti. P. D, Harianto, and Suryana. A, Permintaan Pangan Sumber Karbohidrat di Indonesia [Carbohydrate Food Demand in Indonesia], Analisis Kebijakan Pertanian. 13-26, 2019.
[21]. Yuliana. R, Permintaan Pangan dan Perubahan Tingkat Kesejahteraan Rumah Tangga di Indonesia [Food Demand and Changes in Household Welfare Levels in Indonesia], Bogor: Institut Pertanian Bogor (IPB), 2018.
[22]. Moeis. Jossy P, Indonesian Food Demand System: An Analysis of the Impact of the Economic Crisis On Household Consumption and Nutritional Intake, Washington DC (US): The George Washington Univ, 2003.
[23]. Ackah. C and Appleton. S, Food price changes and consumer welfare in Ghana in the 1990s, University of Nottingham: CREDIT Research Paper No 07/03, 2007.
[24]. Teklu. T and Johnson. S. R, Demand Systems from Cross Section Data: An Experiment for Indonesia, CARD Working Papers, 1987.

[25]. Yafiz. I and Pahlevi. A, “Minyak goreng melimpah setelah harga eceran tertinggi dicabut, tapi sekarang harganya mahal” [“Cooking Oil Abundance After the Removal of the Highest Retail Price Cap, But Now the Price is High”], Maret 19. 2022, BBC News Indonesia: https://www.bbc.com/indonesia/indonesia-60754619. [Accessed Mei 16, 2024].
[26]. Ferdiansyah. Nur Muhammad, Fauzi. Nurul Fathiyah, and Prayuginingsih. H, Permintaan dan Penawaran Buah di Wilayah Kota Kabupaten Jember Pada Masa Pandemi Covid-19 [Demand and Supply of Fruits in the Jember Regency Area During the COVID-19 Pandemic], National Multidisciplinary Sciences. 262-267, 2022.
[27]. Kementerian Kesehatan, Hasil Utama RISKESDAS 2018 [Key Findings of RISKEDAS 2018], Jakarta, 2018.

Appendix 1. Marshallian price elasticity of 12 food groups for the period March 2021 and March 2022

Food Group	Rice	Non rice and tubers	Fresh fish	Meat, eggs, and milk	Vegetables	Nuts	Fruits	Oil and coconut	Drink ingredients	Spices	Other foods	Food, beverages, and tobacco cigarettes
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
March-2021												
Rice	**-0.4303**	0.0224	-0.1440	-0.0321	-0.0476	-0.0214	-0.0487	0.0134	0.0249	-0.0085	-0.0183	-0.0419
Non rice and tubers	0.0816	**-0.9939**	0.5263	0.0626	-0.2300	0.0833	-0.0692	0.0885	0.0179	0.0038	-0.0977	-0.1197
Fresh fish	-0.2712	0.1628	**-0.7803**	-0.0517	-0.0024	0.0406	0.0116	0.0053	-0.0271	-0.0011	-0.0256	0.0392
Meat, eggs, and milk	-0.0836	-0.0004	-0.0645	**-0.6953**	-0.0359	-0.0190	-0.0068	-0.0256	-0.0234	-0.0184	-0.0302	-0.1341
Vegetables	-0.0561	-0.0347	0.0244	0.0121	**-0.7735**	0.0145	-0.0053	0.0286	0.0075	0.0099	0.0242	-0.0237
Nuts	-0.2136	0.1104	0.1643	0.0107	0.0740	**-0.9027**	-0.0226	0.0094	0.0925	0.0670	-0.0804	0.0039
Friuts	-0.2438	-0.0732	-0.0006	0.0077	-0.0389	-0.0259	**-0.6484**	-0.0360	-0.0237	0.0081	-0.0576	0.0560
Oil and coconut	0.0039	0.0671	0.0063	-0.0225	0.0836	0.0131	-0.0172	**-1.0361**	0.0375	0.0174	0.0691	-0.0171
Drink ingredients	0.0434	0.0038	-0.0679	-0.016	0.0114	0.0540	-0.0045	0.0289	**-0.8006**	0.0181	-0.0409	-0.0931
Spices	-0.1355	-0.0182	-0.0265	-0.0143	0.0191	0.0542	0.0350	0.0103	0.0438	**-0.8521**	0.0215	-0.0221
Other foods	-0.1408	-0.0975	-0.0839	-0.0489	0.0468	-0.0582	-0.0557	0.0466	-0.0684	0.0215	**-0.4960**	-0.0786
Food, beverages, and tobacco cigarettes	-0.0376	-0.0261	-0.0089	-0.0459	-0.0239	-0.0161	-0.0089	-0.0184	-0.0259	-0.0175	-0.0221	**-1.0834**
March-2022												
Rice	**-0.4108**	0.0338	-0.1460	-0.0337	-0.0451	-0.0017	-0.0464	-0.0385	0.0229	-0.0087	-0.0274	-0.0533
Non rice and tubers	0.1344	**-1.1402**	0.5491	0.0557	-0.2378	0.0937	-0.1342	0.0597	0.0549	0.0000	-0.0787	-0.0785
Fresh fish	-0.2519	0.1621	**-0.7672**	-0.0599	0.0598	0.0312	-0.0254	-0.0054	-0.0323	-0.0019	-0.0421	0.0383
Meat, eggs, and milk	-0.0819	-0.0024	-0.0754	**-0.6973**	-0.0347	-0.0193	-0.0130	-0.0304	-0.0237	-0.0191	-0.0260	-0.1280
Vegetables	-0.0501	-0.0362	0.0752	0.0118	**-0.7845**	-0.0201	-0.0226	0.0319	0.0026	0.0070	0.0253	-0.0273

Nuts	-0.0748	0.1228	0.1221	0.007	-0.0950	**-0.9038**	0.0193	-0.0318	0.0907	0.0667	-0.0567	0.0114
Friuts	-0.1886	-0.0953	-0.0721	-0.0049	-0.0701	-0.0036	**-0.6099**	-0.0456	-0.0176	-0.0012	-0.0653	0.0831
Oil and coconut	-0.1654	0.0393	-0.0174	-0.0246	0.0770	-0.0090	-0.0248	**-0.7845**	0.0067	0.0107	0.0846	-0.0174
Drink ingredients	0.0427	0.0362	-0.0848	-0.0131	0.0016	0.0561	0.0044	0.0074	**-0.8096**	0.0211	-0.0458	-0.0659
Spices	-0.1209	-0.0169	-0.0335	-0.0151	0.0064	0.0489	0.0227	0.0038	0.0350	**-0.8561**	0.0203	-0.0117
Other foods	-0.1625	-0.0753	-0.1262	-0.0347	0.0498	-0.0432	-0.0714	0.0799	-0.0642	0.0157	**-0.5120**	-0.0776
Food, beverages, and tobacco cigarettes	-0.0385	-0.0208	-0.0072	-0.0425	-0.0229	-0.0140	-0.0024	-0.0169	-0.0210	-0.0150	-0.0199	**-1.0730**

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.732

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2 December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

The Utilization of Model Output Statistic (MOS) in Improving Weather Prediction Model Accuracy of Integrated Forecasting System (IFS)

Isnaini Anjelina Ramadhan[1*], Deni Septiadi [2]

[1,2]*State College of Meteorology Climatology and Geophysics, Tangerang, Indonesia*
**Corresponding Author: E-mail address:* isnainianjelinar@gmail.com

ARTICLE INFO

Article history:
Received 11 March, 2024
Revised 20 April, 2024
Accepted 7 May, 2024
Published 31 December, 2024

Keywords:
Accuracy; IFS; MOS; Prediction; Regression; Weather

Abstract

Introduction/Main Objectives: Integrated Forecasting System (IFS) is one of the most accurate numerical weather prediction (NWP) model for Indonesia region. **Background Problems:** However, in fact, each model always has bias potential against observation which causes inaccuracy in weather prediction. **Novelty:** This research intends to overcome this problem by building a weather prediction model based on Model Output Statistic (MOS) to minimize bias and improve NWP accuracy. **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:** Analysis result states that compared to IFS, MOS fluctuation pattern is more relevant to observation. MOS has higher correlation to observation and lower error. However, the variance of observation value tends to be better represented by IFS. The test result of heavy rain cases prove that the application of MOS is able to provide fairly accurate prediction. This weather prediction will be able to be the basis for decision-making and preventive measure in dealing with extreme condition that may occur.

1. Introduction

Weather prediction is one of the crucial needs to support the smooth operation of the public sector. According to *Badan Nasional Penanggulangan Bencana (BNPB) or National Disaster Management Agency* [1], in 2022 there is 3,544 natural disasters that hit Indonesia. As many as 99.2% of events is dominated by hydrometeorological disasters. Flooding is the disaster with the highest frequency at 1,531 events, followed by extreme weather at 1,068 events. These disasters are closely related to meteorological condition. Therefore, weather prediction is important for decision-making and preventive measures in dealing with extreme condition that may occur [2].

Badan Meteorologi Klimatologi dan Geofisika (BMKG) or Meteorology Climatology and Geophysics Agency, as a weather service provider in Indonesia, applies numerical weather prediction (NWP) method in making weather prediction. The model used by BMKG is integrated in the workstation named 'Synergie' which contains four models, including the Global Forecast System (GFS), Integrated Forecasting System (IFS), ARPEGE, and Weather Research and Forecasting (WRF). The model that has the best performance so far is IFS. According to Kiki and Alam [3], IFS is proved to be able to predict 24-hour accumulated precipitation in various classifications, including per year, per month, per season, per province, to the average percentage per month better than the other three models mentioned.

However, it should be noted that every weather model has the potential to produce bias against observation. The input of observation and assimilation data has the potential to cause uncertainty in the estimation of atmospheric condition [4]. Therefore, a processing method is needed to optimize the work of weather prediction model.

The optimization of NWP can be done by statistical post-processing method, one of which is through model output statistic (MOS). MOS is a method that relates between weather observation as predictand and NWP parameter as predictor using regression model [5] [6]. The first MOS research is developed by National Weather Service (NWS) Oceanic and Atmospheric Administration (NOAA) which is published through the research of Glahn and Lowry [5]. Brunet et al. [7] also conducts research to compare the results of perfect-prog (PP) prediction with MOS prediction. The result is that PP prediction tends to be more suitable for short-term prediction and more sensitive in displaying extreme weather. In contrast, MOS is more suitable for longer period and more reliable because it can overcome model limitation. The German meteorological agency, Deutscher Wetterdienst (DWD), also claims that MOS has high weather prediction accuracy [8]. In operational, DWD launches a MOS-based weather forecast product that combines the Integrated Forecasting System (IFS) and Icosahedral Nonhydrostatic (ICON) models, which is named MOSMIX.

Based on these superiorities, this research intends to build MOS-based weather prediction to improve the accuracy of IFS model. Predicted parameters include temperature, relative humidity, and QFF pressure. MOS will be tested to predict the weather in several cases of rain that had occurred in DKI Jakarta. Three cases of heavy rain (intensity > 50 mm/day) during January to February 2023 has been selected as samples, including January 1, 2023; January 4, 2023; and February 24, 2023. This research is relatively new because the application of MOS in Indonesia is still minimum. The application of MOS to the IFS model has been carried out by DWD with research location in Europe, especially Germany [8]. However, the use of IFS in several MOS studies in Indonesia [6] [9] [10] has not been found yet. Hence, through this research, the application of MOS to IFS for weather prediction in Indonesia is a new matter.

The regression model used to build MOS prediction is stepwise regression with forward selection type. The reason for this selection is because stepwise regression can select predictors with the highest correlation to be included in the model equation. Stepwise regression has been used in making MOS predictions at NWS based on the research of Glahn and Lowry [5] and has been proofed to improve NWP result. Although the predictand is correlated with hundreds of predictors, a regression equation containing only a few predictors can also approximate the observation. An equation that contains too many predictors have the potential to produce worse prediction. Other studies that also utilize stepwise regression in building MOS prediction include Bocchieri et al. [11], Klein and Glahn [12], and Kuligowski and Barros [13].

The location selection of DKI Jakarta is based on the vulnerability that may be obtained when a disaster occurs. According to Badan Penanggulangan Bencana Daerah (BPBD) or Local Disaster Management Agency of DKI Jakarta [14], in 2021 there are 375 hydrometeorological disaster events including floods, strong winds, fallen trees, landslides, and flooded roads. Based on this number, there are 75 floods that affected 51,294 people in 118 sub-districts. As for the 12 landslide incidents, the losses are estimated at 420 million rupiah. In addition, Indonesia's multi-sectoral activities are centered in DKI Jakarta. DKI Jakarta is a fairly dense province with a projected population of 10,679,951 in 2022 [15]. This target is expected to get benefits form this research. This effort to improve the accuracy of IFS are expected to provide positive results in starting the step of accurate objective weather prediction service.

The purpose of this research is to utilize MOS to minimize the bias produced by prediction against observation. This research intends to analyze the most suitable MOS regression model for weather prediction, analyze the performance test between IFS and MOS prediction against observation, and analyze the ability of MOS predictions of heavy rain cases in DKI Jakarta.

2. Material and Methods

2.1. Literature Review

2.1.1. Model Output Statistic (MOS)

Model output statistic (MOS) is an objective weather prediction method expressed by statistical relationship between predictors and predictands using numerical method at a certain time projection [5]. This method utilizes weather observation as a predictand and NWP output as a predictor based on regression [9] [10]. In general, the function of MOS can be written in the following equation.

$$\hat{y}_t = f_{MOS}(x_t)$$

Description:
$\hat{y}_t$: weather forecast at time t
x_t : NWP output variable at time t

2.1.2. Numerical Weather Prediction (NWP)

Numerical weather prediction (NWP) is a system of equations that describes the essential physical rules governing motion and processes in the atmosphere [16]. NWP calculation basically uses partial integral equation. There are three components that need to be taken into account, including observation, diagnostic or analysis, and prognostic [17].

Palmer [18] mentions that there are three uncertainties that cause NWP to deviate and cause bias, including initial uncertainty, model uncertainty, and external parameter uncertainty.

2.1.3. Integrated Forecasting System (IFS)

The Integrated Forecasting System (IFS) is NWP model developed by the European Center for Medium-Range Weather Forecast (ECMWF) in collaboration with Météo-France. IFS is obtained by applying the semi-implicit semi-Lagrangian (SL) method to solve dynamic equations [19]. Currently, NWP calculations are performed by supercomputers that simultaneously predict weather. IFS routinely performs data assimilation by adding the latest observational data to produce model output [20]. These include atmospheric, oceanic, and physical land surface parameters [21].

IFS is a global model that includes surface level and elevation data for all regions on Earth. There are two types of IFS models, namely high-resolution forecasts (HRES) and ensemble forecasts (ENS) [22]. HRES is a single-forecast model consisting of only one model configuration. Meanwhile, ENS is an ensemble model that consists of several model combinations. This research will focus on single-forecast HRES.

2.1.4. Stepwise Regression

Stepwise regression or screening regression is a regression model that uses independent variables as a reference for model processing result. This research uses the forward selection type. According to Glahn and Lowry [5], as well as Kuligowski and Barros [13], the first step in this procedure is to select the variable that is most highly correlated with the prediction (either positive or negative). Next, selecting the variable that together with the first variable increases the reduction of the highest variance. Selection can continue in this way until cut off criterion based on p-value is met. The requirement for a variable to enter the model is that the p-value must be less than α.

2.2. Location and Time Research

The research location focuses on province of DKI Jakarta. In this location there are three meteorological stations, namely Tanjung Priok Maritime Meteorological Station, Kemayoran Meteorological Station, and Halim Perdana Kusuma Meteorological Station. The three were chosen in order to represent the distribution of observation locations as shown in Figure 1.

The research focuses from January 2022 to February 2023. The year 2022 is used as training period, while the year 2023 as testing period. In the testing period, three cases of heavy rain (intensity > 50 mm/day) that hit DKI Jakarta is selected, for instances January 1, 2023; January 4, 2023; and February 24, 2023. The amount of rainfall is presented in Table 1.

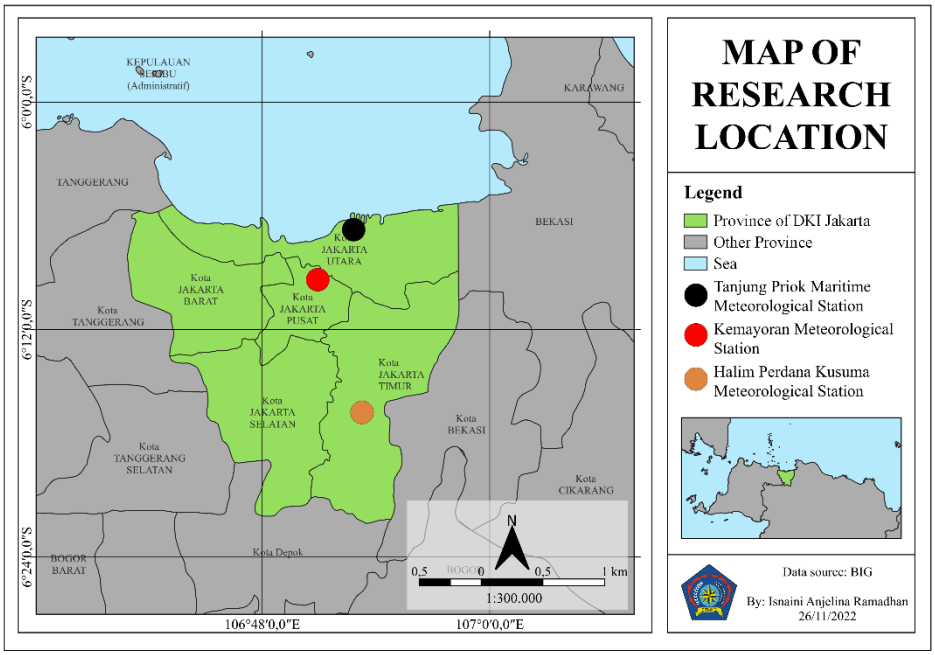


Figure 1. Map of research location.

Table 1. Sample cases of heavy rain in DKI Jakarta during January to February 2023

Date	Rainfall (mm)		
	Tanjung Priok	Kemayoran	Halim Perdana Kusuma
January 1, 2023	134.4	31.5	6.9
January 4, 2023	31.3	35.3	79.6
February, 24 2023	54.7	69.0	84.0

Source: BMKG (2023)

2.3. Data

2.3.1. IFS Data

The IFS data used in this study is HRES which is a single-forecast model. HRES data has spatial resolution of 0.08º x 0.08º or 9 km x 9 km [23]. The temporal resolution is 3 hours with a cycle every 12 hours at 00 and 12 UTC. This research uses single level forecast data consisting of 44 single level and pressure level meteorological parameters.

2.3.2. Synoptic Observation Data

Synoptic observation data is obtained from weather observation at the meteorological station tool park. The measurement tool used is a conventional tool. The data used includes air temperature, relative humidity, and QFF pressure with a range of every three hours. The selection of these parameters is because they are the basic weather parameters that are always observed every hour so that their fluctuations can be observed in detail (in this study a three-hour time span is used to adjust the temporal resolution of IFS).

2.4. Flowchart

The research steps in a coherent and structured manner are presented in Figure 2. Based on the flowchart, more detailed explanation of the research steps is as follows.

1. Performing pre-processing step, including:
   a. Extracting IFS model data from GRIB file into CSV format.
   b. Managing missing data using curve fitting method

c. Performing stationary test using correlogram based on autocorrelation (ACF) and partial autocorrelation (PACF) values. ACF is the correlation or relationship of a time series data for different lags. The ACF value can be obtained using the following equation [24].

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{cov(z_t, z_{t+k})}{\sqrt{Var(z_t)}\sqrt{Var(z_{t+k})}} \quad (1)$$

Description:

ρ_k : ACF function at lag k
γ_k : auto covariance of z_t and z_{t+k}
t : time
$Var(z_t)$: constant variance

The value of the PACF function is the development of the ACF by removing linear dependencies on the variables of Zt+1, Zt+2, dan Zt+k-1

d. Performing data normalization to homogenize the data range according to the following equation.

$$x_{norm} = \frac{x - x_{min}}{x_{maks} - x_{min}} \quad (2)$$

2. Dividing the data into two groups, as training data and testing data. The training data is from January to December 2022, while the testing data is from January to February 2023.
3. Building MOS prediction using stepwise regression. The cut off criteria for the selection of predictors is limited to p-value 0,05. If a parameter has $\alpha < 0.05$, it is considered to be included in the regression model. This process produces a regression equation that becomes the basis for building MOS prediction.
4. Calculating MOS weather prediction for testing period derived from the calculation of IFS model data by regression equation. The results then go through denormalization process to restore the actual value of weather parameters.
5. Testing the performance of IFS and MOS prediction against observation data for the group of testing data. The performance tests include graph analysis and Taylor diagram. Verification uses several calculations, including:
   a. Correlation coefficient (r)

$$r = \frac{\sum_{i=1}^{N}(F_i - \bar{F})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(F_i - \bar{F})^2 \cdot \sum_{i=1}^{N}(O_i - \bar{O})^2}} \quad (3)$$

Description:

F_i : predicted value
$\bar{F}$: average prediction
O_i : observation value
$\bar{O}$: average observation
N : number of data

The value of correlation coefficient is interpreted in the following categories.

Table 2. Correlation coefficient interpretation

Value of Correlation Coefficient	Interpretation
< 0,20	Data relationship is considered non-existent
0,20—0,40	Low relationship
>0,40—0,70	Moderate relationship
>0,70—0,90	High relationship
>0,90—1,00	Very high relationship

Source: Sarwono, 2006

b. Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(F-O)^2} \quad (4)$$

c. Standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N-1}} \quad (5)$$

6. Conducting cases test on three heavy rain events. The MOS prediction results for each case were compared with the observed value and residual.

$$Residual = F - O \quad (6)$$

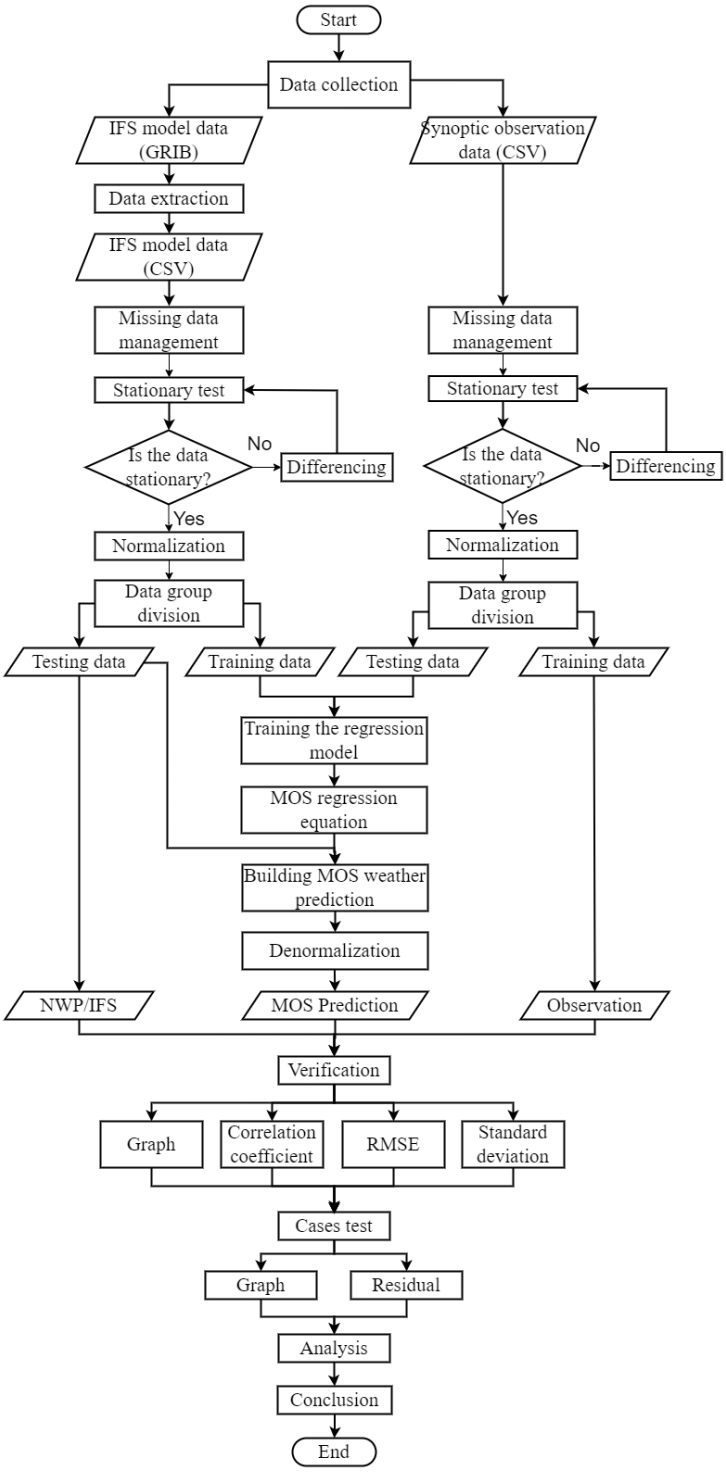


Figure 2. Research flowchart

3. Result and Discussion

3.1. Regression Configuration of Model Output Statistic (MOS)

3.1.1. Air Temperature

Table 3 shows the regression configuration for air temperature at Tanjung Priok. The adjusted R-squared value is 0.730, which means that 73.0% of the independent variables can explain the dependent variable. Parameter 2t (surface temperature) is the most influential variable with the highest coefficient value among others, which is 0.517.

Table 3. Configuration of MOS regression for air temperature at Tanjung Priok

Variable	Coefficient
2t	0.517333
C	0.223572
papt1000	-0.190083
Msl	-0.210207
r925	0.088743
w500	-0.103157
t1000	0.260527
w1000	0.111158
v10	-0.210126
t200	0.055796
r500	0.075744
Tp	0.185290
w700	-0.069637
r1000	0.050890
v1000	0.142059
t0	-0.054935

Next, the configuration of MOS regression for air temperature at Kemayoran is described Table 4. The adjusted R-squared value is 0.738, which is 73.8% of the independent variables can explain the dependent variable. Almost the same as Tanjung Priok, parameter 2t is the variable that mostly affects the temperature with the highest coefficient value, which is 0.638.

Table 4. Configuration of MOS regression for air temperature at Kemayoran

Variable	Coefficient
2t	0.637556
Tp	0.238667
Sp	-0.131850
papt1000	-0.110720
r925	0.119823
papt200	0.043699
w1000	0.113805
v10	-0.281386
t1000	0.146022
v1000	0.204717
w700	-0.089568
r500	0.072898
u700	0.057743
t850	0.045713
t200	0.040517
v500	-0.048726

Furthermore, Table 5 shows the configuration of MOS regression for air temperature at Halim Perdana Kusuma. The adjusted R-squared value is 0.735, which means 73.5% of the independent variables can explain the dependent variable. Again, parameter 2t is the parameter that mostly affects the observed temperature that the coefficient value is 0.641.

Table 5. Configuration of MOS regression for air temperature at Halim Perdana Kusuma

Variable	Coefficient
2t	0.641155
t1000	0.295131
tp	0.227148
v10	-0.149102
t925	-0.110649
papt1000	-0.113882
w1000	0.273773
sp	-0.069180
w925	-0.093330

t200	0.047487
u700	0.075918
papt850	0.077945

Overall, the dependent variable can be explained by 73.0% to 73.8% of the independent variables. Parameter 2t is the parameter that has the greatest influence in producing MOS temperature prediction. This is because parameter 2t is actually an IFS model for surface temperature at 2 meters in height.

3.1.2. Relative Humidity

Table 6 shows the configuration of MOS regression for relative humidity at Tanjung Priok. The adjusted R-squared is 0.644, it means that 64.4% of the independent variables can explain the dependent variable. Parameter 2t (surface temperature) is the parameter that mostly affects the relative humidity of Tanjung Priok, that is proved by the largest coefficient value among others, which is -0.633. Negative value has an effect by reducing the predicted value so that it is more appropriate with the observation.

Table 6. Configuration of MOS regression for relative humidity at Tanjung Priok

Variable	Coefficient
c	0.415155
2t	-0.632871
r1000	0.329955
w1000	-0.189518
msl	0.250727
t925	0.101252
t1000	0.200183
papt1000	0.074940
w500	0.100016
v850	0.041084
t0	0.056962
u500	-0.055033
r200	-0.029158

Table 7. Configuration of MOS regression for relative humidity at Kemayoran

Variable	Coefficient
r1000	0.365408
msl	0.183062
r2	0.040282
2t	-0.571388
c	0.527001
t1000	0.268587
w1000	-0.120337
r925	-0.085770
u700	-0.053809
r500	-0.057680
w500	0.065190

The configuration of MOS regression for relative humidity at Kemayoran is explained in Table 7. Adjusted R-squared value is 0.648, it explains that 64.8% of the independent variables are able to explain the dependent variable. Among the other parameters, parameter 2t has the largest coefficient value, which is -0.571. Just like MOS prediction for relative humidity at Tanjung Priok, parameter 2t also has negative coefficient that reduce the predicted value.

Next, **Error! Not a valid bookmark self-reference.** explains the MOS configuration for relative humidity at Halim Perdana Kusuma. The adjusted R-squared value is 0.675, which means that 67.5% of the independent variables can explain the dependent variable. Parameter 2t is the parameter with the highest coefficient of -0.464. Similar to the previous two locations, the negative coefficient has an influence by providing reduction in value so that the MOS prediction becomes more appropriate.

Table 8. Configuration of MOS regression for relative humidity at Halim Perdana Kusuma

Variable	Coefficient
r1000	0.330640
t925	0.087774
2t	-0.464465
C	0.736266
Sp	0.081185
papt1000	0.082203
Tp	-0.123675
papt700	-0.082258
w1000	-0.265422
v1000	0.107520
w925	0.078613
u700	-0.078555

Overall, the dependent variable can be explained by 64.4% to 67.5% of the independent variables. Regressions at the three locations show that the most influential parameter on the MOS relative humidity prediction is parameter 2t (surface temperature) or temperature at 2 meters in height. This is because temperature and relative humidity are interrelated. Higher air temperature will cause water vapor composition in the air to increase, so relative humidity also increases. Practically, relative humidity is calculated through equations: $RH = 100 - 5(T_{BK} - T_d)$ or $RH = 100 - 7(T_{BK} - T_{BB})$. Relative humidity (RH) is obtained from temperature measurement of the dry bulb thermometer (T_{BK}), wet bulb thermometer (T_{BB}), and dew point (T_d). Temperature of dry bulb thermometer (T_{BK}) is another name for the measurement of air temperature at height of 2 meters. Therefore, it is understandable that parameter 2t provides the largest regression coefficient for predicting relative humidity.

Besides that, parameter r1000 (relative humidity at 1000 mb) is the most influential variable after parameter 2t. The 1000 mb altitude is considered as the near-surface geopotential altitude (sea level pressure measurement). Although the definition of relative humidity observation in operational is different from the measurement at 1000 mb, this value is quite representative to surface relative humidity.

3.1.3. QFF Pressure

Table 9 shows the configuration of MOS regression at Tanjung Priok. The adjusted R-squared value is 0.845, it states that 84.5% of the independent variables can explain the dependent variable. The greatest predictive influence is determined by msl (mean sea level pressure) with coefficient of 2.549.

The MOS configuration at Kemayoran for QFF pressure is shown in Table 10. The adjusted R-squared value is 0.905. Hence, 90.5% of the independent variables are able to explain the dependent variable. So far, this value is the highest among the three locations and other parameters. Similar to Tanjung Priok, msl is the most influential parameter on QFF pressure, which is 2.653.

Furthermore, Table 11 shows the configuration of MOS regression for QFF Pressure at Halim Perdana Kusuma. The adjusted R-squared value is 0.508, which means that 50.8% of the independent variables can explain the dependent variable. This number is not good enough to prove the performance of the training data, because it means that 49.1% of the independent variables fail to explain the dependent variable. This value is lower than the other two locations. Meanwhile, the parameter that is most influential to the regression is msl, its coefficient is 0.261.

Table 9. Configuration of MOS regression for QFF pressure at Tanjung Priok

Variable	Coefficient
Sp	-0.586013
C	-0.405385
msl	2.549298
t0	-0.180187
t925	0.075094
t200	-0.060315
papt1000	0.088071
t850	0.053665
r500	-0.067441
t700	0.035540

v1000	0.125643
v10	-0.105918
u925	-0.031156
u700	-0.030579
v500	-0.041657
w925	-0.038344

Table 10. Configuration of MOS regression for QFF pressure at Kemayoran

Variable	Coefficient
Sp	-0.476642
C	-0.850700
msl	2.652981
t925	0.165593
t700	0.107852
t0	-0.056510
t500	0.058274
2t	-0.139103
t1000	0.058935
papt1000	0.058935
papt925	0.055624
t850	0.031138
u925	-0.027976
v500	-0.033466
w200	-0.053179
w500	0.055966
r700	0.035132

Table 11. Configuration of MOS regression for QFF pressure at Halim Perdana Kusuma

Variable	Coefficient
c	0.317989
msl	0.260788
2t	-0.116825
papt1000	0.080734
t1000	0.086738
t925	0.037626
t200	-0.031259
u200	-0.035565
v700	-0.035796
u1000	0.026117
tp	-0.055914
v500	-0.037715

For QFF pressure prediction, the dependent variable can be explained by 50.8% to 90.5% independent variables. These ranges are the lowest and highest values in this research. In all three locations, msl (mean sea level) has the largest regression coefficient value. Therefore, this parameter is the most influential in predicting QFF pressure. This is suitable because msl is the IFS model for QFF pressure so the two are interrelated.

3.2. MOS Performance Test

3.2.1. Air Temperature

Figure 3 shows the comparison graph between observation, IFS, and MOS for air temperature. It can be seen that both IFS and MOS can follow the observation fluctuation pattern. However, most of IFS predictions are underestimated. Meanwhile, MOS has more reliable ability because most of the values are not much different from the observation.

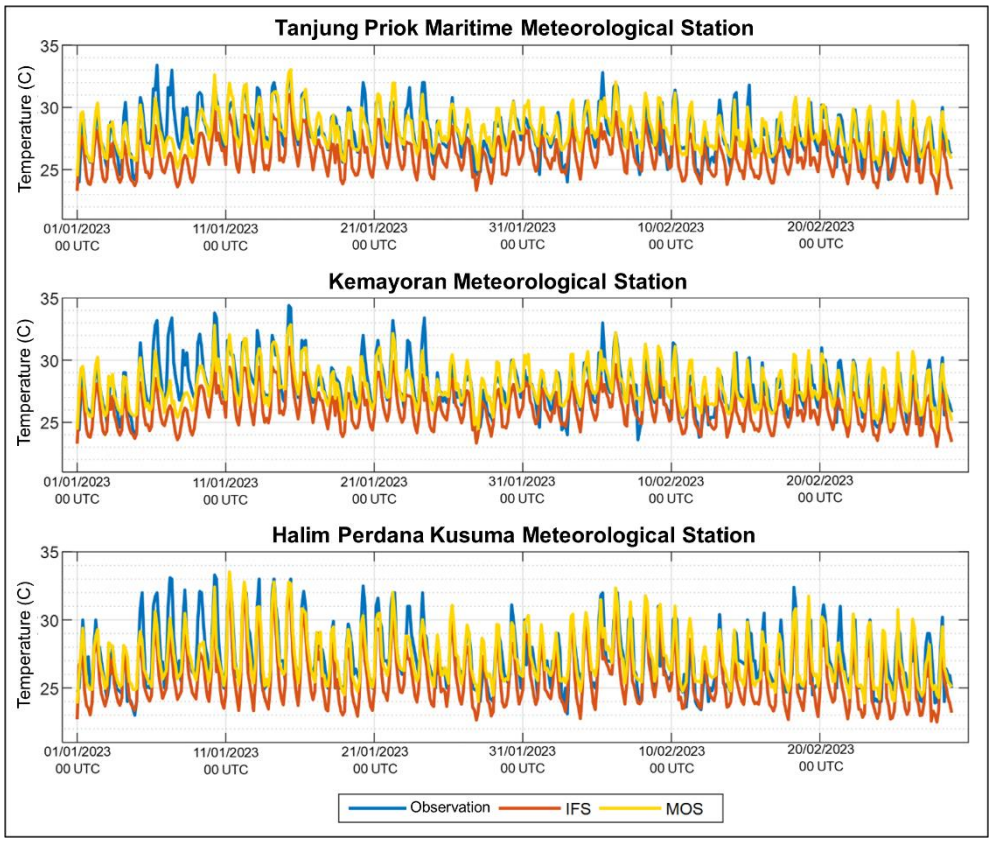
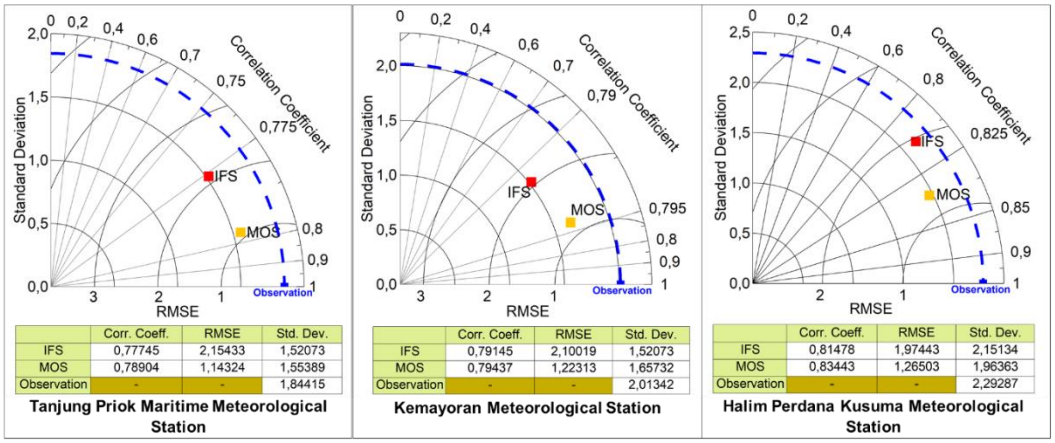


Figure 3. Comparison graph between observation, IFS, and MOS for air temperature



	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,77745	2,15433	1,52073
MOS	0,78904	1,14324	1,55389
Observation	-	-	1,84415

	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,79145	2,10019	1,52073
MOS	0,79437	1,22313	1,65732
Observation	-	-	2,01342

	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,81478	1,97443	2,15134
MOS	0,83443	1,26503	1,96363
Observation	-	-	2,29287

Figure 4. Taylor Diagram between observation, IFS, and MOS for air temperature

Taylor diagram in Figure 4 shows that IFS and MOS correlation values are not much different. Based on table 2, the correlation coefficient categories are all in the high relationship category and MOS are always higher than IFS. It means MOS has better relationship with the observation than IFS. The lower RMSE of MOS indicates that MOS tends to produce lower error than IFS. At Tanjung Priok and Kemayoran, standard deviation of observation is closer to MOS than IFS. It explains that the difference distribution between the observed value and the average is better represented by MOS. IFS is not better because the range of deviation is too small. In contrast, at Halim Perdana Kusuma, standard deviation of the observation is more relevant with IFS.

3.2.2. Relative Humidity

Based on figure 5, it appears that both IFS and MOS have similar fluctuation pattern to the observation. However, in some examples, IFS produces zero value that is actually impossible to obtain in observation. IFS also tends to much overestimate compared to MOS. Based on the graph analysis, the ability of MOS to predict humidity is better than IFS.

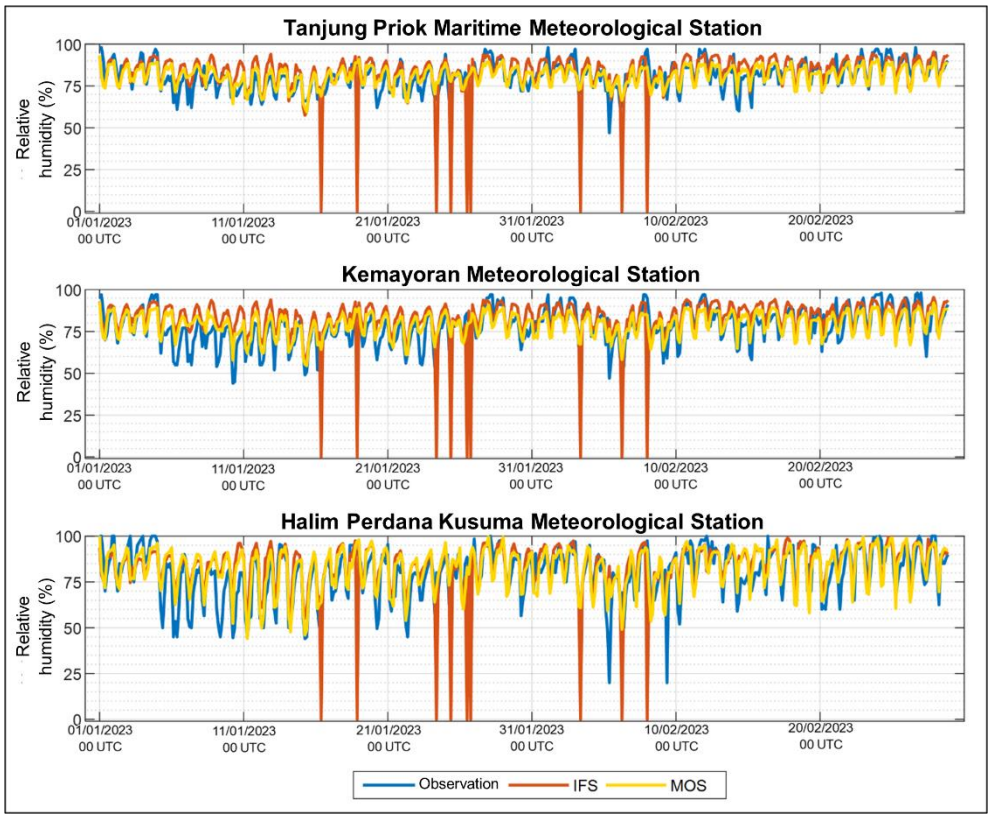
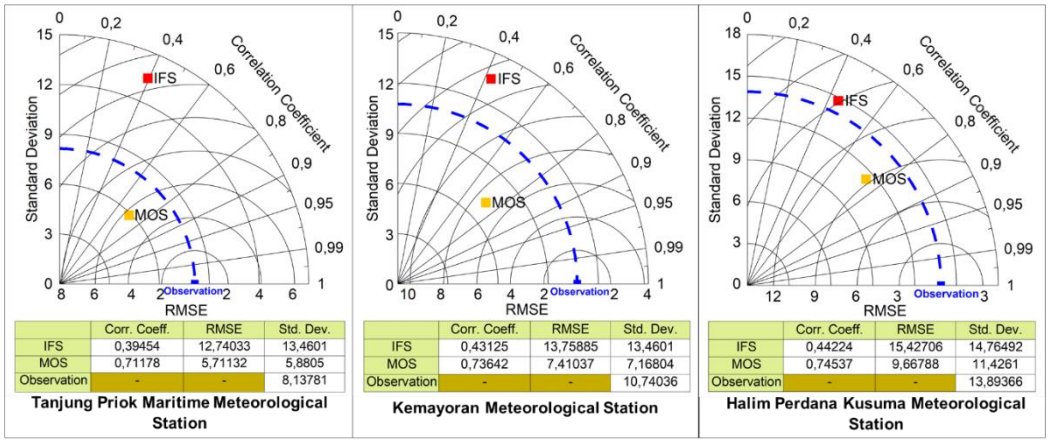


Figure 5. Comparison graph between observation, IFS, and MOS for relative humidity



	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,39454	12,74033	13,4601
MOS	0,71178	5,71132	5,8805
Observation	-	-	8,13781

	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,43125	13,75885	13,4601
MOS	0,73642	7,41037	7,16804
Observation	-	-	10,74036

	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,44224	15,42706	14,76492
MOS	0,74537	9,66788	11,4261
Observation	-	-	13,89366

Figure 6. Taylor Diagram between observation, IFS, and MOS for relative humidity

Moreover, Taylor diagram in figure 6 shows that the IFS and MOS correlation values are quite far apart. In all three locations, the correlation coefficients of MOS are much higher than IFS. Based on the categories in table 2, MOS correlation coefficients are in the high relationship category. Meanwhile, IFS correlation coefficients are in the category of existing but low relationship (Tanjung Priok Maritime Meteorological Station) and moderate relationship (Kemayoran Meteorological Station and Halim Perdana Kusuma Meteorological Station). In terms of RMSE, MOS are lower so it tends to produce smaller error than IFS. At Tanjung Priok Maritime Meteorological Station, standard deviation of observation is closer to MOS. In contrast, at Kemayoran Meteorological Station and Halim Perdana Kusuma Meteorological Station, the standard deviations of observation are closer to IFS. It means that IFS is not able enough to produce prediction with variance close to the observation.

3.2.3. QFF Pressure

The comparison between observation, IFS, and MOS of QFF Pressure is shown in figure 7. IFS and MOS have similar values and even their graphs seem to overlap. Both are good enough to represent the actual QFF pressure fluctuation. Based on the analysis of the graph, IFS and MOS both have superior performances so it cannot be determined yet which ones are better.

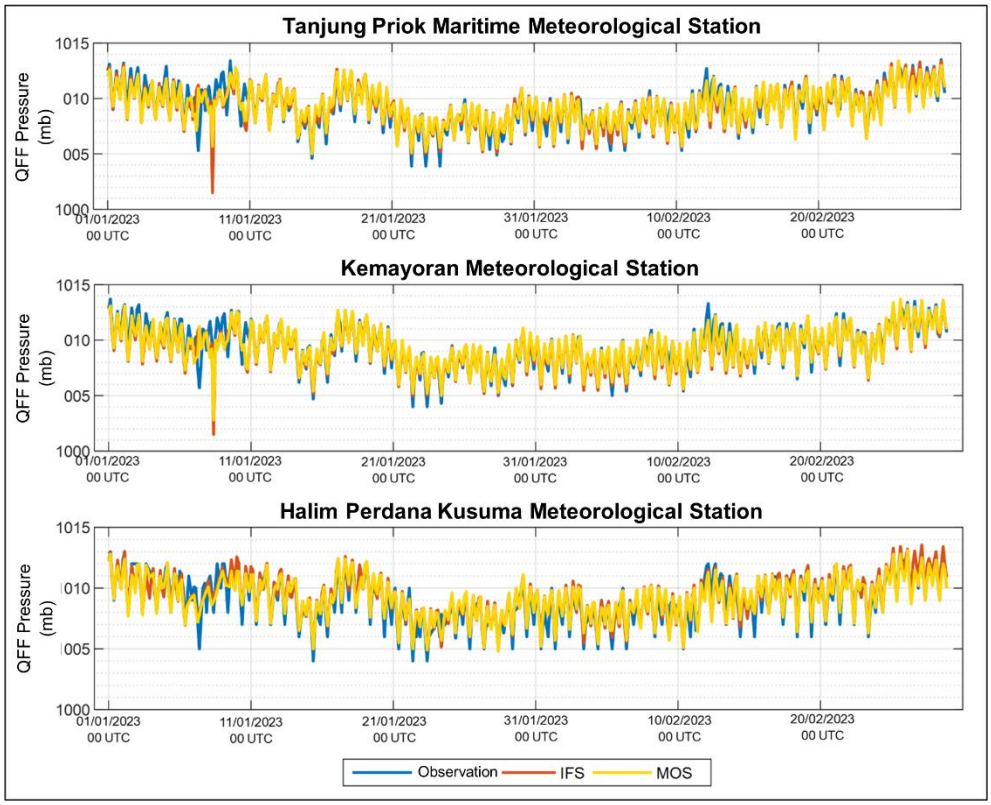
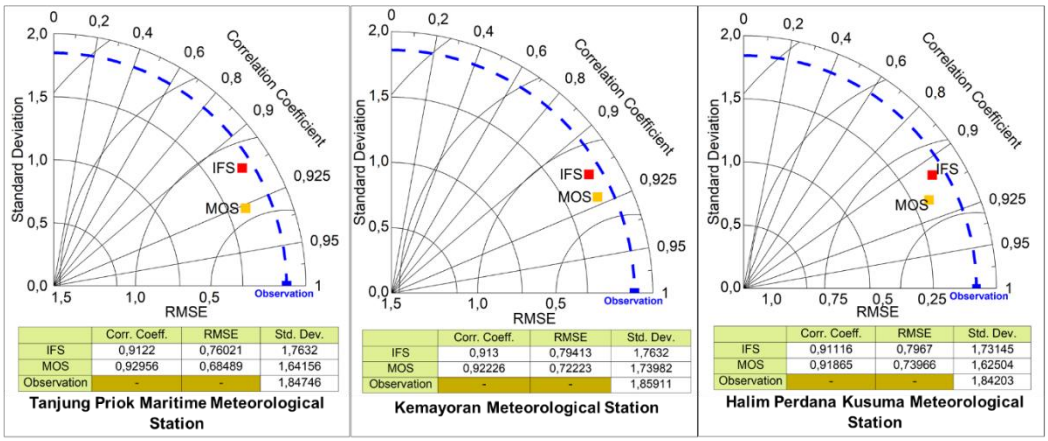


Figure 7. Comparison graph between observation, IFS, and MOS for QFF Pressure



	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,9122	0,76021	1,7632
MOS	0,92956	0,68489	1,64156
Observation	-	-	1,84746

	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,913	0,79413	1,7632
MOS	0,92226	0,72223	1,73982
Observation	-	-	1,85911

	Corr. Coeff.	RMSE	Std. Dev.
IFS	0,91116	0,7967	1,73145
MOS	0,91865	0,73966	1,62504
Observation	-	-	1,84203

Figure 8. Taylor Diagram between observation, IFS, and MOS for QFF Pressure

Based on Taylor diagram in figure 8, MOS has better relationship with observation than IFS, which is characterized by larger correlation coefficient of MOS. In terms of RMSE, MOS tends to produce smaller error than IFS because the RMSE values of MOS are lower. At three locations, the values of observation standard deviation are closer to the IFS standard deviation. It means that MOS is not able enough to produce prediction with variance that is close to the observation.

3.3. Cases Test

3.3.1. Case of January 1, 2023

Based on figure 9, MOS temperature prediction tends to produce more stable and smoother value than observation. The residual graph also proves that the difference between MOS prediction and observation is between 0 ºC to ±3.4 ºC. When applied in real life, the temperature prediction provides quite suitable result although it is necessary to pay attention to the possibility of residual value.

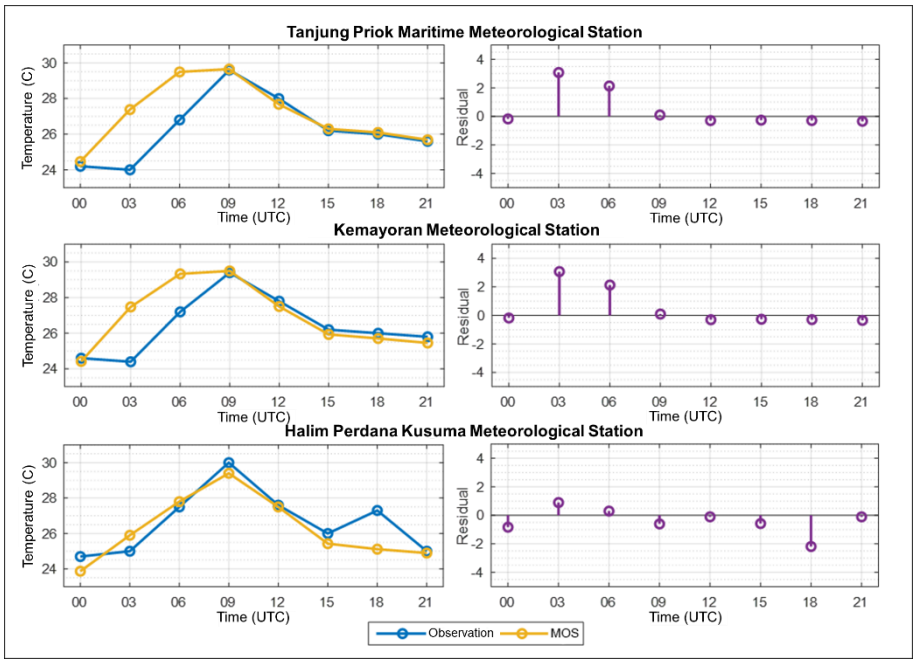


Figure 9. Graph of observation and MOS for temperature on January 1, 2023

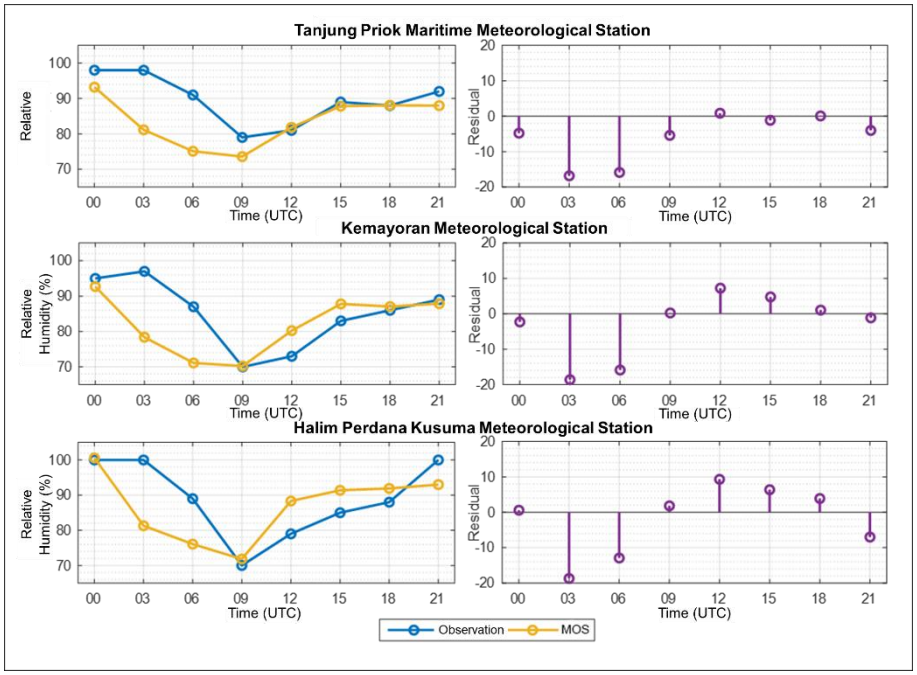


Figure 10. Graph of observation and MOS for relative humidity on January 1, 2023

The relative humidity pattern in figure 10 is also quite representative. However, MOS does not provide the precise value because the predicted value tends to be more sloping and smoother. Both lower and higher values potentially overestimate and underestimate. The residual value ranges from 0% to ±19%. MOS is able to predict relative humidity quite well.

Next, QFF pressure result described in figure 11. MOS prediction pattern is quite precise to observation. The graphs appear to almost overlap which means that the values are not much different. The residual value is the lowest among the other two cases, from 0 mb to ±0.6 mb. The application of the MOS prediction for QFF pressure in this case is considered quite appropriate. However, it should be noted that differences within this range remain crucial for aviation meteorology.

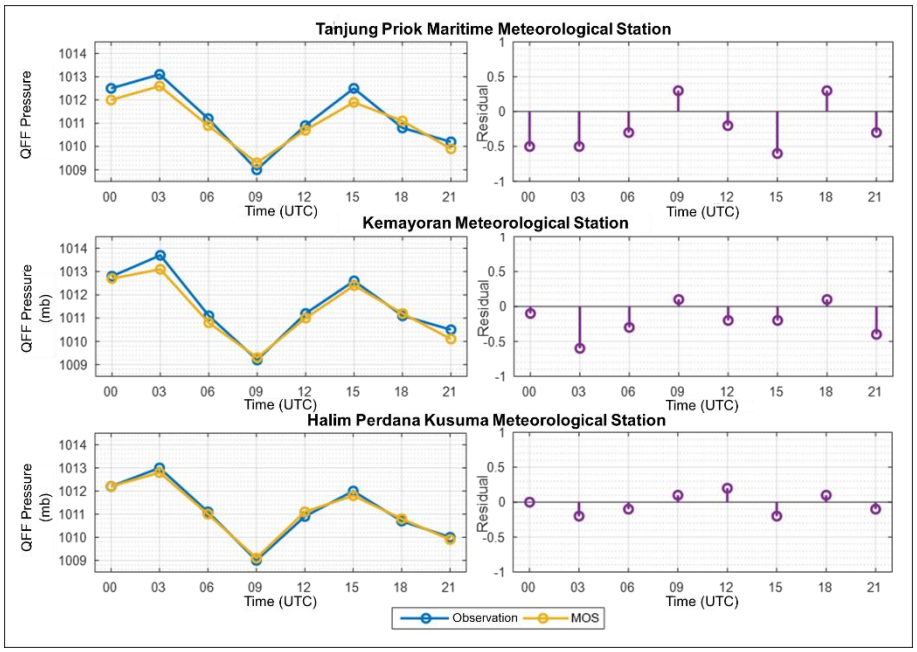


Figure 11. Graph of observation and MOS for QFF Pressure on January 1, 2023

3.3.2. Case of January 4, 2023

Based on figure 12, MOS tends to produce temperature value that is more sloping and smoother than observation. The residual graph also proves that the difference between MOS prediction and observation is between 0 ºC to ±3.3 ºC, the lowest of the two cases. When applied in real life, the temperature prediction is quite appropriate, although considering the possible residual value.

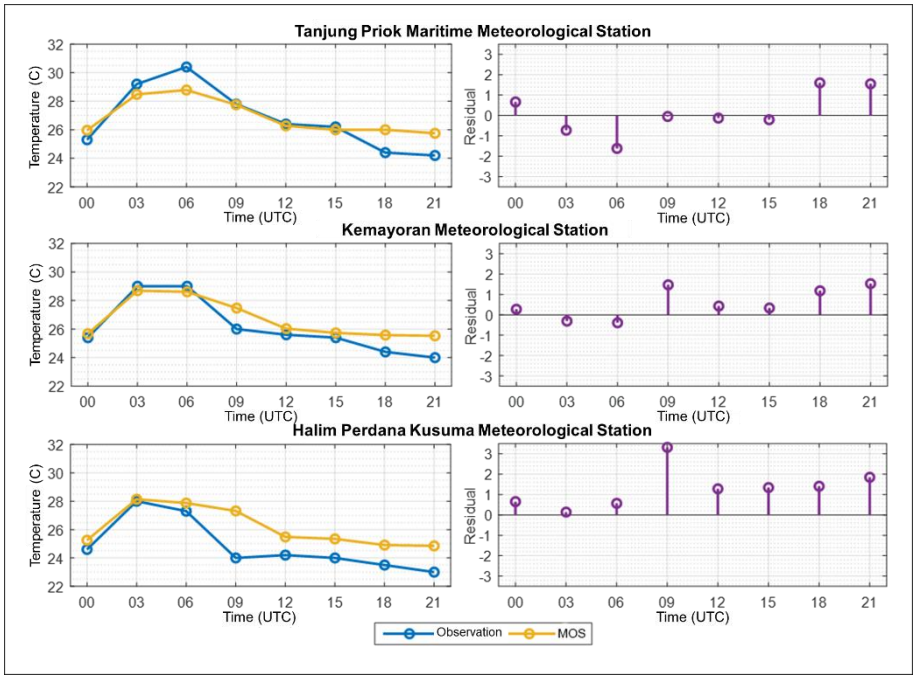


Figure 12. Graph of observation and MOS for temperature on January 4, 2023

. Relative humidity pattern that is described by figure 13 is also quite representative. However, MOS is not able enough to provide the right value because the prediction value tends to be sloping. This case has the lowest residual, which the value between 0% and ±9%. MOS is able to predict the relative humidity quite well.

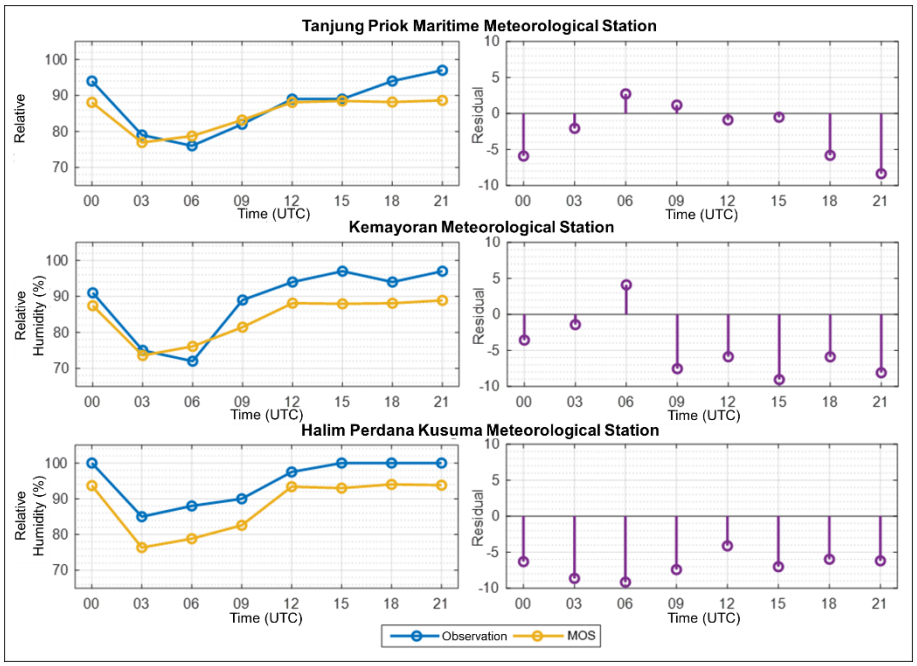


Figure 13. Graph of observation and MOS for relative humidity on January 4, 2023

Next, there is QFF pressure which is described in figure 14. MOS prediction pattern is quite consistent with the observation. The graphs appear to almost overlap, which means that the values are not much different. The residual value ranges from 0 mb to ±1 mb. The application of MOS prediction for QFF pressure in this case is considered quite appropriate.

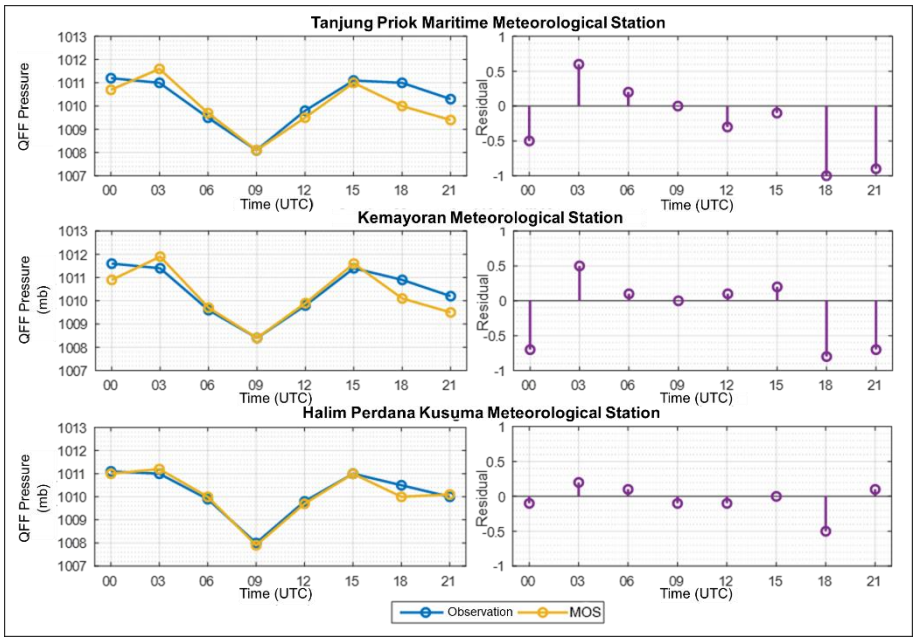


Figure 14. Graph of observation and MOS for QFF Pressure on January 4, 2023

3.3.3. Case of February 24, 2023

Based on figure 15, MOS tends to produce temperature value that is more sloping and smoother than observation. The residual graph also proves that the difference between MOS prediction and observation is between 0 ºC to ±3.5 ºC, the highest among the other two cases. When applied in real life, the temperature prediction is quite suitable although it is necessary to pay attention to the possible residual value.

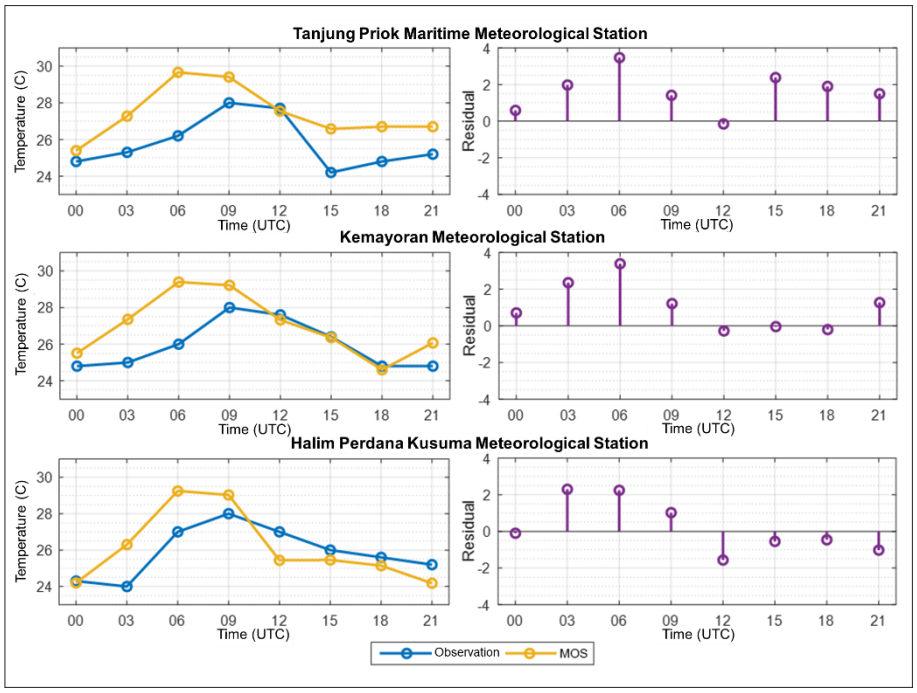


Figure 15. Graph of observation and MOS for temperature on February 24, 2023

The relative humidity that is described by figure 16 is also quite representative in pattern. However, MOS does not provide the right value because the prediction value tends to be stable. The residual value ranges from 0% to ±25%, the highest among the other two cases. The MOS prediction result is quite appropriate although it is necessary to pay attention to the possible deviation.

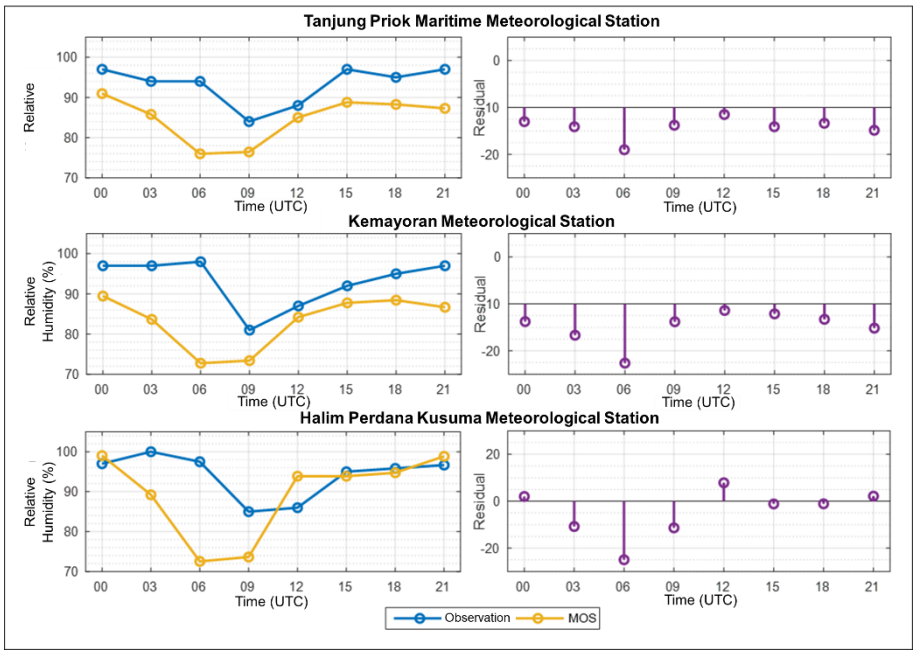


Figure 16. Graph of observation and MOS for relative humidity on February 24, 2023

Next, figure 17 describes QFF parameter result. MOS prediction gives a pattern that is quite consistent with observation. The graphs appear to almost overlap which means that the values are not much different. This case has residual value between 0 mb to ±1 mb just like the case of January 4, 2023. The application of MOS prediction for QFF pressure in this case is considered quite appropriate.

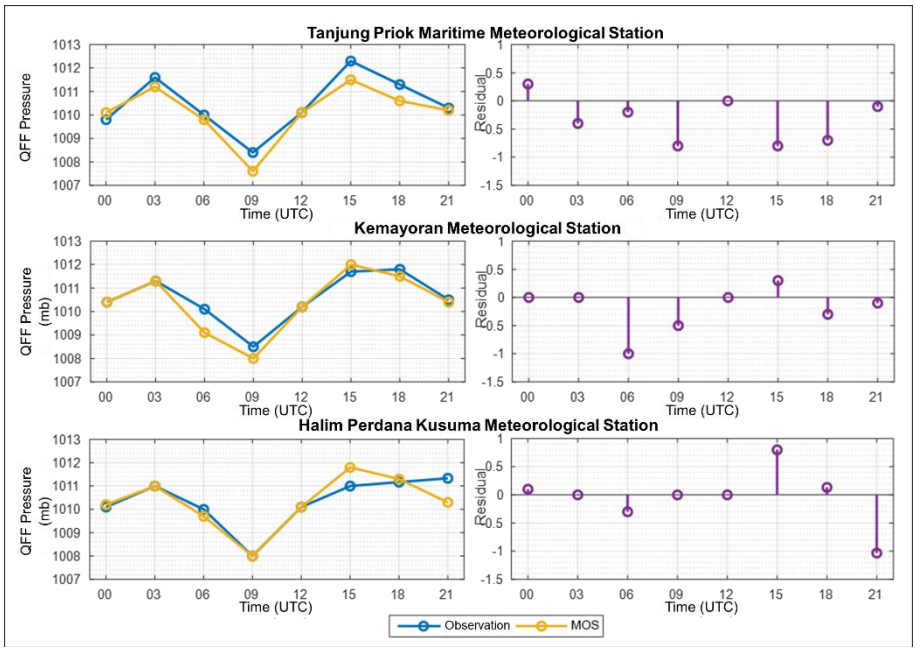


Figure 17. Graph of observation and MOS for QFF Pressure on February 24, 2023

4. Conclusion

Based on the explanation of the research within the section of results and discussion, the conclusions are as follows.

1. Stepwise regression produces MOS regression model containing the influential parameters for prediction based on the p-value with the highest correlation coefficient. The most influential parameter for temperature is parameter 2t because it is a model of the temperature measurement itself. Relative humidity is most influenced by 2t because temperature and relative humidity is directly proportional. Meanwhile, msl is the most influential parameter for QFF pressure prediction because it is a model of the measurement itself.
2. The MOS performance test generally has superior result and is able to improve the accuracy of IFS prediction. Based on the graphs, MOS prediction results are closer to observation than IFS. Verification of the correlation coefficient and RMSE proves that MOS has higher closeness relationship and lower error rate. However, the standard deviation of MOS for most parameters is not better than IFS. It is because MOS prediction result tends to be more stable with narrower range of value.
3. The result of the heavy rain cases test show that the application of MOS is able to provide fairly accurate prediction while still considering the residual value. The highest residual for temperature is ±3.5 oC, relative humidity is ±25%, and QFF pressure is ±1 mb.

Moreover, there are several suggestions that can be applied to further research, including:

1. Updating data regularly to get the most suitable regression equation and MOS prediction result.
2. Conducting trial on cases with different locations and times to prove the accuracy of MOS prediction.

Ethics approval

Not required

Acknowledgments

We would like to express our heartfelt thanks to everyone who contributed to this research. We are especially grateful to Pusat Meteorologi Publik Badan Meteorologi Klimatologi dan Geofisika (BMKG), Tanjung Priok Maritime Meteorological Station, Kemayoran Meteorological Station, and Halim Perdana Kusuma Meteorological Station for supplying the crucial data necessary for this study.

Our appreciation also extends to the reviewers and proofreaders for their thorough work and insightful feedback, which have greatly improved the quality of this publication.

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

Isnaini Ramadhan: Conceptualization, Methodology, Validation, Formal Analysis, Writing – Original Draft, Visualization. **Deni Septiadi:** Supervision, Writing- Reviewing and Editing.

References

[1] Badan Nasional Penanggulangan Bencana, "Infografis Bencana Tahun 2022," ["Disaster Infographic 2022"], *Badan Nasional Penanggulangan Bencana*, 2023, [Online]. Available: https://www.bnpb.go.id/infografis/infografis-bencana-tahun-2022. [Accessed: Mar. 7, 2023].

[2] S. Rahmstorf and D. Coumou, "Increase of extreme events in a warming world," In Proc. National Academy of Sciences, 2011, pp. 17905–17909.

[3] Kiki and F. Alam, "Verifikasi parameter presipitasi akumulasi 24 jam pada model cuaca numerik tahun 2017—2020," ["Verification of 24-hour accumulated precipitation on numerical weather prediction model year 2017—2020"], *Megasains*, no. 12, vol. 2, pp. 11—16, Aug. 2021.

[4] J. A. Garcia-Moya, J. L. Casado, I. Martínez, A. Manzao, A. Martín, C. Fernández-Peruchena, M. Gastón. "Deterministic and probabilistic weather forecasting," Agencia Estatal de Meteorología, Spain, Report PREFLEXMS_DEL_D4.3_20160531_v4, 30 Jul. 2016.

[5] H. R. Glahn and D. A. Lowry, "The use of model output statistics (MOS) in objective weather forecasting," *Journal of Applied Meteorology and Climatology*, vol. 11, no. 8, pp. 1203—1211, Dec. 1972.

[6] Nurhayati, "Model output statistic dengan CART dan random forest untuk prakiraan curah hujan harian," ["Model output statistics in daily rainfall forecasting with CART and random forest"], B. S. thesis, Institut Teknologi Sepuluh November, Surabaya, 2017.

[7] N. Brunet, R. Verret, and N. Yacowar, "An objective comparison of model output statistics and "perfect prog" systems in producing numerical weather element forecasts," *Weather and Forecasting*, vol. 3, no. 4, pp. 273—283, Dec. 1988.

[8] Deutscher Wetterdienst, "Model Output Statistics-MIX (MOSMIX)," *Deutscher Wetterdienst: Wetter und Klima aus einer Hand*, 2015, [Online]. Available: https://www.dwd.de/EN/ourservices/met_application_mosmix/met_application_mosmix.html. [Accessed: Apr. 7, 2024].

[9] N. Qona'ah, H. Pratiwi, and Y. Susanti, "Model output statistic dengan principal component regression, partial least square regression, dan ridge regression untuk kalibrasi prakiraan cuaca jangka pendek" ["Model output statistics with principal component regression, partial least square regression, and ridge regression in short-term weather forecast callibration"], *Jurnal Matematika UNAND*, vol. 10, no. 3, pp. 355—368, Jul. 2021.

[10] R. Safitri and Sutikno, "Model output statistics dengan projection pursuit regression untuk meramalkan suhu minimum, suhu maksimum, dan kelembapan" ["Model output statistics with

projection pursuit regression to minimum temperature, maximum temperature, and humidity prediction"], *Jurnal Sains dan Seni ITS*, vol. 1, no. 1, pp. 296–301, Sep. 2012.

[11] J. R. Bocchieri, R. L. Crisci, H. R. Glahn, F. Lewis, and F. T. Globokar, "Recent developments in automated prediction of ceiling and visibility," *Journal of Applied Meteorology*, vol. 13, no. 2, pp. 277—288, Mar. 1974.

[12] W. H. Klein and H. R. Glahn, "Forecasting local weather by means of model output statistics," *Bulletin of the American Meteorological Society*, vol. 55, no. 10, pp. 1217—1227, Oct. 1974.

[13] R. J. Kuligowski and A. P. Barros, "Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks," *Weather and Forecasting*, vol. 13, no. 4, pp. 1194—1204, Dec. 1998.

[14] Badan Penanggulangan Bencana Daerah DKI Jakarta, "Infografis Kejadian Bencana Provinsi DKI Jakarta Tahun 2021," ["Disaster Infographic in DKI Jakarta Province Year 2021"], *Badan Penanggulangan Bencana Daerah DKI Jakarta*, 2022, [Online]. Available: https://bpbd.jakarta.go.id/infografis/30/hasil-rekapitulasi-kejadian-bencana-yang-ada-di-dki-jakarta-pada-tahun. [Accessed: Mar. 3, 2023].

[15] Badan Pusat Statistik Jakarta, "Jumlah Penduduk Menurut Kabupaten/Kota di Provinsi DKI Jakarta (Jiwa), 2020-2022," ["Number of Population by Regency/Municipality in DKI Jakarta Province, 2020-2022"], *Badan Pusat Statistik Jakarta*, 2021, [Online]. Available: https://jakarta.bps.go.id/indicator/12/1270/1/jumlah-penduduk-menurut-kabupaten-kota-di-provinsi-dki-jakarta-.html. [Accessed: Mar. 1, 2023].

[16] J. R. Holton and G. J. Hakim, "An *Introduction to Dynamic Meteorology*, 5th ed., Amsterdam, Netherlands: Academic Press, 2013.

[17] V. Bjerknes, "The problem of weather prediction, considered from the view points of mechanics and physics," *Meteorologische Zeitschrift*, vol. 18, no.6, pp. 663—667, Dec. 2009.

[18] T. N. Palmer, "Predictability of weather and climate: from theory to practice - from days to decades," In Proc. ECMWF Workshop on the Use of High Performance Computers in Meteorology '10, 2003, pp. 1—29.

[19] J. Flemming, V. Huijnen, J. Arteta, P. Bechtold, A. Beljaars, A. M. Blechschmidt, M. Diamantakis, R. J. Engelen, A. Gaudel, A. Inness, L. Jones, B. Josse, E. Katragkou, V. Marecal, V. H. Peuch, A. Richter, M. G. Schultz, O. Stein, and A. Tsikerdekis, "Tropospheric chemistry in the integrated forecasting system of ECMWF." *Geoscientific Model Development*, vol. 8, no. 4, pp. 975—1003, Apr. 2015.

[20] N. Wedi, P. Bauer, W. Deconinck, M. Diamantakis, M. Hamrud, C. Kuehnlein, S. Malardel, K. Mogensen, G. Mozdzynski, and P. Smolarkiewicz, "The modelling infrastructure of the integrated forecasting system: recent advances and future challenges," European Centre for Medium-Range Weather Forecasts, Reading, UK, Tech. Memorandum. 760, Nov. 2015.

[21] F. M. Lopes, R. Conceição, H. G. Silva, T. Fasquelle, R. Salgado, P. Canhoto, and M. Collares-Pereira, "Short-term forecasts of dni from an integrated forecasting system (ECMWF) for optimized operational strategies of a central receiver system," *Energies*, vol. 12, no. 7, pp. 1—18, Apr. 2019.

[22] European Centre for Medium-Range Weather Forecasts, "ECMWF IFS CY41r2 High-Resolution Operational Forecasts," *UCAR/NCAR - Research Data Archive*, 2016, [Online]. Available: http://rda.ucar.edu/datasets/ds113.1/. [Accessed: Feb. 15, 2023].

[23] European Centre for Medium-Range Weather Forecasts, "Operational Configurations of the ECMWF Integrated Forecasting System (IFS)," *UCAR/NCAR - Research Data Archive*, 2016, [Online]. Available: https://confluence.ecmwf.int/pages/viewpage.action?pageId=324860211. [Accessed: Feb. 15, 2023].

[24] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, G. M. Ljung. *Time Series Analysis: Forecasting and Control*, 5th ed., Hoboken, NJ: Wiley, 2015.

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.789

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

Development of a Hybrid Fuzzy Geographically Weighted K-Prototype Clustering and Genetic Algorithm for Enhanced Spatial Analysis: Application to Rural Development Mapping

Agung Budi Santoso[1*], Arya Candra Kusuma[2], Rani Nooraeni[3], Arie Wahyu Wijayanto[4]

[1]BPS-Statistics Bima City, Indonesia, [2]Department of Statistics, Politeknik Statistika STIS, Jakarta, Indonesia, [3,4]Department of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address:* agung.budi@bps.go.id

ARTICLE INFO

Article history:
Received 16 August, 2024
Revised 29 October, 2024
Accepted 9 November, 2024
Published 31 December, 2024

Keywords:
Clustering; Geographically Weighted Cluster; Mixed-type data; Village Development; Genetic Algorithm

Abstract

Introduction/Main Objectives: Clustering methods are crucial for geodemographic analysis (GDA) as they enable a more accurate and distinct characterization of a region. This process facilitates the creation of socio-economic policies and contributes to the overall advancement of the region. **Background Problems:** The fuzzy geographically weighted clustering (FGWC) method, which is a GDA technique, primarily handles numerical data and is prone to being stuck in local optima. **Novelty:** This study proposed two novel clustering methodologies: fuzzy geographically weighted k-prototypes (FKP-GW) and its hybrid clustering model, which combines genetic algorithm-based optimization (GA-FKP-GW). **Research Methods:** This research conduct simulation study comparing two of the proposed clustering method. For the empirical application, this study applied clustering technique using the official Village Potential Survey of Temanggung, Indonesia. **Finding/Results:** The evaluation results of experiments conducted on simulated data and study cases indicate that the proposed method yields distinct clustering results compared to the previous method while being comparably efficient. The empirical application identifies four distinct groups from the clustered villages, each displaying unique characteristics. The results of our research have the potential to benefit the development of the GDA method and assist the local government in formulating more effective development policies.

1. Introduction

Clustering methods are currently crucial for analyzing an area. Employing appropriate clustering techniques enables a more precise and distinctive description of the characteristics of an area and distinguishes between different groups. Hence, it is imperative to conduct a thorough investigation into selecting appropriate techniques to achieve optimal clustering outcomes. Geo-demographic analysis (GDA) is a commonly employed method for analyzing a specific area. Geodemography is the study and examination of the attributes or trends of individuals or inhabitants, specifically in relation to their geographical location [1]. The fuzzy geographically weighted clustering (FGWC) method is an appropriate clustering technique for geo-demographic analysis (GDA) as it takes into account the spatial variations in population size and distance between regions [2]. However, the exclusive restriction of the

current FGWC method to numerical data presents a challenge when clustering mixed-type geographic data.

The primary goal of GDA is to create clusters or groups based on socioeconomic status within a particular region. This process aids in the formulation of socio-economic policies and the overall development of the region [3]. Geographic data analysis uses Fuzzy Geographically Weighted Clustering (FGWC), a highly effective clustering method. Mason and Jacobson [4] developed this method by combining the fuzzy C-means (FCM) method with the neighborhood effect (NE). This integration allows the method to give greater consideration to the geographical impact of each data element [4]. However, it is similar to the previous approach, known as FCM, which is exclusively suitable for numerical data [5], [6], [7], [8]. This approach typically employs the Euclidean distance as its cost function, which is suitable for measuring distances between numerical data but not for categorical data [9], [10], [11].

The K-prototype algorithm, developed by Huang [12], manages mixed-type data clustering efficiently. This algorithm integrates the K-means algorithm with K-modes [12]. The K-Prototype is a clustering technique that makes use of partial clustering [13]. For clustering mixed data types, the K-prototype algorithm is an efficient and scalable algorithm [12]. Nevertheless, the K-Prototype method still has limitations when it comes to determining the centroid for categorical data [14]. The centroid determination for categorical attributes in the K-Prototype (KP) algorithm still depends on the mode value of those attributes. Because the mode value is used as the centroid, it is unable to provide an accurate description of objects [11], [13].

In 2012, Ji, et al. [14] introduced a technique called fuzzy K-prototype (FKP) that combines fuzzy clustering or soft clustering with the K-prototype algorithm. Unlike the previous method, FKP does not use the mode value as the centroid. Instead, it incorporates a fuzzy centroid into the algorithm [14].The Fuzzy K-Prototype method, a partition-based clustering technique, also encounters issues with the randomly selected initialization centroid. This will result in the method being stuck in the local optimum solution [12], [15].

To address the issue of a local optimum solution, one effective approach is to employ the metaheuristic method [15], [16], [17], [18], [19], [20], [21]. A metaheuristic method that can be utilized is the genetic algorithm, which operates based on Charles Darwin's theory of natural selection [16], [22], [23]. A genetic algorithm is a metaheuristic algorithm that has undergone extensive development and modification to optimize its performance by effectively balancing the trade-off between exploitation and exploration [16]. This algorithm is commonly referred to as a search algorithm that involves fewer mathematical computations compared to other algorithms [24].

Several previous studies have improved the methodology for geodemographic clustering techniques and clustering mixed data. From the FGWC development that consider spatial aspect but limited to numerical data, prior study have improved optimization by using numerous metaheuristic optimization, namely Artificial Bee Colony (ABC) [25], Ant Colony Optimization (ACO) [26], Gravitation Search Algorithm (GSA) [27], Intelligent Firefly Optimization (IFO) [19], and Chaotic Flower Pollination Algorithm (CFPA) [28]. From the clustering mixed data, previous research also developed the K-Protype method. Because the determination of the initial centroid has a great influence on the clustering solution, Nooraeni et al. were conducted to optimize the cluster center initialization using the K-Prototype method with the GA algorithm (GA-KP) [29], [30]. Nooraeni et al. have also enhanced the GA-KP algorithm by resolving the issue of using the mode value as a centroid for categorical data, as well as addressing the problem of finding the local optimum solution by utilizing a fuzzy centroid [31]. However, the development of a hybrid between FGWC and FKP methods using genetic algorithm optimization remains unexplored.

Therefore, this research will develop a hybrid method between FGWC and FKP to overcome the weakness of mixed-type geographic data and optimizing them using a Genetic Algorithm. This integrated approach enables the analysis of mixed data clusters with spatial considerations, effectively overcoming local optima. Furthermore, this study is organized as follows: The Methodology section explains the proposed clustering methods, namely FKP-GW and GA-FKP-GW. To assess the efficiency of the proposed clustering, the methodology section also explains the evaluation through simulation data and empirical data. The Results section explains the application of the FKP-GW and GA-FKP-GW clustering methods. Finally, the Conclusion section summarizes the results of this research, including the development of geodemographic clustering methods and their application for regional development.

2. Material and Methods

2.1 Data

The study utilizes both simulated and empirical data to evaluate the hybrid fuzzy geographically weighted K-prototype clustering and genetic algorithm method. Simulated data are generated to assess the algorithm's performance in a controlled environment, allowing for a clear examination of clustering behavior. Real-world data, on the other hand, is sourced from Indonesia's Village Development Index (VDI) [32] and provides practical context by applying the method to rural development mapping.

2.1.1 Simulated Data

Two sets of simulated data are generated to test the clustering algorithm. The first set consists of random data with seven attributes, including four categorical and three numerical variables. The second set expands on this with eight attributes (four categorical and four numerical). Both of these sets of simulated data are generated with 50 observations and 100 observations, resulting in 4 scenarios of simulation. These simulated datasets allow for an exploration of the algorithm's ability to manage and cluster mixed data types, testing its robustness in assigning clusters effectively under varied attribute structures and sample sizes.

2.1.2 Empirical Data

For implementing the algorithm in the real-world case, the proposed method will be applied to clustering the villages of Temanggung Regency based on indicators of the village development index (VDI). The BPS-Statistic Indonesia aims to find out the potential of each village based on several indicators of socio-economic development. In this study, the term "village" includes the nagari, Transmigration Settlement Unit, and Entity of Transmigration Settlement, which are still fostered by the relevant ministries and government agencies of Indonesia. The Village Development Index (VDI), also known as Indeks Pembangunan Desa (IPD), describes the development progress of a village at a specific time [32]. The VDI calculation is obtained from the results of the Village Potential Statistics (PODES) data collection. The calculation of VDI from PODES 2018 data involves 5 dimensions and 42 indicators that document the availability of infrastructure and service accessibility [33].

The BPS publication on the Village Development Index 2018 reveals that there are 5,606 independent villages, 55,369 developing villages, and 14,461 underdeveloped villages [32]. The publication also explained that the island of Java—Bali became the island with the highest VDI. Despite being the island with the highest VDI, Central Java, one of the provinces, has an VDI that is lower than the average VDI for the island of Java-Bali. Temanggung Regency, one of Central Java's regencies, is actively involved in the design of the Village Development Work Plan (RKPD). It is proven by his achievement of winning a second-place award in National District Level Development 2019. However, the VDI value of Temanggung Regency in 2018 was 66.05, indicating that it remains below the average VDI of Central Java Province. This indicates that Temanggung Regency has the potential to enhance its development plan by identifying the potential of each village within its territory. Thus, clustering was carried out for mapping villages in the Temanggung Regency area to determine the characteristics of each village so that development planning could be more targeted.

The indicators of VDI were collected from PODES 2018, which consists of 289 villages with 88 numerical attributes and 94 categorical attributes. The study then uses the population of each village as the geographic effect variable, and the centroid point of the shapefile for each village as the distance between village areas.

2.1.3 Data Preprocessing

The beginning step of the pre-processing. For the numerical attributes $\boldsymbol{x}^r$, standardized using the min–max method. For the function min-max that used in this research can be written in equation (1).

$$x_{ij}'^{\,r} = \frac{x_{ij}^r - \min(x_j^r)}{\max(x_j^r) - \min(x_j^r)} \tag{1}$$

Where $x_{ij}'^{\,r}$ and x_{ij}^r are the new numerical attribute data standardized and the real old numerical attribute data, min and max x_j^r are the minimum and maximum value in j-th numerical attributes. Then, change the numeric value to be categorical or called discretized as much as the specified number of T intervals. This approach also used by Ji., et al. [14] to develop fuzzy k-prototype algorithm.

This process pre-processing is carried out for all data used in the FKP, FKP-GW, and GA-FKP-GW. For variables distance in geographically weighted compute the distance of each area observation are using function spDists() from package "sp" in Rstudio [34], which can return the result as matrix distance every area observation.

2.2 Clustering Methods

The clustering methods are divided into two types, hierarchical clustering and partitional clustering. In hierarchical clustering, data is grouped hierarchical or stratified. While the partitional clustering methods data is grouped into several clusters without any hierarchical structure [35], [36]. So, the number of clusters must be determined from the beginning step of processing. Determining the number of clusters can be done in various ways, drawing a scree plot to see the significant decrease of the total cost function for each cluster can be one alternative of determination [37], [38].

2.2.1 Fuzzy Geographically Weighted Clustering

FGWC is a method developed by Mason & Jacobson [4] that integrates fuzzy clustering with Neighborhood Effects by taking into account the distance between regions and their populations which makes them more sensitive to neighboring effects and the results clusters are more geographically aware. The FGWC method performs a new calculation for its membership value in each iteration using the equation (2) [4].

$$\boldsymbol{\mu}_i' = \alpha \boldsymbol{\mu}_i + \beta \frac{1}{A} \sum_{j=1}^{n} w_{ij} \boldsymbol{\mu}_j \tag{2}$$

Where $\boldsymbol{\mu}_i'$ is the new membership degree value of the i-th region, while $\boldsymbol{\mu}_i$ is the old membership degree value of the i-th region and $\boldsymbol{\mu}_j$ is the value of the j-th region old membership degree. α and β are the weights of the old membership value and new membership of other objects whose sums are equal by 1. A is the value that ensures the weighting of the membership value in the range of zero and one [0,1]. w_{ij} is the weights between the two geographic areas can be written in equation (3).

$$w_{ij} = \frac{(m_i m_j)^b}{d_{ij}^a} \tag{3}$$

With the m_i and m_j is the number of population in the region of the i and j, while d_{ij} is the distance between regions i and j. a is the magnitude of the interaction effect of the distance between regions and b is the magnitude of the population interaction between the two regions. Both parameters can be determined by the researcher.

2.2.2 K-Prototype

This method is a partitional clustering developed from the K-means method, which previously could only be used for numeric data types but is still maintaining its efficiency. The KP algorithm is still classified as hard clustering with each object can be grouped into one cluster only. The cost function for mixed-type data can be written in equation (4) [14].

$$\begin{aligned} E &= \sum_{l=1}^{k} (E_l^r + E_l^c) \\ &= \sum_{l=1}^{k} \left(\sum_{i=1}^{n} \mu_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{i=1}^{n} \mu_{il} \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right) \right) \end{aligned} \tag{4}$$

Where E_l^r and E_l^c is total cost from every numeric and categorical attribute from cluster l, respectively. x_{ij}^r is the value of object i in j-th numeric attributes set, x_{ij}^c is value of object i in j-th categorical attributes set, and γ_l is a weight for categorical attribute in the cluster l. q_{lj}^r and q_{lj}^c is the prototype or

centroid of every numeric and categorical attribute, respectively. This algorithm cover in function kproto() from package "clustMixType" [39] in Rstudio.

2.2.3 Fuzzy K-Prototype

Ji et al. [14] developed the Fuzzy K-Prototype clustering method for mixed data, utilizing the KP algorithm as their foundation. This method uses fuzzy clustering to find the centroid of the numeric and categorical attributes of the cluster. It also uses new dissimilarity measures that look at the significance level of each numeric attribute and the distance between the values of the categorical attributes. The cost function equation for Fuzzy K-Prototype can be written in equation (5) [14].

$$E(\boldsymbol{\mu}, \mathbf{Q}) = \sum_{l=1}^{k}\left(\sum_{i=1}^{n}\mu_{il}^{\alpha}\left(\sum_{j=1}^{m_r}\left(\omega_l\left(x_{ij}^{r} - v_{jl}^{r}\right)\right)^2 + \sum_{j=1}^{m_c}\varphi\left(x_{ij}^{c}, \tilde{v}_{jl}^{c}\right)^2\right)\right) \tag{5}$$

Notation $\boldsymbol{\mu}$ represents the matrix of cluster degree membership and $\mathbf{Q}$ represents the matrix of cluster centroid. Where centroid for numeric value v_{jl}^{r} and fuzzy centroid categorical value $\tilde{v}_{jl}^{c}$ can written in equation (6) and equation (7), respectively [14].

$$v_{jl}^{r} = \frac{\sum_{i=1}^{n}(\mu_{il})^{\alpha}x_{ij}^{r}}{\sum_{i=1}^{n}(\mu_{il})^{\alpha}} \tag{6}$$

$$\tilde{v}_{jl}^{c} = \frac{a_{jl}^{1}}{\omega_{jl}^{1}} + \frac{a_{jl}^{2}}{\omega_{jl}^{2}} + \cdots + \frac{a_{jl}^{\kappa}}{\omega_{jl}^{\kappa}} + \cdots + \frac{a_{jl}^{t}}{\omega_{jl}^{t}} \tag{7}$$

Fuzzy matrix partition then updated using equation (8).

$$\mu_{il} = \frac{1}{\sum_{z=1}^{k}\left(\frac{d(x_i, q_l)}{d(x_i, q_z)}\right)^{\frac{1}{\alpha-1}}} \tag{8}$$

Where $\sum_{j=1}^{m_r}(\omega_l(x_{ij}^{r} - v_{jl}^{r}))^2$ is the calculation of the distance i-th observation to the l-th cluster centroid for numeric attributes, while $\sum_{j=1}^{m_c}\varphi(x_{ij}^{c}, \tilde{v}_{jl}^{c})^2$ is the calculation of the distance i-th observation to the l-th cluster centroid for the categorical attribute [14]. μ_{il}^{α} is a matrix membership value for i-th observation in the l-th cluster, with parameter fuzziness α whose value is determined by the researcher, while ω_l is the significance value of each numeric attribute in the l-th cluster. The FKP method used in this study employs the following algorithms [14]:

1. Determine the maximum iteration, number of clusters k, values of fuzziness parameter (α), and threshold (ε).
2. Initialization value of partition matrix membership by generating random value.
3. Calculate the centroid for numeric attributes using equation (6) and centroid for categorical attributes using equation (7).
4. Calculate the distance between observation to the centroid.
5. Update the fuzzy matrix partition membership using equation (8).
6. If the difference between the previous cost function is less than the threshold or has reached the maximum iteration, then stop process clustering. If not, return to step 3.

2.2.4 The Proposed Method: Fuzzy K-Prototype Geographically Weighted (FKP-GW)

This research develops a hybrid method to cluster mixed-type geographic data while considering spatial effects. The proposed method, Fuzzy K-Prototype Geographically Weighted (FKP-GW), uses the following algorithm:

1. The algorithm requires the input of several parameters, including data, the number of clusters k, the fuzziness coefficient m, the maximum iteration, and the threshold ε. Additionally, it requires the input of several geographically weighted parameters, including the matrix population, the matrix distance for each observation, α, β, a, and b.
2. Generate partition matrix membership according to the predetermined number of clusters k with random values.
3. Calculate the centroid for numerical and categorical attributes using equations (6) and (7).
4. Determine the observation's distance to the obtained centroid, then add the distances for both numeric and categorical attributes.
5. Using equation (8), update the fuzzy matrix partition membership based on the previously obtained matrix distance.
6. Modify the previous matrix membership with geographical weights according to equation (2).
7. If the difference between the previous cost function is less than the threshold or has reached the maximum iteration, then stop process clustering. If not, return to step 3.

2.2.5 The Proposed Method: Genetic Fuzzy K-Prototype Geographically Weighted (GA-FKP-GW)

This research proposes the Genetic Fuzzy K-Prototype Geographically Weighted (GA-FKP-GW) as the next hybrid method. The Genetic Algorithm is a search algorithm that incorporates Charles Darwin's theory of natural selection [16], [22], [23]. It appears to operate similarly to a natural selection process, where the individuals that survive are those that can endure throughout the evolutionary process. In this method, there are several terms, such as chromosome, which consists of the candidate best solution for matrix membership; genes, which is the value of matrix membership; individual, population, generation, selection, crossover, mutation, and elitism. The proposed method, known as GA-FKP-GW, uses the following algorithm:

1. Input Parameters: Parameters that need to be inputted for GA-FKP-GW are mutation rate, maximum generation, and numbers of population, in addition to several parameters that are used in the FKP-GW algorithm,
2. Initialization individual as much as the numbers of population that contain candidate chromosomes; this chromosome contains candidate matrix membership. The pseudocode used in initialization is:

> **For** i = 1 to n **do**
>
> Generate random numbers from [0,1];
>
> For the i-th point of chromosome;
>
> Calculate
>
> **End for**

Source: Gan, et al. [17]

3. Evaluation of the fitness value of each individual by counting the cost function and evaluating it using equation (9).

$$f = \frac{1}{(h + a)} \tag{9}$$

The lowest value of the cost function shows this individual has a good potential solution.

4. Create new populations by repeating a few steps for as many as maximum generations until the new population produces the most optimal solution:
   a. Selection: The selection process involves the selection of chromosomes to facilitate the subsequent processes of crossovers and mutations. The mutation method used by the researcher is the roulette wheel. This method is done by calculating the probability value of each chromosome based on its fitness value with equation (10).

$$P_i = \frac{f_i}{f_{\text{total}}} \; i = 1,2, \ldots n_{\text{population}} \tag{10}$$

After obtaining the probability value, compute the cumulative probability value using the obtained value. To select individuals, generate a random number R in interval [0,1]. Individuals will be selected if the cumulative probability $\geq R$.

b. Crossover: This crossover process aims to enhance the diversity of chromosomes within a population. In this study, the researcher employed a one-step fuzzy K-prototype, which was weighted geographically. The pseudocode used in Crossover is:

For i=1 to N **do**

Let $\boldsymbol{\mu}_i$ be the fuzzy membership represented by s_i;

Obtain the new set of cluster centers $\boldsymbol{q}_i$ given $\boldsymbol{\mu}_i$ according to formula no 7 & 8;

Obtain the fuzzy membership $\boldsymbol{\mu}_i$ given $\boldsymbol{q}_i$ according to formula no 5;

Replace s_t with the chromosome representing $\boldsymbol{\mu}_i$;

End for

Source: Gan, et al. [17]

c. Mutation: This process involves altering one or more genes within a chromosome, resulting in the creation of a new chromosome and avoiding the issue of a local optimum. Mutations can occur when the probability of a gene being mutated is less than the probability of a mutation rate in a chromosome, and then the gene will be replaced or mutated. The pseudocode used in Mutation is:

For t=1 to N **do**

Let $(a_1, a_2, \cdots, a_{n,k})$ denote the chromosome s_t;

For i=1 to N **do**

Generate k random real number $v \in [0, 1]$;

If $v \leq pm$ **then**

Generate k random numbers $v_{i1}, v_{i2}, \ldots, v_{ik}$ from [0,1] for the i-th point of chromosome;

Replace

End if

End for

Source: Gan, et al. [17]

d. Elitism: This process stores the chromosomes with the best fitness value, with the aim of causing these chromosomes to decrease in their fitness values during the crossover and mutation processes.

5. The best candidate individual or chromosomes that give the best solution will be used as a matrix membership in algorithm FKP-GW for providing a good cluster result.

2.2.6. Indicators Evaluation Clustering Methods

To evaluate the clustering result by using some clustering methods, there are some indicators that can be used to evaluate it. The indicators that can be used are the cost function of its methods, index PC, index SC, and index CVC. By comparing these indicators, it can be used to know how good clustering methods provide clustering results. Several indicators were employed in this study.

1. Partition Coefficient (PC)
   The PC index is an index that measures the amount of overlap between clusters. The PC index value is measured using equation (11).

$$PC = \frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n}\mu_{il}^{2} \tag{11}$$

   The greater the PC index value, the better the cluster result.

2. Classification Entropy (CE)
   The CE index is an index that measures the fuzziness between clusters. The CE index value is measured using equation (12).

$$CE = -\frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n}\mu_{il}\log(\mu_{il}) \tag{12}$$

   The smaller the CE index value, the better the cluster results.

3. Categorical Variance Criterion (CVC)
   CVC is a combination method Category Utility (CU) for categorical attributes and measurement variance σ^2 for numeric attributes. The CVC formula can be written in equation (13) as follows [40].

$$CVC = \frac{CU}{1+\sigma^2} \tag{13}$$

   Where:

$$CU = \sum_{l=1}^{k}\left(\frac{|C_l|}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[P(X_j = V_{ij}|C_l)^2 - P(X_j = V_{ij})^2\right]\right)$$

$$\sigma^2 = \sum_{l=1}^{k}\frac{1}{|C_l|}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(V_{i,j}^{l} - V_{i,avg}^{l}\right)^2$$

   The higher the CVC value, the better the cluster results

3. Results and Discussion

3.1 Comparison Previous Clustering Method with Proposed Clustering Method (FKP-GW and GA-FKP-GW)

The comparison between the proposed clustering method and previous clustering methods is done using simulated data. In this research, there are two types of simulation data, namely data type 1 and data type 2. Data type 1 comprises seven attributes, while data type 2 comprises eight. Each dataset

contains 50 and 100 observations. Some parameters were determined to evaluate the results of the proposed method and compare them with the previous method. Table 1 provides details on the number cluster, parameter fuzziness, parameter alpha, and beta for each data set used in this research.

Table 1. Number of cluster and Parameter Fuzziness for Simulated Data

Data		Number of Cluster (k)	Fuzziness Parameter	Geographically Weighted Parameters	
				Alpha (α)	Beta (β)
(1)		(2)	(3)	(4)	(5)
Type 1	$N = 50$	3	1.3	0.9	0.8
(7 attributes)	$N = 100$	3	1.9	0.9	0.8
Type 2	$N = 50$	3	1.3	0.9	0.8
(8 attributes)	$N = 100$	4	1.4	0.9	0.8

To determine the number of clusters on each dataset, this research is considering a scree plot of decreasing the value of the total cost function for each cluster. In this study, the maximum iteration is 100, and the threshold is 0.00005. The genetic parameters used in this research include a maximum generation of 20, a population of 20, and a mutation rate of 0.00005. Table 2 displays the value of the cost function as the basis for the clustering evaluation, with the smallest value yielding the best result.

Table 2. Cost Function Values of Each Clustering Methods for Simulated Data

Data		Cost Function Value of each Clustering Methods				
		KP	FKP	GA-FKP	FKP-GW	GA-FKP-GW
(1)		(2)	(3)	(4)	(5)	(6)
Type 1	$N = 50$	10.86	0.915	0.915	0.990	0.984
(7 attributes)	$N = 100$	17.18	1.403	1.403	1.531	1.531
Type 2	$N = 50$	13.30	0.944	0.944	1.011	1.006
(8 attributes)	$N = 100$	20.19	1.556	1.556	1.675	1.672

Table 2 compares the cost function values of five clustering methods applied to simulated data with varying sample sizes and attributes. It is clear that the proposed methods, FKP-GW and GA-FKP-GW, consistently produce lower cost function values than the standard KP method. This shows the benefit of adding geographical weighting and fuzziness to the clustering process. Although GA-FKP achieves the smallest cost function values across all data types, the proposed methods perform competitively, particularly for larger datasets. The cost function values show minimal differences, indicating that while GA-FKP excels in cost reduction, FKP-GW and GA-FKP-GW still identify significant spatial features that could aid in the clustering of geographically dispersed data.

For smaller datasets ($N = 50$), the proposed methods show slight improvements in cost function compared to the traditional methods, and for larger datasets ($N = 100$), all fuzzy-based methods tend to converge in performance. Despite the marginally higher cost values in some cases, the spatially weighted methods offer the potential for more nuanced insights due to their ability to incorporate geographical variability. This highlights the trade-off between optimizing a cost function and obtaining clusters that reflect the spatial context of the data, making FKP-GW and GA-FKP-GW valuable for applications where spatial relationships are critical.

3.2 Implementing GA-FKP and GA-FKP-GW For Village Development Data

Before proceeding with the clustering process, it is essential to first determine the appropriate number of clusters and the level of fuzziness. Choosing the correct number of clusters guarantees the model captures the data's structure without overfitting or underfitting. Additionally, defining the degree of fuzziness is crucial in a fuzzy clustering approach, as it allows for flexible membership of data points

across clusters, better reflecting the complexity and ambiguity in real-world data. The determination of the number of clusters used is done using the scree plot shown in the Figure 1.

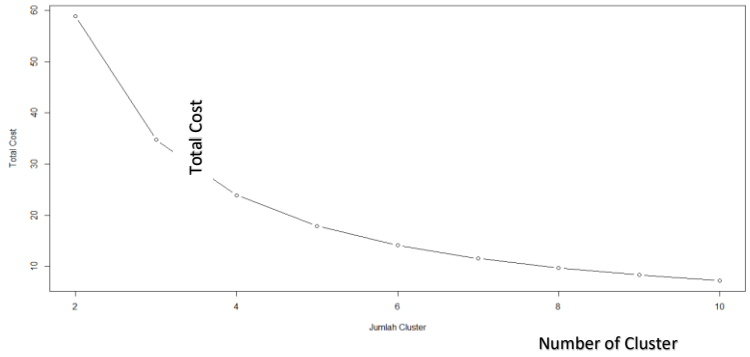


Figure 1. Scatter plot total cost function for each number of clusters

The scree plot suggests that four clusters are a suitable choice for the analysis using the GA-FKP-GW method. The "elbow" at 4 clusters indicates a significant reduction in the cost function, after which the improvements become marginal. This gives four clusters a reasonable balance between complexity and capturing meaningful distinctions in the data. Given that the Village Development Index (VDI) traditionally used 3 categories, expanding to 4 clusters allows for greater nuance in representing village development patterns while maintaining interpretability within the cluster formed.

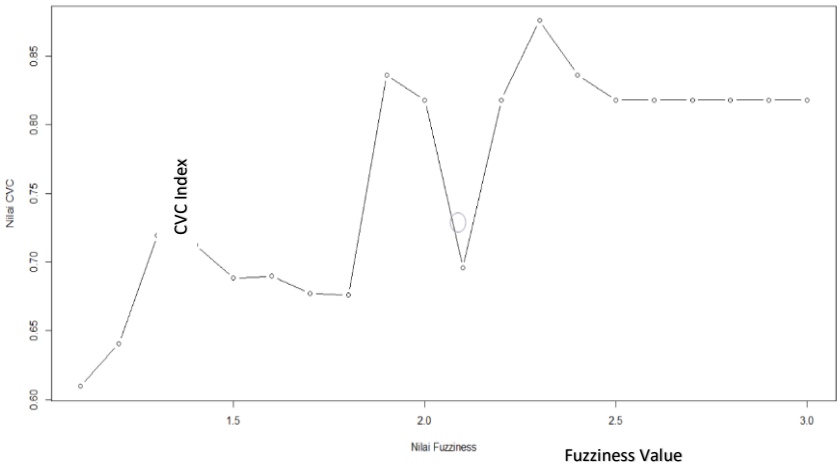


Figure 2. Plot index CVC for every parameter fuzziness

Next, determination of the fuzziness hyperparameter is conducted using the CVC plot in Figure 2. The CVC plot reveals the optimal value of fuzziness for the fuzzy K-prototype clustering method. Observing the graph, it is clear that as the fuzziness parameter increases, the CVC values fluctuate, indicating varying cluster validity across different levels of fuzziness. Approximately 2.4 is the optimal value for producing the best balance between cluster compactness and separation, resulting in the highest CVC value. Consequently, a fuzziness value of 2.4 provides the most appropriate configuration for capturing the characteristics of the data set under analysis.

The hybrid clustering method proposed, GA-FKP-GW, has been successfully built. This proposed method can be used for clustering mixed-type data and adding the spatial effect to it. Table 3 displays the results of the membership matrix for both the previous clustering method (GA-FKP) and the proposed method (GA-FKP-GW).

Table 3. Matrix Membership Degree from Previous Method Clustering Method (GA-FKP) and Proposed Method (GA-FKP-GW)

Clustering Method	Observation	Membership Degree of Each Cluster			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
(1)	(2)	(3)	(4)	(5)	(6)
GA-FKP	1	0.2408	0.2404	0.2537	0.2648
	2	0.2456	0.2454	0.2518	0.2570
	3	0.2443	0.2439	0.2524	0.2592
	4	0.2444	0.2441	0.2524	0.2588
	…	…	…	…	…
GA-FKP-GW	1	0.2435	0.2432	0.2526	0.2606
	2	0.2471	0.2470	0.2511	0.2546
	3	0.2460	0.2458	0.2516	0.2564
	4	0.2462	0.2460	0.2516	0.2560
	…	…	…	…	…

Table 3 presents the performance of the GA-FKP and GA-FKP-GW methods in clustering mixed-type data, demonstrating the incorporation of spatial effects in the latter approach. The differences in membership degrees between the two methods highlight the influence of the geographically weighted component in GA-FKP-GW. By applying this spatial weighting, the geographic location of each observation influences its likelihood of belonging to a specific cluster, leading to shifts in membership values compared to the traditional GA-FKP method.

These subtle differences in membership degrees reflect the intended design of GA-FKP-GW, which adjusts for spatial heterogeneity across the data. This effect is especially important when mapping rural development because the results of the clustering now take into account local geographic factors. This could lead to clusters that better reflect the data's underlying spatial structure and characteristics. To sum up, Table 3 shows how well GA-FKP-GW works at improving the clustering process by adding spatial factors. This makes the analysis more relevant for data that is spread out geographically.

3.3 Profiling villages in Temanggung Regency Based on GA-FKP-GW Algorithm

The proposed method GA-FKP-GW is successfully implemented to analyze clustering data study cases using indicators of the Village Development Index (VDI) of Temanggung Regency 2018. This study cases VDI indicators were included in PODES. The previous section provides explanations for all the parameters used in this clustering. The result from the analysis cluster to data study case using GA-FKP-GW, where from 88 numeric and 94 categorical attributes with 289 villages included, resulted in 4 clusters. To see the detail of mapping distribution villages grouped by GA-FKP-GW and VDI, see Figure 3.

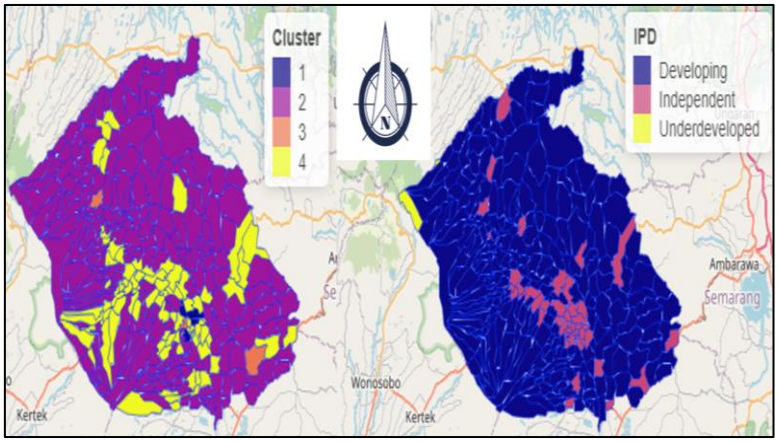


Figure 3. Mapping villages grouped by GA-FKP-GW and VDI

Figure 3 displays a mapping of the distribution of villages according to their cluster and VDI. The villages included in cluster 1 are located and gathered in the central area of Temanggung Regency. Meanwhile, the villages included in cluster 2 are seen scattered throughout the region. Cluster 3's villages appear dispersed, not concentrated in a single adjacent area. Despite the presence of several scattered villages on the area's edge, the villages included in cluster 4 appear to gather in the center of the Temanggung Regency area. The comparison of mapping villages by cluster and VDI original category reveals that certain developing villages share the same color with cluster 2, while independent villages align with cluster 4. Table 4 provides further details.

Table 4. Total villages grouped by Cluster and VDI

VDI	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
(1)	(2)	(3)	(4)	(5)	(6)
Independent	4	10	2	25	41
Developing	0	192	1	53	246
Underdeveloped	0	2	0	0	2
Total	4	204	3	78	289

In order to identify the distinct features of each cluster, researchers chose to examine the attributes that significantly differed between them. A statistical test is used to test whether there is a difference in average between clusters of each attribute [41], [42]. For numerical attributes, researchers used the statistic test one-way Anova. Every attribute was significant, then compute the average and interpret it using descriptive analysis. Some attributes are grouped into a few variables that were included in the calculation of VDI. For detailed profiling villages, each cluster in every variable VDI can be seen below.

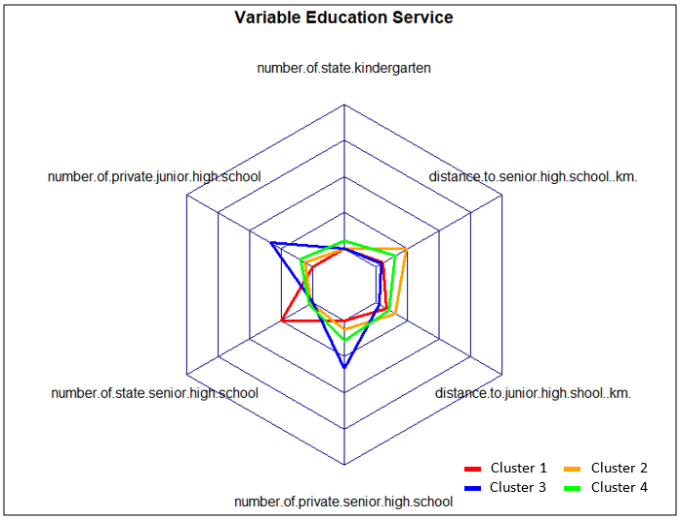


Figure 4. Spider plot variable education service

Figure 4 displays the characteristics of each cluster within the education service variable. Figure 4 shows the average indicators on the education service variable. In this variable, cluster 1 becomes the cluster with the highest average number of state senior high schools, the cluster 3 becomes the cluster with the highest average number of private junior high schools and state high schools, and the cluster 4 becomes the cluster with the highest average number of kindergartens. Meanwhile, Cluster 2 is the one with the farthest average distance to educational services among other clusters.

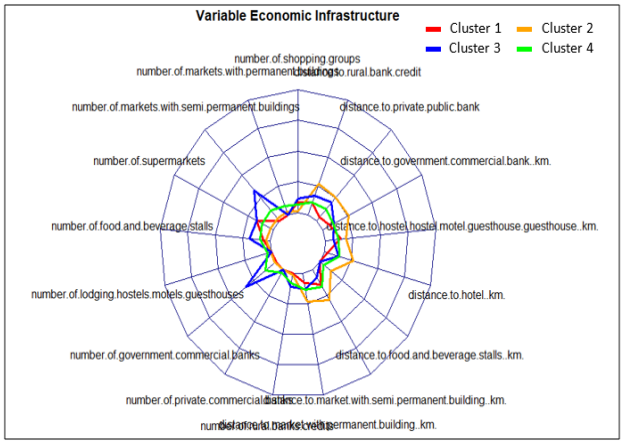


Figure 5. Spider plot variable economic infrastructure

Figure 5 displays the characteristics of each cluster in terms of variable economic infrastructure. Cluster 3 stands out as the cluster with the highest average number of infrastructures, with Cluster 4 and Cluster 1 following closely behind. In contrast to other clusters, Cluster 2 has the longest average distance to the closest economic infrastructure.

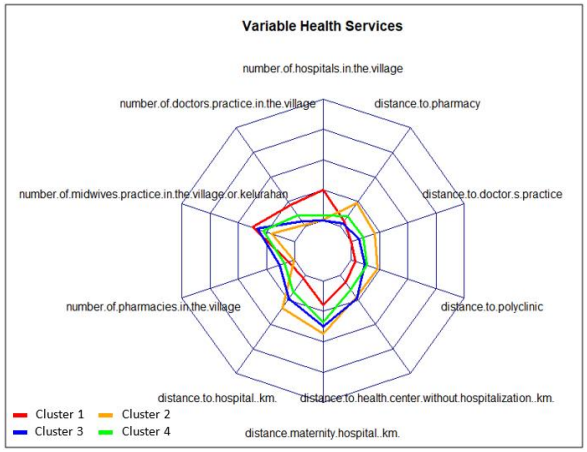


Figure 6. Spider plot variable health services

Figure 6 displays the characteristics of each cluster in terms of variable health services. Compared to other clusters, cluster 1 has the highest average number of health services and the closest average distance to isolated health services, while cluster 3 has the highest average number of pharmacies in the village. Meanwhile, cluster 2 is the cluster with the least and farthest average number and distance to health services compared to other clusters.

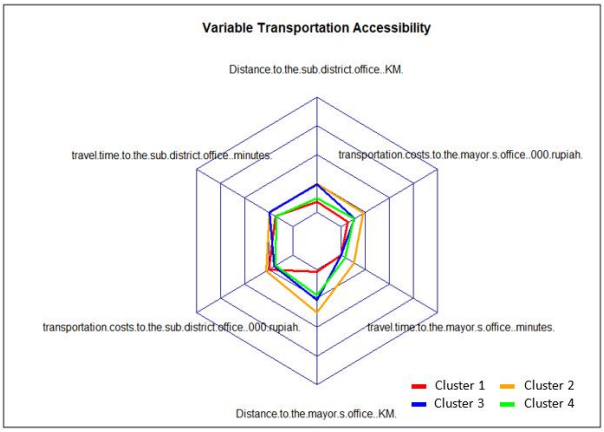


Figure 7. Spider plot variable transportation accessibility

Figure 7 illustrates the characteristics of each cluster in the transportation accessibility variable. Compared to other clusters with the least transportation time and cost, cluster 1 in the transportation accessibility variable has the closest average distance to the sub-district office and district head, while cluster 2 has the average distance and time and typically travels to the district and district offices with the highest costs.

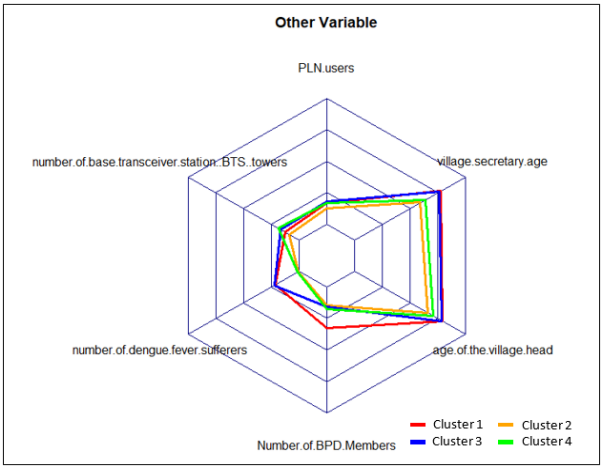


Figure 8. Spider plot of other variable

Figure 8 is a spider plot that displays the average values of other variables for each cluster. In this plot, some variables were combined, including energy infrastructure, communication and information infrastructure, public health, independence, and quality of human resources. The plot image reveals that cluster 3 has the highest average indicator of PLN users, while cluster 2 has the lowest. For the number of BTS towers, cluster 4 is the cluster with the highest average number of towers, while cluster 2 is the cluster with the fewest average BTS towers. For indicators of the number of patients with dengue fever, cluster 3 becomes the cluster with the most sufferers on average, while cluster 2 becomes the average patient with the least. For indicators of the number of BPD members, cluster 1 is the cluster with the highest average number of members, while cluster 2 is the cluster with the least average number of members. For the HR quality variable, cluster 1 is the cluster with the youngest average age of the village head and secretary and cluster 2 with the oldest average age.

For categorical attributes, researchers count the frequency of every categorical attribute within each cluster that is significant and has a different frequency of each cluster using the statistic test chi-square. By counting the frequency of each cluster, the researchers discovered that cluster 2 differs from other clusters in that it has access to the nearest infrastructure, which is easily accessible by road. In addition, cluster 2 was found to have fewer sports facilities than other clusters.

The descriptive analysis of numerical and categorical attributes, which is significant, shows that cluster 1 is the cluster with the best average health service but has the least average number of economic infrastructure. On the other hand, Cluster 3 has the best average economic infrastructure and educational services. Cluster 4 is the cluster with the most stable and above average infrastructure facilities and

infrastructure among other clusters. Meanwhile, cluster 2 is the cluster with the average number of facilities and infrastructure that is less than the other clusters, and the average distance to other infrastructure is the farthest. Cluster 2 also features easy road access to the nearest infrastructure, and its villages have fewer sports facilities than those in other clusters.

After the descriptive analysis, the researcher calculated the distribution of villages in each cluster based on the subdistrict.Table 5 displays the distribution table, which divides the number of villages in each cluster into several sub-districts in Temanggung Regency.

Table 5. Distribution villages each cluster by sub-district

Sub-district	Cluster 1	Cluster 2	Cluster 3	Cluster 4
(1)	(2)	(3)	(4)	(5)
Bansari	0	12 (92.3%)	0	1 (7.7%)
Bejen	0	12 (85.7%)	0	2 (14.3%)
Bulu	0	13 (68.4%)	0	6 (31.6%)
Candiroto	0	11 (78.6%)	1 (7.1%)	2 (14.3%)
Gemawang	0	9 (90%)	0	1 (10%)
Jumo	0	13 (100%)	0	0
Kaloran	0	9 (64.3%)	0	5 (35.7%)
Kandangan	0	16 (100%)	0	0
Kedu	0	7 (50%)	0	7 (50%)
Kledung	0	6 (46.2%)	0	7 (53.8%)
Kranggan	0	12 (92.3%)	0	1 (7.7%)
Ngadirejo	0	15 (75.0%)	0	5 (25%)
Parakan	0	3 (18.8%)	0	13 (81.3%)
Pringsurat	0	11 (78.6%)	1 (7.1%)	2 (14.3%)
Selopampang	0	7 (58.3%)	0	5 (41.7%)
Temanggung	4 (16%)	8 (32%)	1 (4%)	12 (48%)
Tembarak	0	10 (76.9%)	0	3 (23.1%)
Tlogomulyo	0	6 (50%)	0	6 (50%)
Tretep	0	11 (100%)	0	0
Wonoboyo	0	13 (100%)	0	0
Total	**4 (1.4%)**	**204 (70.6%)**	**3 (1%)**	**78 (27%)**

Table 5 is presented in order to find out how well the local government develops or ensures equal distribution of villages in each subdistrict. Almost all sub-districts group several villages into distinct clusters, indicating that the characteristics of the villages within a sub-district are not uniform or equal. However, there are several sub-districts, namely District Jumo, District Kandangan, District Tretep, and District Wonoboyo, of which all of the villages are included in Cluster 2.

4. Conclusion

This study contributes to the development of geodemographic clustering methodology. The study successfully develops and applies two methods, FKP-GW and GA-FKP-GW, for cluster analysis of mixed-type geographic data. Experiments comparing the proposed method with the previous method in clustering simulation and study case data have shown that the proposed method yields different clustering results. However, the evaluation results indicate that the proposed method is still less efficient than the previous method. This proposed method is highly effective and can be utilized to conduct clustering analysis of geographic data that contains a combination of different types.

The empirical implementation of the proposed method GA-FKP-GW in clustering analysis for data PODES, which serves as a composition indicator for calculating VDI of Temanggung Regency 2018, successfully grouped 289 villages into four clusters. Villages with an independent status dominate clusters 1 and 3, while villages with a developing status dominate clusters 2 and 4. The result of descriptive analysis for each cluster has its own characteristics: cluster 1 is good in health services, cluster 3 is good in economic infrastructure, cluster 4 has the most stable and adequate infrastructure among all clusters, and cluster 2 is the cluster with the least infrastructure with the distance farthest.

This study still has limitations that can be improved in further research. This study repeatedly tests several parameters, including alpha and beta, potentially leading to less than optimal results. Therefore, further research is required to identify suitable parameters. On the other hand, some genetic algorithm

operators used in this study only focus on reducing the processing time, so it is possible to apply other operators to get more optimal results. Nevertheless, this study can serve as a reference for the development of the geodemographic analysis method and clustering analysis.

Ethics approval

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

Agung Budi Santoso: Conceptualization, Methodology, Software, Data Curatiom, Writing – Original Draft, Writing – Review and Editing, Visualization. **Arya Candra Kusuma**: Conceptualization, Writing – Original Draft, Writing – Review and Editing. **Rani Nooraeni:** Conceptualization, Methodology, Validation. **Arie Wahyu Wijayanto:** Validation, Writing – Review and Editing.

References

[1] P. Sleight, *Targeting customers : how to use geodemographic and lifestyle data in your business / Peter Sleight.*, Second edi. Henley-on-Thames: NTC, 1997.

[2] R. Harris, P. Sleight, and R. Webber, *Geodemographics, GIS and Neighbourhood Targeting*. in Mastering GIS: Technol, Applications & Mgmnt. Wiley, 2005. [Online]. Available: https://books.google.co.id/books?id=Z8K25AxTjDcC

[3] L. H. Son, B. C. Cuong, P. L. Lanzi, and N. T. Thong, "A novel intuitionistic fuzzy clustering method for geo-demographic analysis," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 9848–9859, Aug. 2012, doi: 10.1016/j.eswa.2012.02.167.

[4] G. A. Mason and R. D. Jacobson, "Fuzzy Geographically Weighted Clustering," in *Proceedings of the 9th International Conference on Geocomputation*, Sep. 2007, pp. 1–7.

[5] 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.

[6] H. Izakian and A. Abraham, "Fuzzy C-means and fuzzy swarm for fuzzy clustering problem," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 1835–1838, Mar. 2011, doi: 10.1016/j.eswa.2010.07.112.

[7] J. Wu, H. Xiong, C. Liu, and J. Chen, "A generalization of distance functions for fuzzy c-means clustering with centroids of arithmetic means," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 557–571, 2012, doi: 10.1109/TFUZZ.2011.2179659.

[8] Z. Feng and R. Flowerdew, "Fuzzy geodemographics: a contribution from fuzzy clustering methods," in *Innovations In GIS 5*, CRC Press, 1998, pp. 141–149. doi: 10.1201/b16831-20.

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. 1981.

[10] L. Hunt and M. Jorgensen, "Clustering mixed data," *WIREs Data Min. Knowl. Discov.*, vol. 1, no. 4, pp. 352–361, Jul. 2011, doi: 10.1002/widm.33.

[11] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1263–1271, Aug. 2004, doi: 10.1016/j.patrec.2004.04.004.

[12] Z. Huang, "Clustering Large Data Sets With Mixed Numeric And Categorical Values," Proceedings Of 1st Pacific-Asia Conference on Knowledge Discouvery And Data Mining," 1997, *Singapore*.

[13] X. Zhong, T. Yu, and H. Xia, "A new partition-based clustering algorithm for mixed data," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2017.

[14] J. Ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Syst.*, vol. 30, pp. 129–135, Jun. 2012, doi: 10.1016/j.knosys.2012.01.006.

[15] J. Ji, Y. Chen, G. Feng, X. Zhao, and F. He, "Clustering mixed numeric and categorical data with artificial bee colony strategy," *J. Intell. Fuzzy Syst.*, vol. 36, no. 2, pp. 1521–1530, Mar. 2019, doi: 10.3233/JIFS-18146.

[16] W. Alomoush and A. Alrosan, "Review: Metaheuristic Search-Based Fuzzy Clustering Algorithms," *CoRR*, vol. abs/1802.0, 2018, [Online]. Available: http://arxiv.org/abs/1802.08729

[17] G. Gan, J. Wu, and Z. Yang, "A genetic fuzzy <math altimg="si218.gif" overflow="scroll"> <mrow> <mi>k</mi> </mrow> </math> -Modes algorithm for clustering categorical data," *Expert Syst. Appl.*, vol. 36, no. 2, pp. 1615–1620, Mar. 2009, doi: 10.1016/j.eswa.2007.11.045.

[18] W. Min and Y. Siqing, "Improved K-means clustering based on genetic algorithm," in *2010 International Conference on Computer Application and System Modeling (ICCASM 2010)*, 2010, pp. V6-636-V6-639. doi: 10.1109/ICCASM.2010.5620383.

[19] B. I. Nasution, R. Kurniawan, T. H. Siagian, and A. Fudholi, "Revisiting social vulnerability analysis in Indonesia: An optimized spatial fuzzy clustering approach," *Int. J. Disaster Risk Reduct.*, vol. 51, Dec. 2020, doi: 10.1016/j.ijdrr.2020.101801.

[20] B. S. Hadi, "Pendekatan Modified Particle Swarm Optimization dan Artificial Bee Colony pada Fuzzy Geographically Weighted Clustering (Studi Kasus pada Faktor Stunting Balita di Provinsi Jawa Timur) [Modified Particle Swarm Optimization and Artificial Bee Colony Approach on Fuzzy Geographically Weighted Clustering (Case Study on Stunting Factors of Toddlers in East Java Province)]," *Inst. Teknol. Sepuluh Nop.*, 2017.

[21] R. Gupta, S. K. Muttoo, and S. K. Pal, "Meta-Heuristic Algorithms to Improve Fuzzy C-Means and K-Means Clustering for Location Allocation of Telecenters Under E-Governance in Developing Nations," *Int. J. FUZZY Log. Intell. Syst.*, vol. 19, no. 4, pp. 290–298, Dec. 2019, doi: 10.5391/IJFIS.2019.19.4.290.

[22] M. Gen and R. Cheng, "Genetic algorithms and engineering design, Canada," 1997, *John Wiley & Sons, Inc*.

[23] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*. in Wiley InterScience electronic collection. Wiley, 2004. [Online]. Available: https://books.google.co.id/books?id=k0jFfsmbtZIC

[24] E. Wirsansky, *Hands-On Genetic Algorithms with Python: Applying genetic algorithms to solve real-world deep learning and artificial intelligence problems*. Packt Publishing, 2020. [Online]. Available: https://books.google.co.id/books?id=A0vODwAAQBAJ

[25] 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," *Appl. Intell.*, vol. 44, no. 2, pp. 377–398, Mar. 2016, doi: 10.1007/s10489-015-0705-7.

[26] 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.

[27] S. Pramana and I. H. Pamungkas, "Improvement Method of Fuzzy Geographically Weighted Clustering using Gravitational Search Algorithm," *J. Ilmu Komput. dan Inf.*, vol. 11, no. 1, p. 10, Feb. 2018, doi: 10.21609/jiki.v11i1.580.

[28] 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," *Int. J. Disaster Risk Reduct.*, vol. 83, Dec. 2022, doi: 10.1016/j.ijdrr.2022.103407.

[29] R. Nooraeni, "Cluster Method Using A Combination of Cluster K-Prototype Algorithm and Genetic Algorithm for Mixed Data," *J. Apl. Stat. Komputasi Stat.*, vol. 7, no. 2 SE-Articles, p. 17, Dec. 2015, doi: 10.34123/jurnalasks.v7i2.23.

[30] R. Nooraeni, N. P. Yudho, and S. Pramana, "Mapping the socio-economic vulnerability in Aceh to reduce the risk of natural disaster," 2018, p. 030012. doi: 10.1063/1.5062736.

[31] R. Nooraeni, M. I. Arsa, and N. W. Kusumo Projo, "Fuzzy Centroid and Genetic Algorithms: Solutions for Numeric and Categorical Mixed Data Clustering," *Procedia Comput. Sci.*, vol. 179, pp. 677–684, 2021, doi: 10.1016/j.procs.2021.01.055.
[32] BPS, "Indeks Pembangunan Desa 2018 [Village Development Index 2018]," Jakarta, 2019.
[33] BPS, "Statistik Potensi Desa Indonesia (Village Potential Statistics Of Indonesia) 2018," Jakarta, 2018.
[34] E. J. Pebesma and R. Bivand, "Classes and methods for spatial data in {R}," *R News*, vol. 5, no. 2, pp. 9–13, Nov. 2005, [Online]. Available: https://cran.r-project.org/doc/Rnews/
[35] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. in Pearson International Edition. Pearson Addison Wesley, 2006. [Online]. Available: https://books.google.co.id/books?id=_XdrQgAACAAJ
[36] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2011.
[37] S. Pramana, B. Yuniarto, I. Santoso, R. Nooraeni, and L. H. Suadaa, *Data Mining dengan R, Konsep dan Implementasi [Data Mining with R, Concepts and Implementation]*. 2023.
[38] S.-H. Jun, "An Optimal Clustering using Hybrid Self Organizing Map," *Int. J. Fuzzy Log. Intell. Syst.*, vol. 6, no. 1, pp. 10–14, Mar. 2006, doi: 10.5391/IJFIS.2006.6.1.010.
[39] G. Szepannek, "clustMixType: User-Friendly Clustering of Mixed-Type Data in R," *R J.*, vol. 10, no. 2, p. 200, 2019, doi: 10.32614/RJ-2018-048.
[40] C.-C. Hsu and Y.-P. Huang, "Incremental clustering of mixed data based on distance hierarchy," *Expert Syst. Appl.*, vol. 35, no. 3, pp. 1177–1185, Oct. 2008, doi: 10.1016/j.eswa.2007.08.049.
[41] W. Johnson and R. Wichern, "Applied Multivariate Statistical Analysis Sixth Edition," 2007.
[42] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*, vol. 5. Macmillan New York, 1993.

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.790

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Aspect-Based Sentiment Analysis of Transportation Electrification Opinions on YouTube Comment Data

Rahmi Elfa Adilla[1*], Muhammad Huda[2], Muhammad Aziz[3], Lya Hulliyyatus Suadaa[4]

[1]BPS-Statistics Indonesia, Jakarta, Indonesia, [2]PT PLN (Persero), Indonesia, [3]The University of Tokyo, Tokyo, Japan, [4]Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address: [1]rahmielfa1010@gmail.com*

ARTICLE INFO

Article history:
Received 27 August, 2024
Revised 29 October, 2024
Accepted 09 November, 2024
Published 31 December, 2024

Keywords:
Transportation Electrification; Electric Vehicles; Aspect-Based Sentiment Analysis; Machine Learning; Transfer Learning

Abstract

Introduction/Main Objectives: This research aims to conduct an aspect-based sentiment analysis of transportation electrification opinions on YouTube comment data. **Background Problems:** It is difficult to summarize the sentiment of many YouTube user comments related to electric vehicles (EVs) based on their aspects; therefore, aspect-based sentiment analysis is needed to conduct further analysis. **Novelty:** This study identifies five aspects of EV and their sentiments at the same time. The aspects are usefulness, ease of use, comfort, cost, and incentive policies. One of this study's methods is the transfer learning model. This model can be a solution to overcome the shortcomings of deep learning in classifying aspect-based sentiment classification on small datasets. **Research Methods:** The sentiment classification model used is a machine learning model, namely support vector machine (SVM) and transfer learning models from pre-trained IndoBERT and mBERT. **Finding/Results:** Based on the experimental results, transfer learning from the IndoBERT model achieved the best performance with accuracy and F1-Score of 89.17% and 52.66%, respectively. Furthermore, the best IndoBERT model was developed with input in the form of a combination of aspects and comment sentences. Experimental results show that there is an improvement in performance with accuracy and F1-Score of 90% and 60.70%, respectively.

1. Introduction

Motor vehicles have a direct effect on global environmental challenges and the depletion of natural resources. The transportation sector plays a crucial role in contributing to air pollution and climate change, particularly in urban regions, due to emissions of greenhouse gases (GHGs). This sector is responsible for approximately 25% of total global GHG emissions, with projections indicating that this figure could rise from 23% to 50% by 2030 [1]. As a result, transportation is identified as a significant obstacle to achieving a sustainable economy [2], [3].

Based on historical data from IQAir, Indonesia in 2022 ranked 26th out of 131 most polluted countries in the world. Meanwhile, according to IQAir real-time data, accessed on August 30, 2023 at 15:40 WIB, Jakarta is ranked second as the city with the poorest air quality globally, with 152 AQI US air indicators in red, which means that the air quality in Jakarta is unhealthy [4]. The order of cities with the poorest air quality globally is illustrated in Figure. 1.

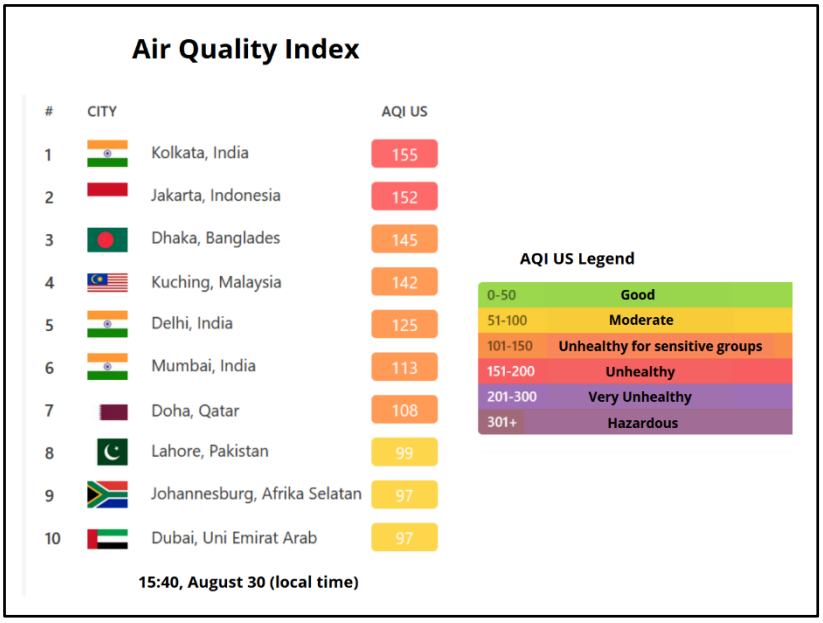


Figure. 1. Ranking of cities with the poorest air quality globally according to IQAir real-time data on August 30, 2023 at 15.40 WIB

Indonesia's poor air quality index cannot be ignored, so it needs special attention from the government, the community, and various other parties. Reducing greenhouse gas emissions from vehicles is one of the solutions [5].This encourages the electrification of road transportation, namely by replacing conventional vehicles with alternative energy options, such as electric vehicles (EVs). EVs are described as eco-friendly vehicles powered by electric motors that utilize energy from batteries and can be recharged from outside sources.[6]. EVs produce lower emissions, so they are environmentally friendly vehicles that can reduce air pollution. The Indonesian government established an EV development program through Presidential Regulation Number 55 of 2019 was issued on August 12, 2019, focusing on the acceleration of the Battery-Based Electric Motor Vehicle Program for road transportation. This regulation is the initial rule as a legal law for EVs in Indonesia.

The electrification of road transportation continues to be socialized by the government, both central and regional, as one way to reduce worsening air pollution. However, there are many parties who disagree with the opinion of transportation electrification. Many argue that switching from fossil fuel vehicles to EVs will not solve the problem and that electrification is a “false solution” because the energy source used for EVs still comes from coal-fired power plants, which are also a contributor to air pollution. In addition, one of the main reasons many parties disagree with the opinion of transportation electrification is the very expensive purchase cost, usage co14st, and maintenance cost of EVs [7].

Based on this background, the author will analyze public opinion regarding EVs. The data analyzed are public comments on transportation electrification opinions obtained from the YouTube platform. The YouTube platform was selected as the data source for this study due to its status as the second most visited website globally. [8], [9]. While there are several video-sharing platforms like Vimeo and Dailymotion, YouTube stands out as the site with the highest volume of videos shared among billions of users around the world [10]. It allows individuals to express their thoughts, ideas, and emotions through video content, for example, related to EVs, and respond by submitting comments. Comments on YouTube videos tend to be more relevant and appropriate to the context of the video shown, so that predetermined aspects of EVs can be extracted more easily and the analysis process can be carried out [11]. The analysis process to find out the public's response to the electrification of transportation is referred to as sentiment analysis. Sentiment analysis is typically used to determine whether a particular opinion conveys a positive, negative, or neutral sentiment. This shows that sentiment analysis in general has not been able to identify sentiment for certain aspects of EVs. If the opinion given by a person on an aspect of an EV has a positive sentiment, it does not necessarily mean that the person also gives positive sentiments to all aspects of EVs, and vice versa. For example, in the sentence “EVs produce low emissions, but the price is very high,” the emission aspect of electric vehicles has a positive sentiment, but gives a negative sentiment towards the price. Therefore, a more complete analysis is needed to determine the sentiment towards EVs based on their aspects. This analysis is called aspect-

based sentiment analysis. Constructing a model to understand the sentiment of transportation electrification is essential, as various stakeholders can benefit from this analysis. For instance, the government can utilize the insights to support the development of incentive policies for electric vehicles. The state electricity company (PLN) can gain valuable information regarding the needs for transportation electrification, including the establishment of charging stations and the necessary voltage or wattage for home setups. Additionally, EV suppliers can benefit from insights related to user interests and complaints, which can inform product improvements and customer service strategies. Overall, this model provides crucial information that can guide decision-making for multiple parties involved in the electrification of transportation.

In this study, aspect-based sentiment analysis can be utilized to analyze public responses regarding EVs in various aspects. According to the technology acceptance model (TAM) of EVs and extensions of the model [5], [12], [13], the aspects used in this study are usefulness, ease of use, comfort, cost, and incentive policies. The classification model used to solve the aspect-based sentiment analysis task in this study is a machine learning model, namely the support vector machine (SVM), which provides the best performance in the research of Mustakim and Priyanta [14]. Mustakim and Priyanta [14] conducted aspect-based sentiment analysis to extract aspects of KAI Access user reviews in the review column on the Google Play Store. The methods used are a naïve bayes classifier (NBC) and an SVM. The aspects used are learning ability, memory, efficiency, and errors. The results showed that the majority of user sentiments were negative in each aspect, with the most discussed aspect of errors indicating high system errors. The test results provide the best model in SVM with hyperparameter tuning with an average accuracy score of 91.63%, F1-Score of 75.55%, and recall of 74.47%.

Jeong has carried out an aspect-based sentiment analysis of EVs [15]. The study utilized user satisfaction data from automotive forums (Edmunds.com, Cars.com, Cargurus.com, Carfax.com, and Carbuyers.co.uk) and YouTube reviews. Sixteen main aspect categories were used, namely eight main components of EVs and eight main characteristics of human factors. As a result of the study, it was found that users had positive sentiments on the aspects of acceleration, space, interior, power, safety, ergonomics, price, and power. In addition, users had negative sentiments about seats, battery, charging, noise, winter, and ice.

Related research has also been conducted by Anwar et al. [16] regarding aspect-based sentiment analysis on car reviews. The data used is car review data sourced from the Edmunds website (www.edmunds.com). The results showed that most of the positive sentiments were in the aspects of comfortable driving, excellent fuel efficiency, dependability, comfort, great value, a useful rear camera, a smooth and quiet ride, impressive acceleration, an appealing design, a quality sound system, and a robust build. Some negative sentiment elements share similarities with those in the positive category, although they appear less frequently. Jena [17] also conducted research related to consumer sentiment towards EVs with a big data approach. The data in this study used data for two years (2016 to 2018) collected from various social media platforms by extracting the data into aspects of price, maintenance, and safety. The classification analysis shows that price and maintenance have mostly negative sentiments, while security aspects have mostly neutral sentiments.

In addition, researchers also use transfer learning models from pre-trained models used in the research conducted by Tao and Fang [18]. Tao and Fang [18], who proposed a transfer learning-based approach. First, the proposed approach extends the aspect-based sentiment analysis method with multi-label classification capability. Second, it proposes an advanced sentiment analysis method for categorizing text into sentiment categories while considering the aspects of entities. Third, it extends two transfer learning models, namely Bidirectional Encoder Representations from Transformers (BERT) and XLNet, as analysis tools for aspect-based and multi-label sentiment classification. The data used are three datasets from different domains, namely reviews related to Yelp restaurants, wine, and movies. The findings indicate that the suggested transfer learning model consistently surpasses the baseline SVM and CNN models across all three datasets. This evidence supports the notion that the proposed method is better suited for managing multi-label classification tasks. Liu and Zhao proposed the utilization of the transfer learning approach in aspect-based sentiment analysis using Amazon product review data [19]. The classification models used are CNN and BERT, which use a combination of corpus at the aspect level and corpus at the sentence level to create sequential sentence pairs as input. The results show that the BERT model, which combines aspects and sentences, performs better than the CNN model. The transfer learning model is a solution to overcome the shortcomings of deep learning in performing aspect-based sentiment classification on small datasets [20]. Both types of models are evaluated using accuracy, precision, recall, and F1-Score values. Furthermore, a model comparison is carried out to determine which model performs better at performing aspect-based sentiment classification on public responses to transportation electrification opinions on YouTube comment data.

Based on the described background, this research has several objectives. First, it aims to construct a dataset for sentiment classification focused on aspects of electric vehicles (EVs) using YouTube comment data. Second, the study intends to perform a descriptive analysis of this data to gain insights into public opinion regarding transportation electrification. Third, it seeks to develop an aspect-based sentiment analysis model by applying machine learning and transfer learning techniques to predict public sentiment towards opinions on transportation electrification as reflected in YouTube comments. Finally, the research will evaluate the performance of these aspect-based sentiment analysis models.

The limitations of this research conducted by the researcher are as follows: First, the data utilized in this research was collected through scraping YouTube comments using the YouTube Data API. Second, the data was collected from comments on the top 11 recommended videos based on a YouTube search conducted on September 9, 2023, using the keyword "electric vehicles." Third, the aspects of electric vehicles examined in this research include usefulness, ease of use, comfort, cost, and incentive policies. Fourth, the machine learning algorithm employed for aspect-based sentiment analysis is the Support Vector Machine (SVM). Finally, the transfer learning algorithms used for aspect-based sentiment analysis are IndoBERT and mBERT.

2. Material and Methods

The analytical approach employed in this study is aspect-based sentiment analysis, utilizing machine learning and transfer learning classification models. The research process follows the stages outlined in the Cross Industry Standard Process for developing Machine Learning applications with a Quality assurance methodology (CRISP-ML(Q)). This framework includes steps such as understanding business and data, preparing data, engineering models, evaluating models, and managing model operations [21]. The stages of this research are depicted in Figure. 2.

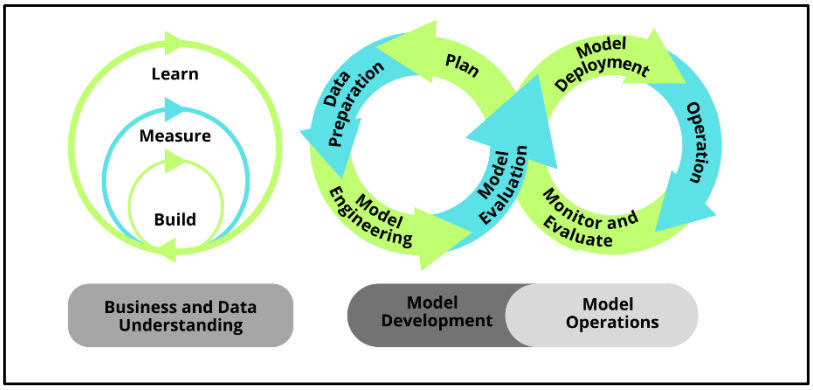
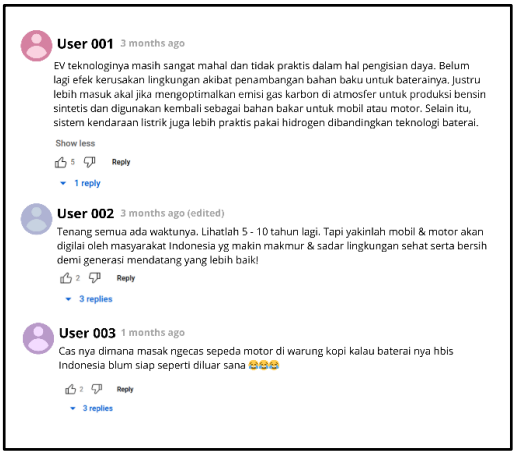


Figure. 2. Life cycle process of research stage

2.1. Business and data understanding

This research starts from the importance of supporting information about the level of public acceptance of transportation electrification, which is a government effort to decarbonize and reduce dependence on fossil fuels in Indonesia. YouTube user comments on EVs can be used to conduct aspect-based sentiment analysis, which can be used as input to overcome shortcomings and add advantages to the use of EVs in the future. It is difficult to summarize the sentiment of many YouTube user comments related to EVs based on their aspects; therefore, aspect-based sentiment analysis is needed to conduct further analysis.

a



b

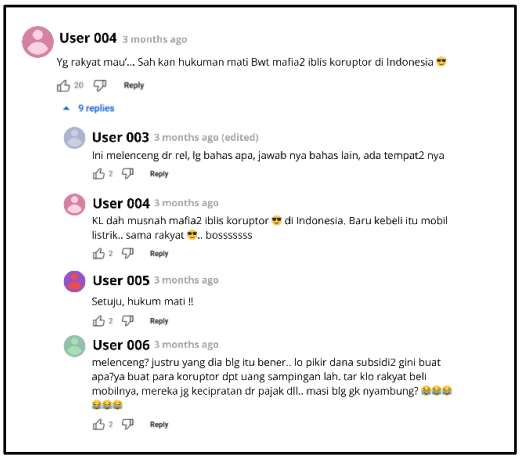


Figure. 3. (a) example of YouTube user comments on EVs; (b) examples of irrelevant YouTube user "reply" comments

The dataset utilized in this research is data collected through scraping YouTube comments using YouTube Data API, which is relevant to the topic of EVs. The dataset used in this research is data obtained from scraping YouTube comments using YouTube Data API, which is relevant to the topic of EVs. The data is collected from the comments of the top 11 recommended videos from YouTube search results. Limiting the analysis to comments from the top 11 recommended videos from the YouTube search results because it will compare machine learning and transfer learning model, with machine learning as the base model. Machine learning requires a lot of training data to build a good model, while transfer learning can overcome the availability of small datasets with better model result [20]. So to validate that transfer learning can overcome this, only the top 11 recommended videos from the YouTube search results were used. The keyword used in video searches is "kendaraan listrik". The keyword was chosen so that the video search results obtained cover the topic of EVs in general so that they can be classified into their aspects. An example of comments on a YouTube video is illustrated in Figure 3 (a). In the data collection process, comments in the form of "replies" were not collected in this study. This is because "replies" tend to deviate from the observed keywords, which would make the data collected irrelevant to EVs. An example of an irrelevant reply is illustrated in Figure. 3 (b).

2.2. Data preparation

2.2.1. Data labelling

After collecting data, the next step is data labeling. Data labeling is performed for each comment based on specific aspects. Based on the technology acceptance model (TAM) of EVs and extensions of the model [5], [12], [13], the aspects used in this study are usefulness, ease of use, comfort, cost, and incentive policies. Usefulness is defined as the extent to which consumers believe that electric vehicles can improve their lives, particularly in environmental performance. Uses of electric vehicles include reducing carbon emissions, controlling engine fuel consumption, improved health quality due to protection from air pollution/smog, and so on. Ease of use is how an item can be understood, used, or operated. The ease of use of electric vehicles includes easy operation, automatic transmission, easy access to charging infrastructure, speed of charging time, battery life, and so on. Electric vehicle comfort reflects how users overall perception of the exterior, interior, and technical electric vehicles such as seat comfort, sound noise level, space, electric vehicle model, and safety. Meanwhile, the cost of electric vehicles includes purchase costs, usage costs, and maintenance costs. To overcome the problem of the high price effect, the government provides incentive policy support in the form of purchase subsidies, tax deductions in the use of EVs, the use of special number plates, and so on.

To reduce subjectivity, data labeling is performed by no fewer than three annotators, and the final outcomes is determined based on majority voting. The sentiment labels used for each aspect consist of "positive," "negative," or "neutral" labels. For comments that do not contain predefined aspects, the sentiment label given is "none." An illustration of data labeling can be seen in Figure. 4.

Comment				
EV teknologinya masih **sangat mahal** dan **tidak praktis dalam hal pengisian daya**. Belum lagi efek **kerusakan lingkungan akibat penambangan bahan baku untuk baterainya**. Justru lebih masuk akal jika mengoptimalkan emisi gas karbon di atmosfer untuk produksi bensin sintestis dan digunakan kembali sebagai bahan bakar untuk mobil atau motor. Selain itu, sistem kendaraan listrik juga lebih praktis pakai hidrogen dibandingkan teknologi baterai. *EV technology is still **very expensive** and **cumbersome in terms of charging**. Not to mention the **environmental damage caused by mining the raw materials for the batteries**. Instead, it makes more sense to optimize carbon gas emissions in the atmosphere for the production of synthetic gasoline and reuse it as fuel for cars or motorcycles. In addition, EV systems are also more practical with hydrogen than battery technology.*				
Aspects				
Usefulness	Ease of Use	Comfort	Cost	Incentive Policies
negative	negative	none	negative	none

Figure. 4. Example of data labeling

To see the level of consistency of several annotators in labeling data, an Inter-Rater Reliability test was conducted using Krippendorff's alpha. Krippendorff's alpha's calculation can be done using the following formula [22]. D_o represents observed disagreement, and D_e represents expected disagreement that occurs by chance. The value of α can range from 0 to 1, with 0 representing an indication of no agreement (no reliability), and 1 representing perfect agreement among raters (perfect reliability).

$$\alpha = 1 - \frac{D_0}{D_e} \quad (1)$$

2.2.2. Data preprocessing

Before Preprocessing
Terbukti subsidi mobil listrik tidak hanya untuk orang yg mampu..pupuk atau yg lain lebih bermanfaat. #pakai akal sehat *Evidently electric car subsidies are not only for the well-off..fertilizer or something else is more useful. #use common sense.*
After Preprocessing
terbukti subsidi mobil listrik tidak hanya untuk orang yg mampu pupuk atau yg lain lebih bermanfaat *it is proven that electric car subsidies are not only for people who can afford fertilizer or other more useful things*

Figure. 5. Example of data preprocessing

Following the data collection and labeling, the next step is preprocessing. Preprocessing is a technique for cleaning data. The preprocessing procedures applied in this study are outlined below.

1) Cleaning, removing irrelevant characters including usernames, hashtags, and URLs.
2) Removing excess punctuation marks, letters, or whitespace.
3) Case folding, the process of converting text into entirely lowercase letters.

In this research, the stopword removal step is not used because the process can remove some important words or phrases that are relevant to keywords in sentiment classification. For example, the word “tidak” can be a determinant of a negative sentiment, so if removed, it can classify the sentiment as positive. An example of the results from preprocessing is depicted in Figure. 5.

2.2.3. Feature extraction

In this study, the model classification relies on the Term Frequency-Inverse Document Frequency (TF-IDF) word weighting method for feature extraction. TF-IDF is a technique used to assess the significance of a word within a collection of documents. Term frequency (TF) indicates that a word's weight increases with its frequency in a document, while inverse document frequency (IDF) suggests that a word's weight decreases as it appears more frequently across the document set. Thus, the TF-IDF score rises when a word occurs more frequently in a specific document and is less common in the overall collection of documents [23]. Feature extraction is applicable solely to machine learning models. In transfer learning, the extraction process does not start from scratch; rather, it leverages token vector weights from models that have previously been trained on a more extensive dataset. The following is the formula for calculating TF-IDF [24].

$$TF - IDF = tf_{ij} idf_i = tf_{ij} \times log \frac{D}{df_i} \quad (2)$$

2.3. Model engineering

In this research, the classification model is developed using machine learning techniques, specifically SVM and transfer learning models derived from pre-trained IndoBERT and mBERT.

1. Machine Learning

   The machine learning model used is SVM. SVM is a technique in machine learning based on vector space that aims to identify the decision boundary between two classes, ensuring it is as far away as possible from any point in the training dataset. In essence, SVM seeks to determine the optimal hyperplane by maximizing the separation between the classes. [25].

2. Transfer Learning

   - IndoBERT

     IndoBERT is a bidirectional Transformer model that has been pre-trained on an extensive Indonesian corpus (Indo4B), which includes a mix of formal and informal language sources like Indonesian Wikipedia, websites, news articles, video subtitles, blogs, and social media. [26].

   - mBERT

     Multilingual BERT, also known as mBERT, is a bidirectional Transformers model trained on 102 (uncased) and 104 (cased) languages in Wikipedia.

2.4. Model evaluation

In this research, the evaluation of the model was conducted by comparing the precision, recall, accuracy, and F1-Score metrics of both the machine learning and transfer learning models [25].

- Precision

  Precision is the proportion of true positive data to the total predicted positive data.

  $$precision = \frac{TP}{TP+FP} \quad (3)$$

- Recall

  Recall is the proportion of true positive data compared to the total amount of actual positive data.

  $$recall = \frac{TP}{TP+FN} \quad (4)$$

- Accuracy

  Accuracy is the proportion of the amount of data that is predicted correctly from all data.

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (5)$$

- F1-Score

  F1-Score is a combination measure that combines both precision and recall measures (weighted harmonic mean).

  $$F1 - Score = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (6)$$

Description:

True Positive (TP):	Actual is in the class and predicted is also in the class
False Negative (FN):	Actual is in the class and predicted is not in the class
False Positive (FP):	Actual is not in the class and predicted to be in the class
True Negative (TN):	Actual is not in the class and predicted is not in the class either

The model evaluation stage in this study utilizes the cross-validation method. Cross-validation is a technique employed to identify an suitable model for making predictions. The dataset is split into two segments: one segment is utilized to train the model, while the other is utilized for evaluation the model's predictive performance. The most fundamental form of cross-validation is called k-fold cross-validation [27]. This research uses 5-fold cross-validation.

2.5. Model operation

Based on the best model obtained, the model was implemented to predict aspect-based sentiment labels on a new unlabeled dataset. The new dataset used is obtained from scraping the comments of a YouTube video about the revision of EV regulations. Based on the implementation results, the prediction results are checked by manually calculating the evaluation metric, which compares the aspect-based sentiment label by the annotator along with the results of the predictions of the aspect-based sentiment label by the best model.

3. Result and Discussion

3.1. Business and data understanding

The scraping process was carried out on September 9th, 2023, using the YouTube Data API on each video from the top 11 YouTube video search results with the keyword “electric vehicle” and obtained a dataset of 3,669 comments. The dataset consists of comments that are relevant and irrelevant to EVs. After manual filtering of 3,669 comments, 1,881 comments are relevant and 1,788 comments are not relevant to EVs. In this study, only relevant comments were used in the analysis process.

The relevant comment data consists of Indonesian, English, and other languages. The percentage of languages used by YouTube users in commenting on transportation electrification is shown in Figure. 6. Language identification is done using the langdetect library. The results obtained state that Indonesian is the most widely used language in commenting, with a percentage of 91%. Furthermore, followed by English by 1% and the rest is categorized in other languages with a percentage of 8%. In identifying languages, there are some errors because of the constraints of the model employed by langdetect. For instance, in Indonesian comments that use informal words, the langdetect library identifies the comment as a non-Indonesian comment. Figure. 7 shows a visualization of the most frequently occurring words in the dataset of comments on EVs. The most frequent words are mobil listrik (electric car), kendaraan listrik (EV), subsidi (subsidies), mahal (expensive), infrastruktur (infrastructure), and harga (price).

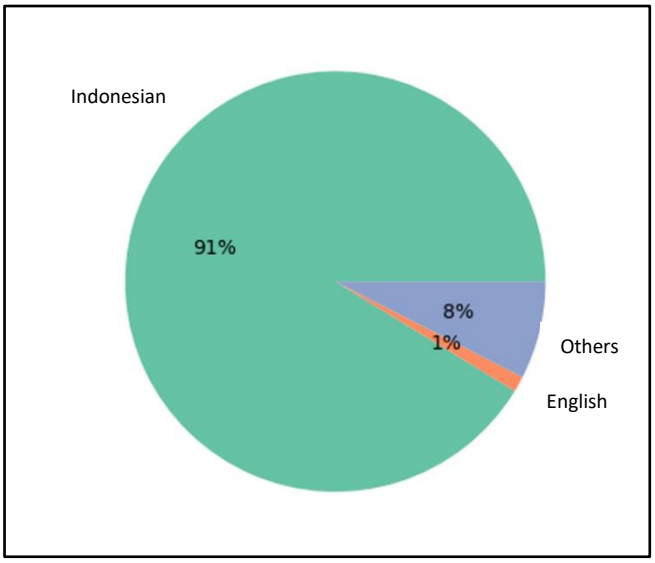


Figure. 6. Percentage of language in YouTube comment based on the langdetect library

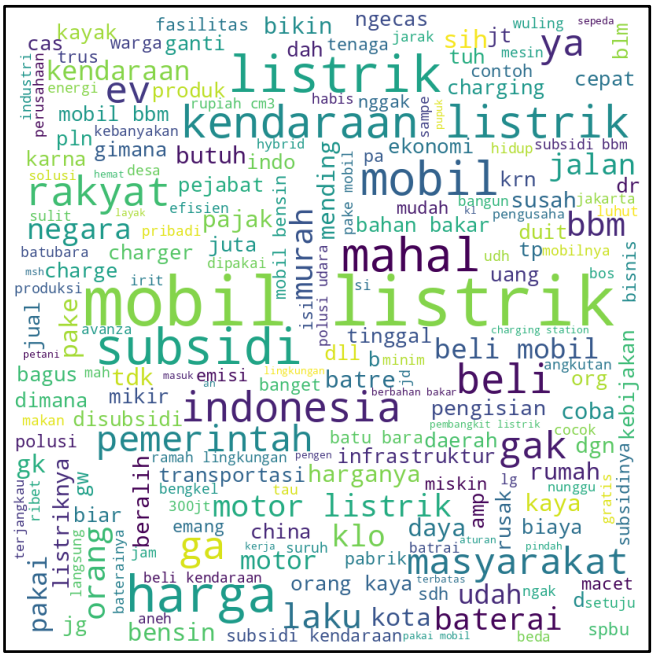


Figure. 7. Word cloud of relevant data set to EV

3.2. Data preparation

3.2.1. Data labeling

Labeling of comment data was done manually by three undergraduate students as annotators, and the final result was determined by majority voting. Labels that received disagreement from at least two annotators were excluded from the analysis process. From 1,881 datasets relevant to EVs, 1839 labeled data were obtained. There is a reduction in the dataset of 42 data due to disagreement by at least two annotators on the labels given. The distribution of labeling results after applying majority voting is shown in Figure. 8.

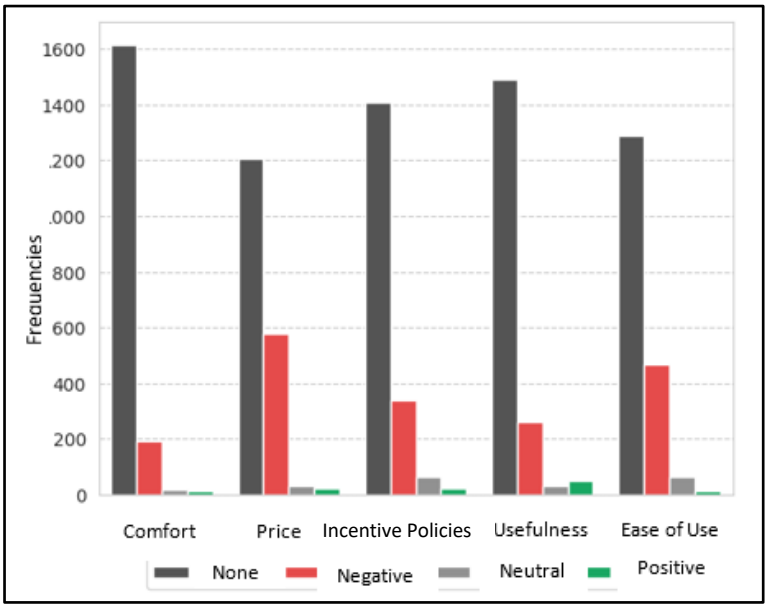


Figure. 8. Distribution of labeling results based on sentiment per aspect

Based on the distribution graph of labeling results for each aspect of EVs in Figure. 8, it is clear that the most commented aspects are the cost aspect and the ease of use aspect. To see the level of consistency of the three annotators in labeling the data, an inter-rater reliability test was conducted using Krippendorff's alpha. The higher the alpha value, the more consistent the annotators are in labeling the data and vice versa. The alpha value for each aspect of EVs is shown in Table 1. According Table 1, each aspect demonstrates a relatively high alpha value, nearing 1. This means that the three annotators have a high level of agreement in labeling, resulting in a reliable dataset.

Table 1. Inter-rater reliability test after majority voting

Aspect	Alpha
Usefulness	0.734
Ease of use	0.823
Comfort	0.776
Cost	0.852
Incentive policies	0.781

3.2.2. Dataset statistics

The characteristics of the obtained dataset are shown in Table 2. According to Table 2, it is clear that on average, there are 23 words per comment, 20 unique words per comment, and two sentences per comment.

Table 2. Dataset statistics

Characteristics	Minimum	Maximum	Average
Word count per comment	1	411	23
Number of unique words per comment	1	257	20
Number of sentences per comment	1	63	2

3.2.3. Data preprocessing

The preprocessing stages carried out in this research are cleaning, removing punctuation, letters or excess spaces, and case folding. In this research, the stopword removal step is not used because the process can remove some important words or phrases that are relevant to keywords in sentiment classification.

3.2.4. Feature extraction

Feature extraction is performed using TF-IDF and the TfidfVectorizer library from the Sklearn module to assess the significance of a word within the document. The TF-IDF hyperparameters for each machine learning model are determined from the results of a grid search, namely $min\_df = 1$, $max\_df = 0.25$, and $max\_features = 2000$.

3.3. Model Engineering

In this study, a classification model using a machine learning model, namely SVM, is constructed using a machine learning model, which is already available in the Scikit-learn library, also known as sklearn. Hyperparameters for each machine learning model are derived from the outcomes of the grid search. The SVM model is built with cost 10, gamma 0.1, and radial basis function (rbf) kernel parameters. To build transfer learning models from pre-trained models, the PyTorch and transformer libraries are utilized, employing Adam as the optimization algorithm, with a batch size of 10 and a specified learning rate of $1\ x\ 10^{-5}$ which are the default parameters.

A representation of the architecture of the pre-trained model used is shown in Figure. 9 and Figure. 10. Text classification with IndoBERT and mBERT is done by tokenizing or converting sentences into tokens, then converted into vector representations in high-dimensional space through token embeddings, segment embedding, and position embedding. The representation is passed to the transformer layer which allows the model to understand the context of the sentence in depth. The output from the Transformer layer is subsequently sent to the classification layer, which generates probabilities to determine which class the sentence belongs to in terms of aspect and sentiment.

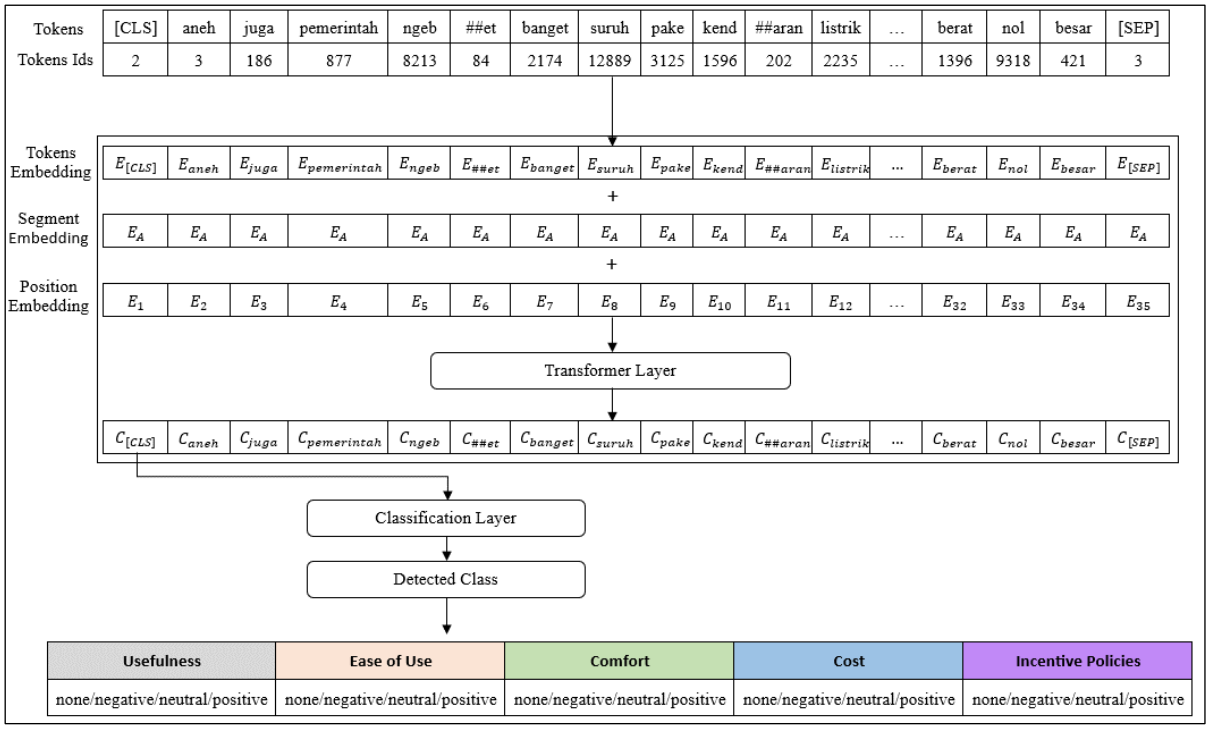


Figure. 9. Illustration of fine-tuned IndoBERT and mBERT for aspect-based sentiment classification

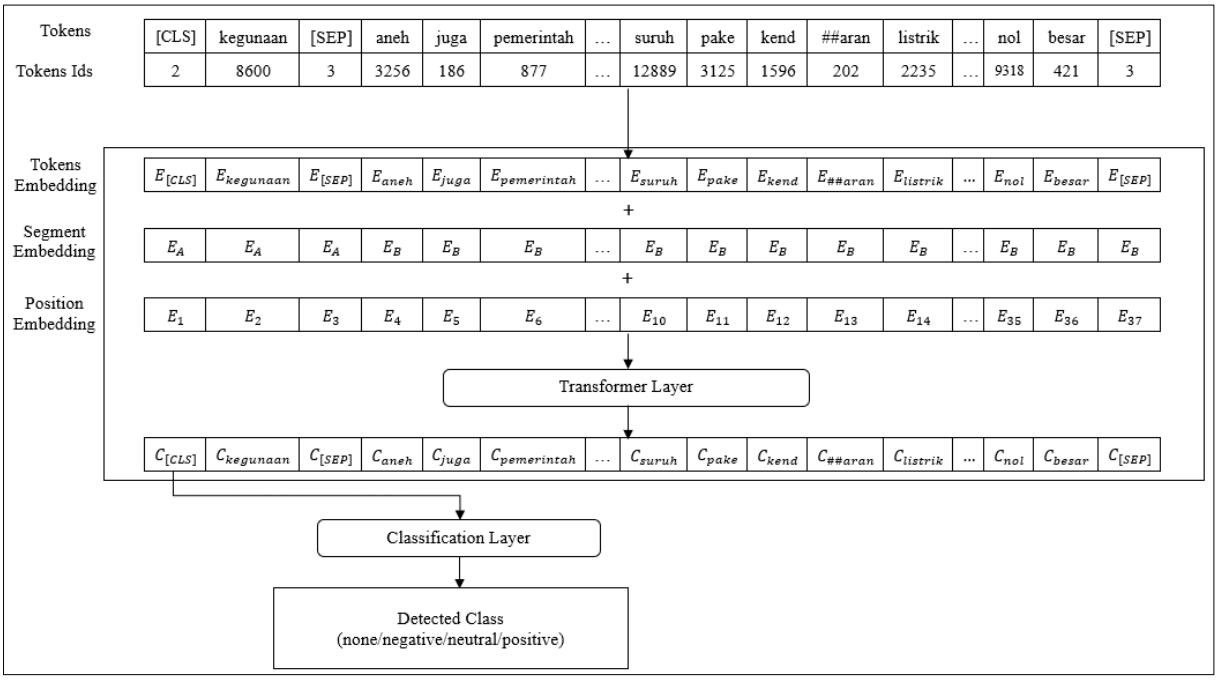


Figure. 10. Illustration of fine-tuned IndoBERT (with aspect and comment combination inputs) for aspect-based sentiment classification

3.4. Model evaluation

Experimental results of the aspect-based sentiment classification model using 5-fold cross-validation is presented in Table 3. Based on Table 3, it is clear that the transfer learning model from pre-trained IndoBERT exhibits the highest accuracy and F1-Score in comparison to the other classification models. This is because the transfer learning model has been trained in advance on a vast language corpus and adjusted its model parameters through a fine-tuning stage with training data for aspect-based sentiment analysis tasks. While the machine learning model learns solely from the training data provided in this study.

Table 3. Experimental results of aspect-based sentiment classification model using 5-fold cross-validation

Model	Accuracy	Precision	Recall	F1-Score
Machine Learning				
SVM	0.8765	**0.7798**	0.4454	0.4818
Transfer Learning				
IndoBERT	**0.8917**	0.6588	**0.4975**	**0.5266**
mBERT cased	0.8722	0.4311	0.4244	0.4188
mBERT uncased	0.8722	0.4843	0.4248	0.4258

Of all the fine-tuning models, the IndoBERT classification model records the highest evaluation scores, achieving accuracy and F1-Score values of 89.17% and 52.66%, respectively. Regarding the language corpus for each model, the mBERT model was trained on 102 languages (uncased) and 104 languages (cased), which includes Indonesian. However, in this study, the IndoBERT model provides better evaluation results. This can happen because multilingual BERT (mBERT) is better suited for multilingual datasets compared to the IndoBERT model. IndoBERT was developed using a substantial Indonesian corpus (Indo4B) that encompasses both formal and informal languages, which makes it more appropriate for the data utilized in this research. In addition, YouTube user comment data generally uses non-formal language and has a percentage of Indonesian language usage of 91%, which is much greater than the utilization of English along with other languages. As a result, the IndoBERT model is more appropriate for this study compared to the uncased or cased mBERT models.

Table 4. Aspect classification results for each label on IndoBERT using 5-fold cross-validation

Aspect	Sentiment	Precision	Recall	F1-Score	Support
Usefulness	None	0.9114	0.9626	0.9362	146
	Negative	0.6843	0.6508	0.6655	29
	Neutral	0.0000	0.0000	0.0000	4
	Positive	0.4000	0.2429	0.2927	6
Ease of Use	None	0.9127	0.9398	0.9260	129
	Negative	0.7678	0.8096	0.7874	46
	Neutral	0.5200	0.1857	0.2480	6
	Positive	0.2000	0.0667	0.1000	2
Comfort	None	0.9409	0.9818	0.9606	162
	Negative	0.7240	0.5517	0.6156	20
	Neutral	0.0000	0.0000	0.0000	1
	Positive	0.0000	0.0000	0.0000	1
Cost	None	0.9280	0.9337	0.9304	123
	Negative	0.8282	0.8565	0.8411	57
	Neutral	0.0000	0.0000	0.0000	2
	Positive	0.4000	0.1400	0.2000	3
Insentive Policies	None	0.9424	0.9619	0.9518	142
	Negative	0.7590	0.8072	0.7729	35
	Neutral	0.4500	0.1836	0.2452	6
	Positive	0.0000	0.0000	0.0000	2

The distribution of sentiment classification scores for every aspect of the IndoBERT model is presented in Table 4. Based on Table 4, it is evident that aspect-based sentiment classification utilizing the IndoBERT model provides varying F1-Score results. There are models that achieve very high scores for each aspect with the sentiment label “none,” each of which has an F1-Score of more than 90%. In addition, there is an F1-Score that exceeds 60% on each aspect with a “negative” sentiment label. There is also the lowest F1-Score reaching 0% on the “neutral” and “positive” sentiment labels in some aspects. This can occur due to the uneven distribution of datasets for each aspect and sentiment label, as indicated in the "support" column of Table 4. In addition, this also occurs due to the large number of “none” sentiment labels for each aspect in the dataset due to the limited types and number of predetermined aspects, which are only limited to the usefulness aspect, ease of use aspect, comfort aspect, cost aspect, and incentive policies aspect.

In the machine learning classification model, SVM provides better F1-Score results than the transfer learning model of pre-trained mBERT. In general, the F1-Score for each aspect and sentiment label in SVM has a similar pattern to the results of the previous IndoBERT model due to dataset imbalance. Following Liu and Zhao [19], the best IndoBERT model was developed for aspect-based sentiment classification with input in the form of a combination of aspects and comment sentences. The results of the experiment are presented in Table 5.

Table 5. Experimental results of aspect-based sentiment classification model on IndoBERT using 5-fold cross-validation

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT	0.8917	0.6588	0.4975	0.5266
IndoBERT (combination of aspect and sentence)	**0.9000**	0.6444	0.5921	**0.6070**

Based on Table 5, it can be seen that the IndoBERT model with a combination of aspects and comment sentences has higher accuracy and F1-Score than the IndoBERT model without a combination of aspect and comment sentences. This is because the presence of aspects in the input can help the model better classify comment sentences into their sentiment labels. Of all the aspect-based sentiment classification models that have been built, the IndoBERT classification model with a combination of aspect and comment sentences has the highest evaluation score, with accuracy and F1-Score values of 90% and 60.70%, respectively. The distribution of sentiment classification scores for each aspect in the IndoBERT model (combination of aspect and sentence) is shown in Table 6. In general, the F1-Score

for each aspect and sentiment label in IndoBERT (combination of aspect and sentence) increases and has a similar pattern to the results of the previous IndoBERT without combination input of aspect and sentence.

Table 6. Aspect classification results for each label on IndoBERT (combination of aspect and sentence) using 5-fold cross-validation

Aspect	Sentiment	Precision	Recall	F1-Score	Support
Usefulness	None	0.9220	0.9574	0.9392	146
	Negative	0.7203	0.6690	0.6930	29
	Neutral	0.3333	0.1400	0.1809	4
	Positive	0.4333	0.3262	0.3661	6
Ease of Use	None	0.9263	0.9522	0.9388	129
	Negative	0.8306	0.7909	0.8093	46
	Neutral	0.4655	0.5048	0.4594	6
	Positive	0.2000	0.0667	0.1000	2
Comfort	None	0.9622	0.9660	0.9639	162
	Negative	0.6951	0.6524	0.6652	20
	Neutral	0.1167	0.3000	0.1667	1
	Positive	0.5000	0.4167	0.5470	1
Cost	None	0.9452	0.9491	0.9471	123
	Negative	0.8648	0.8692	0.8662	57
	Neutral	0.0667	0.0667	0.0667	2
	Positive	0.6000	0.2700	0.3633	3
Insentive Policies	None	0.9500	0.9617	0.9555	142
	Negative	0.7975	0.8011	0.7902	35
	Neutral	0.4050	0.3569	0.3520	6
	Positive	0.0000	0.0000	0.0000	2

Comment				
Kndaraan listrik cm di ousat kl di kmpg giln ngecas di mn itu yg jd maslh trs model motor jg jlk jlk modelnya beat smua *EVs are only in the city center if in the village where to charge, that's the problem, and the motorcycle model is also ugly, all beat models.*				
SVM Prediction				
Usefulness	Ease of Use	Comfort	Cost	Incentive Policies
none	none	negative	none	none
IndoBERT Prediction				
Usefulness	Ease of Use	Comfort	Cost	Incentive Policies
none	negative	negative	none	none
IndoBERT (combination aspect and sentence) Prediction				
Usefulness	Ease of Use	Comfort	Cost	Incentive Policies
none	negative	negative	none	none
Gold Label				
Usefulness	Ease of Use	Comfort	Cost	Incentive Policies
none	negative	negative	none	none

Figure. 11. Example of comment 1, gold label and predicted label using SVM and IndoBERT

An example of a comment with the actual label given by the annotator (gold label), as well as the prediction results from the IndoBERT and SVM classification models, are shown in Figure. 11. The prediction example of the comment shows that the IndoBERT model is able to provide the same

prediction results as the gold label. Meanwhile, the SVM model did not succeed in identifying sentiment labels on the ease of use aspect. This example leads to the conclusion that the IndoBERT model is better at capturing sentence context than the SVM model.

3.5. Model operation

Based on the best model obtained, namely the IndoBERT model with input combinations of aspects and comment sentences, the implementation of the model is carried out to predict aspect-based sentiment labels on a new unlabeled dataset. The new dataset used is 69 comment data obtained from scraping the comments of a YouTube video about the revision of EV regulations.

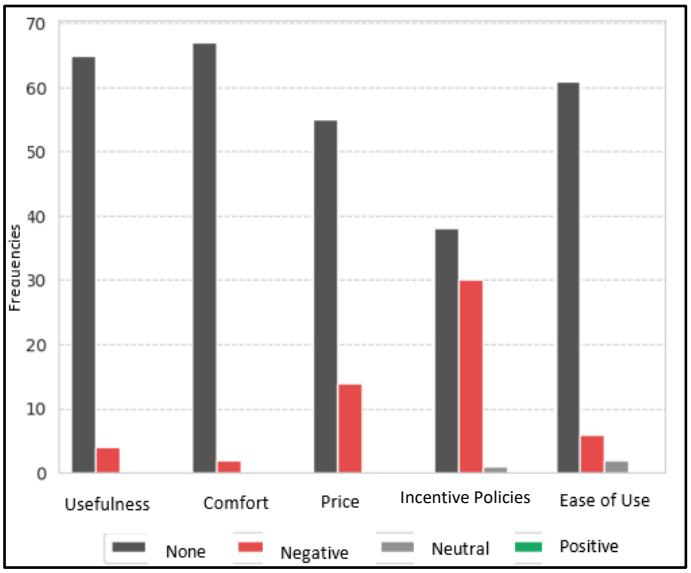


Figure. 12. Distribution of sentiment label prediction results per aspect

The results of the aspect-based sentiment label prediction of the video comments are depicted in Figure. 12. Based on Figure. 12, it can be seen that the comments on the revision of the EV rules are overall negative. The most commented aspects are the incentive policies aspect and the cost aspect.

Based on Figure. 13, it can be seen that the most frequently occurring words in the usefulness aspect are the words “mobil”, “listrik”, “menguntungkan”, “polusi”, and “lingkungan” (“car”, “electricity”, “favourable”, “pollution”, and “environment”). The words that appear most often in the ease of use aspect are the words “listrik”, “kendaraan”, “ekosistem”, “pengisian”, and “insfrastruktur” (“electricity”, “vehicles”, “ecosystem”, “charging”, and “infrastructure”). The most frequently occurring words in the comfort aspect are the words “mobil”, “listrik”, “aneh”, and “bentuk” (“car”, “electricity”, “weird”, and “shape”.). The words that appear most often in the cost aspect are the words “mobil”, “listrik”, and “harga” (“car”, “electricity”, and ’price’). Meanwhile, the words that appear most often in the incentive policies aspect are the words “pajak”, “mobil”, “listrik”, and “subsidi” (“tax”, “car”, “electricity”, and “subsidy”).

Furthermore, the prediction results are checked manually and the evaluation metrics of the aspect-based sentiment label prediction results on IndoBERT are obtained with the input of a combination of aspects and comment sentences from the dataset on the revision of EV rules. The experimental results are presented in Table 7. Based on Table 7, it is clear that the IndoBERT model with the input of a combination of aspects and comment sentences from the dataset on the revision of EV rules has an accuracy and F1-Score that is almost similar to the dataset on EVs. This means that the best model obtained has been able to perform aspect-based sentiment classification well on other EV-related datasets.

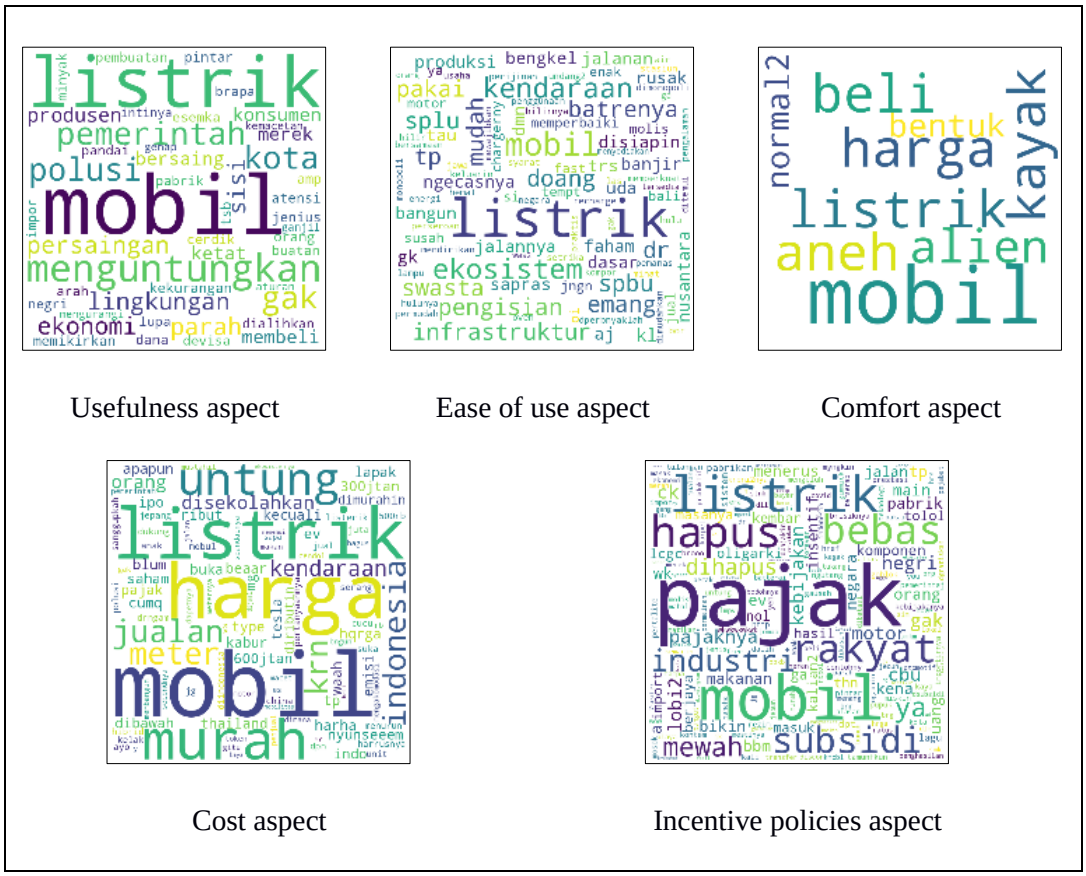


Figure. 13. Word cloud of each aspect of the new dataset

Table 7. Aspect-based sentiment label prediction result evaluation metrics on IndoBERT (combination of aspect and sentence)

Model	Accuracy	Precision	Recall	F1-Score
IndoBERT (combination of aspect and sentence)	**0.9275**	0.5994	0.6166	**0.5878**

4. Conclusion

According to the findings of the research described above, a number of conclusions can be drawn. A dataset of YouTube user comments on transportation electrification has been constructed, although it exhibits imbalanced labels; this study focuses solely on aspect-based sentiment classification. The analysis of the dataset reveals that the sentiment label "none" is the most frequently encountered across all aspects, followed by the "negative" label. The most commented aspects include cost and ease of use, with the predominant language used being Indonesian at 91%, followed by English at 1% and other languages at 8%. For the machine learning models, the hyperparameters identified through grid search include a cost of 10, a gamma of 0.1, and an RBF kernel. Meanwhile, the transfer learning models constructed from pre-trained models utilize the Adam optimization algorithm, with a batch size of 10 and a learning rate of $1\ x\ 10^{-5}$. The top-performing model generated in this research is the fine-tuned IndoBERT model, which combines aspect and comment sentences, achieving an accuracy of 90% and an F1-Score of 60.70%.

Looking ahead, this research offers several suggestions for future studies. While this work focused on five aspects related to transportation electrification—usefulness, ease of use, comfort, cost, and incentive policies—subsequent research could explore additional aspects such as security and performance to help reduce the prevalence of the "none" sentiment label. This expansion could provide a more comprehensive understanding of public opinions and preferences. By addressing these additional factors, researchers may be able to capture a broader range of sentiments, thereby helping to reduce the prevalence of the "none" sentiment label. This would enhance the overall analysis and contribute to

more informed discussions about the future of transportation electrification. Additionally, applying text augmentation techniques, such as an oversampling approach using back translation techniques to overcome the imbalance of datasets for each sentiment in each aspect. Furthermore, while this study utilized the $IndoBERT_{Base}$ model, future research might consider larger models like the $IndoBERT_{Large}$. Lastly, developing a dashboard or application could enhance user experience by facilitating aspect-based sentiment classification regarding opinions on transportation electrification.

Ethics approval

Not required.

Acknowledgments

We are very grateful for the support and contributions of all parties involved in 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

Rahmi Elfa Adilla: Conceptualization, Data Collection, Formal Analysis, Writing-Original Draft, Visualization. **Muhammad Huda**: Methodology, Writing-Review & Editing, Supervision. **Muhammad Aziz:** Writing-Review & Editing, Supervision. **Lya Hulliyyatus Suadaa:** Writing - Review & Editing, Supervision.

References

[1] IEA, "World Outlook," *Econ. Outlook*, vol. 11, no. 4, pp. 1–8, 2016, doi: 10.1111/j.1468-0319.1987.tb00425.x.

[2] H. Turton, "Sustainable global automobile transport in the 21st century: An integrated scenario analysis," *Technol. Forecast. Soc. Change*, vol. 73, no. 6, pp. 607–629, 2006, doi: 10.1016/j.techfore.2005.10.001.

[3] A. Ajanovic, "The future of electric vehicles: Prospects and impediments," *Wiley Interdiscip. Rev. Energy Environ.*, vol. 4, no. 6, pp. 521–536, 2015, doi: 10.1002/wene.160.

[4] IQAir, "Live most polluted major city ranking." Accessed: Aug. 30, 2023. [Online]. Available: https://www.iqair.com/world-air-quality-ranking

[5] A. Vafaei-Zadeh, T. K. Wong, H. Hanifah, A. P. Teoh, and K. Nawaser, "Modelling electric vehicle purchase intention among generation Y consumers in Malaysia," *Res. Transp. Bus. Manag.*, vol. 43, no. January, p. 100784, 2022, doi: 10.1016/j.rtbm.2022.100784.

[6] US Department of Energy, "Alternative Fuels Data Center: Electric Vehicle (EV) Definition." Accessed: Sep. 01, 2023. [Online]. Available: https://afdc.energy.gov/laws/12660

[7] BBC, "Kendaraan listrik disebut 'solusi palsu' untuk perbaiki kualitas udara di Indonesia." [Electric vehicles deemed a 'false solutions' for improving air quality in Indonesia] (in Indonesia). Accessed: Aug. 31, 2023. [Online]. Available: https://www.bbc.com/indonesia/articles/c51qrg47241o

[8] J. Eagle, "The Most Popular Websites by Web Traffic (1993 to 2022)." Accessed: Oct. 26, 2023. [Online]. Available: https://www.visualcapitalist.com/cp/most-popular-websites-by-web-traffic/

[9] R. Gomez, "25 YouTube Stats_ Users, Marketing, Demographics [2023] _ Sprout Social." Accessed: Oct. 26, 2023. [Online]. Available: https://sproutsocial.com/insights/youtube-stats/

[10] A. U. R. Khan, M. Khan, and M. B. Khan, "Naïve Multi-label Classification of YouTube Comments Using Comparative Opinion Mining," *Procedia Comput. Sci.*, vol. 82, no. March, pp. 57–64, 2016, doi: 10.1016/j.procs.2016.04.009.

[11] K. M. Kavitha, A. Shetty, B. Abreo, A. D'Souza, and A. Kondana, "Analysis and classification of user comments on YouTube videos," *Procedia Comput. Sci.*, vol. 177, no. 2018, pp. 593–598, 2020, doi: 10.1016/j.procs.2020.10.084.

[12] D. Jaiswal, V. Kaushal, R. Kant, and P. Kumar Singh, "Consumer adoption intention for electric vehicles: Insights and evidence from Indian sustainable transportation," *Technol. Forecast. Soc. Change*, vol. 173, no. November 2020, p. 121089, 2021, doi: 10.1016/j.techfore.2021.121089.

[13] Z. Yang, Q. Li, Y. Yan, W. L. Shang, and W. Ochieng, "Examining influence factors of Chinese electric vehicle market demand based on online reviews under moderating effect of subsidy policy," *Appl. Energy*, vol. 326, no. September, p. 120019, 2022, doi: 10.1016/j.apenergy.2022.120019.

[14] H. Mustakim and S. Priyanta, "Aspect-Based Sentiment Analysis of KAI Access Reviews Using NBC and SVM," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 16, no. 2, p. 113, 2022, doi: 10.22146/ijccs.68903.

[15] S. Jeong, "Aspect-Level Analysis and Predictive Modeling for Electric Vehicle Based on Aspect-Based Sentiment Analysis Using Machine Learning," 2020.

[16] M. T. Anwar, D. Trisanto, A. Juniar, and F. A. Sase, "Aspect-based Sentiment Analysis on Car Reviews Using SpaCy Dependency Parsing and VADER," *Adv. Sustain. Sci. Eng. Technol.*, vol. 5, no. 1, p. 0230109, 2023, doi: 10.26877/asset.v5i1.14897.

[17] R. Jena, "An empirical case study on Indian consumers' sentiment towards electric vehicles: A big data analytics approach," *Ind. Mark. Manag.*, no. November 2018, pp. 0–1, 2020, doi: 10.1016/j.indmarman.2019.12.012.

[18] J. Tao and X. Fang, "Toward multi-label sentiment analysis: a transfer learning based approach," *J. Big Data*, vol. 7, no. 1, pp. 1–26, 2020, doi: 10.1186/s40537-019-0278-0.

[19] N. Liu and J. Zhao, "A BERT-Based Aspect-Level Sentiment Analysis Algorithm for Cross-Domain Text," *Comput. Intell. Neurosci.*, vol. 2022, 2022, doi: 10.1155/2022/8726621.

[20] G. Xu, Z. Zhang, T. Zhang, S. Yu, Y. Meng, and S. Chen, "Aspect-level sentiment classification based on attention-BiLSTM model and transfer learning," *Knowledge-Based Syst.*, vol. 245, p. 108586, 2022, doi: 10.1016/j.knosys.2022.108586.

[21] S. Studer *et al.*, "Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology," *Mach. Learn. Knowl. Extr.*, vol. 3, no. 2, pp. 392–413, 2021, doi: 10.3390/make3020020.

[22] K. Krippendorff and R. Craggs, "The Reliability of Multi-Valued Coding of Data," *Commun. Methods Meas.*, vol. 10, no. 4, pp. 181–198, 2016, doi: 10.1080/19312458.2016.1228863.

[23] V. N. Gudivada, D. L. Rao, and A. R. Gudivada, "Information Retrieval: Concepts, Models, and Systems," in *Handbook of Statistics*, 1st ed., vol. 38, Elsevier B.V., 2018, pp. 331–401. doi: 10.1016/bs.host.2018.07.009.

[24] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," *Procedia Eng.*, vol. 69, pp. 1356–1364, 2014, doi: 10.1016/j.proeng.2014.03.129.

[25] A. Shiri, *Introduction to Modern Information Retrieval (2nd edition)*. Cambridge University Press, 2008. doi: 10.1108/00242530410565256.

[26] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," pp. 843–857, 2020, [Online]. Available: http://arxiv.org/abs/2009.05387

[27] P. Reafeilzadeh, L. Tang, and H. Liu, "Cross Validation," in *Contemporary Interventional Ultrasonography in Urology*, 2009, pp. 1–6. doi: 10.1007/978-1-84800-217-3_1.

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.791

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Available online at https://jurnal.stis.ac.id/

The Impact of ICT on Regional Economy in Indonesia Through MSEs as Mediators: Application of Causal Mediation Analysis in Instrumental-variable Regressions

Luthfio Febri Trihandika[1*], Ribut Nurul Tri Wahyuni[2], Meilinda Fitriani Nur Maghfiroh[3]

[1]BPS-Statistics Indonesia, Jakarta, Indonesia, [2]Politeknik Statistika STIS, Jakarta, Indonesia, [3]Muscat University, Muscat, Oman
**Corresponding Author: E-mail address:* febrifio22@gmail.com

ARTICLE INFO

Article history:
Received 29 August, 2024
Revised 11 November, 2024
Accepted 11 November, 2024
Published 31 December, 2024

Keywords:
Causal Mediation Analysis, Instrumental Variable, ICT, Local Economy, Nighttime Light, SMEs

Abstract

Introduction/Main Objectives: The development of information and communication technology (ICT) and small and micro enterprises (SMEs) can encourage regional economic growth. **Background Problems:** Studies on the impact of ICT on the rural economy at the village level are very limited. Furthermore, the Indonesian study neglects to tackle the endogeneity issue associated with this variable and the indirect effects of ICT on the regional economy. **Novelty:** Using SMEs as a mediator, this study examines the impact of ICT (internet signal strength) on the village's local economy (nighttime light), both directly and indirectly. ICT is considered to be endogenous. **Research Methods:** This study employs causal mediation analysis in instrumental-variable (IV) regressions at the village level in 2018 and 2021, using lightning strike intensity as IV. **Finding/Results:** Internet signal strength can increase the number of SMEs, and this increase can positively and significantly impact the local economy. In addition, the direct impact of internet signal strength on the local economy is significantly negative. However, the total impact of internet signal strength is significantly positive.

1. Introduction

Development is critical to a country's sustainability. The objectives of this development are high economic growth, poverty alleviation, and improving the quality of human resources (HR) [1]. In developing countries, such as Indonesia, the primary goal of development implementation is economic growth, as it can bridge Indonesia's gap with developed countries. The realization of this condition can occur when the processing industry sector demonstrates satisfactory performance [2].

The manufacturing industry has been the most significant contributor to the Indonesian economy, as evidenced by data from BPS-Statistics Indonesia. The manufacturing industry's contribution to Indonesia's Gross Domestic Product (GDP) in 2022 was 18.34 percent, which was significantly higher than the other two highest sectors (agriculture at 12.4 percent and trade at 12.85 percent) [3]. The manufacturing industry sector also has an impact on job creation [4]. In 2023, this sector was capable of accommodating 14.17 percent of the total working population in Indonesia. Consequently, this sector ranks third in labor absorption, following the agriculture and trade sectors [5].

BPS-Statistics Indonesia classifies manufacturing industry enterprises into four categories: large industry (100 or more workers), medium industry (20-99 workers), small industry (5-19 workers), and micro industry (1-4 workers). Small and micro enterprises (SMEs) are the most significant of the four

categories, as they account for 99 percent of manufacturing industry enterprises and employ 60 percent of the industry's workforce [6]This demonstrates that SMEs can be a solution to economic issues, including poverty and unemployment. To optimize SMEs' potential, it is imperative that they are provided with the necessary ICT infrastructure.

Many previous studies have shown that ICT plays an important role in the manufacturing industry [7-9] because ICT can increase productivity, encourage innovation, and create new jobs [10-14]. Furthermore, ICT can assist enterprises in identifying strategic locations for economic activities [15]. Therefore, we suppose that the increase in ICT stimulates the growth of SMEs and indirectly boosts the regional economy.

Multiple studies have examined the impact of ICT on the regional economy [41–43]. Nonetheless, these studies have just examined the direct effects of ICT. In actuality, ICT can impact the regional economy through both direct and indirect means [16–20]. Mediation analysis can elucidate these effects; however, it neglects to address the endogeneity of ICT. The strength of the internet signal, serving as a proxy for the ICT variable, is not wholly arbitrary. This variable may exhibit endogeneity and non-randomness [21-22].

Regions with higher populations, superior infrastructure, and income levels may exhibit enhanced signal strength. This could make signal strength endogenous, as the decision to develop communications infrastructure could correlate with the error term, potentially influencing the outcome variables in our study. Indonesia frequently encounters several natural disasters. Anticipating such catastrophes would diminish the motivation for telecommunications firms to invest in infrastructure in regions with a higher likelihood of natural disasters, resulting in worse signal strength and related consequences. Reverse causality also skews the results since regions with superior social development circumstances may exhibit robust mobile phone signals [28]. Neglecting these problems may result in bias within the mediation analysis [27]. This study will utilize an instrumental variable approach to mitigate these issues by using the fluctuation in lightning strike intensity. Reduced connectivity is attributable to a rise in lightning occurrences. This study forecasts that the intensity of lightning strikes may affect signal quality in Indonesia, a tropical area marked by considerable geographical diversity [28]. Furthermore, previous studies did not utilize village-level data. Given these limitations, this study will utilize causal mediation analysis within instrumental-variable (IV) regressions to investigate the direct and indirect impacts of ICT on the village economy. This study employs nighttime light (NTL) data as an indicator of the village economy [23] and utilizes the number of SMEs as a mediating variable.

Economic growth theory and location theory can elucidate the relationship between ICT, SMEs, and regional economies. Location theory functions as a conceptual framework for examining the factors that influence the location decisions of enterprises [24]. Neoclassical, institutional, and behavioral theories are the three theories that Hayter discusses regarding the location of economic activities [25]. Neoclassical theory aims to maximize profits or minimize costs by selecting the optimal location based on economic agglomeration factors, technology, and human resources. The network of economic relations determines profit and cost functions in institutional theory, which is the basis for location determination. Conversely, behavioral theory underscores the significance of individual preferences. Dicken observed the influence of ICT on enterprise location decisions, specifically the necessity for technology-based labor, internet connectivity, and access to technological infrastructure [26].

Meanwhile, the Solow model explains the relationship between ICT and regional economies. The model explains how capital stock, labor force, and technological progress interact with the economy, as well as how they affect the total output of goods and services in a country [27]. However, the Solow Model still has gaps because it assumes that technological growth is an exogenous variable, meaning that technology is considered something that occurs outside the economic system. The endogenous growth theory serves as a complementary theory to address the shortcomings of the Solow model. In this theory, technology is considered an endogenous variable where economic factors, such as investment, human resource development, and economic policy, directly affect the rate of technological growth [27]. This theory also links the role of industrial innovation with economic growth. The industrial sector's innovation development through technology adoption can accelerate economic growth. The difference between the Solow Model theory and the Endogenous Growth theory reflects the different approaches to the role of technology in explaining long-term economic growth.

2. Material and Methods

2.1. Scope of Study

This study focuses on analyzing the impact of ICT on the development of SMEs and regional economies at the village level in Indonesia in 2018 and 2021. The strength of the Internet signal is a proxy for ICT; the number of SMEs is a proxy for SMEs' development; and NTL data is a proxy for the village economy (known as the Gross Regional Domestic Product (GRDP)). The idea that villages with higher levels of illumination tend to exhibit greater economic dynamism is the foundation of the correlation between NTL and economic output. Higher population density, improved infrastructure, and increased industrial activity typically attribute this association. We overcome the endogeneity problem in the internet signal strength variable by using the instrument variable, lightning strike intensity, at the village level [28]. The study also includes several control variables, such as demographic, government, and geographic characteristics.

2.2. Data dan Data Sources

The majority of data used in this study are secondary data sourced from BPS-Statistics Indonesia. In addition, this study also uses satellite imagery data obtained from Earth Observation Group (EOG) [29], NASA's Global Hydrometeorology Resource Center (GHRC) [30], WorldPop [31], and WorldClim [32]. Table 1 provides details of the data used.

Table 1. Data and Data Sources

Variable	Definition	Data Level	Data Source
Main variables			
T	Internet signal strength (1 = signal is strong, 0 = otherwise)	Village	BPS-Statistics Indonesia
M	The number of SMEs (unit)	Village	BPS-Statistics Indonesia
Y	NTL data (nW/cm^2/sr)	Village	BPS-Statistics Indonesia
Z	Lightning strike intensity (flash per village)	Village	GHRC NASA
Demographic and government characteristics (X_1)			
$agri$	Main income sources (1 = The majority of the population is employed in agriculture, 0 = otherwise)	Village	BPS-Statistics Indonesia
$muslim$	Muslim dummy (1 = the majority of the population is Muslim, 0 = otherwise)	Village	BPS-Statistics Indonesia
$christian$	Christian dummy (1 = the majority of the population is Christian, 0 = otherwise)	Village	BPS-Statistics Indonesia
$ethnic$	Ethnic dummy (1 = the population consists of several ethnic groups, 0 = otherwise)	Village	BPS-Statistics Indonesia
age	Age of the village head (year)	Village	BPS-Statistics Indonesia
sex	Sex of the village head (1 = if the village head is male, 0 =otherwise)	Village	BPS-Statistics Indonesia
$educ$	Education of the village head (1 = if the village head's education is high school or below, 0 = otherwise)	Village	BPS-Statistics Indonesia
Demographic characteristics in the previous year (X_2)			
$lagpop$	The number of populations in the previous year (persons)	Village	WorldPop
RLS	Years of schooling in the previous year (years)	District	BPS-Statistics Indonesia
Geographic characteristics (X_3)			
$temp$	Average temperature (ºC)	Village	WorldClim
$elevation$	Average elevation (m)	Village	WorldClim
$rainfall$	Average rainfall (mm)	Village	WorldClim
$flatland$	Topography (1 = if the village topography is flatland, 0 = otherwise)	Village	BPS-Statistics Indonesia

2.3. Causal Mediation Analysis in Instrumental-variables Regressions

In many contexts, researchers often seek to comprehend the process that explains the estimated effect of a treatment on an outcome. In this study, we are interested in the hypothesis that ICT, as a treatment variable (T), affects the regional economy as an outcome variable (Y), through the number of SMEs as an intermediate variable (M). Since the distribution of SMEs across villages is probably not random, we present an IV (Z) to prove that ICT contributes to an increase in the number of SMEs, thereby stimulating economic growth. In IV regressions, this method is known as causal mediation analysis.

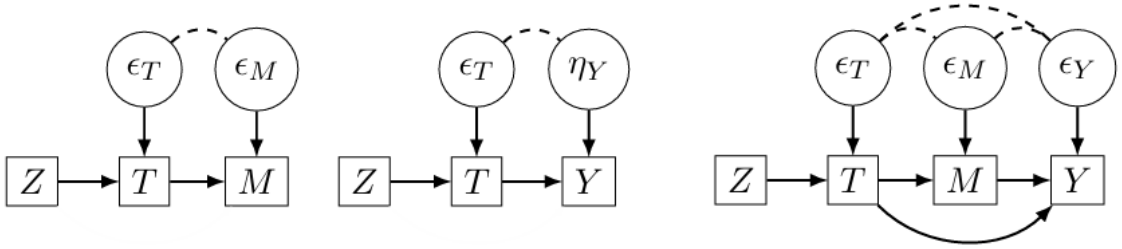
Figure 2 illustrates the identification challenge described above. Model A is a standard instrumental variable two-stage least squares (IV-2SLS) model. This model allows us to identify the causal effects of T on M. Model B is the standard IV-2SLS model that enables the identification of T's causal effects on Y. Model C is the IV mediation model, with an instrumental variable Z. It estimates three separate 2SLS regressions: the effect of T on M, the effect of T on Y, and the effect of M on Y conditional on T. Model C can decompose the total effect of T on Y into an indirect effect of T on Y that operates through M and a direct effect that does not work through M [33].

Prior to employing the IV-2SLS model, we conduct endogeneity tests to ascertain whether an explanatory variable in a regression model is endogenous. The Wu-Hausman test is one of the most commonly used tests for endogeneity. It checks explicitly whether the difference between the OLS and IV-2SLS estimates is statistically significant. The test employs the following hypothesis:

H_0: The difference between the OLS and IV-2SLS estimates is zero (endogeneity is not present).

H_1: The difference between the OLS and IV-2SLS estimates is non-zero (endogeneity is present).

If the test statistic is significant, it suggests that the OLS estimates are inconsistent due to endogeneity, and IV-2SLS estimation is necessary. Even though the test result is not statistically significant, this study maintains the assumption that ICT is endogenous, as other variables can indeed affect it. Previous studies that used the IV mediation method did not conduct the Hausman test [28, 44].



(a) Model A (b) Model B (c) Model C

Figure 2. The Identification Problem of Causal Mediation Analysis

Under linearity, we can write the causal relations in Model C in Figure 2 as

$$\begin{bmatrix} Z \\ T \\ M \\ Y \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_T^Z & 0 & 0 & 0 \\ 0 & \beta_M^T & 0 & 0 \\ 0 & \beta_Y^T & \beta_Y^M & 0 \end{bmatrix} \cdot \begin{bmatrix} Z \\ T \\ M \\ Y \end{bmatrix} + \begin{bmatrix} \epsilon_Z \\ \epsilon_T \\ \epsilon_M \\ \epsilon_Y \end{bmatrix} \quad (1)$$

Parameter β_M^T is identified by the standard IV-2SLS model, described by the following equations:

First stage: $T = \beta_T^Z . Z + \epsilon_T$ (2)

Second stage: $M = \beta_M^T . \hat{T} + \epsilon_M$ (3)

$\hat{T}$ are the estimated values of T in the first stage. β_Y^M and β_Y^T can be fit by the following IV-2SLS model:

First stage: $M = \gamma_M^Z . Z + \gamma_M^T . T + \epsilon_M$ (4)

Second stage: $Y = \beta_Y^M . \hat{M} + \beta_Y^T . T + \epsilon_Y$ (5)

$\hat{M}$ stands for the estimated values of M in the first stage.

There are two first stages (Equation 2 and Equation 4). Causal mediation analysis assesses weak identification by reporting the corresponding F statistics on the excluded instrument. A rule of thumb is that an F test of the excluded instrument in the first stage should yield an F statistic of 10 or more [34]. When using a robust standard error, the regression output presents the F statistic calculated by Kleibergen-Paap Wald (KPW) [35]. There is a link between equations 2-5 and the direct estimation of the total effect in Model B of Figure 2. We obtain Model B from Model C by substituting Equation 3 into Equation 5.

$$Y = \beta_Y^M.(\beta_M^T.T + \epsilon_M) + \beta_Y^T.T + \epsilon_Y = (\beta_Y^M.\beta_M^T + \beta_Y^T)T + \beta_Y^M.\epsilon_M + \epsilon_Y \quad (6)$$

Equation 6 shows that the estimate of the total effect produced by Model B is identical to the product of estimates $\beta_Y^M.\beta_M^T + \beta_Y^T$ produced by Model C (equations 2-5). In Model C, the direct effect is shown by β_Y^T, while the indirect effect is shown by $\beta_Y^M.\beta_M^T$. By incorporating the control variable vectors $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, and $\boldsymbol{X_3}$ into Equations 2-6, Equations 2-6 is transformed as follows:

$$T = \beta_T^Z.Z + \boldsymbol{\beta_T^1 X_1} + \boldsymbol{\beta_T^2 X_2} + \boldsymbol{\beta_T^3 X_3} + \epsilon_T \quad (7)$$

$$M = \beta_M^T.\hat{T} + \boldsymbol{\beta_M^1 X_1} + \boldsymbol{\beta_M^2 X_2} + \boldsymbol{\beta_M^3 X_3} + \epsilon_M \quad (8)$$

$$M = \gamma_M^Z.Z + \gamma_M^T.T + \boldsymbol{\beta_M^1 X_1} + \boldsymbol{\beta_M^2 X_2} + \boldsymbol{\beta_M^3 X_3} + \epsilon_M \quad (9)$$

$$Y = \beta_Y^M.\widehat{M} + \beta_Y^T.T + \boldsymbol{\beta_Y^1 X_1} + \boldsymbol{\beta_Y^2 X_2} + \boldsymbol{\beta_Y^3 X_3} + \epsilon_Y \quad (10)$$

$$Y = \beta_Y^M.\left(\beta_M^T.T + \boldsymbol{\beta_M^1 X_1} + \boldsymbol{\beta_M^2 X_2} + \boldsymbol{\beta_M^3 X_3} + \epsilon_M\right) + \beta_Y^T.T + \boldsymbol{\beta_Y^1 X_1} + \boldsymbol{\beta_Y^2 X_2} + \boldsymbol{\beta_Y^3 X_3} + \epsilon_Y = (\beta_Y^M.\beta_M^T + \beta_Y^T)T + \left(\boldsymbol{\beta_M^1} + \boldsymbol{\beta_Y^1}\right)\boldsymbol{X_1} + (\boldsymbol{\beta_M^2} + \boldsymbol{\beta_Y^2})\,\boldsymbol{X_2} + (\boldsymbol{\beta_M^3} + \boldsymbol{\beta_Y^3})\,\boldsymbol{X_3} + \epsilon_M + \epsilon_Y \quad (11)$$

Equation 11 shows the same result. The indirect effect is β_Y^T, while the indirect effect is $\beta_Y^M.\beta_M^T$.

3. Results and Discussion

3.1. NTL as a Proxy for Regional Economy

As previously explained, researchers use NTL data to approximate the regional economy (GRDP) in regions with limited or inaccurate economic data [23]. BPS-Statistics Indonesia has not yet provided GRDP data at the village level. This study demonstrates the relationship between GRDP and NTL data at the district level using thematic maps, scatter plots, and Pearson correlation calculations. Figure 3 and Figure 4 are thematic maps showing the distribution of GRDP and NTL data in each district in Indonesia in 2022. The redder the color, the higher the GRDP and NTL values. Based on these two figures, the majority of districts in western Indonesia are red. This means that the majority of districts in this region have high GRDP and NTL values. On the other hand, the majority of districts in eastern Indonesia are green, indicating that the GRDP and NTL values in this region are relatively low. The similar distribution of GRDP and NTL data demonstrates that NTL data can represent GRDP data.

Table 2. Pearson Correlation between GRDP and NTL Data at the District Level in Indonesia by Year

Year	Pearson Correlation	p-value
2017	0.747***	0.000
2018	0.758***	0.000
2019	0.750***	0.000
2020	0.744***	0.000
2021	0.733***	0.000
2022	0.717***	0.000

Note: *** denotes 1% significance level.

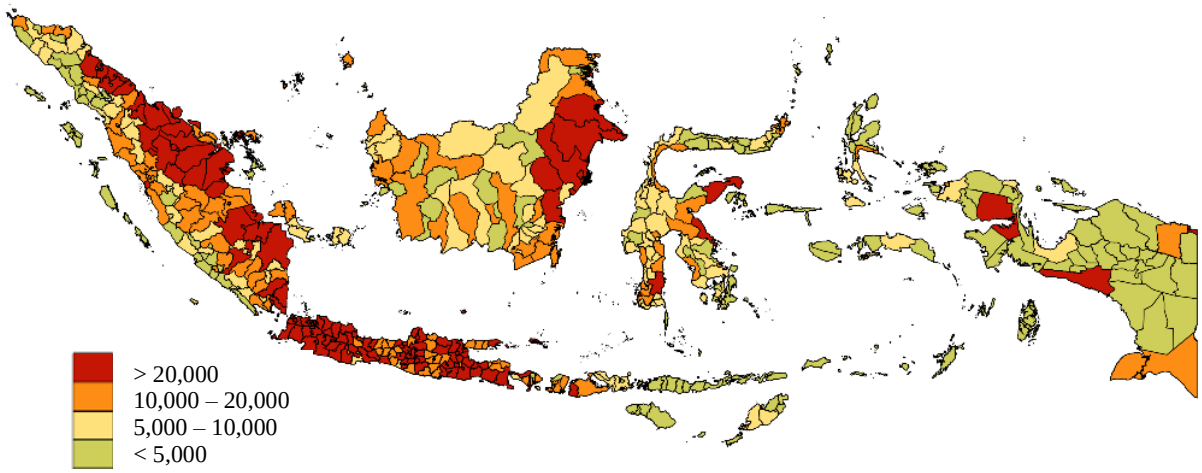


Figure 3. GRDP in Indonesia in 2022 by District (Billion Rupiah)

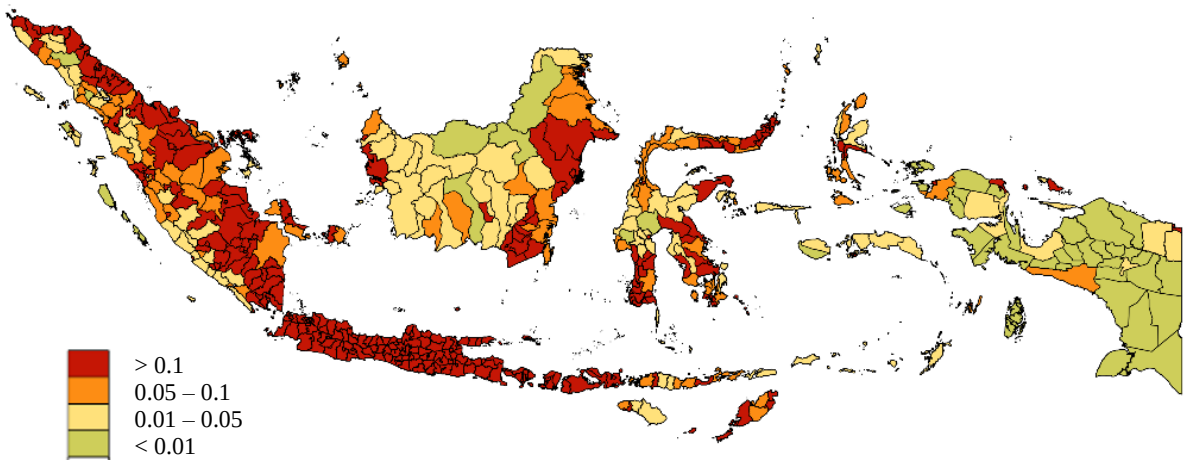


Figure 4. NTL data in Indonesia in 2022 by District (nW/cm^2/sr)

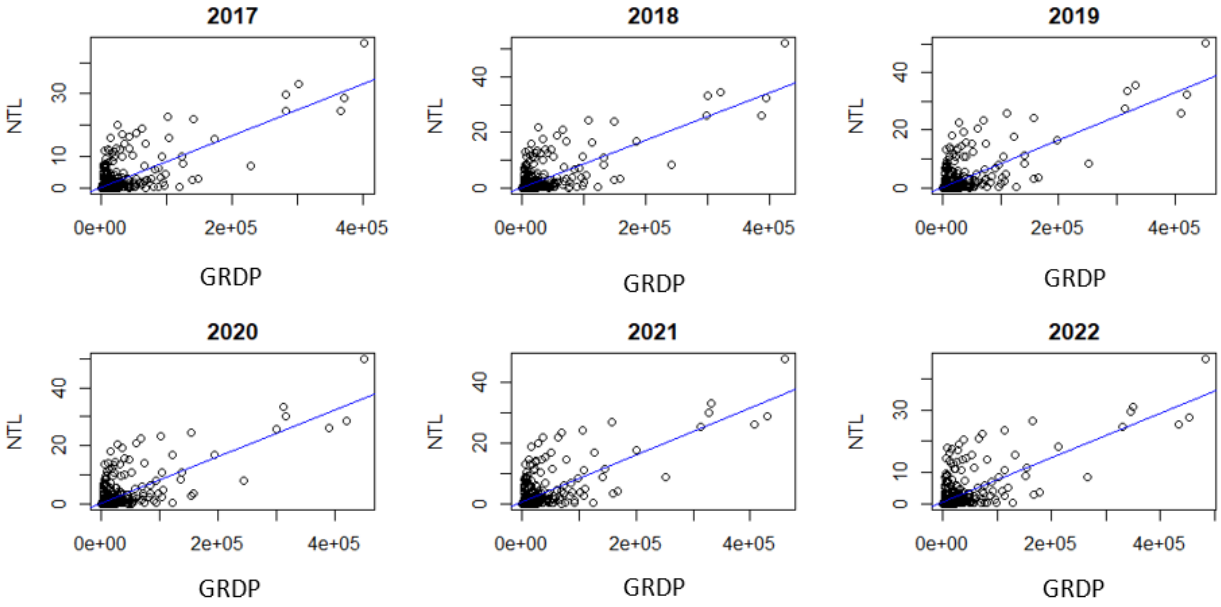
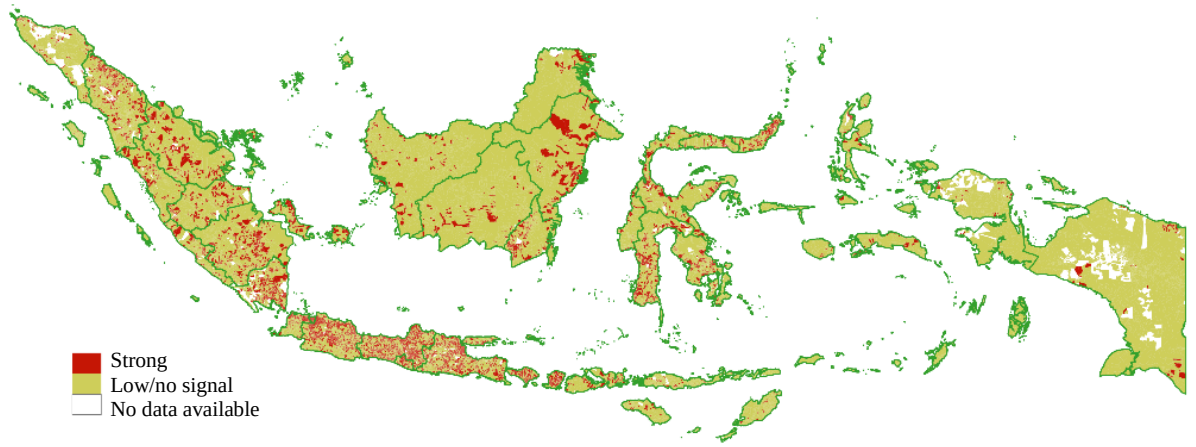


Figure 5. Scatter Plot between GRDP and NTL Data in Indonesia in 2017-2022

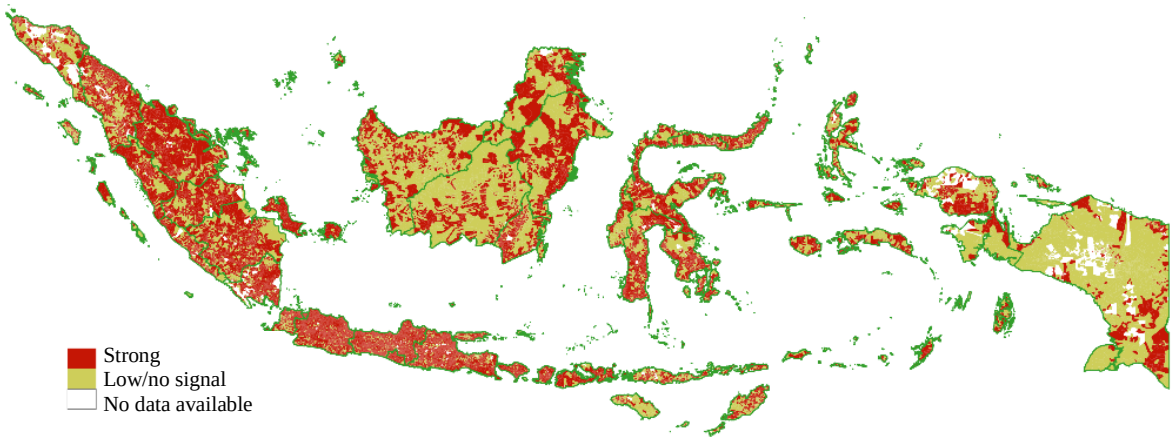
The relationship between NTL data and GRDP is clearly visible in the scatter plot in Figure 5. During the 2017-2022 period, NTL data and GRDP have a positive relationship. The NTL data increases in tandem with the district's GRDP value. In line with this argument, the Pearson correlation between PDRB and NTL data at the district level also provides similar results, as shown in Table 2. During 2017–2022, the Pearson correlation value of these variables was always greater than 0.7 and had a p-value of less than 1%. It shows that at a significance level of 1%, PDRB data is positively and significantly correlated with NTL data. Therefore, we can use NTL data as a proxy for PDRB data.

3.2. Lightning Strike Intensity, Internet Signal Strength, the Number of SMEs, and Regional Economy in Indonesia by Regions

Figure 6 presents the results of mapping the internet signal strength in each village in 2018 and 2021. The red color indicates areas with strong internet signals, while the green color indicates areas with weak or no internet signals. The map reveals that the distribution of strong signals remained uneven in 2018. However, very rapid development occurred in 2021. The distribution of strong internet signals is much more even, especially in areas located on the islands of Java and Sumatra. However, some areas, such as Papua Island, continue to experience weak or no internet signals. This is likely due to the conflict in the Papua region, which has made it difficult for the government to carry out development. Topographically, Papua also has many mountainous areas. Furthermore, the government may still excessively focus on advancing development in Western Indonesia, causing the Eastern region to lag behind. Several areas on the island of Kalimantan still lack a strong internet signal. However, the distribution is still more evenly distributed than on Papua Island.



2018



2021

Figure 6. Internet Signal Strength at the Village Level in Indonesia in 2018 and 2021

A map of the distribution of NTL values in each village in Indonesia in 2018 and 2021 is showcased in Figure 7. In general, the distribution pattern of NTL in 2018 and 2021 does not exhibit any substantial differences. The thematic map in Figure 7 suggests that the distribution of NTL is more extensive in Sumatra and Java. The NTL value decreases as one moves further east. The East and South Kalimantan

regions of Kalimantan Island are the primary locations for NTL distribution. Consequently, the distribution of NTL in Sulawesi Island is confined to the North and South Sulawesi regions. In the interim, the vast majority of Papua Island is devoid of any NTL value. The NTL value can offer a glimpse into the economic activity of a given region. Therefore, it is possible to infer that a high level of economic activity characterizes Western Indonesia. Economic activity decreases as one moves eastward. The distribution pattern of NTL data is quite similar to the distribution pattern of internet signal intensity data. Consequently, there is a suggestion of a unidirectional relationship between the NTL's value and the internet signal's intensity.

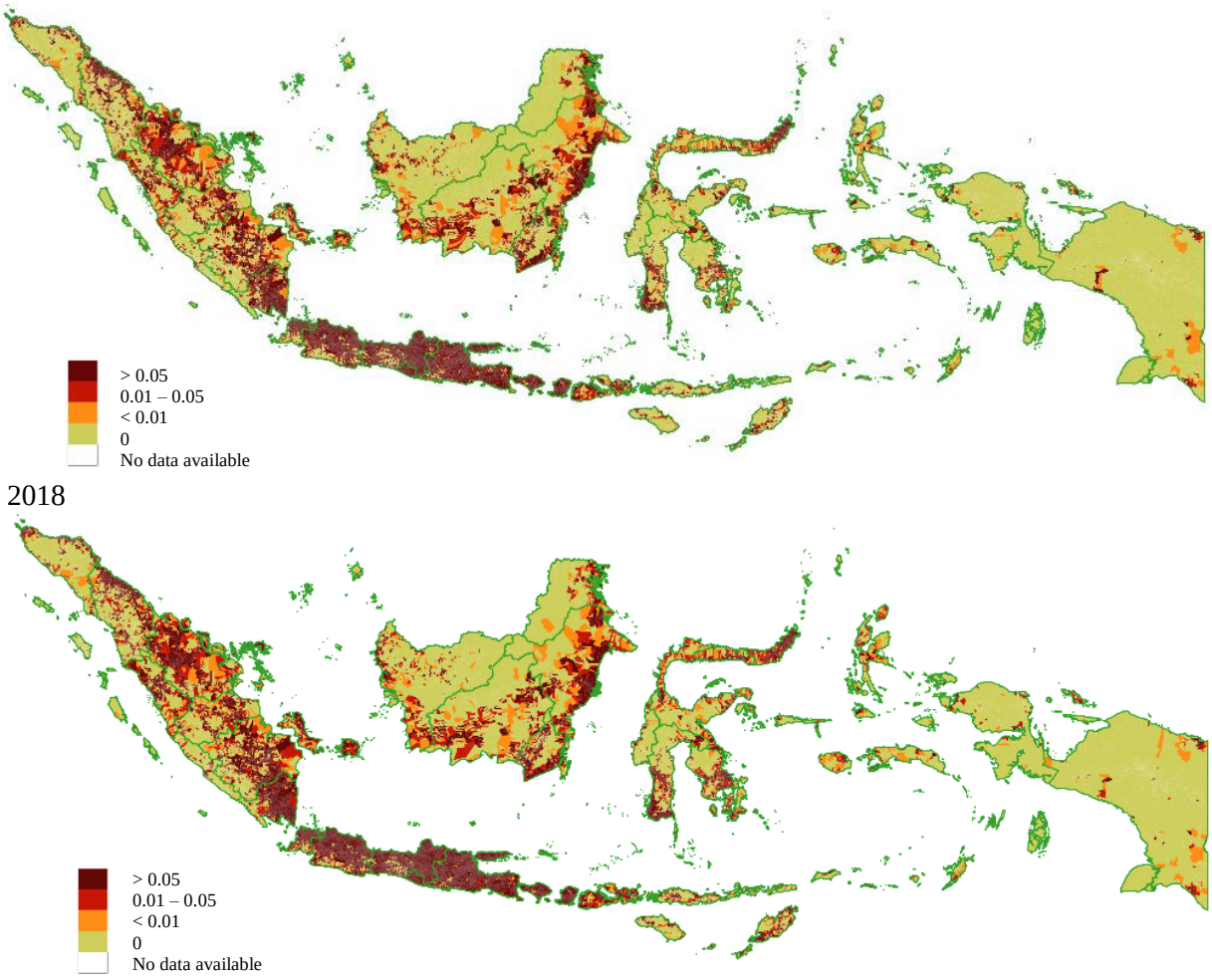


Figure 7. NTL at the Village Level in Indonesia in 2018 and 2021

The results of the mapping of the number of SMEs in each village in 2018 and 2021 are depicted in Figure 8. The data distribution pattern regarding the number of SMEs generally remains consistent between 2018 and 2021. Nevertheless, the coverage area of SMEs in 2021 appears to be more extensive than in 2018. The majority of SMEs are situated in the western regions of Indonesia, specifically in Java and Sumatra, similar to the two preceding variables. Nevertheless, the Kalimantan and Sulawesi islands have a more extensive coverage. In the interim, the Papua Island region maintains a minimal distribution, which is consistent with the two preceding variables. The similarity in the data distribution pattern between the number of SMEs, internet signal intensity, and NTL demonstrates the unidirectional relationship between the three variables.

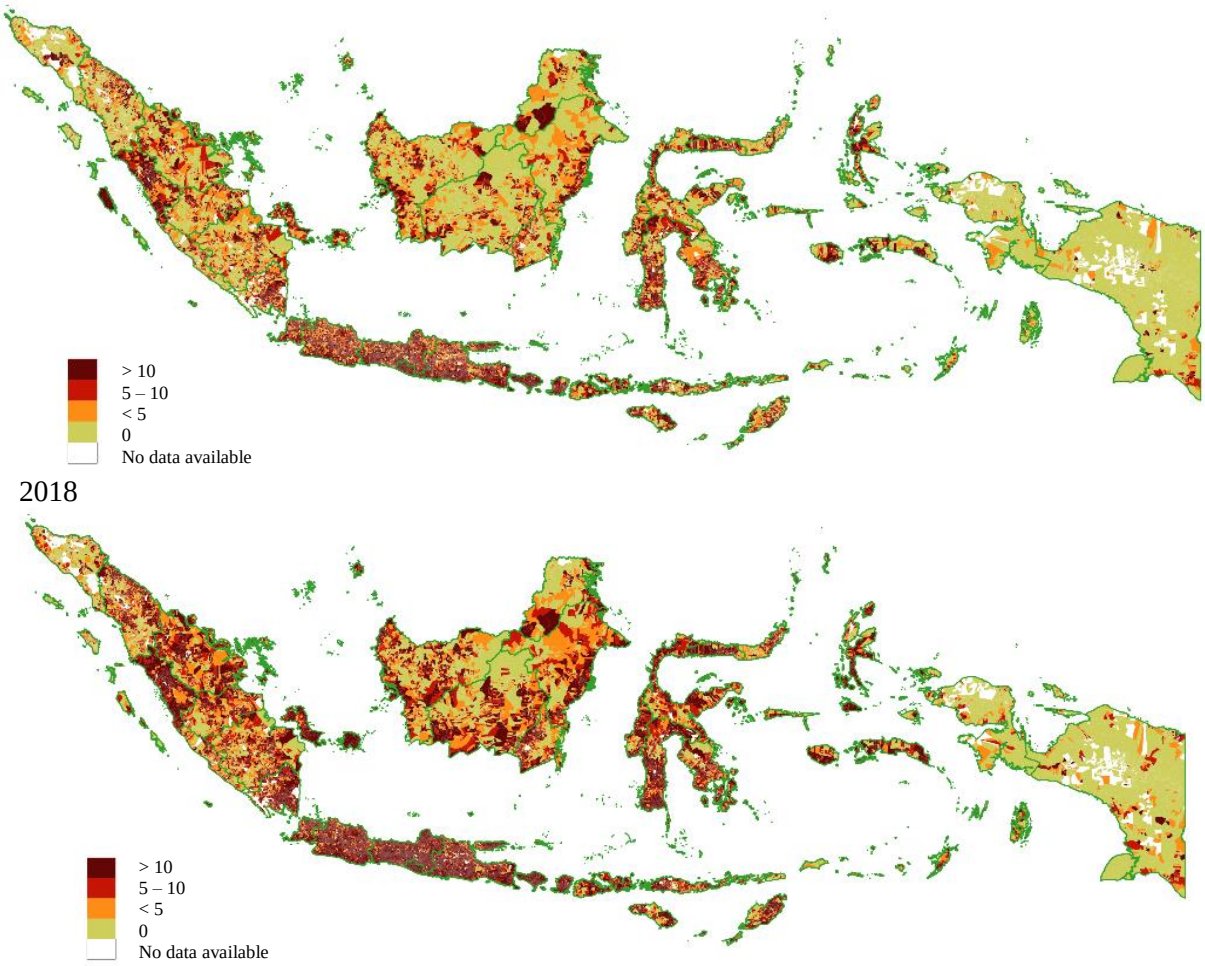
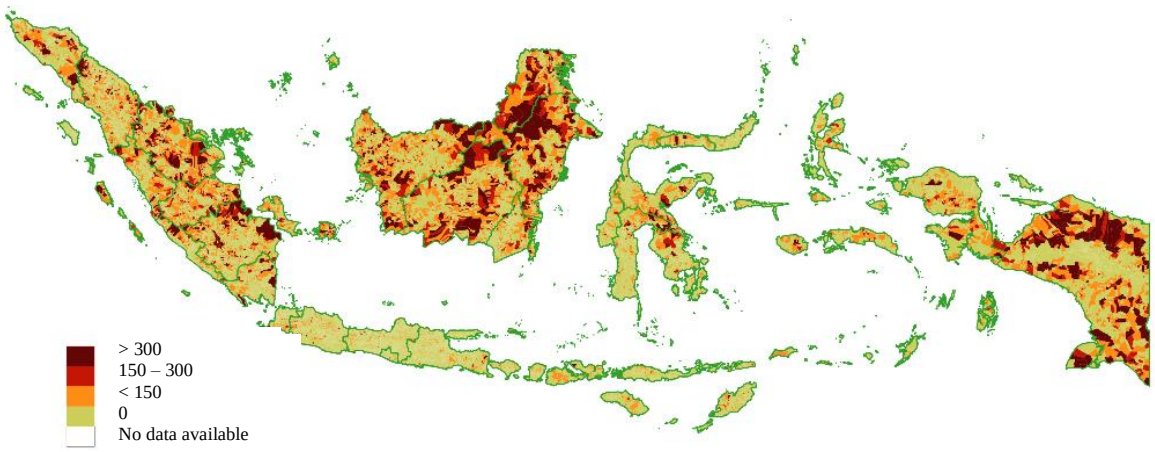
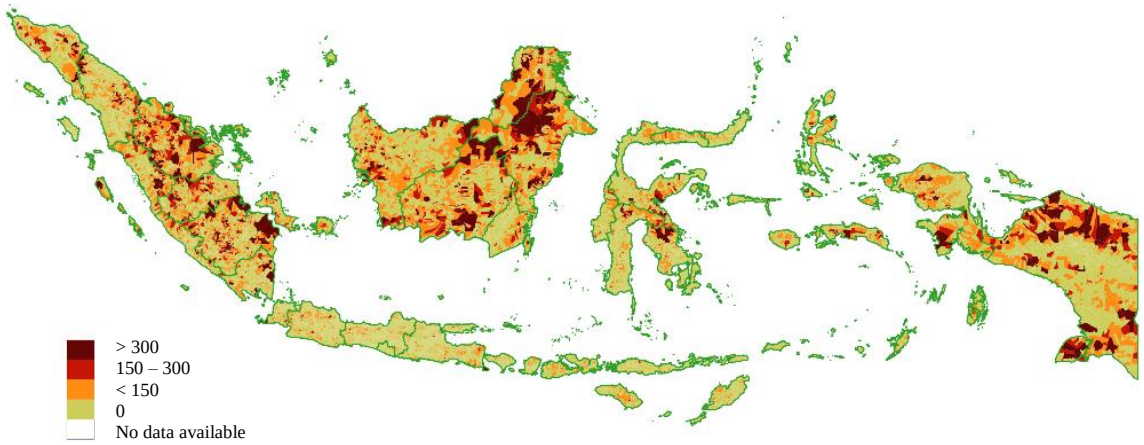


Figure 8. The Number of SMEs at the Village Level in Indonesia in 2018 and 2021

Lastly, Figure 8 illustrates a depiction of the number of lightning flashes in each village in Indonesia between 2018 and 2021. There are no substantial distinctions between the data patterns of 2018 and 2021. The distribution pattern of lightning flashes is distinct from that of the preceding three variables, particularly in Java. The situation is reversed in this instance, as Java had the highest distribution in the preceding three variables. Compared to other regions, Java has the lowest distribution of lightning bolts. In contrast, the distribution of lightning bolts is particularly high in Kalimantan and Papua. The distribution of lightning flashes is more pronounced in regions with reduced internet signal strength when correlated with this variable. This suggests that the intensity of the internet signal is inversely proportional to the number of lightning flashes.



2018



2021

Figure 8. Lightning Strike Intensity at the Village Level in Indonesia in 2018 and 2021

3.3. Correlation between Lightning Strike Intensity, Internet Signal Strength, the Number of SMEs, and Regional Economy

Figure 9 presents a descriptive analysis of the relationship between the variables of lightning strike intensity, internet signal strength, the number of SMEs, and NTL. Because internet signal strength data is categorical, descriptive analysis uses bar graphs. Villages with strong internet signals have higher NTL values. There are intriguing results. In 2021, the average NTL value was smaller than in 2018. This is likely due to the uneven distribution of strong internet signals in 2018. Furthermore, the majority of villages in urban areas have strong internet signals. These regions tend to have higher levels of economic activity than rural areas, as indicated by high NTL values.

Both in 2018 and 2021, villages with strong internet signals tend to have a greater number of SMEs compared to villages with weak or no internet signals. In 2018, villages with weak or no internet signals had an average number of 18 SMEs, and in 2021, the value dropped to 16 units. Then, in 2018 and 2021, villages with strong internet signals had the same average number of SMEs, namely 32. Based on these results, there is an indication of a positive relationship between internet signal strength and the number of MSMEs. A stronger internet signal will correspond to an increase in the number of SMEs.

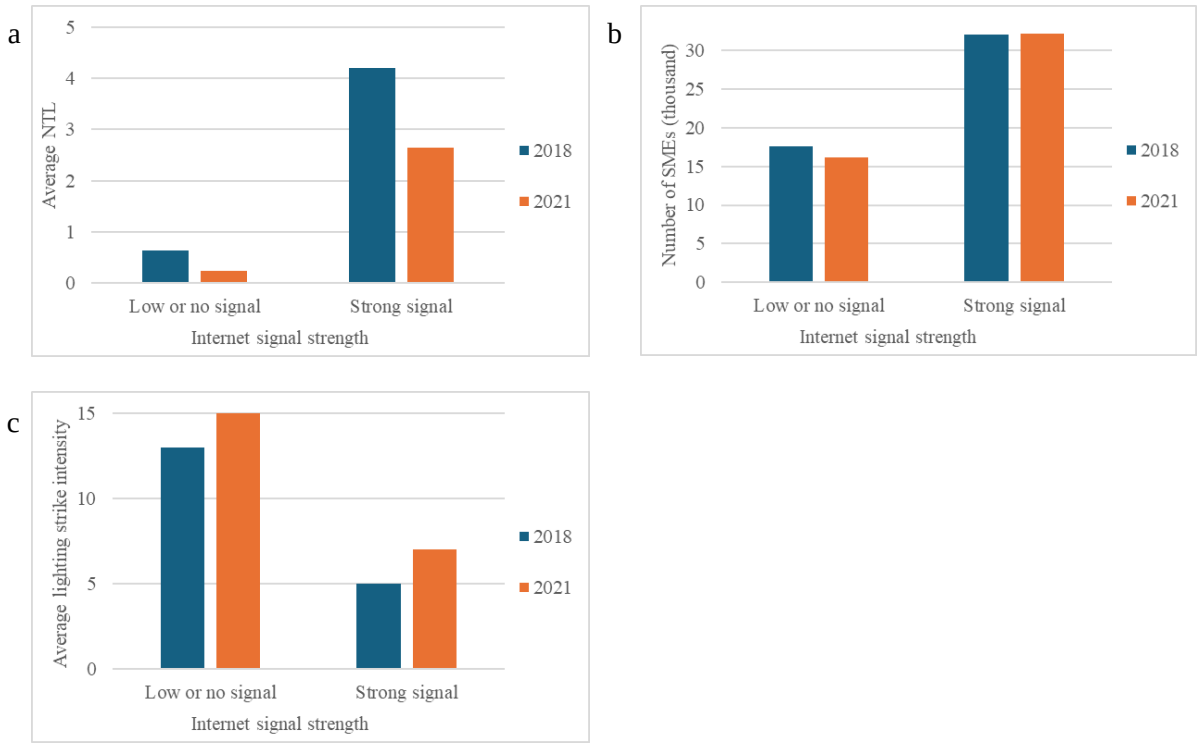


Figure 9. (a) Average NTL by Internet Signal Strength Category and Year; (b) the Number of SMEs per Village by Internet Signal Strength Category and Year; (c) Lightning Strike Intensity by Internet Signal Strength Category and Year.

The next analysis looks at the relationship between internet signal strength and the instrument variable, namely the number of lightning flashes. Based on the visualization results in Figure 9c, villages with strong signal strength tend to have a lower average number of lightning flashes compared to villages with weak or no internet signals. Both in 2018 and 2021, the relationship between internet signal strength and the number of lightning flashes has the same pattern. In 2018, villages with strong signals had an average number of lightning flashes that was 59.34% lower than villages with weak or no internet signals. Meanwhile, in 2021, the difference in the number of lightning flashes was 52.47%. It indicates a negative relationship between the number of lightning flashes and internet signal strength.

3.4. Endogeneity Test

Before modeling, we first used the Wu-Hausman test to conduct an endogeneity test of the internet signal strength variable. The null hypothesis of the test states that the internet signal strength variable is not endogenous. After the calculation, the test statistic value was 733.799, with a p-value of less than 1%. As a result, we can conclude that the internet signal strength variable is endogenous. This is in line with Rezki's research in 2023 [28]Therefore, to overcome this endogeneity, we need an analysis method involving instrument variables, such as causal mediation analysis in instrumental-variable regressions.

3.5. The Impact of Internet Signal Strength on the Regional Economy

We use the IV-2SLS estimation method to demonstrate how internet signal strength impacts NTL. The first stage involves regressing the number of lightning strike intensity, an instrument variable, and other exogenous variables against the internet signal strength, an endogenous variable. This study formed four models using different control variables for the robustness check. We use this method to verify the consistency of the estimated parameter values generated [13, 28, 36]. Table 3 presents the results of the first stage regression estimation.

The coefficient estimation results in Table 3 show that the number of lightning flashes has a significant negative effect on internet signal strength. The estimated value is also consistent across all models. In model 4, the lightning flash coefficient estimate is -0.00034. This means that a 1-point increase in lightning flash intensity will reduce the village's chance of having a strong internet signal by

0.034 percentage points. This result aligns with Rezki's study, which shows that higher lightning strike intensity can reduce the strength of mobile phone signals [28]. To see how strong the instrument variable can explain the strength of the internet signal, we conducted an identification test by looking at the value of the KPW F test statistic. For testing, the critical point limit is 16.38 [34]. Model 4's KPW F test statistic yields a value higher than 16.38, indicating the strength and relevance of the instrument variable in predicting internet signal strength.

Table 3. The Results of the First Stage Regression Estimation using IV-2SLS Method

Dependent Variable: Internet Signal Strength	Model 1	Model 2	Model 3	Model 4
Lightning strike intensity	-0.00043*** (0.00002)	-0.00032*** (0.00002)	-0.00037*** (0.00002)	-0.00034*** (0.00002)
Demographic and government characteristics	No	Yes	Yes	Yes
Demographic characteristics in the previous year	No	No	Yes	Yes
Geographic characteristics	No	No	No	Yes
KPW F-statistic	315.329	216.837	290.187	263.269
Observations	160,680	160,680	160,680	160,680

Note: *** denotes 1% significance level. Clustered standard errors at village level are in parentheses.

Table 4 presents the estimation results for the impact of internet signal strength on NTL as a proxy for economic growth. We obtained these results from stage 2 regression using the IV-2SLS method. Table 4 also presents four models with different control variables for robustness checks. All models show the same results, namely that internet signal strength has a positive and significant effect on NTL, with a value of 3.863-4.959. For example, in Model 4, the estimated coefficient of the internet signal strength variable is 4.224. This means that the NTL value in villages with strong internet signals is 422.4 percent higher than in villages with weak or no internet signals. Therefore, a village's economy becomes more developed as its internet signal strength increases.

Table 4. The Results of the Second Stage Regression Estimation using IV-2SLS Method

Dependent Variable: ln NTL	Model 1	Model 2	Model 3	Model 4
Internet signal strength	4.441*** (0.3680)	4.959*** (0.4753)	3.863*** (0.3657)	4.224*** (0.4065)
Demographic and government characteristics	No	Yes	Yes	Yes
Demographic characteristics in the previous year	No	No	Yes	Yes
Geographic characteristics	No	No	No	Yes
Observations	160,680	160,680	160,680	160,680

Note: *** denotes 1% significance level. Clustered standard errors at village level are in parentheses.

3.6. The Number of SMEs as a Mediator

This section presents the results of causal mediation analysis in instrumental-variable regressions. This study adopted the method from Dippel et al. in 2020 [33]. The number of SMEs is the mediating variable, while the number of lightning flashes is the IV. This study divides the total impact of internet signal strength on NTL (see Table 4) into two categories: a direct effect from other unspecified pathways and an indirect effect from the number of SMEs. The results of this decomposition are presented in Table 5.

Table 5, column 3, shows the total effect of internet signal strength on NTL. It has the same value as the estimated results of the IV-2SLS method in Table 4, column 5 (model 4). Table 5, column 2, displays the effect of internet signal strength on the number of SMEs. The results show that internet signal strength has a positive and significant effect on the number of SMEs, namely 3.857. This means villages with strong internet strength have 385.7% more SMEs than villages with weak or no internet signals. These results are in accordance with the theory of industrial location. According to this theory,

technological advances are one of the factors that influence the location of an industry's economic activities. Mack and Grubesic in 2009 [37] and Duvivier in 2021 [38] both demonstrated similar results.

Table 5, column 4, second row, shows the effect of the number of SMEs on NTL. We conclude that the number of SMEs acting as mediators between internet signal strength and NTL has a positive and significant effect on NTL, with a significance level of 1% and a coefficient value of 1.168. This means that increasing the number of SMEs in a village by 1% will increase the village economy by 1.168%, assuming other variables are constant.

Table 5. The Results of the Causal Mediation Analysis in Instrumental-variable Regressions

	Ln SME	Ln NTL	Ln NTL
Internet signal strength	3.857*** (0.1944)	4.224*** (0.4065)	-0.280*** (0.0313)
Ln SME			1.168*** (0.1004)
KPW F-Statistics in the first stage regression estimation	263.269	263.269	263.269
KPW F-Statistics in the first second regression estimation			452.403

Note: *** denotes 1% significance level. Clustered standard errors at village level are in parentheses.

The coefficient estimate in the first row of column 4 in Table 5 is equivalent to the direct influence value in the second row of column 5 in Table 6. The calculated value is -0.280. Villages with strong internet signals have a 28% lower NTL than villages with weak or no internet signals. Therefore, the internet signal's strength can potentially reduce the regional economy directly. This discovery is consistent with the research findings of Atkinson and McKay in 2007 [39] and Bakari and Tiba in 2019 [40]. Atkinson and McKay attributed this phenomenon to the disparity in technology. The unequal distribution of technology and information adoption can worsen income inequality, which has the potential to cause a deterioration in the economy [39]. Meanwhile, Bakari and Tiba argued that the detrimental impact of the internet on the economy arises from its utilization in non-productive pursuits, such as engaging in social media and playing online games [40].

Table 6. The Total, Direct, and Indirect Effects of Internet Signal Strength on the Regional Economy using Causal Mediation Analysis in Instrumental-variable Regressions

Dependent Variable: Ln NTL	Model 1	Model 2	Model 3	Model 4
Total effect	5.087*** (0.4229)	5.795*** (0.5653)	3.863*** (0.3657)	4.224*** (0.4065)
Direct effect	-1.055*** (0.1147)	-0.882*** (0.0845)	-0.255*** (0.0302)	-0.280*** (0.0313)
Indirect effect	6.142*** (0.5938)	6.677*** (0.7451)	4.118*** (0.4008)	4.504*** (0.4487)
Demographic and government characteristics	No	Yes	Yes	Yes
Demographic characteristics in the previous year	No	No	Yes	Yes
Geographic characteristics	No	No	No	Yes

Note: *** denotes 1% significance level. Clustered standard errors at village level are in parentheses.

Additionally, Table 6 displays the indirect impact. The findings indicate that the number of MSEs has a positive and statistically significant effect on NTL, indirectly influencing internet signal strength. To determine this impact, for instance, we obtain the result by multiplying the coefficient representing the influence of internet signal intensity on the number of SMEs in column 2 of Table 5, with the coefficient representing the effect of the number of SMEs on NTL in column 4 of the same table. The calculated coefficient value is 4.504, indicating that villages with strong internet signals have an NTL value of 450.4 percent greater than villages with weak or no internet signals.

3.7. The Impact of Internet Signal Strength on the Number of SMEs

This section estimates the impact of internet signal strength on the number of SMEs using causal mediation analysis in instrumental-variable regressions. The first-stage modeling yields the same results as those listed in Table 3. At a significance level of 1%, the number of lightning strikes has a positive and significant effect on internet signal strength. The identification test results with the F KPW test statistic are also greater than 16.38, indicating that the instrumental variable is considered strong and relevant.

Table 7. The Impact of Internet Signal Strength on the Number of SMEs using Causal Mediation Analysis in IV Regressions

Dependent Variable: Ln SME	Model 1	Model 2	Model 3	Model 4
Internet signal strength	2.756*** (0.1244)	2.773*** (0.1671)	3.972*** (0.1823)	3.857*** (0.1944)
Demographic and government characteristics	No	Yes	Yes	Yes
Demographic characteristics in the previous year	No	No	Yes	Yes
Geographic characteristics	No	No	No	Yes
KPW F-statistic	315.329	216.837	290.187	263.269
Observations	160,680	160.680	160.680	160.680

Note: *** denotes 1% significance level. Clustered standard errors at the village level are in parentheses.

Table 7 presents the effect of internet signal strength on the number of SMEs. The estimation result in column 5 of Table 7 is the same as the estimation result in column 2 of Table 5, both of which are 3.857. Models 1-4 provide the same results, with internet signal strength having a positive and statistically significant effect on the number of SMEs at a significant level of 1%. These results suggest that entrepreneurs will tend to determine the location of their enterprises in areas with strong internet signal strength. This is consistent with the theory of industrial location, in which ICT is one of the determining factors in enterprise location decisions, as previously explained.

4. Conclusion

Based on the research results, villages with strong internet signals, as an indicator of ICT improvement, tend to have higher NTL values and the number of SMEs compared to villages with weak internet signals. ICT's indirect role through SMEs has the potential to enhance the regional economy, whereas its direct impact actually diminishes it. The greater indirect influence leads to a positive total effect of ICT. The disparity in ICT adoption and utilization, which often results in non-productive activities, has the potential to directly reduce the regional economy [39-40]. However, by increasing the number of SMEs, we can mitigate this negative impact, which also contributes to improving the regional economy. The increase in the number of SMEs occurs because the development of ICT in a region can encourage new industries in the region.

Improving ICT infrastructure is crucial, particularly in underdeveloped regions, as it can potentially increase the regional economy. Building more high-speed internet networks, expanding cellular telecommunications coverage, and improving accessibility can accomplish this. Training and mentoring activities for SMEs should accompany the increase in SMEs, enabling them to innovate effectively and market products using ICT, such as e-commerce. Thus, increasing the number of SMEs can boost the regional economy.

Unfortunately, this study has limited information about SME data. We hope that additional studies will shed light on how ICT can enhance MSE performance and how this enhancement could potentially contribute to regional economic development. Additionally, ICT can affect the labor market by enabling the entrance of new firms. Indonesia has not yet conducted studies on this subject.

Ethics approval

Not required.

Acknowledgments

We express their gratitude for the assistance and participation of all those participating in this work. We express our gratitude to BPS-Statistics Indonesia for supplying the vital data that enabled our investigation. We also express our gratitude to the reviewers and proofreaders for their diligent endeavors and important input, which significantly improved the quality of this work.

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

Luthfio Febri Trihandika: Conceptualization, Data Collection, Writing – Original Draft, Visualization. **Ribut Nurul Tri Wahyuni**: Methodology, Formal Analysis, Writing – Review and Editing, Supervision. **Meilinda Fitriani Nur Maghfiroh:** Writing – Review and Editing, Supervision.

References

[1] E. A. Hanushek, "Economic Growth in Developing Countries: The Role of Human Capital," *Econ Educ Rev*, vol. 37, pp. 204–212, Dec. 2013, doi: 10.1016/j.econedurev.2013.04.005.

[2] A. N. Rahmah and S. Widodo, "The Role of the Manufacturing Industry Sector in the Indonesian Economy with an Input-Output Approach in 2010-2016," *Economie: Jurnal Ilmu Ekonomi*, vol. 1, no. 1, p. 14, Jun. 2019, doi: 10.30742/economie.v1i1.819.

[3] BPS-Statistics Indonesia, "Gross Regional Domestic Product of Provinces in Indonesia by Industry 2017-2021," Jakarta, 2022.

[4] A. M. Hilman and A. M. Ester, "The Role of the Manufacturing Sector in the Indonesian Economy: Input-Output Model," *Media Ekonomi*, vol. 26, no. 1, pp. 63–76, Aug. 2019, doi: 10.25105/me.v26i1.5210.

[5] BPS-Statistics Indonesia, "Indonesia Labor Market Indicators February 2023," Jakarta, 2023.

[6] BPS-Statistics Indonesia, "Micro and Small Industry Profile 2021," Jakarta, 2023.

[7] R. Richardson and A. Gillespie, "Advanced Communications and Employment Creation in Rural and Peripheral Regions: A Case Study of the Highlands and Islands of Scotland," *Ann Reg Sci*, vol. 30, no. 1, pp. 91–110, Mar. 1996, doi: 10.1007/BF01580539.

[8] E. A. Mack, L. Anselin, and T. H. Grubesic, "The Importance of Broadband Provision to Knowledge Intensive Firm Location," *Regional Science Policy & Practice*, vol. 3, no. 1, pp. 17–35, Mar. 2011, doi: 10.1111/j.1757-7802.2011.01026.x.

[9] World Bank, "World Development Report 2016: Digital Devidens," Washington, DC, 2016.

[10] E. Brynjolfsson and A. McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, 1st ed. New York: W. W. Norton & Company, 2014.
[11] M. Arntz, T. Gregory, and U. Zierahn, "The Risk of Automation for Jobs in OECD Countries: A Comparative Analysis," Paris, 189, 2016. doi: 10.1787/5jlz9h56dvq7-en.
[12] J. Brodny and M. Tutak, "Analyzing the Level of Digitalization among the Enterprises of the European Union Member States and Their Impact on Economic Growth," *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8, no. 2, p. 70, Jun. 2022, doi: 10.3390/joitmc8020070.
[13] A. T. Falentina, B. P. Resosudarmo, D. Darmawan, and E. Sulistyaningrum, "Digitalisation and the Performance of Micro and Small Enterprises in Yogyakarta, Indonesia," *Bull Indones Econ Stud*, vol. 57, no. 3, pp. 343–369, 2021, doi: 10.1080/00074918.2020.1803210.
[14] M. Ivanović-Đukić, T. Stevanović, and T. Rađenović, "Does Digitalization Affect the Contribution of Entrepreneurship to Economic Growth?," *Zbornik radova Ekonomskog fakulteta u Rijeci: časopis za ekonomsku teoriju i praksu/Proceedings of Rijeka Faculty of Economics: Journal of Economics and Business*, vol. 37, no. 2, Dec. 2019, doi: 10.18045/zbefri.2019.2.653.
[15] P. Dicken, *Global Shift: Mapping the Changing Contours of the World Economy*, 7th ed. London: Sage Publication, 2015.
[16] A. Awad and M. Albaity, "ICT and Economic Growth in Sub-Saharan Africa: Transmission Channels and Effects," *Telecomm Policy*, vol. 46, no. 8, p. 102381, Sep. 2022, doi: 10.1016/j.telpol.2022.102381.
[17] A. Skorupinska and J. Torrent-Sellens, "ICT, Innovation and Productivity: Evidence Based on Eastern European Manufacturing Companies," *Journal of the Knowledge Economy*, vol. 8, no. 2, pp. 768–788, Jun. 2017, doi: 10.1007/s13132-016-0441-1.
[18] D. Stamopoulos, P. Dimas, and A. Tsakanikas, "Exploring the Structural Effects of the ICT Sector in the Greek Economy: A Quantitative Approach Based on Input-Output and Network Analysis," *Telecomm Policy*, vol. 46, no. 7, p. 102332, Aug. 2022, doi: 10.1016/j.telpol.2022.102332.
[19] K. Vu, P. Hanafizadeh, and E. Bohlin, "ICT as A Driver of Economic Growth: A Survey of the Literature and Directions for Future Research," *Telecomm Policy*, vol. 44, no. 2, p. 101922, Mar. 2020, doi: 10.1016/j.telpol.2020.101922.
[20] A. Prakash, "The Role of Industrialisation and ICT in Africa's Growth and Integration into Global Value Chains," *Working Papers PB-2019-01*, Economic Research Institute for ASEAN and East Asia (ERIA).
[21] M. Manacorda and A. Tesei, " Liberation Technology: Mobile Phones and Political Mobilization in Africa," *Econometrica*, vol. 88, no. 2, pp. 533–567, doi: 10.3982/ECTA14392.
[22] J. C. Aker and I. M. Mbiti, "Mobile Phones and Economic Development in Africa," *Journal of Economic Perspectives*, vol. 24, no. 3, pp. 207–232, doi: 10.1257/jep.24.3.207.
[23] J. V. Henderson, A. Storeygard, and D. N. Weil, "Measuring Economic Growth from Outer Space," *American Economic Review*, vol. 102, no. 2, pp. 994–1028, Apr. 2012, doi: 10.1257/aer.102.2.994.
[24] A. M. Rugman and A. Verbeke, "A Perspective on Regional and Global Strategies of Multinational Enterprises," *Journal of International Business Studies*, vol. 35, no. 1, pp. 3–18, doi: 10.1057/palgrave.jibs.8400073.
[25] R. Hayter, *The Dynamics of Industrial Location: The Factory, the Firm and the Production System*. Burnaby: Wiley, 2004.
[26] P. Dicken, *Global Shift: Mapping the Changing Contours of the World Economy*, 7th ed. Manchester: Sage Publication, 2015.
[27] G. N. Mankiw, *Macroeconomics*, 10th ed. New York: Worth Publishers, 2019.
[28] J. F. Rezki, "Does the Mobile Phone Affect Social Development? Evidence from Indonesian Villages," *Telecomm Policy*, vol. 47, no. 3, Apr. 2023, doi: 10.1016/j.telpol.2023.102503.
[29] C. D. Elvidge, M. Zhizhin, T. Ghosh, F.-C. Hsu, and J. Taneja, "Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019," *Remote Sens (Basel)*, vol. 13, no. 5, pp. 922, Mar. 2021, doi: 10.3390/rs13050922.
[30] D. J. Cecil, "LIS/OTD Gridded Lightning Climatology Data Collection," 2015, *NASA EOSDIS Global Hydrology Resource Center Distributed Active Archive Center, Huntsville, Alabama, U.S.A.* doi: http://dx.doi.org/10.5067/LIS/LIS-OTD/DATA311.

[31] C. T. Lloyd *et al.*, "Global Spatio-temporally Harmonised Datasets for Producing High-resolution Gridded Population Distribution Datasets," *Big Earth Data*, vol. 3, no. 2, pp. 108–139, Apr. 2019, doi: 10.1080/20964471.2019.1625151.

[32] S. E. Fick and R. J. Hijmans, "WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas," *International Journal of Climatology*, vol. 37, no. 12, pp. 4302–4315, Oct. 2017, doi: 10.1002/joc.5086.

[33] C. Dippel, A. Ferrara, and S. Heblich, "Causal Mediation Analysis in Instrumental-variables Regressions," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 20, no. 3, pp. 613–626, Sep. 2020, doi: 10.1177/1536867X20953572.

[34] J. Stock and M. Yogo, *Testing for Weak Instruments in Linear IV Regression*. In: Andrews DWK Identification and Inference for Econometric Models. New York: Cambridge University Press, 2005, doi: 10.3386/t0284.

[35] F. Kleibergen and R. Paap, "Generalized Reduced Rank Tests using the Singular Value Decomposition," *Journal of Econometrics*, vol. 133, no. 1, pp. 97-126, doi: 10.1016/ j.jeconom.2005.02.011.

[36] M. H. Yudhistira, W. Indriyani, A. P. Pratama, Y. Sofiyandi, and Y. R. Kurniawan, "Transportation Network and Changes in Urban Structure: Evidence from the Jakarta Metropolitan Area," *Research in Transportation Economics*, vol. 74, pp. 52–63, May 2019, doi: 10.1016/j.retrec.2018.12.003.

[37] E. A. Mack and T. H. Grubesic, "Broadband Provision and Firm Location in Ohio: An Exploratory Spatial Analysis," *Tijdschrift voor Economische en Sociale Geografie*, vol. 100, no. 3, pp. 298–315, Jul. 2009, doi: 10.1111/j.1467-9663.2008.00487.x.

[38] C. Duvivier, "Broadband and Firm Location: Some Answers to Relevant Policy and Research Issues using Meta-analysis," *Canadian Journal of Regional Science*, vol. 42, no. 1, pp. 24–45, Nov. 2021, doi: 10.7202/1083638ar.

[39] R. D. Atkinson and A. S. McKay, "Digital Prosperity: Understanding the Economic Benefits of the Information Technology Revolution," *SSRN Electronic Journal*, 2007, doi: 10.2139/ssrn.1004516.

[40] S. Bakari and S. Tiba, "The Impact of Internet on Economic Growth: Evidence from North Africa," *Munich Personal RePEc Archive*, 2019.

[41] Dimitrie-Daniel Plăcintă, A. Toma, L. Bătăgan, Corina-Marina Mirea, Florin-Valeriu Pantelimon, "The Impact of ICT Education on GDP Evolution at Local Level. A Romanian Case Study," *Business, Management and Economics Engineering*, vol. 22, no. 1, pp. 156-173, May 2024, doi:10.3846/bmee.2024.19762.

[42] S. Min, M. Liu, J. Huang, "Does the Application of ICTs Facilitate Rural Economic Transformation in China? Empirical Evidence from the Use of Smartphones among Farmers," *Journal of Asian Economics*, vol. 70, 2020, doi:10.1016/j.asieco.2020.101219.

[43] G. A. Untura, "The Knowledge Economy and Digitalization: Assessing the Impact on Economic Growth of Russian Regions," *Regional Research of Russia*, vol. 13, no. 3, pp. 397-406, September 2023, doi: 10.1134/S2079970523700909.

[44] P. W. Handayani, R. Nasrudin, J. F. Rezki, "Reliable Electricity Access, Micro-Small Enterprise and Poverty Reduction in Indonesia," *Bulletin of Indonesian Economic Studies*, vol. 60, no. 1, pp. 35-66, February 2023, doi: 10.1080/00074918.2023.2175782.

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.796

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Implementation of a RESTful API-Based Evolutionary Algorithm in a Microservices Architecture for Course Timetabling

Zuhdi Ali Hisyam[1*], Farid Ridho[2], Arbi Setiyawan[3]

[1,2]Politeknik Statistika STIS, Jakarta, Indonesia, [3]School of Management, Jiangsu University, China
**Corresponding Author: E-mail address:* 222011642@stis.ac.id

ARTICLE INFO

Article history:
Received 3 September, 2024
Revised 29 October, 2024
Accepted 9 November, 2024
Published 31 December, 2024

Keywords:
Course Timetabling; Evolutionary Algorithm; (1+1) Evolutionary Strategies; RESTful API; Microservices; Cost Function; Black Box Testing

Abstract

Introduction/Main Objectives: Implement an evolutionary algorithm within a RESTful API for a course timetabling system that employs a microservices architecture. **Background Problems:** The current course timetabling at Politeknik Statistika STIS uses the third-party application (aSc Timetables), which lacks a generator as a service, resulting in its inefficiency due to the lack of integration with SIPADU NG. **Novelty:** The evolutionary algorithm is built as a service (RESTful API) within a microservices architecture and supports custom constraints for timetables. **Research Methods:** One of the evolutionary algorithm families, the (1+1) evolutionary strategy, is implemented and used to create a course timetable 1000 times. Each course timetable created will have its cost calculated to assess the goodness of the algorithm implementation. The developed RESTful API is also evaluated through black box testing. **Finding/Results:** For the odd semester data, 40.5% of the trials yielded a cost value between 4 and 5, while for the even semester, all trials produced a cost value below 1. The resulting cost value is close to 0, which indicates that the timetable created has minimal violations. Additionally, black box testing concluded that the service operates as expected, delivering the anticipated output.

1. Introduction

Timetabling or scheduling is one of many issues institutions deal with, and the academic field is no exception [1]. This issue is usually called the University Course Timetabling Problem (UCTTP). In a big institution like a university, timetabling is a worthwhile task that needs to be resolved efficiently. In the context of lectures (courses), timetabling is a way to combine lecturers, groups of students, subjects, time of the lectures, and classroom availability with some sort of rules, which results in a flexible and conflictless course timetable. Flexible means the timetable can accommodate its resources, for example, a classroom preference from some lecturer, a time preference of the lecturers, or a forbidden time for lectures.

Politeknik Statistika STIS is a college in the Badan Pusat Statistik (BPS) environment. While Politeknik Statistika STIS is functionally built by the Head of BPS, it is still technically built by the Indonesian Ministry of Research, Technology and Higher Education [2]. Politeknik Statistika STIS, as a college institution, faces the same problem as many other colleges/universities, which is related to course scheduling or timetabling.

Courses timetabling problems can be resolved manually or automatically by using a system. However, making a manual timetable can take much time and effort, making it difficult, especially for a larger institution with many entities like Politeknik Statistika STIS. That is why the utilization of an algorithm-based scheduling system can be done to solve timetabling problems.

At this moment, Politeknik Statistika STIS is utilizing a third-party application (aSc TimeTables) to build its timetables. On aSc TimeTables official site, this application does not provide the generator as a service and resulted in a problem in which the timetables result of this application cannot be integrated with SIPADU NG (*Sistem Informasi Perkuliahan Terpadu Next Generation*). Building timetables using a third-party application can impact inefficiency because of the need to manually input timetable results into SIPADU NG. Also, users need to define constraints themselves, whether lecturer or general constraints. Lecturer constraint contains the availability of teaching time and lecture room (classroom) preferences desired by the lecturer. These constraints must be defined one by one, which, of course, takes a long time. General constraint is a limitation that is inherent in many entities, be it a few lecturers, some classes of students, or all entities. These, too, must be defined one by one. After completing the schedule, the timetable must still be inputted into SIPADU NG again. This procedure raises the opportunity to create a timetabling system that is integrated with SIPADU NG to be more efficient.

Creating a timetabling system requires allocating each lecture containing lecturers, students, and lecture rooms to a specific time so there is no conflict between the three. In addition, timetables are also required to follow the rules (constraints) by the needs of the Politeknik Statistika STIS. As a result, a timetabling algorithm must be implemented to address this need since manually creating timetables is both time-consuming and inefficient. Studies [3], [4], [5] demonstrate that the timetabling problem can be solved using algorithms or formulas.

The course timetabling system that will be created is divided into three modules: the front-end web development module, the constraints or back-end rule provider module, and the timetabling module. This research will focus solely on the timetabling module, namely how to manage the available resources to produce the best timetable possible. These resources include the availability of time and preferences for lecturers' rooms, lecture rooms, lecture meetings, and other constraints.

Course timetabling problems are classified as non-polynomial time (NP) and combinatorial optimization problems (COP), indicating that they can be addressed using optimization algorithms to generate the desired optimal timetable [1]. Many algorithms can be used to address course timetabling problems, one of which is the evolutionary algorithm (EA). Evolutionary Algorithms are a subset of Evolutionary Computations (EC) and fall within modern heuristic-based search methods. Thanks to their flexibility and robust characteristics derived from Evolutionary Computation, they are a practical problem-solving approach for a wide range of global optimization challenges. EAs have been successfully applied in a wide range of problems [6], [7], [8].

Additionally, studies [9], [10], [11] implement evolutionary algorithms to address course timetabling, demonstrating that timetabling problems can be solved quickly, efficiently, and effectively. However, as done in this study, only some studies implement the algorithm into a service. Moreover, the algorithm in this study accommodates custom constraints, making it more flexible when creating course timetables.

The algorithm applied to the timetabling system of the Politeknik Statistika STIS in this study is one type of evolutionary algorithm, namely (1+1) evolutionary strategies (ES). Because this algorithm works randomly, several solutions will be made by multiprocessing in creating schedules. One of the best solutions will be selected from several solutions due to this service.

In developing a timetabling system, the concept of microservices is used. Therefore, the timetable module will later be added to the service. This RESTful API-based service will generate timetables based on existing data. Other modules will later use the results, especially the front-end module, to display to users. That way, users can see the lecture schedules that have been created and allow users to change some lecture schedules if needed.

2. Material and Method

2.1. Data

Several methods were used to collect data for this study.

1. Interview

The interview method involves direct communication with relevant parties to gather data. Interviews were conducted with the IT Unit of the Politeknik Statistika STIS, the department responsible for creating the semester timetable. These interviews aimed to discuss the current course timetabling business process at Politeknik Statistika STIS, identify any existing problems or challenges, and understand the specific needs of the subject matter experts.

Additionally, interviews were conducted with the *Bagian Administrasi, Akademik, dan Kemahasiswaan*, or Academic Affairs Division, who created the semester timetable. These interviews aimed to identify the usual constraints imposed each semester, providing a guideline for developing a system to address these limitations. Furthermore, these interviews were conducted to request course data that would be used for testing purposes.

2. Observation

   Observation is a data collection method that involves directly observing the research object. This method involves observing the requirements for creating a timetable, such as data on all student classes, lecturers, classrooms, available time slots, and, most importantly, observing the constraints of the course timetabling at the Politeknik Statistika STIS. This activity is necessary for the researcher to translate these findings into a system.

3. Literature Study

   The objective of this method is to conduct an in-depth study of theories related to course timetabling and system development by gathering literature from relevant research papers. The references collected are related to timetabling algorithm theories, system implementation theories, and so on. The sources obtained are varied, ranging from scientific journals, books, the Internet, and other valid sources.

2.2. Timetabling

The word "scheduling" is regarded as a broad term that includes a range of issues, such as timetabling, sequencing, and rostering. Research [1] defines timetabling as the process of allocating resources to activities over time and space, subject to a set of constraints. This problem has gained prominence in various domains, including education. Research [12] further classifies timetabling problems into three categories: school, exam, and course timetabling. This study focuses on the course timetabling.

Course timetabling entails assigning courses, lecturers, and students to specific time slots and rooms, subject to a multitude of constraints. As identified by research [3], [12], these constraints can be categorized into hard and soft constraints. Hard constraints are mandatory requirements that must be met, while soft constraints are desirable but not strictly necessary, aiming to optimize the overall timetable. The specific constraints for course timetabling at the Politeknik Statistika STIS may vary from semester to semester.

At the Politeknik Statistika STIS, the academic timetable is divided into time blocks. A standard day consists of 9-time blocks, and classes are held 5 days a week, totaling 45 blocks per week. The duration of each course is determined by the number of credits associated with the course. For instance, a 3-credit course will occupy three consecutive time blocks, while a 2-credit course will occupy 2. Using the same concept as [13], we can define a set of working days from Sunday to Friday as $D = \{1, 2, 3, 4, 5\}$, where Sunday is one and Friday is five. Each day $d \in D$ consist of P block time which $P = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Course assignment is represented as x $x_{d,p,l,c,r,s}$, where it equal to one if a subject $c \in C$, taken by student group $s \in S$, is scheduled to be taught by a lecturer $l \in L$ on a day $d \in D$, at a block time $p \in P$ in a room $r \in R$.

The main purpose of timetabling is to create a timetable so that there is no conflict between lecturers, student groups, and classes. At most, one subject must be assigned to only one lecturer in one room at each time block. Hence, a lecturer could only attend one course at a time, as shown in (2). A subject is taught once per time block/room, as shown in (1), and could not be taught in two different rooms simultaneously, as shown in (3). More models can be seen in [13].

$$\sum_{l \in L} \sum_{c \in C} \sum_{r \in R} x_{d,p,l,c,r,s} \leq 1 \tag{1}$$

$$\sum_{c\in C}\sum_{r\in R}\sum_{s\in S} x_{d,p,l,c,r,s} \leq 1 \quad (2)$$

$$\sum_{l\in L}\sum_{c\in C}\sum_{s\in S} x_{d,p,l,c,r,s} \leq 1 \quad (3)$$

The timetabling system created in this study will implement an evolutionary algorithm. The Academic, Administration, and Student Affairs Division of Politeknik Statistika STIS will use this system to create a timetable every semester (six months).

2.3. Evolutionary algorithm: (1+1) Evolutionary strategies

Evolutionary strategies are a type of search algorithm classified as evolutionary algorithms. These algorithms, in general, continuously evolve solutions through mutation until an optimal solution is found or the maximum number of iterations is reached [14]. Research [6] mapped several search algorithms, as shown in Figure 1.

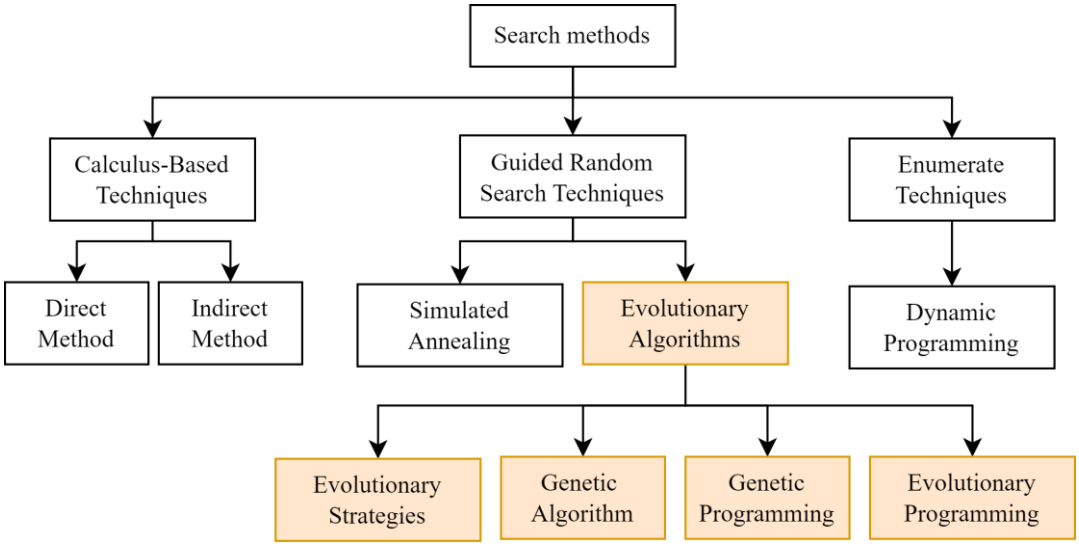


Figure. 1. Search algorithms and their family.

Generally, evolutionary algorithms operate on a population of individuals. These individuals undergo crossover, a process similar to genetic recombination, as one of the mechanisms of evolution. However, the algorithm used in this study differs. This research employs (1+1) evolutionary strategies, the simplest form of evolutionary strategies. As the name suggests, it operates on only two individuals: one parent and one child produced by mutating the parent. If the child is better than the parent, it replaces the parent in the next generation [14]. According to Fernandes [9], a cost function determines whether the child is better than the parents. The cost function will be further explained in the following equation.

The algorithm implementation is built on a microservices architecture, and a RESTful API is implemented. The service's functionality will be tested using black-box testing to verify if it can provide the expected responses. Additionally, the implementation of the evolutionary algorithm will be evaluated based on the resulting cost value using the Politeknik Statistika STIS's course data. The algorithm implementation is summarized in Figure 2.

2.4. RESTful API

An API (Application Programming Interface) is a mechanism for exchanging data between two software applications, adhering to specific rules and protocols. A RESTful API is an API that implements the REST (Representational State Transfer) architectural style. The REST architecture was first introduced by Roy Fielding in 2000. Derived from network-based architectures, REST incorporates

concepts such as client-server and layered systems [15]. This procedure allows each component of the architecture to be isolated while still enabling communication. Communication between the client and server in REST utilizes HTTP and employs request methods such as GET, PUT, POST, DELETE, PATCH, and others.

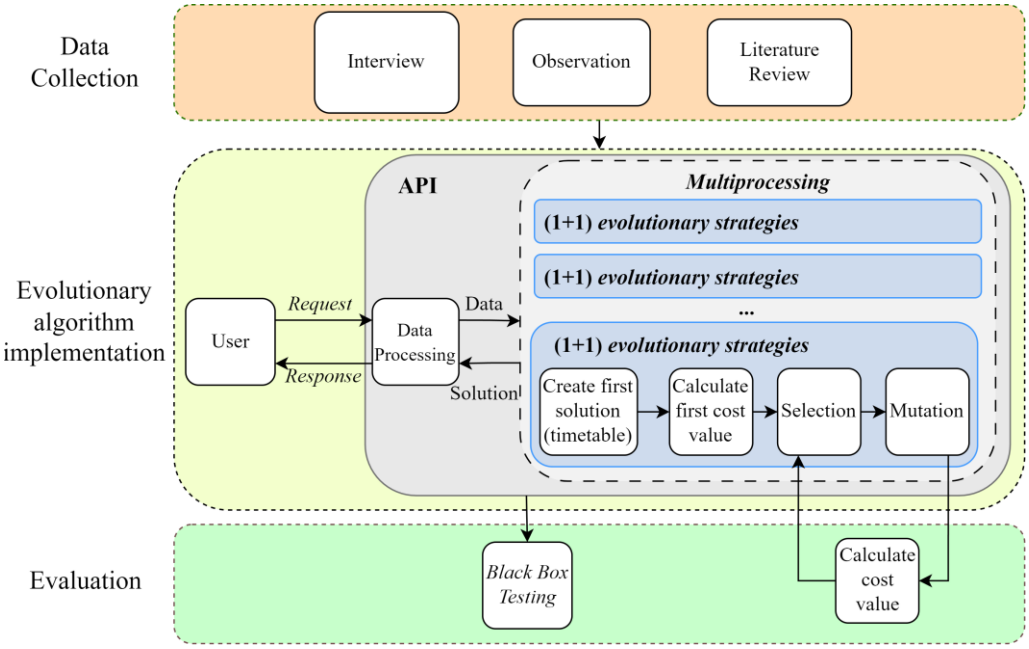


Figure. 2. Research method.

Research [16] indicates that a service created using RESTful API is easier to use without relying on the same operating system with another service, programming languages, or databases because the service communicates using a standard data format using JSON. Besides, REST architecture has several advantages compared to one of the well-known network communication architectures, namely SOAP [17]. This architecture is why evolutionary algorithms are implemented into a RESTful API in this study.

After the RESTful API is created, it is evaluated using testing. Testing is used to ensure the RESTful API's quality in general. The importance of testing has led developers to develop methodologies and approaches that support the testing process. The RESTful API is tested using black box testing, which ignores the internal structure and implementation of the component being tested, focusing solely on its inputs and outputs [18].

2.5. Evaluation Method

This research will be evaluated using two tests. First, we tested the RESTful API using black-box testing, checking if the system gives the correct output for every input. Second, we tested the (1+1) evolutionary algorithm by creating 1,000 different course timetables using data from the Politeknik Statistika STIS for the 2023/2024 academic year. We will then calculate a cost value for each timetable. A lower cost means a better timetable. The cost function used in this study will be explained later.

2.6. Microservices Architecture

Microservices is an architectural and organizational approach to software development in which software is composed of small, independent services that communicate through well-defined APIs [19]. Microservices architecture have several benefits, including increased modularity, flexible configuration, easier development, easier maintenance, and increased productivity [20]. The adoption of microservices simplifies system development as each service can be built using different tools (flexible), ensuring that errors in one service do not impact others. Microservices also enable faster system development. Multiple services can be developed simultaneously by different teams without having to wait for other services to be completed.

3. Result and Discussion

3.1. Data Preparation

Before implementing (1+1)-ES, several data are needed to create a course timetable. The data are:

1. Room data that can be used for courses. Room data is categorized based on building or floor.
2. Lecturer data contains information about the lecturer's name as well as their preferred teaching time and room.
3. Course meeting data combines lecturer, students, and subjects. It also contains information about the duration of the course (two blocks or three blocks) and the type of course (theory or practical).
4. The user defines constraints data or limitations that must be applied to the course timetable.
5. Configuration data contains information about the maximum number of iterations in the implementation of (1+1) evolutionary strategies, the maximum number of courses in a day for lecturers and students along with their priorities, the number of timetabling trials to be used in multiprocessing, the number of candidate solutions to be returned in the response, and the list of blocks that can be used for the start of courses, both with a duration of 2 blocks and three blocks.

Data preparation involves processing the data into a specific format to facilitate running the (1+1)-ES algorithm. The initial data preparation focused on time availability for lecturers, students, and classrooms, which were stored in an array. The possible value of the time availability array is: '0' meaning available, '99' meaning not available, and '999' meaning optional. This array has a length of 45, corresponding to the courses at the Politeknik Statistika STIS, which consist of 5 days with nine blocks each day. The representation of the array index with the day and course block is shown in Table 1.

Table 1. Representation of Day and Time block into index array

Day	Time blocks of course								
	1	2	3	4	5	6	7	8	9
Monday	0	1	2	3	4	5	6	7	8
Tuesday	9	10	11	12	13	14	15	16	17
Wednesday	18	19	20	21	22	23	24	25	26
Thursday	27	28	29	30	31	32	33	34	35
Friday	36	37	38	39	40	41	42	43	44

The next step in data preparation is to apply constraints related to prohibited blocks, mandatory blocks, and required rooms as defined by the user. Specifically, the required room constraint is limited to courses only. Block-related constraints are applied by modifying the values in the time availability array. Meanwhile, room-related constraints are applied by creating a list of courses and possible rooms.

The following data preparation step involves listing the possible time blocks for all lecturers, both for 2-credit and 3-credit courses. This step matches the lecturer's availability with the template for 2-credit and 3-credit courses. Figure 3 illustrates the process of obtaining possible time blocks with a duration of 3 credits for lecturer Ach001. The exact process is applied to 2-credit courses and all lecturers.

In Figure 3, the green color means the time block is free and can be filled with a course, while the red color means the time block is prohibited for courses, and the yellow color means that a course can start from this block because there will be no violation of the constraints if it starts from this block. Other data preparations include handling courses with theory and practical components, team teaching (courses taught by more than one lecturer), and creating course lists for lecturers and students.

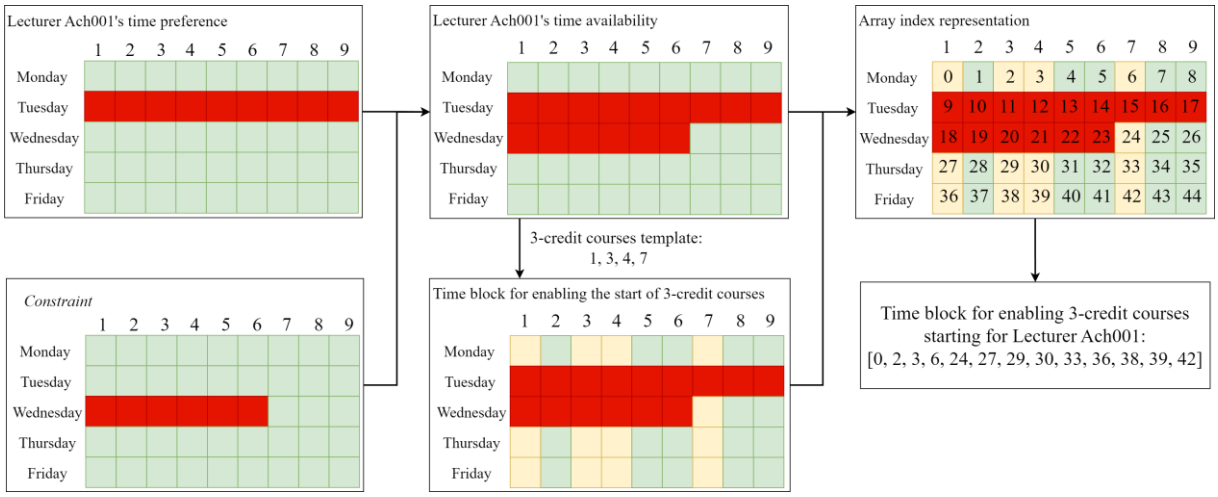


Figure. 3. An example calculation of representing possible course start time blocks into an array index.

3.2. Implementation of (1+1) Evolutionary Strategies

The implementation of (1+1)-ES works using a chromosome. The chromosome contains all the resources needed to create a timetable. In this study, the chromosome is defined as an array of length 7. Each array index stores information about:

1. All of the course meetings (Figure 5a) are an array.
2. The array contains all constraints, time preference, time availability (Figure 5b), time blocks for enabling 3-credit courses and 2-credit courses starting, and a list of courses (Figure 5b) belonging to each lecturer.
3. The array contains all-time availability (Figure 5d) belonging to each classroom.
4. The array contains all constraints, time availability (Figure 5c), and a list of courses (Figure 5c) belonging to each group class.
5. The array contains all plotting of subjects and time blocks.
6. Configuration of timetable, including maximal iteration ("iterasi"), number of multiprocessing to be performed ("jml_percobaan"), number of timetable candidate solution to be returned ("jml_kandidat_solusi"), number of maximal courses in one day ("maks_matkul"), weight penalty of maximal courses in one day ("maks_matkul_prioritas"), list of courses that needs special rooms ("ruang_wajib"), and time blocks for starting lectures for three credits ("blok_mulai_3_sks") and two credits ("blok_mulai_2_sks"). An example of configuration is below in Table 4.
7. An array of custom constraints is defined by the user (Explained later in Table 4).

The flowchart of (1+1)-ES implementation can be seen in Figure 4. The implementation process of (1+1)-ES involves the following steps:

1. Creation of initial chromosome as the first solution

   Initial chromosome generation is conducted by selecting a room and a starting block time for each course meeting. Room selection is naturally limited to rooms included in the lecturer's room preferences. This selection begins by randomly choosing a building. Afterward, a course room is randomly selected from the chosen building. The selected building and room are stored in the course meeting with the keys "lantai_terpilih" (selected building) and "ruang_terpilih" (selected room).

   Meanwhile, the selection of the starting block uses the index list that has been created, as shown in Figure 3. This selection of time and room aims to avoid selecting times or rooms that violate constraints or are outside the lecturer's room preferences. After a course meeting has been assigned a time and room, changes will occur in the time availability and course list array. An illustration of the changes that occur in each selection of time and room for each course meeting can be seen in Figure 5.

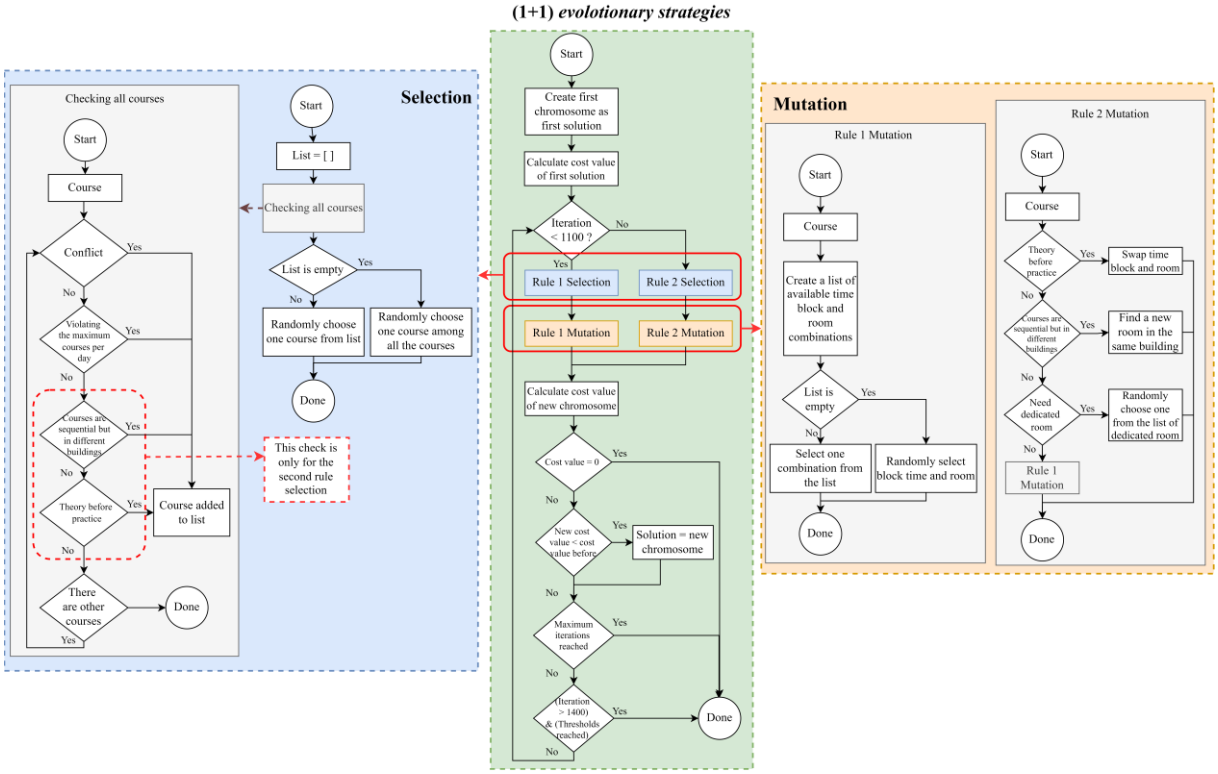


Figure. 4. Flowchart (1+1) Evolutionary strategies implementation.

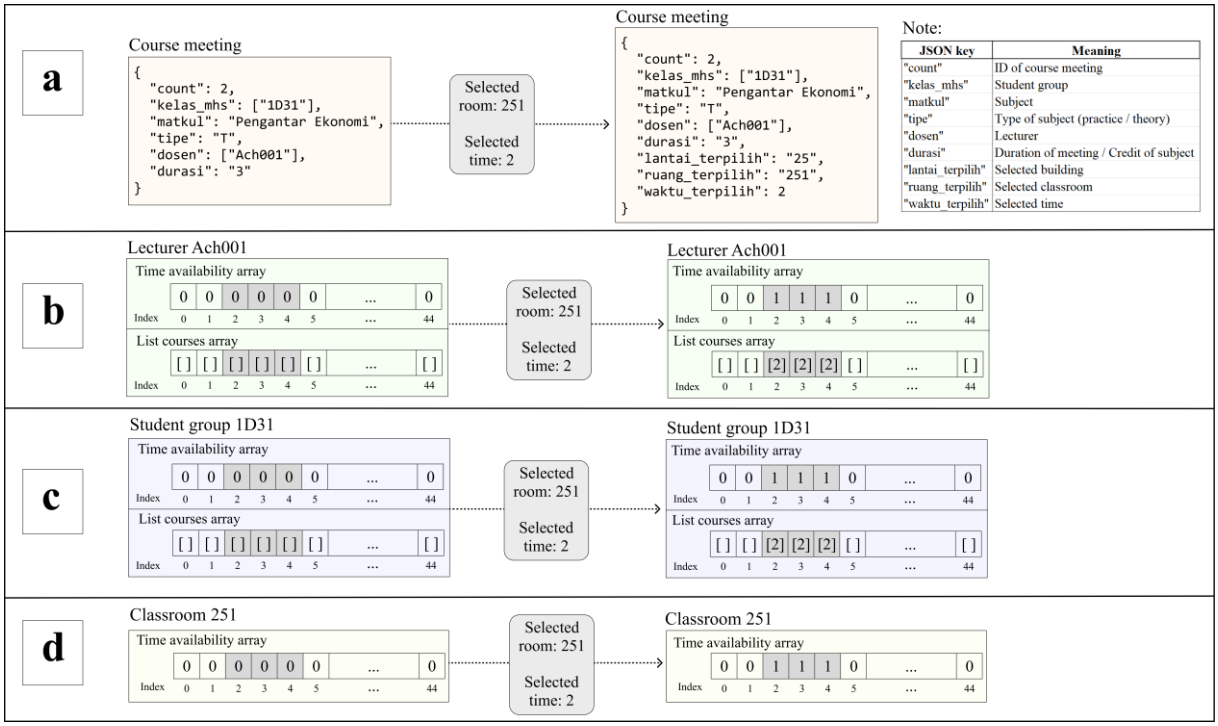


Figure. 5. Modifications to the scheduled time and room for course meetings that result in alterations to the array of: (a) course meetings, (b) lecturers, (c) student groups, and (d) classrooms. The alternations are visually represented by a grey background.

2. Cost assessment

The cost value is calculated using the following cost function:

$$Cost(chromosome) = \sum_{i=1}^{R} C_i w_i \quad (4)$$

The variable R signifies the total quantity of constraints. At the same time, C_i denotes the number of times the i-th constraint is violated, and w_i represents the penalty (weight) associated with each violation of the i-th constraint. Based on equation (4), a lower cost value implies a more optimal timetabling solution.

The preparation of previous data has made the cost calculation easier. The cost value will increase as violations of certain constraints occur. Due to the well-designed data preparation and the selection of time and room for each course meeting that avoids prohibited times, violations of user-defined hard constraints or violations of the lecturer's time and room preferences will never occur. The remaining possible violations are course conflicts, the maximum number of courses in one day, and soft constraint violations.

The time availability arrays of lecturers, students, and classrooms can be checked to identify conflicting timetables. If any value in the time availability array is within the range of 2-98, there is a course conflict. The course list array is examined to determine the number of courses in one day. If there are more course IDs in one day than the maximum number of courses allowed in one day, then there is a violation of the maximum number of courses in one day. Finally, the time availability arrays of lecturers, students, and classrooms are checked to identify soft constraint violations. If a value in the array is more significant than 999, it means there is a soft constraint violation.

The constraint weights used in this study refer to [9] with some modifications. In addition, the weights for user-defined constraints can be adjusted as needed. A higher weight indicates that a constraint has a higher priority than another constraint with a lower weight. If the user does not define the constraint weight, the value will be set to the default weight. Table 2 details the list of constraints and their default weights used in this study.

Table 2. Default weight of each constraint

Violation	Type	Default weight
Timetabling conflicts among lecturers, students, or classrooms	Hard	1.00
Violation of lecturer’s time preferences	Hard	1.00
Violation of lecturer’s room preferences	Hard	1.00
Maximum number of courses per day for a lecturer or student	Hard	1.00
	Soft	0.02
Theoretical courses precede practical courses.	Hard / Soft	0.02
Other violation hard constraints	Hard	1.00
Other violation soft constraints	Soft	0.02

3. Selection

The selection process uses two different rules, but both begin by creating a list of courses that need improvement. The first rule is used when the iteration is still below 1100 because it focuses more on handling conflicting courses. This study selects the number 1100 because, after several trials using real course data from two semesters, the selection and mutation rules following the first rule no longer produce better chromosomes after 1100 iterations.

In the selection process using the first rule, a course is categorized as a course that needs to be improved when there is a conflicting course or when there is a violation of the number of courses in a single day, while the selection process using the second rule also considers theory courses that precede practical sessions and if there are consecutive courses but in different buildings from the lecturer's perspective. This checking process is done by looking at the time availability array and the course list array. From the created list, one course will be randomly selected for mutation. If the list is empty, one will be randomly selected from all courses.

4. Mutation

Mutation is done by changing the time and/or room of the selected course in the selection process. The mutation process also has two different rules, but both start by removing the relationship between the time availability array and the course list array. After that, a list of

combinations of available time blocks and rooms will be created. One of the combinations from the list created will be used as a replacement for the time and/or room of the course. If no combinations meet the requirements, then the time and room of the course will be randomly selected again and separately.

In mutations with the second rule, after the relationship between the course, the time availability array and the course list array is removed, several checks will be performed on the course. If this course has a pair (theory-practice) and violates the theory preceding practice, it will be swapped in time and room with its pair. If this does not happen, the check continues by looking at whether it is a consecutive course in a different building from the lecturer's point of view. If the course meets this condition, a new room in the same building as the previous course will be randomly selected. The next check for the course is whether the subject is a subject that must be placed in a specific room only. If it falls under this condition, then the course meeting will be assigned one of the rooms required to be occupied by this subject. The course-selected time will also be updated by selecting one of the block times where the course can start for the lecturer concerned (visualized in Figure 3). If the course passes all checks, the mutation that will be applied is the mutation with the first rule.

After the new time and room have been determined, the last step of the mutation with the first and second rules is to connect them to the time availability and course list array.

5. Last checking

The next step of implementation of (1+1)-ES is checking the threshold and iteration. The threshold is used to determine whether the algorithm can produce better chromosomes during a number of iterations. The algorithm will stop if the iteration exceeds 1400 and the threshold is reached. However, the algorithm will continue running if this condition is not met. This study selects the number 1400 because treatment with the second rule selection and mutation is expected to be carried out at least 300 times (from iteration 1100-1400). If, after the maximum iteration is reached, there are still violations, these violations will be reported to the user.

3.3. Multiprocessing

In this research, multiprocessing is employed to create multiple timetable trials simultaneously. The configuration data determines the number of trials. For instance, only two timetables will be generated in parallel if the configuration specifies two trials. The service is limited to 8 CPU cores, so the maximum number of trials is capped at 8. The results of each trial are compared based on their cost values, and the trial with the lowest cost is selected as the final solution.

The 'concurrent.futures' library in Python does the multiprocessing. The library provides a high-level interface for asynchronously executing callables using the 'ThreadPoolExectutor' class or 'ProcessPoolExecutor' class. This study uses the second class.

When multiprocessing is activated, several processes run independently, indicated by high CPU usage on multiple cores. To verify this, we can examine the system monitor. Figure 6 presents a screenshot of the system monitor during multiprocessing.

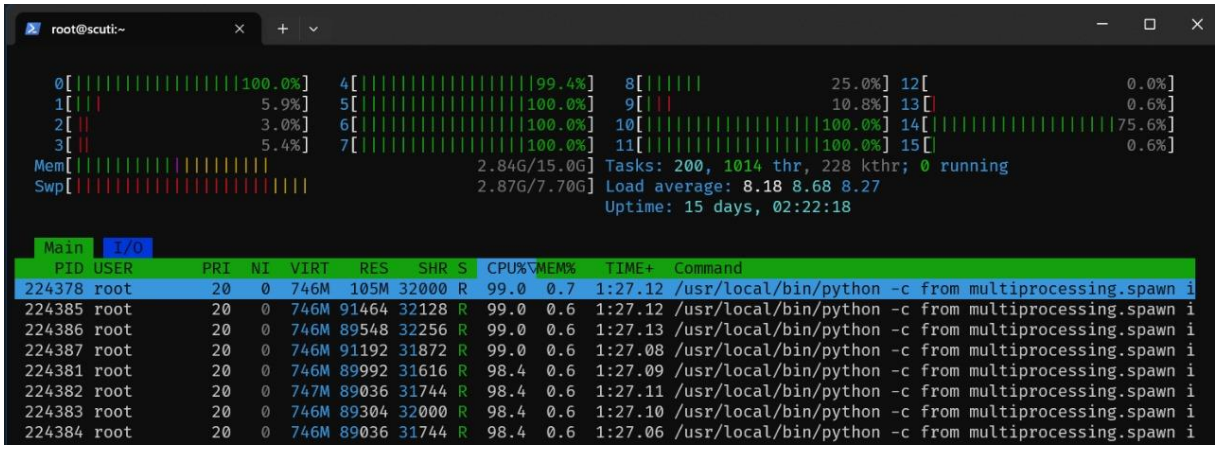


Figure. 6. Server system monitoring while multiprocessing is working

3.4. RESTful API

No specific requirements exist for creating a timetable algorithm as a Representational State Transfer (RESTful) Application Programming Interface (API). The algorithm needs to be written using a programming language that supports the development of RESTful API (native or using library/framework). Utilizing the FastAPI framework facilitates the development of a RESTful API in this research. FastAPI is a modern web framework designed to streamline the creation of APIs in Python. By harnessing the power of Python 3.8+ and the asynchronous capabilities of Uvicorn, FastAPI offers developers a robust, efficient, and the best tool for building high-performance web applications [21], [22]. Benchmark comparisons conducted by TechEmpower have placed FastAPI among the top performers, showcasing its ability to rival the speed and efficiency of Node.js and Go, with Starlette and Uvicorn being the only frameworks to achieve slightly higher benchmarks. Moreover, FastAPI provides

The process commences with the instantiation of an object employing the "fastapi" library. The instantiated object possesses methods that are congruent with the Hypertext Transfer Protocol (HTTP), such as `get()`, `post()`, `put()`, `delete()`, and others. Each method necessitates a string-typed argument that will subsequently represent the endpoint of the constructed RESTful API. After declaring an endpoint, a function must be defined to process any request directed toward the specified endpoint.

This research has developed three endpoints with specific functions. Table 3 shows a detailed list of the implemented endpoints.

Table 3. Endpoint list

Endpoint	Method	Utility
`/generatemultiproc`	POST	Create course timetable
`/validate`	POST	Course data validation
`/update-perkuliahan`	POST	Check timetable results

Despite serving distinct utilities, the endpoints `/generatemultiproc` and `/validate` necessitate an identical request body. These endpoints enforce that each incoming request encapsulates the resources essential for timetabling within the request body. The resources related to the data required for (1+1)-ES, as described in the data preparation (Section 3.1). The enclosed request body must conform to a comprehensible format for the system. A detailed encoding for constraint definition is provided in Table 4.

Table 4. Custom Constraint Specification within the Request Body

Key name	Note	Possible value	Value representation
Keterangan	Information about constraint	Free	-
jenis_target_1	Target constraint	1 2 3 4 5	Lecture Student Subject Classroom All
jenis_constraint	Constraint types	1 2 3	Forbidden block time Mandatory block time Mandatory classroom
prioritas	Priority	1 2	Hard constraint Soft constraint
is_all	Flag to signify universal applicability	True or False	-
data_1	Constrained object	According to jenis_target_1	-
daftar_blok	Constrained block time	0-44	-
daftar_ruang	Constrained classroom	Classroom	-
bobot	Weight constraint	0-1	-

Examples of request body for both endpoints can be seen in the following code snippet.

```
{
 "ruang_kelas": {
  "25": ["251", "252", "253", "254", "255", "256", "257"],,
  ...
 },
 "dosen": [
  {
  "nama": "Ach001",
  "nama_lengkap": "Nama Dosen Ach001",
  "pref_waktu": [
   0, 0, 0, 0, 0, 0, 0, 0, 0, 99, 99, 99, 99, 99, 99, 99, 99, 99, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
  ],
  "pref_ruangan": {
    "25": ["251", "252", "253", "254", "255", "256", "257"]
  }
 },
  ...
 ],
 "perkuliahan": [
  {
   "count": 2,
   "kelas_mhs": ["1D31"],
   "matkul": "Pengantar Ekonomi",
   "tipe": "T",
   "dosen": ["Ach001"],
   "durasi": "3"
  },
  ...
 ],
 "constraints": [
  {
   "keterangan": "Jum'atan",
   "jenis_target_1": 5,
   "jenis_constraint": 1,
   "prioritas": 1,
   "is_all": true,
   "data_1": [],
   "daftar_blok": [39, 40, 41]
  },
  ...
  ],
  "config": {
    "iterasi": 2000,
    "maks_matkul": 2,
    "maks_matkul_prioritas": 1,
    "jml_percobaan": 8,
    "jml_kandidat_solusi": 3,
    "blok_mulai_2_sks": [1, 3, 5, 7],
    "blok_mulai_3_sks": [1, 3, 4, 7]
  }
}
```

For the `/generatemultiproc` endpoint, a valid request body will be forwarded for data processing (data preparation) before finally entering the (1+1)-ES implementation. The response generated by this endpoint is a list of courses with assigned times and rooms, a list of violations, if any, a list of costs generated during the algorithm execution, a list of candidate course timetable solutions, and additional information such as an array of available times and an array of course lists. Meanwhile, for the `/validate` endpoint, a valid request body will only undergo data validation. If the data can be used to create a course timetable, a message indicating the data is ready to be used will appear in the response. Conversely, if the data cannot be used to create a course timetable, a list of error messages in the sent data will appear. The `/update-perkuliahan` endpoint requires a request body similar to the other two. The difference is that in the course definition, the "waktu_terpilih" (selected time) and "ruang_terpilih" (selected room) keys must be added. The response generated by this endpoint is the same as the `/generatemultiproc` endpoint.

3.5. Microservices architecture

The development of the course timetabling system at Politeknik Statistika STIS employs a microservices architecture. This architecture divides the system into four services, each with its specific function. These four services are the front end, back end, data storage, and timetabling service. The focus of this research is on the timetabling service. The microservices architecture and its relationships are shown in Figure 7.

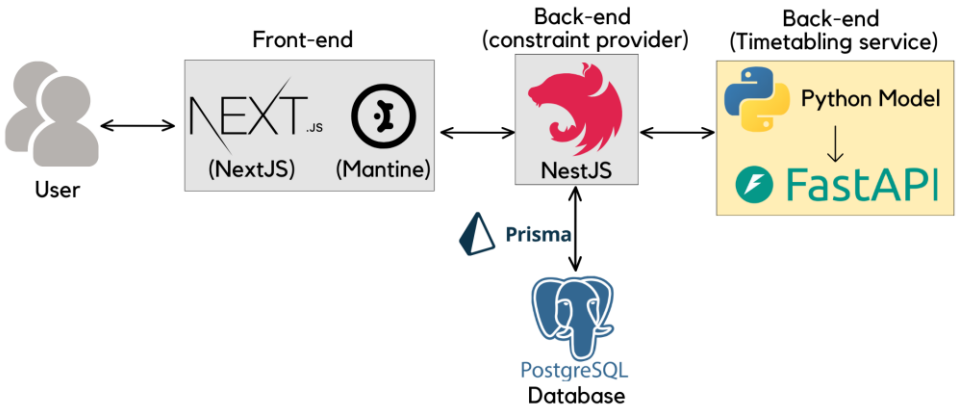


Figure. 7. Microservices architecture

The front-end module is a web page responsible for handling user interactions with the system. Users can interact directly through the webpage to modify timetabling data. Subsequently, this modification request will be sent from the front end to the back end for processing. Once the processing is complete, the back end will respond to the request. This response will be read by the front end and displayed to the user.

The back-end module is a service that interacts with all other services, including data storage. This module handles requests from the front end. If the request is for data modification, the back-end module will modify the data stored in the database using the ORM (Object Relational Mapping) technique and the Prisma framework. However, if the front-end requests to create a timetable, the back-end module will prepare course data from the database and send a request to the timetabling service with the course data as the request body. The response from the timetabling service will be stored in the database and returned to the front-end module.

This research focuses on the timetabling service. This service is built using the Python programming language and utilizes the FastAPI framework. It generates course timetables using the resources sent in the request body. In creating course timetables, the (1+1)-ES implementation is carried out within the service. The resulting timetable, along with any constraint violations, will be sent back to the back-end module.

3.6. Evaluation

As previously explained in the methodology section, the evaluation in this research is conducted through two tests: the cost-value to determine the goodness of the solution generated by the (1+1)-ES implementation and the black box testing to test the RESTful API. All tests were conducted on the service running on the server. The server specifications are as follows:

1) CPU: Intel Xeon 16 cores @ 2.2 GHz
2) RAM: 16GB
3) Storage: 92GB

3.6.1. Cost Value

The course data used in the testing is the course data of Politeknik Statistika STIS for the odd and even semesters of the academic year 2023/2024. The details of the data are presented in Table 5.

The course data was used in an experiment to generate course timetables 1000 times, with each experiment containing 8 different solutions using multiprocessing and aiming for 3 candidate solutions.

The maximum number of iterations used in the (1+1)-ES was 2000, and the cost value of each experiment was calculated.

Table 5. Course data of Politeknik Statistika STIS of the academic year 2023/2024

Note	Semester	
	Odd	Even
Number of student groups	60	50
Number of lecturers	103	88
Number of classrooms	41	41
Number of course meetings	328	281
(2 credit)	39	23
(3 credit)	289	258
Number of constraints defined by user (custom constraints)	9	9

a. Results of the cost value using odd semester course data

The estimated time to generate the odd semester course timetable is 110-160 seconds, averaging 130 seconds (2 minutes and 10 seconds). Table 6 reveals that most of the generated timetables have a cost value between 4 and 5. Given that a hard constraint violation incurs a cost of 1, it can be inferred that most timetables contain hard constraint breaches, specifically four timetable conflicts. However, the (1+1)-ES implementation in this study is designed to strictly adhere to user-defined hard constraints such as forbidden and mandatory blocks. Consequently, the potential violations in the generated timetables are limited to timetable conflicts, exceeding the daily course limit, and soft constraint breaches. Thus, most generated odd semester timetables likely contain four hard constraint violations (timetable conflicts), and the remaining violations are soft constraint breaches. Despite these minor violations, the relatively small cost values compared to the initial iterations suggest that the (1+1)-ES implementation for course timetabling is effective.

Table 6. Percentage distribution of cost values for the odd semester course data.

Range of cost values	Amount	Percentage
$3 \leq Cost < 4$	198	19.8 %
$4 \leq Cost < 5$	405	40.5 %
$5 \leq Cost < 6$	254	25.4 %
$6 \leq Cost < 7$	109	10.9 %
$7 \leq Cost < 8$	29	2.9 %
$8 \leq Cost < 9$	4	0.4 %
$9 \leq Cost < 10$	1	0.1 %
Total	1000	100.0 %

Furthermore, Table 6 indicates that the minimum cost achieved from 1000 timetabling trials is 3. This outcome seems anomalous as the expected minimum cost should be 0. Upon investigation, this anomaly was attributed to three lecturers having ample time availability for their assigned courses. As a result, every generated timetable inevitably violates the daily course limit for these three lecturers. Despite the maximum daily course limit being set at 2, combining these lecturers' time preferences and associated constraints necessitates at least one day with more than two courses.

b. Results of the cost value using even semester course data

In contrast to the odd semester data, the even semester course timetabling experiments yielded significantly better results. Of the 1000 trials conducted, most generated timetables exhibited only two types of violations: consecutive courses in different rooms or buildings and deviations from lecturers' "medium" preferred time blocks. As these violations pertain to soft constraints, the (1+1)-ES algorithm could produce even semester timetables with a minimum cost below 1. Moreover, the average timetabling time was reduced to 110 seconds (1 minute and 50 seconds). These findings strongly suggest

that the (1+1)-ES implementation for timetabling courses at the Politeknik Statistika STIS has been highly successful. A detailed breakdown of cost values is presented in Table 7.

Table 7. Percentage of cost values for the even semester course data.

Cost values	Amount	Percentage
Under 1	978	97.8 %
0.02	594	59.4 %
0.04	211	21.1 %
0.06	134	13.4 %
0.08	26	2.6 %
0.1	6	0.6 %
0.12	7	0.7 %
Above 1	22	2.2 %
1	6	0.6 %
1.02	6	0.6 %
1.04	1	0.1 %
1.06	5	0.5 %
1.08	3	0.3 %
2.04	1	0.1 %
Total	1000	100.0 %

3.6.2. Black Box Testing

Each endpoint underwent black box testing, tailored to its specific request body. Postman, an API platform that streamlines the entire API lifecycle, was employed for these tests. The scenarios and outcomes of the RESTful API black box testing are summarized in Table 8. All test success indicates that the (1+1)-ES has been successfully implemented as a RESTful API.

Table 8. Scenarios and results of black box testing

Tested endpoint	Test scenario	Expected output	Result
`/generatemultiproc`	Assigning values to the JSON attributes of classroom, lecturer, course, constraint, and config.	A course list is presented, including scheduled time and room, violation records, additional information, and a cost value history	Success
`/validate`	Assigning values to the JSON attributes of classroom, lecturer, course, constraint, and config.	A JSON response is generated with a "status" field set to "success" and a message confirming that the course data is prepared for creation.	Success
`/update-perkuliahan`	Assigning values to the JSON fields of classroom, lecturer, course, constraint, and config. Additionally, the course field is populated with both the selected time and the chosen room.	A course list is presented, including scheduled time and room, violation records, additional information, and a cost-value history	Success

4. Conclusion

A (1+1) evolutionary strategy, a specific type of evolutionary algorithm, has been successfully implemented to generate course timetables for the Politeknik Statistika STIS. The service's adaptability

is improved through a user-configurable constraint system. Timetabling conflicts are identified and reported to the user for manual intervention. The algorithm's performance is deemed satisfactory given an average near-zero cost for the even semester data. The slight variation in the lowest achievable cost between semesters suggests that the specific characteristics of the course data have a notable impact on the optimization results. However, the evolutionary algorithm is a standard approach. Currently, there are many advancements in evolutionary algorithms that future researchers can implement into a service.

The RESTful API service has been successfully implemented and is functioning as intended. Black box testing of each endpoint has confirmed that the service is producing the correct outputs for all defined test cases. This finding means that (1+1)-ES has been successfully implemented as a RESTful API.

Ethics approval

Not Required

Acknowledgments

We would like to express our deepest gratitude to all parties who have helped in this research. Special thanks to Farid Ridho, SST., MT., who has provided many new insights and guidance during the 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

Zuhdi Ali Hisyam: Conceptualization, Software Development, Testing, Writing- original draft. **Farid Ridho:** Methodology, System Architecture, writing-review and editing, supervision. **Arbi Setiyawan:** Writing-review and editing, supervision

References

[1] A. Bashab *et al.*, “Optimization Techniques in University Timetabling Problem: Constraints, Methodologies, Benchmarks, and Open Issues,” *Computers, Materials and Continua*, vol. 74, no. 3, pp. 6461–6484, 2023, doi: 10.32604/cmc.2023.034051.

[2] Politeknik Statistika STIS, “A Brief History of the Politeknik Statistika STIS”, (in Indonesian), Accessed: May 18, 2024. [Online]. Available: https://stis.ac.id/hal/16/sejarah-singkat

[3] C. H. Wong, S. L. Goh, and J. Likoh, “A Genetic Algorithm for the Real-world University Course Timetabling Problem,” in *2022 IEEE 18th International Colloquium on Signal Processing and Applications, CSPA 2022 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 46–50. doi: 10.1109/CSPA55076.2022.9781907.

[4] M. V. Rane, V. M. Apte, V. N. Nerkar, M. R. Edinburgh, and K. Y. Rajput, "Automated timetabling system for university course," in *2021 International Conference on Emerging Smart Computing and Informatics, ESCI 2021*, Institute of Electrical and Electronics Engineers Inc., Mar. 2021, pp. 328–334. doi: 10.1109/ESCI50559.2021.9396906.

[5] M. Zunino, S. V. del Valle, and L. Gatti, "An Automated Approach to University Course Timetabling Focused on Professor Assignment," in *2024 L Latin American Computer Conference (CLEI)*, IEEE, Aug. 2024, pp. 1–4. doi: 10.1109/CLEI64178.2024.10700361.

[6] Pradnya A. Vikhar, "Evolutionary Algorithms: A Critical Review and its Future Prospects," *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication*, 2016, doi: 10.1109/ICGTSPICC.2016.7955308.

[7] A. Slowik and H. Kwasnicka, "Evolutionary algorithms and their applications to engineering problems," Aug. 01, 2020, *Springer*. doi: 10.1007/s00521-020-04832-8.

[8] T. Chugh, K. Sindhya, J. Hakanen, and K. Miettinen, "A survey on handling computationally expensive multiobjective optimization problems with evolutionary algorithms," *Soft comput*, vol. 23, no. 9, pp. 3137–3166, May 2019, doi: 10.1007/s00500-017-2965-0.

[9] C. Fernandes, J. P. Caldeira, F. Melicio, and A. Rosa, "HIGH SCHOOL WEEKLY TIMETABLING BY EVOLUTIONARY ALGORITHMS," *Proceedings of the 1999 ACM symposium on Applied computing*, pp. 344–350, 1999, doi: 10.1145/298151.298379.

[10] I. A. Abduljabbar and S. M. Abdullah, "An evolutionary algorithm for solving academic courses timetable scheduling problem," *Baghdad Science Journal*, vol. 19, no. 2, pp. 399–408, 2022, doi: 10.21123/BSJ.2022.19.2.0399.

[11] S. Choudhary, S. Janarthanan, and P. Maurya, "A Study and Analysis of Timetable Generation using a Genetic Algorithm," in *Proceedings - IEEE 2023 5th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 700–703. doi: 10.1109/ICAC3N60023.2023.10541761.

[12] A. Schaerf, "A Survey of Automated Timetabling," *Artif Intell Rev*, vol. 13, pp. 87–127, 1999, doi: 10.1023/A:1006576209967.

[13] H. Algethami and W. Laesanklang, "A mathematical model for course timetabling problem with faculty-course assignment constraints," *IEEE Access*, vol. 9, pp. 111666–111682, 2021, doi: 10.1109/ACCESS.2021.3103495.

[14] T. Bartz-Beielstein, J. Branke, J. Mehnen, and O. Mersmann, "Evolutionary Algorithms," *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 4, no. 3, pp. 178–195, 2014, doi: 10.1002/widm.1124.

[15] R. T. Fielding, "Architectural Styles and the Design of Network-based Software Architectures," 2000.

[16] I. Ahmad, E. Suwarni, R. I. Borman, Asmawati, F. Rossi, and Y. Jusman, "Implementation of RESTful API Web Services Architecture in Takeaway Application Development," in *2021 1st International Conference on Electronic and Electrical Engineering and Intelligent System, ICE3IS 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 132–137. doi: 10.1109/ICE3IS54102.2021.9649679.

[17] A. Soni and V. Ranga, "API features individualizing of web services: REST and SOAP," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 9 Special Issue, pp. 664–671, Jul. 2019, doi: 10.35940/ijitee.I1107.0789S19.

[18] Z. A. Hamza and M. Hammad, "Web and Mobile Applications' Testing using Black and White Box approaches," in *2nd Smart Cities Symposium (SCS 2019)*, 2019, pp. 1–4. doi: 10.1049/cp.2019.0210.

[19] D. Liu, C. Y. Li, Z. Jiang, R. Kong, L. Wu, and C. Ma, "Integrated Power Grid Management System based on Micro Service," in *Proceedings - 2020 8th International Conference on Advanced Cloud and Big Data, CBD 2020*, Institute of Electrical and Electronics Engineers Inc., Dec. 2020, pp. 37–41. doi: 10.1109/CBD51900.2020.00016.

[20] M. Söylemez, B. Tekinerdogan, and A. K. Tarhan, “Challenges and Solution Directions of Microservice Architectures: A Systematic Literature Review,” Jun. 01, 2022, *MDPI*. doi: 10.3390/app12115507.

[21] S. Ramírez, “FastAPI Documentation.” Accessed: Oct. 26, 2024. [Online]. Available: https://fastapi.tiangolo.com

[22] S. J. C. TRAGURA, *BUILDING PYTHON MICROSERVICES WITH FASTAPI build secure, scalable, and structured Python microservices from design concepts to infrastructure*. PACKT PUBLISHING LIMITED, 2022.

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.802

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, December 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Spatial Dependencies in Environmental Quality: Identifying Key Determinants

Omas Bulan Samosir[1], Rafidah Abd Karim[2], M. Irfan Fauzi[3], Sarni Maniar Berliana[4*]

[1]*Lembaga Demografi Faculty of Economics and Business, Universitas Indonesia, Depok, Indonesia*
[2]*Academy of Language Studies, Universiti Teknologi MARA Perak Branch Tapah Campus, Malaysia*
[3]*BPS-Statistics Pesisir Selatan Regency, West Sumatera, Indonesia*
[4]*Politeknik Statistika STIS, Jakarta, Indonesia*
**Corresponding Author:* sarni@stis.ac.id

ARTICLE INFO

Article history:
Received 4 September, 2024
Revised 15 November, 2024
Accepted 19 November, 2024
Published 31 December, 2024

Keywords:
Natural Environment; Queen Contiguity; Spatial Analysis; Spatial Autoregressive Regression

Abstract

Introduction/Main Objectives: Environmental quality is essential to human development because it reflects the condition of our natural surroundings. **Background Problems:** Understanding the determinants of environmental quality is crucial for Indonesia as it helps identify the key factors influencing environmental quality. **Novelty:** Spatial models offer detailed, location-specific insights but require extensive data and computational resources, while non-spatial models provide a broader overview with simpler data requirements but may miss important spatial nuances. This study seeks to identify the determinants of environmental quality in regencies and municipalities on Java Island, incorporating spatial effects into the analysis. **Research Methods:** The dependent variable is environmental quality index. The independent variables are GRDP in industrial sector, GRDP in agricultural sector, urban population rate, population density, and poverty rate. We applied spatially lag regression model using contiguity spatial weight matrix. **Finding/Results:** This study shows the spatially lag regression model outperforms the OLS model. GRDP in the industrial sector, GRDP in the agricultural sector, urban population rate, and population density have negative effects, suggesting the increases in these variables were associated with lower environmental quality. About 40%–44% of each variable's effect on environmental quality is due to spatial spillover effects.

1. Introduction

Environmental quality is essential to human development and well-being because it reflects the condition of our natural surroundings and their capacity to support life [1]. The environment has changed significantly over the past century due to rapid urbanization, industrialization, and population growth. This has resulted in a number of problems, such as deforestation, air and water pollution, and biodiversity loss. In addition to disrupting ecosystems, these changes have put human health at serious risk. For example, air pollution from vehicle and industrial exhausts leads to cardiovascular issues and respiratory problems, while contaminated water sources can cause a variety of waterborne illnesses.

Wildlife is also put at risk by the destruction of natural habitats, which reduces biodiversity—a crucial component of ecological resilience and balance [2].

Several of the Sustainable Development Goals (SDGs), which reflect a global commitment to safeguarding and enhancing the natural environment [3], are closely linked to environmental quality. SDG 6 (Clean Water and Sanitation) emphasizes the importance of ensuring access to clean water and sanitation, as well as their sustainable management for all. Likewise, SDG 11 (Sustainable Cities and Communities) seeks to minimize the adverse environmental effects of urban areas, such as waste management and air pollution. SDG 13 (Climate Action) stresses the need to integrate climate action into national plans and calls for urgent measures to address climate change and its impacts. Additionally, SDGs 14 (Life Below Water) and 15 (Life on Land) focus on the protection of marine and terrestrial ecosystems, respectively. Collectively, these SDGs underscore the critical role that environmental quality plays in achieving sustainable development and improving global living standards.

The Indonesian government places significant importance on environmental quality, as evidenced by a range of key regulations aimed at addressing this issue. Central to this framework is Law No. 32/2009 on Protection and Management of the Environment, which was amended by Law No. 6 of 2023 following the enactment of Government Regulation in Lieu of Law No. 2 of 2022 on Job Creation. This law emphasizes sustainable development and environmental harm reduction, outlining the core principles, objectives, and strategies for environmental governance. It requires the preparation of Environmental Protection and Management Plans (RPPLH) and Environmental Impact Assessments (AMDAL) for projects that may significantly affect the environment. Additionally, the law defines the roles and responsibilities of key government agencies, such as the Ministry of Environment and Forestry (MoEF), in ensuring the enforcement and ongoing maintenance of environmental regulations.

Environmental Protection and Management Regulation No. 22/2021 provides detailed regulations on various aspects of environmental quality, complementing the broader environmental law. This regulation covers topics such as waste management, the preservation of terrestrial and marine ecosystems, and the quality of water and air. It outlines procedures for monitoring and reporting environmental performance and sets requirements for waste treatment, emissions control, and effluent discharge. To ensure adherence to environmental standards, the regulation also stipulates administrative penalties for non-compliance, including fines and the revocation of licenses. Together with other laws and regulations, this framework creates a robust legal structure designed to safeguard Indonesia's natural resources and promote sustainable development.

The MoEF is dedicated to fostering sustainable development that improves the welfare of the Indonesian people and contributes to a more advanced nation. The MoEF's main responsibilities include developing and enforcing policies related to environmental protection, forest conservation, and climate change mitigation. The ministry manages several programs focused on reducing deforestation, conserving peatlands, and promoting biodiversity. Additionally, the MoEF works on encouraging sustainable forest management, controlling pollution, and addressing issues related to hazardous and toxic substances [4]. A key tool used by the MoEF to assess and monitor the environment across Indonesia is the Environmental Quality Index (EQI). This index offers a comprehensive assessment of environmental health by evaluating factors such as air quality, water quality, land conditions, and seawater quality. The EQI is essential for the MoEF to identify areas needing improvement and to guide targeted actions to improve environmental quality [5].

Understanding the factors that determine the Environmental Quality Index (EQI) is essential for Indonesia, as it helps pinpoint the key elements affecting environmental quality. By examining these factors, policymakers and environmental experts can create focused strategies to tackle specific challenges. For example, studies have indicated that variables such as poverty, slum conditions, sanitation, and income inequality (as measured by the Gini ratio) have a significant impact on the EQI [6]. Recognizing these determinants enables more effective interventions, such as strengthening regulations and enforcement for businesses and empowering local communities, which can lead to better environmental management [7]. A deeper understanding of these factors also offers valuable insights for reducing environmental exposures, ultimately helping to mitigate adverse health outcomes [8].

Previous research on the factors influencing environmental quality has been quite extensive. For example, a study by [9] analyzed environmental quality determinants in 198 countries from 1990 to 2018 using panel quantile regression, finding a link between economic activities and carbon emissions. Another study by [10], conducted between 2000 and 2018 in 54 countries part of the Belt and Road Initiative (BRI), employed spatial econometric techniques to explore factors contributing to environmental degradation caused by economic activities. Meanwhile, [11] examined the factors influencing varying levels of environmental quality in Indonesia, particularly comparing Java to other

islands, using canonical discriminant analysis. While the second study applied spatial modeling, the first and third studies utilized non-spatial models. Spatial models offer more detailed, location-specific insights but demand extensive data and computational resources. In contrast, non-spatial models provide a broader overview with simpler data needs, though they may overlook spatial variations and fail to capture spatial dependencies, potentially resulting in less accurate policy recommendations. This study, therefore, seeks to identify the determinants of environmental quality in regencies and municipalities on Java Island, incorporating spatial effects into the modeling of the relationship between independent variables and environmental quality.

2. Material and Methods

2.1. Data

The data used are secondary data sourced from the BPS-Statistics Indonesia and the MoEF. The dependent variable is EQI for regencies/municipalities on Java Island in 2021 except for Kepulauan Seribu. We removed this regency since we applied spatial area approach using contiguity spatial weight matrix. Thus, 118 regencies/municipalities are included in the study. The independent variables are Gross Regional Domestic Product (GRDP) in industrial sector (in natural logarithm), GRDP in agricultural sector (in natural logarithm), percentage of urban population, population density, and poverty rate. We employed several software or packages for data processing, including Rstudio version 2024.04.2 to obtain the regression models, GeoDa version 1.20.0.10 to obtain the spatial weight matrix, and QGIS version 3.36.2 to create thematic maps.

2.2. Model and Analysis Step

Spatial analysis, as defined by [12], refers to the quantitative study of phenomena that occur within a given space. According to [13], it involves examining how human activities and the physical environment vary across space, or, in other words, how these activities change with distance from specific reference points or objects of interest. The concept of "spatial analysis" is broad and includes several key components: (a) processing spatial data through geographic information systems (GIS); (b) conducting descriptive and exploratory analyses of spatial data; (c) applying statistical methods to assess the potential for drawing conclusions; and (d) creating models to predict outcomes and identify relationships within a spatial context [14].

According to Tobler's first law of geography, objects that are located near each other tend to share similar characteristics and are more likely to interact than those that are farther apart. As a result, spatial analysis requires the application of specific values or functions to define what is considered "near," "far," or "neighbor" for a set of spatial objects. In this study, the neighborhoods used to calculate the spatial weights matrix and regression model are defined through the contiguity adjacency matrix, which is the simplest way to define neighbors. The most common contiguity methods include: (1) the rook definition, where neighbors are defined as areal units that share a common edge; (2) the bishop definition, where neighbors are those sharing a common vertex; and (3) the queen definition, which combines both the rook and bishop methods, treating any object that shares either a common edge or vertex as a neighbor [15].

Most spatial analyses follow a standard approach that starts with a non-spatial linear regression model, and then determines whether spatial interaction effects should be included in this baseline model. The non-spatial linear regression model is expressed as [16]

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \varepsilon \tag{1}$$

In equation, $\mathbf{y}$ represents a vector of the dependent variable for each unit in the sample with the size of $n \times 1$ $(i = 1, \mathrm{K}, n)$, $\mathbf{1}_n$ is a vector of ones corresponding to the constant term parameter α to be estimated, $\mathbf{X}$ is a matrix of independent variables with the size of $n \times k$, $\boldsymbol{\beta}$ is a vector of unknown parameters to be estimated with the size of $k \times 1$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \mathrm{K}, \varepsilon_n)^T$ is a vector of error terms, where ε_i is assumed to be independent and identically distributed with a mean of zero and variance σ^2. The linear regression model is often called the OLS model since it is usually calculated using Ordinary Least Squares (OLS).

A comprehensive model that includes all types of interaction effects can be written as

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\theta} + \mathbf{u}, \quad (2)$$

$$\mathbf{u} = \lambda\mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}.$$

In the model , $\mathbf{W}\mathbf{y}$ represents the interaction effects on the dependent variable, $\mathbf{W}\mathbf{X}$ reflects the interaction effects between the independent variables, and $\mathbf{W}\mathbf{u}$ represents the interaction effects between the error terms of the different units, ρ is the spatial autoregressive coefficient, while λ the spatial autocorrelation coefficient. Both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are vectors of unknown parameters to be estimated with a dimension of $k \times 1$. Meanwhle, $\mathbf{W}$ is a nonnegative weight matrix with the size of $n \times n$ that defines the spatial arrangement or structure of the locations in the sample.

When $\lambda = 0$ and $\boldsymbol{\theta} = \mathbf{0}$, the model in simplifies to

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} \quad (3)$$

The simplified model in is known as the spatial lag model or spatially autoregressive model (SAR model), where the spatial lag of y_i for each location i is computed as the weighted sum of the dependent variable values of neighboring locations

$$[\mathbf{W}\mathbf{y}]_i = \sum_{j=1}^{n} w_{ij} y_j = w_{i1} y_1 + w_{i2} y_2 + \mathrm{K} + w_{in} y_n \quad (4)$$

Here w_{ij} represents the weight between the i^{th} and j^{th} location, which is stored in the spatial weights matrix $\mathbf{W}$.

When when $\rho = 0$ and $\boldsymbol{\theta} = \mathbf{0}$, the model in becomes:

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

Where $\mathbf{u}$ is defined in . The model in is referred to as the spatial error model (SEM model). The SEM model addresses spatial dependencies by incorporating of a spatial autoregressive error term [13].

In this case, the values of the dependent variable y at each location are influenced by the stochastic error ε of neighboring locations, as determined by the filter $(\mathbf{I} - \lambda\mathbf{W})^{-1}$.

Before applying model or , it is essential to first assess whether the data show spatial dependence. The first step in a spatial area analysis is to test for spatial dependence without considering independent variables. Moran's I was used to tests for spatial autocorrelation in the data in this study. The formula of Moran's I is presented as [17]

$$I = \left[\frac{n}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right] \times \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}\right] \quad (6)$$

When the data exhibit positive autocorrelation, this means nearby locations will have similar values, resulting in a Moran's I with positive value. Conversely, negative autocorrelation implies that the neighboring locations have dissimilar values, resulting in a negative Moran's I value. A score of 0.3 or greater, or of -0.3 or less, suggests a strong autocorrelation. To assess the significance of Moran's I, a p-value and a z-score are computed based on the null hypothesis of no spatial dependence (i.e., complete spatial randomness). If $p > 0.05$, we fail to reject the null hypothesis, suggesting the spatial distribution of the values could be due to random chance. However, if $p < 0.05$, we reject the null hypothesis and conclude that spatial dependence is present.

When spatial dependence exists in the data, choosing the correct spatial dependence model becomes essential. As outlined in Table 1, we used a set of diagnostic tools known as Lagrange multiplier (LM) test statistics, which include four different tests. If either the LM lag or LM error test is statistically significant, we proceed with the corresponding model. In situations where both tests are

significant, we then check the results of the corresponding robust tests. If only one of the robust tests is significant, we select the model associated with the significant test. However, if both robust tests are significant, we choose the model with the higher value of the robust statistic.

Table 1. Lagrange multiplier diagnostics for spatial dependence.

Name of diagnostic test	Detects	Hypothesis
LM lag	Spatial lag effect	$H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ +
LM error	Spatial error effect	$H_0 : \lambda = 0$ vs. $H_1 : \lambda \neq 0$ +
Robust LM lag	Spatial lag effect	$H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$ +
Robust LM error	Spatial error effect	$H_0 : \lambda = 0$ vs. $H_1 : \lambda \neq 0$ +

+ Reject H_0 when p-value < α

The spatial multiplier, which links the independent variables to the dependent variable, complicates the interpretation of effects in a spatial lag model compared to a non-spatial model (or spatial error model). In a non-spatial model, the effect of a change in an independent variable is consistent across all locations, regardless of which specific location undergoes the change. In contrast, a spatial lag model shows that the impact of independent variables varies by location, due to the differences in neighboring units for each location [18]. The spatial regression model captures the feedback between locations through spatial lag terms, such as $\mathbf{Wy}$, which create interactions where changes in independent variables at one location, say location j, can influence the dependent at another location, i [19]. In a spatial lag model, the effect of an independent variable consists of two components: a direct effect, which reflects the local impact on location i, and an indirect or spillover effect, which is mediated through the spatial multiplier. Thus, the total effect of an independent variable is the sum of both the direct and indirect effects [20].

Table 2. Non-spatial diagnostics.

Name of diagnostic test	Detects	Hypothesis/Value
Jarque-Bera 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 Inflator Factor (VIF)	Multicollinearity	$VIF = 1-4$ (No evidence of collinearity)
		$VIF = 4-10$ (Additional analysis is required)
		$VIF > 10$ (Indicates strong collinearity)

† Reject H_0 when p-value < α

Non-spatial diagnostics were also conducted to check whether the assumptions of normality and homoscedasticity are met, as well as the absence of multicollinearity. A detailed discussion of these is not the focus of this study. Readers can get a detailed discussion of these in [21]. The non-spatial diagnostics carried out in this study are presented in Table 2.

2.3. Model Evaluation

For models' evaluation we used two measures, which are Akaike Information Criterion (AIC) and likelihood ratio test (LRT) to determine the best-fitting model, whether it is the OLS model or the spatial regression model. The AIC is calculated as follows [22]

$$AIC = -2\log\left(L\left(\hat{\boldsymbol{\Theta}}\right)\right) + 2k \tag{7}$$

where $\hat{\boldsymbol{\Theta}}$ is a vector of estimated parameters, $L(\hat{\boldsymbol{\Theta}})$ is the likelihood function of the estimated model parameters, k is the number of estimated parameters, and n is the number of observations. A model is considered to have better performance when its AIC is lower.

We also calculated a likelihood ratio test to compare the spatial regression model as a more complex model to the OLS model as the simpler model. A higher likelihood score is always obtained by adding more parameters. Nevertheless, there comes a time when a model's fit to a given dataset can no longer be significantly improved by adding more parameters. Following the work of [23], the hypotheses for LRT in this study can be defined as follows

H_0 : the OLS model is equivalent to the spatial regression model

H_1 : the spatial regression model outperforms the OLS model

The test statistic for testing the hypotheses stated in is:

$$G^2 = -2\ln\left(\frac{L(\text{spatial regression model})}{L(\text{OLS model})}\right) \tag{8}$$

where $L(\text{spatial regression model})$ is the loglikelihood function under the spatial regression model and $L(\text{OLS model})$ is the loglikelihood function under the OLS model. This test statistic follows a Chi-square distribution with degree of freedom $v_2 - v_1$, where v_1 and v_2 are the number of estimated parameters in the OLS model and in the spatial regression model, respectively. The null hypothesis is rejected when $G^2 > \chi^2_{\alpha, v_2 - v_1}$. If the null hypothesis is rejected, then we should choose the spatial regression model.

3. Result and Discussion

The EQI distribution across Java Island's regencies/municipalities shown in Figure 1. The interval classes were divided based on natural breaks. The EQI value ranges from 41.26 in Bekasi to 72.24 in Batu City, with an average of 59.97. Adjacent regencies/municipalities that have relatively similar EQI scores tend to be clustered together, as shown in Figure 1. In general, the eastern side of Java Island has generally higher environmental quality than the western side, particularly in the regencies/municipalities that are part of Greater Jakarta.

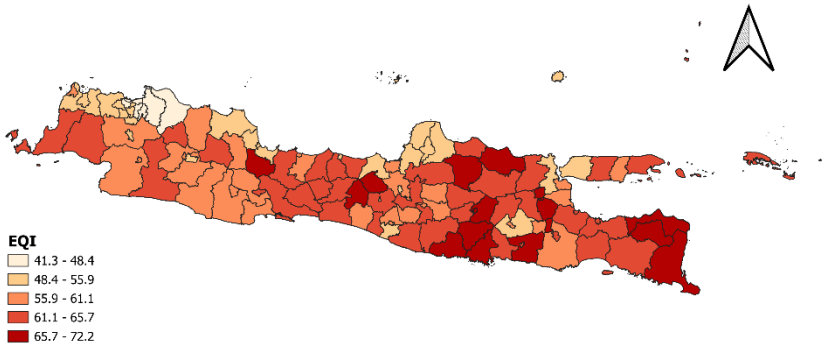


Figure 1. Environmental Quality Index by regencies/municipalities on Java Island.

The non-spatial diagnostics that we performed to assess if our model satisfies the classic assumptions are presented in Table 3. The normality assumption using the Jarque-Bera test gives p=0.3596 that we cannot reject the null hypothesis. Meanwhile, the Breusch-Pagan test for testing the homoscedasticity assumption gives p=0.1255, so we also cannot reject the hypothesis that the errors have constant variance. Each independent variable's *VIF* value is less than five, indicating the lack of

multicollinearity. Therefore, each classic assumption now holds true. The estimated parameters of the OLS model for EQI are presented in Table 4.

All independent variables significantly affect EQI except poverty rate as shown in Table 4. All independent variables have a negative relationship to EQI, meaning that a one-unit increase in the independent variable will decrease EQI by the corresponding regression coefficient. Based on the smallest *p*-value, population density seems to have the strongest effect on EQI, followed by GRDP in industry and urban population rate.

As previously stated, to support the application of spatial regression models, we have to assess if our data reveal spatial autocorrelation. First, we need to compute the spatial weight matrix. In this study, we apply the queen contiguity definition to define a location's neighbors. Since there was the Suramadu bridge that connected Surabaya City and Bangkalan, we modified the neighbor for these two locations, where previously these two locations were not considered as neighbors based on the shapefile used for the analysis. Because there is no connection between Kepulauan Seribu and the Java Island's mainland, we therefore removed this location from the analysis. We have ensured that each location in our study has at least one neighbor, as shown in Figure 2.

Table 3. Results of non-spatial diagnostics.

Name of diagnostic test	Statistic	*p*-value
Jarque-Bera test	2.0458	0.3596
Breusch-Pagan test	8.6135	0.1225
	Independent variable	Value
VIF	GRDP in industrial	1.6346
	GRDP in agricultural sector	4.5700
	Percentage of urban population	3.0646
	Population density	3.2033
	Poverty rate	1.1429

Table 4. Coefficients of the OLS model for EQI on Java Island, 2021.

Coefficients	Estimate	Std. Error	*p*-value
Intercept	88.020337	4.6650	< 2e-16
GRDP in industrial sector	-1.067489	0.3676	0.0044
GRDP in agricultural sector	-1.552635	0.5818	0.0088
Percentage of urban population	-0.067834	0.0253	0.0086
Population density	-0.000861	0.0002	0.0000
Poverty rate	-0.000017	0.0072	0.9982

A location can have as many as 11 neighbors at most, with the average number being 4.42. Bogor has the highest number of neighboring areas, including Bekasi, Bekasi City, Bogor City, Cianjur, Depok City, Karawang, Lebak, Purwakarta, Sukabumi, Tangerang, and South Tangerang. On the other hand, 14 locations have only one neighbor, namely Blitar City, Bogor City, Cilegon City, Cirebon City, Kediri City, Magelang City, Malang City, Mojokerto City, Pasuruan City, Probolinggo City, Salatiga City, Serang City, Sukabumi City, and Sumenep. Except for Sumenep, all of these are cities within larger regencies. Sumenep, located on Madura Island, has Pamekasan as its only neighboring area.

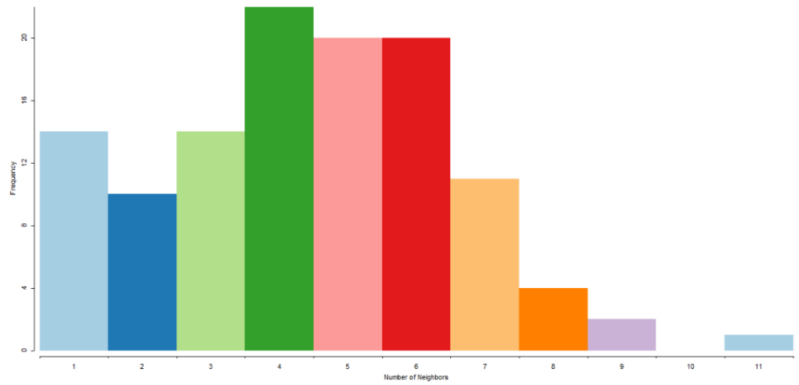


Figure 2. Number of neighbors based on queen contiguity definition.

The Moran's *I* for EQI, computed using the queen contiguity definition is 0.5752 with a *z*-score of 8.7454 (p=0.0000). The value indicates positive spatial autocorrelation, meaning that locations with higher EQI values tend to be surrounded by other locations with high EQI values, and similarly for lower EQI areas. The Moran's *I* distribution, shown in Figure 3, reveals that the observed value indicated by the green line departs significantly from the reference distribution, which contains no Moran's *I* value greater than 0.5752 under 999 permutations. This strongly suggests that the observed spatial autocorrelation is genuine rather than implying that it happened by chance, supporting the use of a spatial regression model over an OLS model for analyzing the determinants of EQI.

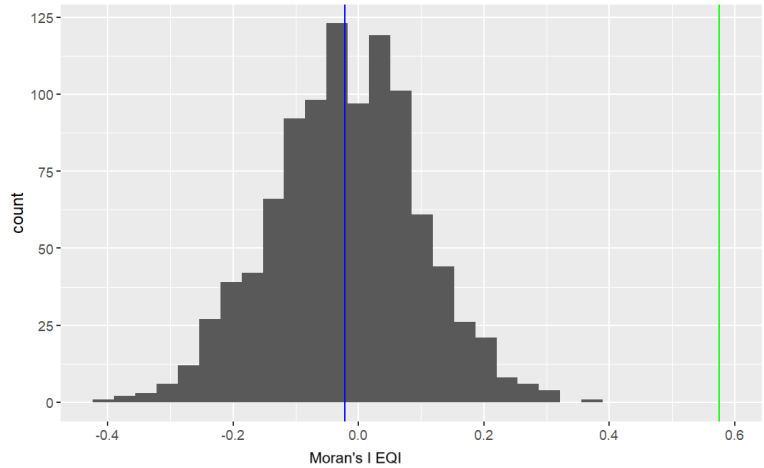


Figure 3. Moran's *I* of EQI under 999 permutations.

The next phase involves choosing the best spatial regression model for the dataset, which requires performing spatial diagnostics as previously mentioned above. The LM diagnostic test results, shown in Table 5, reveal that both the LM lag and the LM error tests are significant at 5%. Consequently, we turn to the robust LM test, which shows that the robust LM lag is markedly significant. This indicates that the correct model to use is the spatial lag model, also referred to as the spatially autoregressive model (SAR model).

Table 5. Results of LM diagnostics for spatial dependence.

Name of diagnostic test	Statistic	*p*-value
LM lag	24.83292	0.0000006252
LM error	20.40707	0.0000062598
Robust LM lag	4.67573	0.03059
Robust LM error	0.24989	0.61716

Table 6. Coefficients of the SAR model for EQI on Java Island, 2021.

Coefficients	Estimate	Std. Error	*p*-value
Intercept	51.4980	7.4930	0.0000
GRDP in industrial sector	-0.6298	0.3151	0.0456
GRDP in agricultural sector	-1.2274	0.5027	0.0146
Percentage of urban population	-0.0636	0.0216	0.0033
Population density	-0.0005	0.0002	0.0018
Poverty rate	-0.0023	0.0061	0.7047
ρ	0.4813	0.0816	0.0000

The results from the SAR model are reported in Table 6. The autoregressive parameter (ρ), is statistically significant, suggesting that spatial spillover effects play a substantial role in the relationship between the independent variables and EQI. The positive value of value is consistent with the positive Moran's I statistic. As shown, all independent variables, except for the poverty rate, have a significant effect on EQI. All variables have negative coefficients, suggesting that increases in these independent variables are associated with a decrease in EQI. As noted earlier, the total effect of each independent variable includes both a direct effect at the location and an indirect spillover effect due to spatial dependencies. The direct, indirect, and total effects of the independent variables along with their significances are presented in Table 7. All independent variables, except for the poverty rate, show significant direct, indirect, and total effects on EQI except for the poverty rate. Approximately 40%–44% of the effect of each independent variable on EQI is attributed to spatial spillover effects.

Table 7. Direct, indirect, and total effectc in SAR model for EQI on Java Island, 2021.

Variable	Direct	*p*-value	Indirect	*p*-value	Total	*p*-value
GRDP in industrial sector	-0.6718	0.0346	-0.5423	0.0702	-1.2141	0.0377
GRDP in agricultural sector	-1.3092	0.0140	-1.0570	0.0432	-2.3662	0.0166
Percentage of urban population	-0.0679	0.0012	-0.0548	0.0246	-0.1227	0.0033
Population density	-0.0005	0.0019	-0.0004	0.0102	-0.0010	0.0015
Poverty rate	-0.0025	0.6587	-0.0020	0.6615	-0.0045	0.6559

Table 8 shows that the SAR model outperforms the OLS model in this study. The positive autocorrelation in the data makes the SAR model an appropriate choice for analyzing the determinants of EQI, as confirmed by the spatial diagnostics in Table 5. The *AIC* value for the SAR model is lower than that of the OLS model, indicating a better fit model. Additionally, the likelihood ratio test (LRT) yields a *p*-value of 0.0000, leading to the rejection of the hypothesis of no difference. The lower *AIC* value and the significant LRT result support the superiority of the SAR model. Furthermore, the standard errors of the SAR model estimators are generally smaller than those of the OLS model. These findings suggest that the SAR model provides a more accurate and reliable fit for analyzing the factors affecting EQI in the regencies and municipalities on Java Island.

Table 8. Model evaluation.

Measure		OLS	SAR
AIC		699.0427	675.3000
Log-likelihood		-342.5213	-329.6500
Likelihood ratio	= -25.743		
Degree of freedom	= 1		
p-value	= 0.0000003901		

This study shows that for every one percent increase in GRDP in the industrial sector at location i will have a direct impact on decreasing the EQI on its own location by 0.6718%. Industrial activities are

often associated with increased pollution and environmental degradation due to emissions from manufacturing processes, waste generation, and resource extraction. Studies have shown that regions with higher industrial GRDP tend to experience a decline in environmental quality, primarily due to the release of pollutants such as carbon emissions (CO2), sulfur air pollution (SO2), and airborne particles [24]. For instance, a study on newly industrialized countries (NICs) found that industrialization-driven economic growth frequently degrades environmental quality, as increased industrial output results in larger ecological footprints [25]. Based on Table 7, it is also known that for every one percent increase in GRDP in the industrial sector of all neighbors of location i will decrease the EQI at location i by 0.5423%. Thus, EQI will decrease by 1.2141% in total for every one percent increase in GRDP in the industrial sector.

The EQI at location i is directly influenced by the GRDP in the agriculture sector, with every one percent increase in this variable resulting in a 1.3092% decrease in EQI. Research indicates that the agricultural sector often has a positive and significant relationship with environmental quality [26, 27]. This is because, when agricultural activities are managed sustainably, practices like crop rotation, organic farming, and conservation farming, which boost soil health and lower pollution, can improve environmental quality. However, the degree of intensity and type of farming practices employed can alter this relationship. The overuse of chemical pesticides and fertilizers, for instance, can cause soil degradation and water contamination, which have an adverse effect on the EQI. The GRDP in the agricultural sector indirectly reduces the EQI at location i by 1.0570% for every one percent increase of all neighbors of location i. So, in total, a one percent increase in this variable will reduce EQI by 2.3662%.

As the urban population rate increases one percent at location i, it impacts directly to decrease the EQI at the same location by 0.0679 units. As the number of people living in cities increases, so does the demand for resources like water, energy, and land, which frequently leads to environmental degradation. High population densities in urban areas can exacerbate pollution of the air and water, increase waste production, and impose a burden on infrastructure and public services. A study on urban growth and population density changes in China shows that rapid urbanization frequently leads to higher levels of air pollution and traffic congestion, which have a substantial negative impact on urban dwellers' quality of life [28]. A different study by [29] found that the unplanned expansion of cities known as "urban sprawl" may result in the loss of natural environments and green spaces, which lowers the quality of the surrounding environment. A study by [30] also shows that the air pollution rises in connection with the urban population. The effect of an increase in urban population from all location i's neighbors will cause a 0.0548-unit decrease in the EQI at location i. This variable's total impact on EQI decline as its increase is 0.1227 units.

There is a statistically significant impact of population density at location i on EQI at that location, both directly and indirectly. The EQI will decrease by 0.0005 and 0.0004 units, respectively, for each one unit rise in population density. This finding is consistent with a study by [30] that shows rising population density will result in higher air pollution. High population density often leads to increased pollution levels as more people generate more waste and emissions. Another study conducted in Indonesia discovered that higher population density correlates with increased air and water pollution, which negatively affects the EQI [31]. High-population density urban areas typically have higher rates of energy consumption, industrial activity, and vehicles—all of which worsen the degradation of the environment. This study shows the total effect of a one-unit increase in population density on the decrease in EQI was 0.001 unit.

4. Conclusion

When spatial dependence is present in the data, using a spatial regression model is essential to appropriately model the spatial structure. This study shows that the SAR model provides a more accurate fit than the OLS model by accounting for spatial dependencies, leading to more reliable parameter estimates and smaller standard errors. The findings indicate that the industrial and agricultural GRDP, urban population rate, and population density all have negative effects on EQI, with increases in these variables being associated with declines in environmental quality. Moreover, around 40%–44% of the effect of each variable on environmental quality is due to spatial spillover effects.

To improve environmental quality, policies should promote sustainable industrial and agricultural practices, manage urban population growth, and address population density issues. Given the significant spatial spillover effects, it is crucial to promote regional collaboration to handle environmental

challenges collectively, such as joint pollution control efforts and coordinated land-use planning to ensure that improvements in one area benefit neighboring regions as well.

Ethics approval

Approval for ethics was not required for this study.

Competing interests

The authors confirm that there are no conflicts of interest.

Funding

No external funding was provided for this study.

Underlying data

The data supporting the results of this study can be requested from the corresponding author.

Credit Authorship

Omas Bulan Samosir: Writing – original draft, Writing – review and editing, Supervision. **Rafidah Abd Karim**: Writing – original draft, Writing – review and editing. **M. Irfan Fauzi**: Data processing. **Sarni Maniar Berliana**: Model conceptualization, Writing – original draft, Writing – review and editing.

References

[1] P. J. Campbell, A. MacKinnon and C. R. Stevens, "The Natural Environment," in *An introduction to global studies*, West Sussex, United Kingdom, Wiley-Blackwell, 2010, pp. 122-160.

[2] United Nations Environment Programme, "Healthy environment, healthy people," UNEP, 2016.

[3] Department of Economic and Social Affairs, "The 17 Goals," United Nations, [Online]. Available: https://sdgs.un.org/goals. [Accessed 25 August 2024].

[4] Ministry of Environment and Forestry, Regulation of the Minister of Environment and Forestry No. 16 of 2020 about the Strategic Plan 2020-2024, Jakarta: Ministry of Environment and Forestry, 2020.

[5] Ministry of Environment and Forestry, Indeks Kualitas Lingkungan Hidup Indonesia 2022 [The 2022 Indonesia Environmental Quality Index], Jakarta: Ministry of Environment and Forestry, 2022.

[6] A. R. Noormalitasari and A. Setyadharma, "Determinants of Environment Quality Index In Indonesia," *Efficient: Indonesian Journal of Development Economics*, vol. 4, no. 2, pp. 1174-1187, 2021.

[7] P. F. Butarbutar, C. F. Ananda and F. Prasetyia, "The determinants of environmental quality in Indonesia," *Journal of Indonesian Applied Economics*, vol. 11, no. 1, pp. 27-39, 2023.

[8] US EPA Research, "EPA Researchers Release Updates to Environmental Quality Index," United States Environmental Protection Agency, 18 May 2021. [Online]. Available: https://www.epa.gov/sciencematters/epa-researchers-release-updates-environmental-quality-index. [Accessed 30 August 2024].

[9] M. Wang, N. Arshed, M. Munir, S. F. Rasool and W. Lin, "Investigation of the STIRPAT model of environmental quality: a case of nonlinear quantile panel data analysis," *Environment, Development and Sustainability,* vol. 23, p. 12217–12232, 2023.

[10] Q. Chen, G. R. Madni and A. A. Shahzad, "The usage of spatial econometric approach to explore the determinants of ecological footprint in BRI countries," *PLoS ONE,* vol. 18, no. 10, pp. 1-14, 2023.

[11] A. Pujiati, T. Nurbaeti and N. Damayanti, "What are the factors that determine differing levels of environmental quality? Evidence from Java and other islands in Indonesia," *Management of Environmental Quality,* vol. 34, no. 2, pp. 290-307, 2022.

[12] L. Anselin, "What is Special About Spatial Data? Alternative Perspectives on Spatial Data Analysis," University of California, Santa Barbara, 1989.

[13] F. Wang, Quantitative Methods and Socio-Economic Applications in GIS, New York: CRC Press, 2014.

[14] D. O'Sullivan and D. J. Unwin, Geographic Information Analysis, 2 ed., Hoboken, New Jersey: John Wiley & Sons, 2010.

[15] L. M. Scott and M. V. Janikas, Spatial Statistics in ArcGIS, M. M. Fischer and A. Getis, Eds., Springer-Verlag Berlin Heidelberg, 2010.

[16] J. P. Elhorst, Spatial Econometrics: From Cross-Sectional Data to Spatial Panels, New York: Springer Heidelberg, 2014.

[17] P. A. P. Moran, "Notes on Continuous Stochastic Phenomena," *Biometrika,* vol. 37, no. 1/2, pp. 17-23, 1950.

[18] M. D. Ward and K. S. Gleditsch, "Location, Location, Location: An MCMC Approach to Modeling the Spatial Context of War and Peace," *Political Analysis,* vol. 10, no. 3, pp. 244-260, 2002.

[19] J. P. LeSage and M. M. Fischer, "Spatial Econometric Methods for Modeling Origin Destination Flows," *SSRN,* p. 25, 2008.

[20] J. P. LeSage and R. K. Pace, "Spatial Econometric Models," in *Handbook of Applied Spatial Analysis*, M. M. Fischer and A. Getis, Eds., Berlin Heidelberg, Springer-Verlag, 2010, pp. 355-376.

[21] D. N. Gujarati, Basic Econometrics, 4 ed., New York: McGraw Hill , 2004.

[22] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Second International Symposium on Information Theory*, Budapest, 1973.

[23] Q. H. Vuong, "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses," *Econometrica,* vol. 57, no. 2, pp. 307-333, 1989.

[24] M. Demiral, Ö. Haykır and E. D. Aktekin-Gök, "Environmental pollution effects of economic, financial, and industrial development in OPEC: comparative evidence from the environmental Kuznets curve perspective," *Environment, Development and Sustainability,* 2023.

[25] C. Karaduman, "The effects of economic globalization and productivity on environmental quality: evidence from newly industrialized countries," *Environmental Science and Pollution Research,* vol. 29, p. 639–652, 2022.

[26] I. Ummi, Rusdarti and H. Y. , "Relationship of the GRDP Sectors with Environmental Quality Index in Indonesia 2012-2017," *Journal of Economic Education,* vol. 8, no. 2, pp. 152 - 158, 2019.

[27] S. Saghaian, H. M. and M. Mohammadi, "The Effects of Agricultural Product Exports on Environmental Quality," *Sustainability,* vol. 14, pp. 1-11, 2022.

[28] H. Lu, Z. Shang, Y. Ruan and L. Jiang, "Study on Urban Expansion and Population Density Changes Based on the Inverse S-Shaped Function," *Sustainability,* vol. 15, no. 13, pp. 1-19, 2023.

[29] H. Zhang, "The Impact of Urban Sprawl on Environmental Pollution: Empirical Analysis from Large and Medium-Sized Cities of China," *International Journal of Environmental Research and Public Health,* vol. 18, no. 16, pp. 1-19, 2021.

[30] L. Han, W. Z. W. Li and Y. Qian, "Urbanization strategy and environmental changes: An insight with relationship between population change and fine particulate pollution," *Science of The Total Environment,* vol. 642, pp. 789-799, 2018.

[31] A. N. Wafiq and Suryanto, "The Impact of Population Density and Economic Growth on Environmental Quality: Study in Indonesia," *Jurnal Ekonomi & Studi Pembangunan*, vol. 22, no. 2, pp. 301-312, 2021.

DOI: https://doi.org/%2010.34123/jurnalasks.v16i2.799

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 2, June 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Small Area Estimation Approaches Using Satellite Imageries Auxiliary Data for Estimating Per Capita Expenditure in West Java, Indonesia

Muhamad Feriyanto[1], Arie Wahyu Wijayanto[2*], Ika Yuni Wulansari[3], Novia Budi Parwanto[4]

[1]BPS-Statistics Barru Regency, Indonesia, [2]Politeknik Statistika STIS, Jakarta, Indonesia, [3]University of Technology Sydney, Sidney, Australia, [4]IsDB Group Chief Economist, Islamic Development Bank, Jeddah, Saudi Arabia
**Corresponding Author: E-mail address:* ariewahyu@stis.ac.id

ARTICLE INFO

Article history:
Received 03 September, 2024
Revised 09 December, 2024
Accepted 09 December, 2024
Published 31 December, 2024

Keywords:
EBLUP; Expenditure per Capita; SAE; Remote Sensing; NTL

Abstract

Introduction/Main Objectives: The economy of a country can determine the welfare of its people. One of the economic indicators in Indonesia is per capita expenditure, which has the lowest estimation at the district level. **Background Problems:** Sub-district level estimates provide detailed information on inequality that cannot be explained at the district level. Unfortunately, sub-district level estimates of per capita expenditure in Indonesia have poor Relative Standard Error (RSE) values. **Research Method**: The Small Area Estimation (SAE) method can improve estimator accuracy on small samples by using auxiliary variable information. **Novelty:** The existence of big geospatial data such as remote sensing provides an advantage in the efficient use of auxiliary variables. **Finding Result**: The Empirical Best Linear Unbiased Prediction (EBLUP) model using Nighttime Light Intensity (NTL) as an auxiliary variable provides the best results of the five proposed models. Remote sensing data can potentially be used in SAE auxiliary variables. It provides opportunities for cheaper, faster, and more efficient data collection compared to conventional data.

1. Introduction

The economy of both developed and emerging nations is vitally important. Each individual has a significant part in the expansion and progress of a nation's economy. As social beings, individuals generally participate in collective transactions such as buying and selling. The economic volatility of a country is predominantly determined by its purchasing and selling operations. Undoubtedly, each household allocates a portion of its earnings towards meeting its daily necessities, including but not limited to education, healthcare, sustenance, taxes, and real estate. Consumers and producers are two economic categories that households might occupy. In addition to providing goods and services, households may also provide capital, land, and entrepreneurialism [1].

As per the classification provided by the BPS-Statistics Indonesia (Badan Pusat Statistik, BPS) [1], food expenditure and non-food expenditure. Food expenditure refers to the monthly amount that each household spends on food to meet its total consumption. Whereas, expenditure for non-food is the total consumption each household spends in a month for non-food, such as education, health, tax, etc. At now, the estimation of per capita expenditure in Indonesia is conducted at the district level, the lowest

attainable level. Per capita gross domestic product in Indonesia amounted to IDR 56 million in 2018, which is equivalent to IDR 4.67 million per month. This economic value increased by 5.17 percent compared to the previous year [2]. Provinces in Java contributed the most to this economic value expansion, which also surpassed the target.

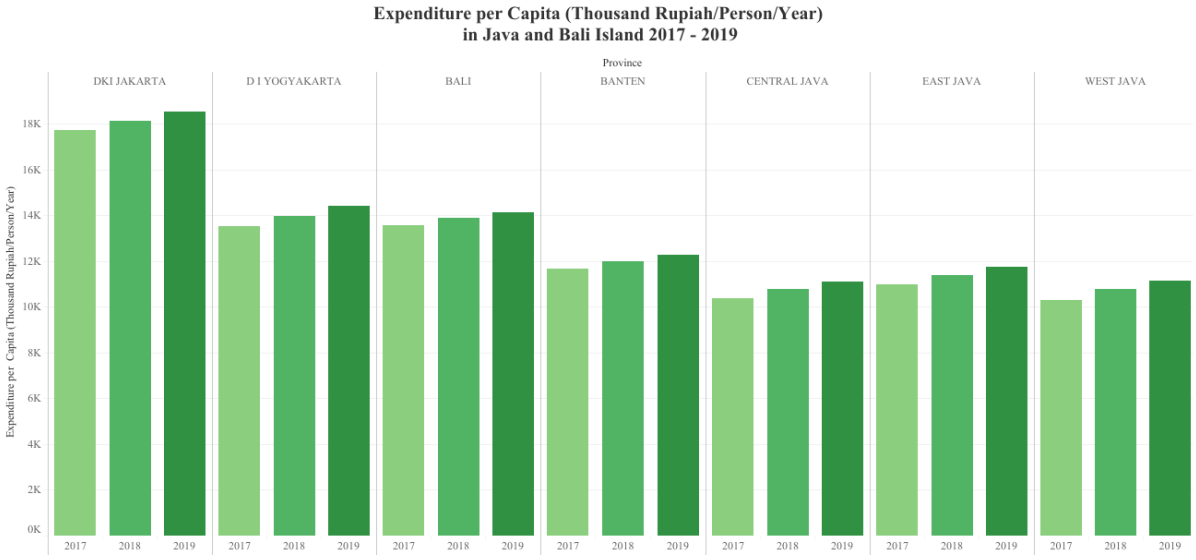


Figure 1. Expenditure per Capita in Java and Bali Island 2017-2019

West Java, situated on Java Island, is the second most populated province in Indonesia, trailing only Jakarta in terms of population [3]. Despite being one of the provinces on Java Island—the administrative heart of Indonesia—West Java ranked fifteenth in Indonesia in terms of per capita expenditure in 2018. Figure 1 shows that West Java ranks lowest out of all the provinces in Java and Bali Island from three years, 2017 until 2019. The province on Java Island should, in theory, be more populous and reflect higher levels of human well-being as depicted in Figure 1 on Expenditure per Capita in Java and Bali Island 2017-2019. As it ought to, the province on the island of Java is capable of accommodating a greater populace and enhancing human welfare.

Estimation of expenditure per capita at the district level only reflects the region as a whole, whereas, in reality, the situation in the district is not homogeneous. The estimated value of per capita expenditure at a lower level (such as the sub-district level) can be used by stakeholders as the foundation for good policy-making. The lower the level, the more detailed the information can be obtained. Unfortunately, per capita expenditure in Indonesia is only assessed at the district level since the lower the administrative level, the lower the estimated value, as demonstrated by the high Relative Standard Error (RSE) value. In Indonesia, a good estimated value is one with an RSE value of less than 25% [4]. Estimation of expenditure per capita at a lower level, which has a minimum sample, can be done using the Small Area Estimation (SAE) model.

Small Area Estimation (SAE) is an indirect estimation method that borrows the strength of auxiliary variables to gain the variation of some variables from the area between [5]. According to Rao and Molina (2015), an explicit model of Small Area Estimation is a model that considers the variability of the area between and combines the linear mixed model and generalized linear mixed model. The SAE model can be classified into two, unit level and area level. Area-level models are often used in various analyses because the ease of data in the level area can be obtained. The feasibility of Small Area Estimation model is determined from the goodness of auxiliary variables used.

By reflecting or sending waves to Earth, a technology known as remote sensing can be used to gather information about the planet [6]. Satellite imagery is the end product of remote sensing, and it contains information that can be extracted. Satellite imagery is the end product of remote sensing, and it contains information that can be extracted. For instance, The NOAA-VIIRS creates a product called Nighttime Light Intensity (NTL), which provides data on the index used to observe the level of electricity in the area [7]. Currently, the results of remote sensing have been used in various fields such as poverty [8], [9], [10], [11], [12], land cover [13], [14], estimation of electricity [7], [15], urban identification [16], and many more.

Nowadays, the application of the Small Area Estimation model is already in various fields, such as estimation in agriculture indicators [17], [18], [19], poverty estimation [20], [21], infant mortality rate (IMR) [22], food insecurity [23], [24], and expenditure per capita [25], [26]. Big data has also been

widely employed in SAE modelling, in addition to data from surveys and censuses. One type of big data that is frequently utilized in Small Area Estimation modelling is remote sensing. In order to represent the whole region down to the smallest grid, auxiliary variables must have error-free coverage of the entire region. Similarly, satellite imagery findings are thought to cover the region extensively. For instance, Singh et al. (2002) modelled crop yield estimation using the Normalized Different Vegetation Index (NDVI) and the Ratio Vegetation Index (RVI), while Kaban et al. (2022) estimated spending per capita using Nighttime Light Intensity (NTL).

This study follows on from the prior research [25] on estimation per capita using Nighttime Light Intensity. The purpose of this study's renewal is to broaden the use of remote sensing for auxiliary variables in SAE. Furthermore, the data from satellite imagery, Nighttime Light Intensity (NTL) and Land Surface Temperature (LST) for auxiliary variables are combined in this study. The use of big data, particularly remote sensing in small area models, allows the government to reduce the Relative Standard Error (RSE), allowing it to make public policy in sub-districts or other level areas with a small sample size. Furthermore, the abundance and low cost of remote sensing data, as well as the comprehensive coverage of remote sensing data, means that estimates can be made more efficiently, both in terms of cost and time. Hence, a combination of remote sensing and Small Area Estimation provides the estimation expenditure per capita in granular or minimum sample areas, such as sub-districts more effectively and efficiently.

2. Material and Methods

2.1. Type of Research

This study assesses per capita expenditure at the sub-district area using quantitative approach. The method Empirical Best Linear Unbiased Predictor (EBLUP) is used to compare the proposed model with several possible auxiliary variables datasets.

2.2. Location and Time Research

Located in the west of Java Island near Banten and Central Java Provinces, West Java Province had 27 districts and 626 sub-districts in 2018. This is a cross-sectional study that makes use of secondary data from Statistics Indonesia (Badan Pusat Statistik, BPS) for conventional data and remote sensing data as the application of the usage of big data. The Indonesian National Socio-Economic Survey (SUSENAS) in March 2018 and Village Potential (PODES) 2018 data were used in this study. SUSENAS is a biannual survey conducted by BPS every March and September. Whereas, PODES is a census conducted by BPS three times every ten years, exactly one year after the main census. Data remote sensing was collected in the Google Earth Engine (GEE) catalogue which provides many sources of remote sensing data from many types of satellites.

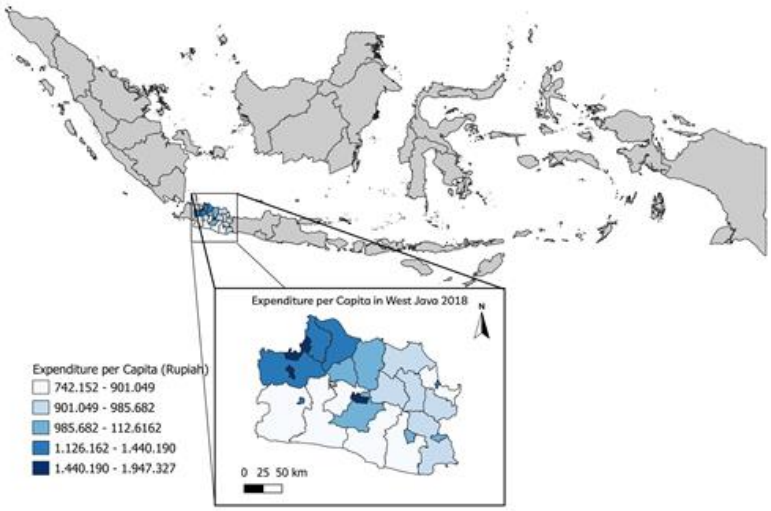


Figure 2. West Java (Area of Research Interest)

2.3. Research Variables

Expenditure per capita was the main estimator collected in SUSENAS in March 2018. Per capita expenditure in SUSENAS is calculated by adding the sum of food and non-food expenditures in a month from the consumption and expenditure questionnaire. The variables obtained by PODES, which are detailed in Table 1, are included as auxiliary variables in the Small Area Estimation model to model per capita expenditure. All variables in Table 1 are aggregated to the sub-district level, beginning with raw data collected at the village and household levels, to facilitate estimations at the sub-district level.

Table 1. Conventional Data

Symbol	Variable	Description	Source
Y	Expenditure per Capita	Expenditure per capita in a month in Indonesian Rupiah (IDR)	SUSENAS 2018
$X1$	University	Total of University	PODES 2018
$X2$	Islamic Boarding School	Total of Islamic Boarding School	PODES 2018
$X3$	Mechanical Training	Total of Mechanical Training Place	PODES 2018
$X4$	Language Training	Total of Language Training Place	PODES 2018
$X5$	Medical Facility	Total of Medical Facility	PODES 2018
$X6$	Mini Market	Total of Mini Market	PODES 2018

Source: Statistics Indonesia (Badan Pusat Statistik, BPS)

The application of remote sensing in this study uses two variables from remote sensing, Nighttime Light Intensity (NTL) and Land Surface Temperature (LST). The Asian Development Bank (2016), ADB used NTL data as an approach to socioeconomics indicators in several developing countries. The NTL data is strongly positively correlated with average household expenditure and GDP growth in developing countries [28]. The occurrence of economic growth can be followed by the population growth, which indicated by increasing luminosity of nighttime light intensity. The increase in population in urban areas has a tendency to increase land surface temperature as one of the effects of urban heat islands [29].

These two kind of variables are used individually and combined to get the best estimate. From Table 2, the NTL data in 2018 were obtained from the Suomi-NPP Satellite with Visible Infrared Imaging Radiometer Suite (VIIRS) instrument which has 750 meters of spatial resolution. Whereas LST data in 2018 were obtained from Terra Satellite with a Moderate Resolution Imaging Spectroradiometer (MODIS) instrument which has 1,000 meters of spatial resolution.

Table 2. Remote Sensing Data

Variable	Satellite	Resolution Spatial	Temporal	Band used
Nighttime Light Intensity (NTL)	Suomi-NPP	750 m	16 days	avg_rad
Land Surface Temperature (LST)	Terra	1,000 m	16 days	LST_Day_1

This research uses remote sensing, a big data application, in conjunction with the statistical method. According to Figure 5, the analysis starts by obtaining raw data SUSENAS and PODES 2018 from BPS. Afterwards, the BPS-specified weighted value for design sampling was used for the calculation of subdistrict-level expenditure per capita. The formula used to calculate the expenditure per capita using weights is as follows:

$$\bar{y}_i = \frac{\sum_{i=1}^{n} w_{ij} y_{ij}}{\sum_{i=1}^{n} w_{ij}} \quad (1)$$

where $\bar{y}_i$ is average expenditure per capita in sub-district i^{th}, $w_{ij}$is a weighted factor in j^{th}household in i^{th} sub-district, and y_{ij} is expenditure per capita a month in j^{th}household in i^{th} sub-district. This estimation is known as direct estimation from subdistrict level expenditure per capita.

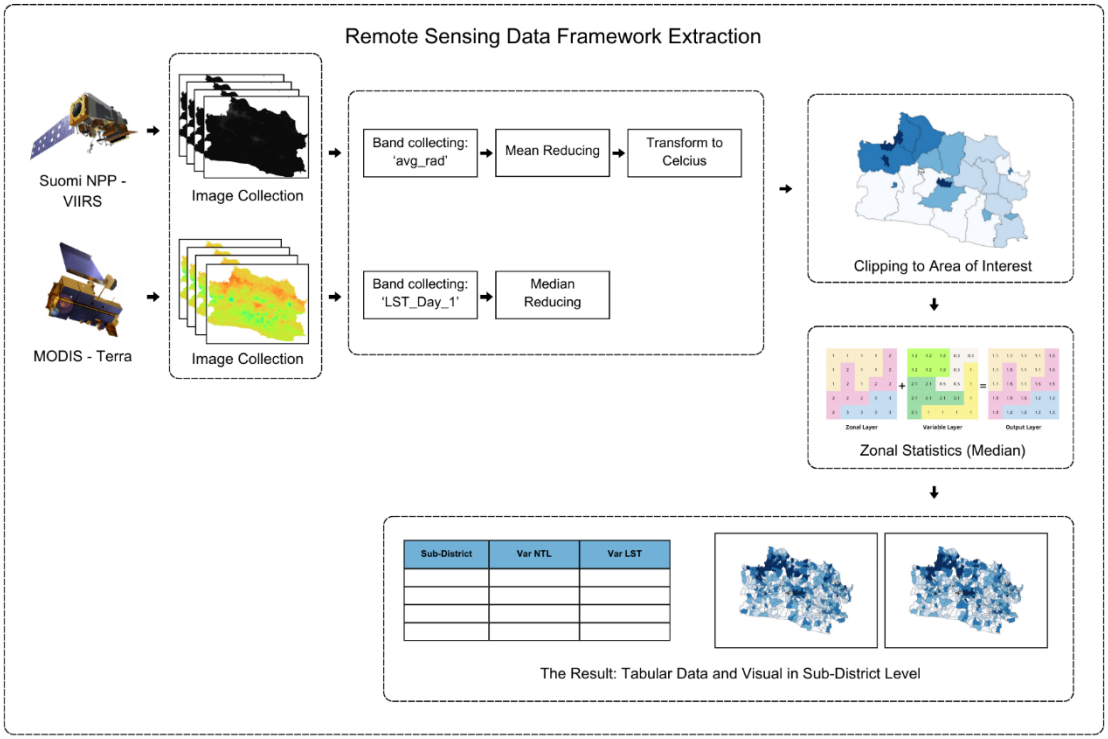


Figure 3. Remote Sensing Data Framework Extraction

All datasets are aggregated into sub-district levels. Data from SUSENAS is aggregated using formula (1), which uses unique weights for every household. PODES data obtained at the village level is aggregated by adding up all the total sub-district variables. Whereas, the remote sensing data was obtained from the image collection throughout 2018 from its satellite and its band. The image collection containing the band used in this study is reduced by mean or median reducer as shown in Figure 3. The result is single layer from NTL or LST in 2018 followed by clipping to area of interest (West Java) for zonal statistics. The zonal statistics concept is to take average or median of all pixel value of the variables within the administrative area (sub-district). Figure 4 shows that all of the variable values in the same zonal statistics administration will be averaged. Hence, the output layer represents the variable value (NTL or LST) in each sub-district.

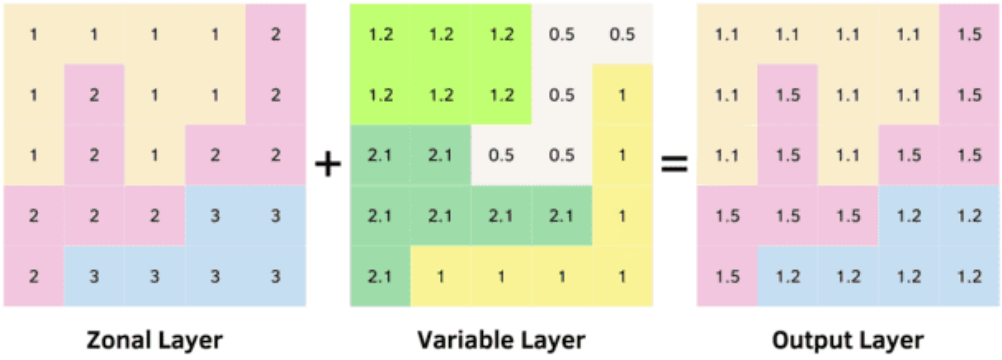


Figure 4. Zonal Statistics Illustration

Small Area Estimation (SAE) is one of the techniques of indirect estimation that borrows strength from auxiliary variables to estimate an estimator that has an inadequate sample [5]. SAE is typically divided into two distinct levels of estimation: area level and unit level. Due to the simplicity of data acquisition, area level is frequently used. Model Empirical Best Linear Unbiased Prediction (EBLUP) is used to estimate expenditure per capita in West Java. This model is an enhanced version of the preceding model BLUP; it estimates σ_v^2 using $\hat{\sigma}_v^2$ since that value σ_v^2 is unknown in reality.

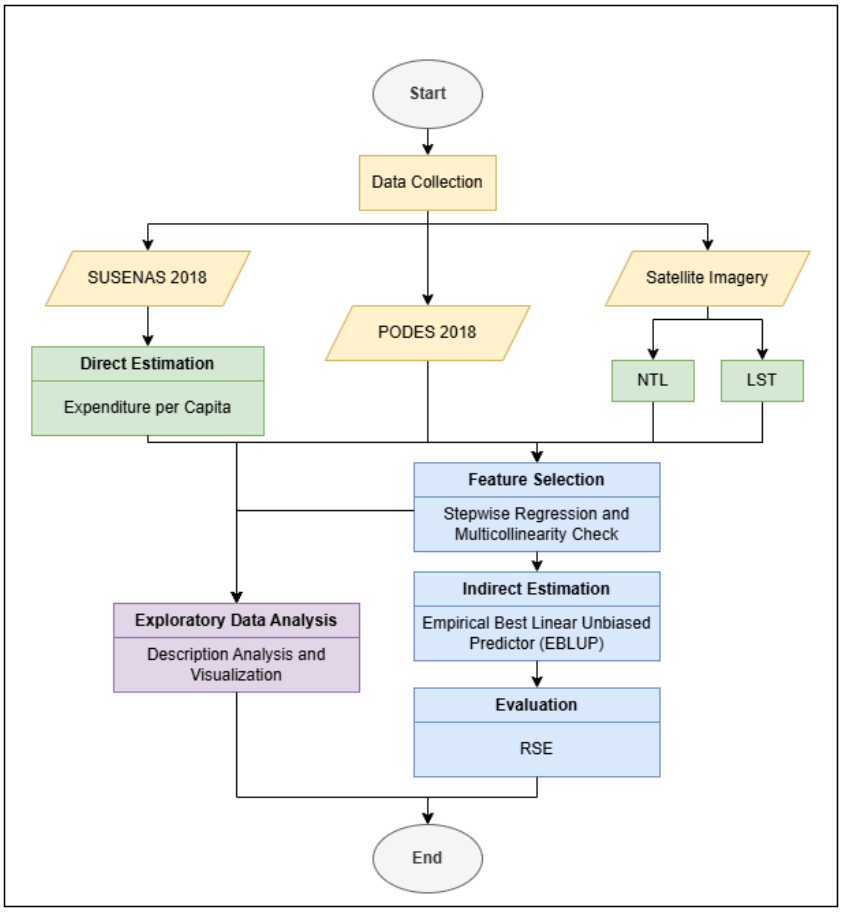


Figure 5. Framework Study

Before the variables are used, the stepwise regression method is used to select which auxiliary variables have a significant effect on the expenditure per capita. Besides stepwise regression, the Pearson correlation is also used to see the relationship between auxiliary variables. The pair of auxiliary variables that have a Pearson correlation value greater than 0.8 is excluded from the model. To ensure that there is no multicollinearity in the auxiliary variables, in addition to the Pearson correlation, the VIF (Variance Inflation Factor) value is also used in feature selection. A good VIF value to use is no more than 10 [30].

The equation of the EBLUP model to estimate expenditure per capita is as follows [5]:

$$\hat{\theta}_i = \boldsymbol{z_i^T \beta} + b_i v_i + e_i \; ; i = 1,2,3, \dots, m \tag{2}$$

with random effect area $v_i \sim iid\ N(0, \hat{\sigma}_v^2)$ and error $e_i \sim iid\ N(0, \sigma_e^2)$

$$\hat{\theta}_i^H = \hat{\gamma}_i \hat{\theta}_i + (1 - \hat{\gamma}_i) \boldsymbol{z_i^T \hat{\beta}} \tag{3}$$

$$\hat{\gamma}_i = \frac{\hat{\sigma}_v^2 b_i}{(\hat{\sigma}_v^2 b_i + \psi_i)} \tag{4}$$

where $\hat{\theta}_i^H$ is an estimator of expenditure per capita in i^{th} sub-district using EBLUP, $\hat{\gamma}_i$ is a measure of uncertainly model, $\boldsymbol{z_i^T}$ is a covariate area level that is an auxiliary variable, $\boldsymbol{\hat{\beta}}$ is an estimator of β that is a vector regression coefficient, and $\hat{\sigma}_v^2$ is an estimator of σ_v^2 using the Restricted Maximum Likelihood (REML) method.

Mean Squared Error (MSE) is the average of differences between the true value and the estimator value [5]. The smaller the MSE value, the resulting estimator is considered good and can be used in estimation. According to Rao and Molina [5], MSE from EBLUP model can be obtained by Taylor Series Expansion. By following formula (5), the MSE value for each model can be compared to obtain an MSE value that tends to be low.

$$MSE\left(\hat{\theta}_i^H\right) = g_{i1}(\hat{\sigma}_v^2) + g_{2i}(\hat{\sigma}_v^2) + 2g_{2i}(\hat{\sigma}_v^2) \tag{5}$$

$$g_{1i}(\hat{\sigma}_v^2) = \frac{\hat{\sigma}_v^2 \psi_i}{(\hat{\sigma}_v^2 + \psi_i)} = \hat{\gamma}_i \psi_i \tag{6}$$

$$g_{2i}(\hat{\sigma}_v^2) = (1 - \hat{\gamma}_i)^2 \boldsymbol{z_i^T} \left[\sum_{i=1}^{\boldsymbol{m}} \frac{\boldsymbol{z_i z_i^T}}{(\hat{\sigma}_v^2 + \psi_i)} \right]^{-1} \boldsymbol{z_i} \tag{7}$$

$$g_{3i}(\hat{\sigma}_v^2) = \psi_i^2 (\psi_i + \hat{\sigma}_v^2)^{-3} \underline{V}(\hat{\sigma}_v^2) \tag{8}$$

Where $\underline{V}(\hat{\sigma}_v^2)$ is the asymptotic variance of $\hat{\sigma}_v^2 = 2m^{-2} \sum_{i=1}^{m} (\hat{\sigma}_v^2 + \psi_i)^2$. In addition to Mean Squares Error (MSE), Relative Standard Error (RSE) is utilized for model evaluation; a model with a lower RSE value is more suitable for estimating per capita expenditure. The RSE calculation formula is as follows:

$$RSE\left(\hat{\theta}\right) = \frac{\sqrt{MSE\left(\hat{\theta}\right)}}{\hat{\theta}} \times 100\% \tag{9}$$

where $\hat{\theta}$ is the estimate of expenditure per capita and $MSE(\hat{\theta})$ is the Mean Squared Error (MSE) of the estimator $\hat{\theta}$.

3. Result and Discussion

3.1. Characteristic of Per Capita Expenditure in West Java

Per capita expenditure in West Java is not evenly distributed at the district level. Figure 7 shows that the distribution still contains the high outlier and has a right-skewed. The distribution of level subdistrict expenditures per capita tends to be concentrated in the northwest, with the highest expenditure group (Figure 6). Besides that, the center of West Java is also clustered highest expenditure per capita. This trend is in big cities in West Java, such as Bandung City, Bekasi City, and Depok. The municipalities that exhibited the highest per capita expenditure in West Java in 2018 are encompassed under this group. Sub-districts in these cities tend to cluster homogeneous, conversely, the clustering of sub-districts differs in other cities, including Karawang, Subang, Indramayu, and Cianjur. It is crucial, then, to obtain more precise estimates of per capita expenditures at the subdistrict level.

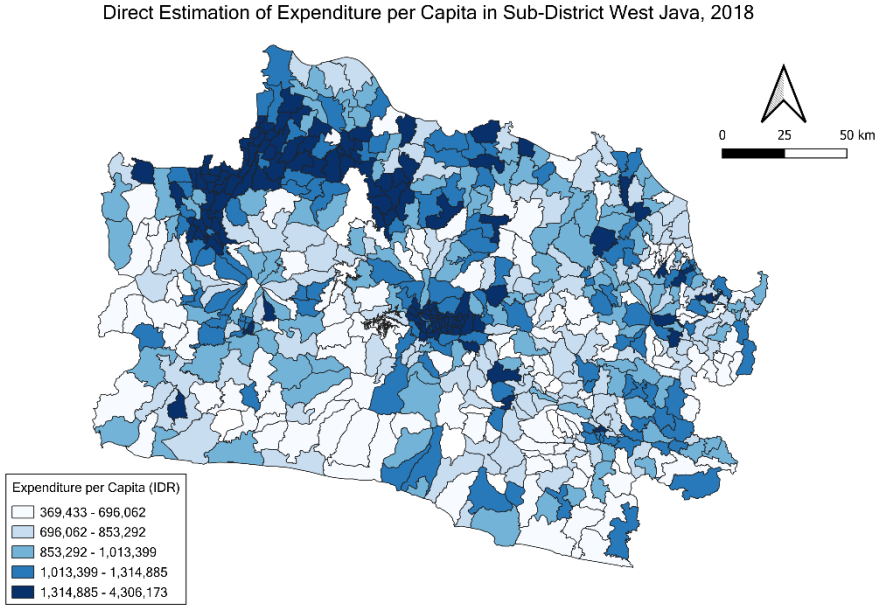


Figure 6. Distribution Expenditure per Capita in Sub-District West Java 2018

Visual mapping is used to find out the general distribution of Nighttime Light Intensity (NTL) and Land Surface Temperature (LST) in West Java. Table 3 shows that the NTL value that has been aggregated in sub-district level is in the index range between 0.24-32.59, whereas the LST value is in the range between 22.62°C – 38.14°C. The subdistrict with the lowest NTL value is Naringgul (0.24) in

Cianjur Regency, whereas Sumur Bandung in Bandung City (32.59) has the highest NTL value in West Java. In West Java, the region characterized by the greatest average surface temperature is Bekasi Barat, Bekasi City (38.14^{o}C), whereas Kudadampit, Sukabumi has the lowest temperature (22.62^{o}C).

Nighttime Light Intensity (NTL) is produced by human activity in the night such as building lighting in the town, house lights, gardens, plantation, and limestone quarries [31]. According to Subash et al. [32], NTL demonstrates the social-economy in a given location and can be used to forecast poverty.

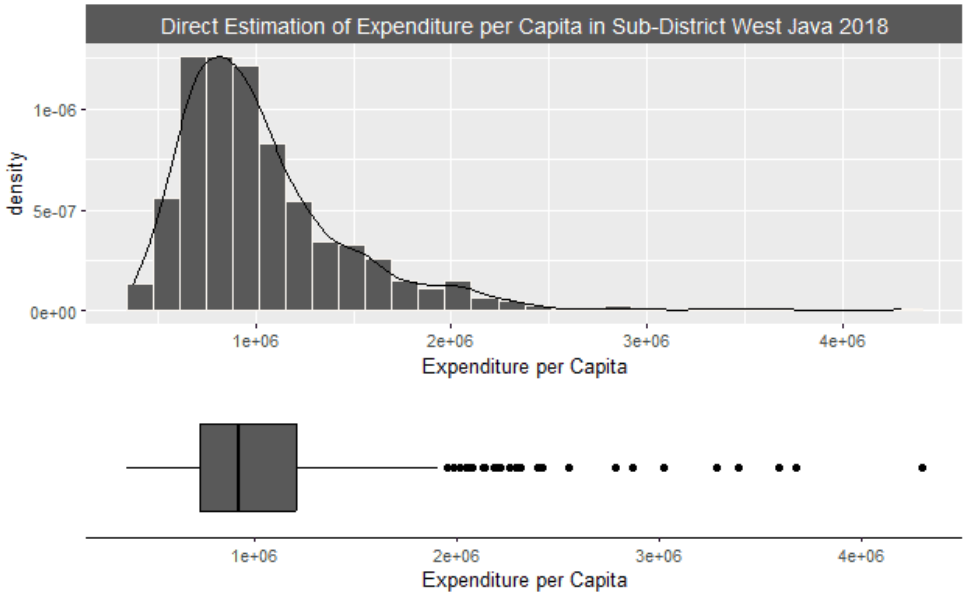
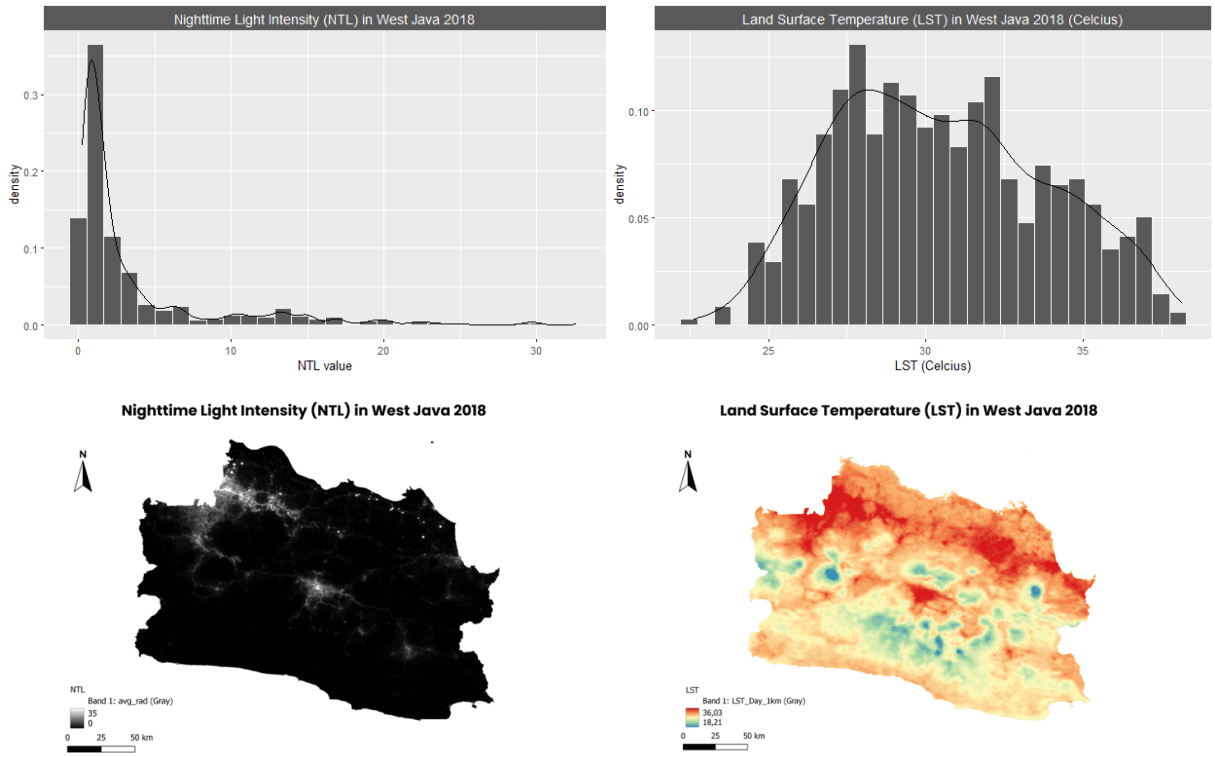


Figure 7. Boxplot and Histogram Expenditure per Capita in Sub-District West Java 2018

Figure 8a shows that the intensity of nighttime light in West Java tends to be high in the northwest and central parts of West Java, Bandung City. This evidence shows that the Nighttime Light Intensity has a high correlation with expenditure per capita. Figure 8b shows that the north part of West Java has a high surface temperature, this is indicated by the red mark in the figure. In contrast with the east side, this area tends to be blue, which indicates that the area has a colder temperature than the north. According to Bodruddoza et al. [33], the higher the Land Surface Temperature (LST) of an area, the more rapid the development and growth of that area will be.

Table 3. Summary of Variables

Variable	Min	Mean	Median	Max	Std Dev	Range
Expenditure per Capita (Y)	369,432.91	1,051,612.26	928,571.89	4,306,173.42	486,660.60	3,936,740.51
University (X1)	0	0	0	7	0.62	7
Islamic Boarding School (X2)	0	50	35	223	41.09	223
Mechanical Training (X3)	0	1	0	48	4.33	48
Language Training (X4)	0	1	0	41	2.67	41
Medical Facility (X5)	0	6	5	27	5.52	27
Mini Market (X6)	0	14	8	190	19.92	190
NTL	0.24	3.72	1.42	32.59	5.43	32.35
LST (Celsius)	22.62	30.49	30.23	38.14	3.31	15.52



(a) (b)

Figure 8. (a) Distribution of Nighttime Light Intensity (NTL) in West Java 2018 Sub-district level; (b) Distribution of Land Surface Temperature (LST) in West Java 2018 Sub-district level

3.2. Feature Selection

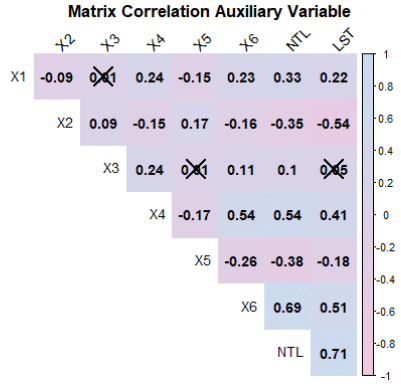


Figure 9. Matrix Correlation Auxiliary Variables

The Pearson correlation between auxiliary variables which is not more than 0.8 (as shown in Figure 9) and the VIF values below 10 for each variable (as shown in Table 4 (2)) indicate that there is no multicollinearity problem. All variables are good predictors because there is no multicollinearity among them, and they have a significant correlation to per capita expenditure.

Table 4. Pearson Correlation and VIF Value

Variable	VIF Value	Pearson Correlation with Y	p-value
X1	1.1323	0.2955	0.0000*
X2	1.5225	-0.3299	0.0000*
X3	1.0788	0.1336	0.0007*
X4	1.6078	0.4520	0.0000*
X5	1.2172	-0.3449	0.0000*
X6	2.0991	0.5432	0.0000*
NTL	3.4628	0.7019	0.0000*
LST	2.6207	0.5648	0.0000*

*significant in 5%

Table 5. Stepwise Regression

Step	AIC
NA	15,610.62
+NTL	15,208.26
+LST	15,199.56
+X5	15,192.03
+X4	15,185.25
+X3	15,181.98
+X1	15,178.88
+X6	15,176.78
+X2	15,174.43

Akaike Information Criterion (AIC) can be used to determine how good a model is. The lower the AIC value, the better the model will be. If all variables are used in this scenario, Table 5 shows that all variables are included in the model, and the minimum AIC is obtained. Therefore, there is no auxiliary variable is dropped in all scenarios.

3.3. Small Area Estimation

The Empirical Best Linear Unbiased Prediction (EBLUP) model is used to estimate the expenditure per capita in West Java, 2018. The scenarios used in this study are: using all PODES variables, using only NTL or LST, using combined remote sensing data, using all variables as auxiliary variables. According to Table 6, every PODES variable is significant at the 5% significant level. This means that all variables from PODES have a significant influence on per capita expenditure in West Java. Similar to Table 6, the next scenarios (shown in Table 7, Table 8, and Table 9) yield significant results in 5%. This implies that the auxiliary variables have done a good job of predicting the per capita expenditure in the four scenarios. In contrast to the previous result, in the scenario where all variables are used in forming the EBLUP model, there is one variable, Language Training (X4) that is not significant at the 5% significant level (Table 10). In SAE modelling, the beta interpretation of each variable is not necessary because the focus of SAE modelling is to improve the estimation accuracy by lowering the Relative Standard Error (RSE) value to as low as possible.

Table 6. EBLUP with PODES as Auxiliary Variables

Variable	Beta	Std Error	t-value	p-value
Intercept	989,954.71	24,506.56	40.40	0.0000*
X1	102,653.68	22,104.02	4.64	0.0000*
X2	-1,887.53	268.85	-7.02	0.0000*
X3	6,254.24	2,669.50	2.34	0.0191*
X4	26,203.50	5,526.00	4.74	0.0000*
X5	-11,027.13	2,081.52	-5.30	0.0000*
X6	7,600.96	690.00	11.02	0.0000*

*significant in 5%

Table 7. EBLUP with NTL as Auxiliary Variable

Variable	Beta	Standard Error	t-Value	p-value
Intercept	780,457.90	12,457.10	62.65	0.0000*
NTL	54,644.50	2,088.44	26.17	0.0000*

*significant in 5%

Table 8. EBLUP with LST as Auxiliary Variable

Variable	Beta	Standard Error	t-Value	p-value
Intercept	-1,105,965.34	120,375.45	-9.19	0.0000*
LST	68512.57	3931.89	17.42	0.0000*

*significant in 5%

Table 9. EBLUP with NTL and LST (Remote Sensing) as Auxiliary Variables

Variable	Beta	Standard Error	t-Value	p-value
Intercept	350,916.79	128,187.91	2.74	0.0062*
NTL	48,070.87	2,845.95	16.89	0.0000*
LST	14,841.07	4,412.08	3.36	0.0008*

*significant in 5%

Table 10. EBLUP with All Variables as Auxiliary Variables

Variable	beta	Standard Error	t-Value	p-value
Intercept	523,226.24	148,806.97	3.52	0.0004*
X1	54,303.92	20,926.54	2.59	0.0095*
X2	-688.22	286.98	-2.40	0.0165*
X3	5,758.84	2,415.87	2.40	0.0166*
X4	9,980.48	5,277.31	1.89	0.0586
X5	-6,234.48	1,962.77	-3.18	0.0015*
X6	2,785.12	751.74	-3.70	0.0002*
NTL	32,851.68	3,689.42	8.90	0.0000*
LST	11,309.70	4,882.52	2.32	0.0205*

*significant in 5%

As shown in Figure 10, certain subdistricts have a high value of direct estimation per capita expenditure. When compared to direct estimation, the outcomes of estimation utilizing the SAE model are comparatively lower. The maximum estimated value obtained using direct estimation is IDR 4,306,173.42, while the maximum estimated value from other models is no more than IDR 2.6 million. The highest value in the direct estimation is in Lengkong Sub-District, Bandung City. In the SAE model, Lengkong is not the sub-district with the highest expenditure per capita.

According to Table 11, there is a tendency for differences in minimum values from direct estimation and indirect estimation models, SAE. The minimum value in direct estimation is IDR 369,433 in Cijati Sub-District, Cianjur. Whereas the minimum value in the indirect estimation with all possibilities dataset is around IDR 390,000. This demonstrates a substantial disparity in the minimum value. Similarly, direct estimating tends to yield a mean figure for expenditure per capita that is greater than that of indirect estimation. At IDR 990,000, the mean value in indirect estimation tends to be uniform. In contrast, direct estimation yields a mean value of IDR 1,051,612.26. The median values tend towards homogeneity. As illustrated in Figure 7's boxplot, the discrepancy between the median and mean values in direct estimation is due to outliers.

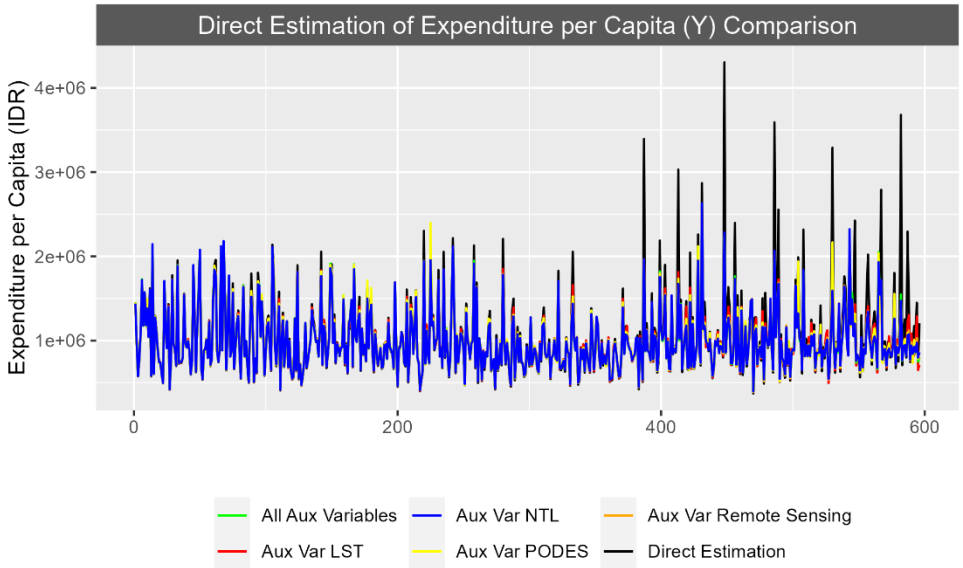


Figure 10. Direct Estimation of Expenditure per Capita Comparison

The SAE model with PODES as the auxiliary variable has a minimum value in Cijati Sub-District (IDR 391,343) Cianjur and a maximum value in Beji Sub-District (IDR 2,400,269). The mean value in this model is IDR 991,298.87. Implementation of remote sensing (NTL) as an auxiliary variable produces the estimation expenditure per capita with the minimum value also in the Cijati Sub-District (IDR 396,827), whereas Sumur Bandung (IDR 2,631,645) in Bandung City is the sub-district with the highest value of expenditure per capita using EBLUP model. The maximum value from the SAE model with Land Surface Temperature (LST) as an auxiliary variable is in East Bekasi Sub-District (IDR 1,977,508), Bekasi City. The maximum estimated value using LST is the minimum value from other models. Meanwhile, the maximum value estimated using a combination of both remote sensing (NTL and LST) and all variables (PODES and remote sensing) is respectively in Sumur Bandung Sub-District. The subdistrict in West Java with the lowest estimated per capita spending is Cijati Sub-District, based on all models employed in the Small Area Estimation.

Table 11. Summary Estimation of Expenditure per Capita in All Model

Indicator	Direct Estimation	Auxiliary Variables Used				
		PODES	**NTL**	LST	Remote Sensing	All Variables
Min	369,432.91	391,343.37	**396,827.26**	389,915.12	397,254.10	397,174.55
Q1	735,262.11	745,966.71	**748,441.52**	742,553.36	747,047.50	747,252.79
Mean	1,051,612.26	991,298.87	**992,055.30**	993,264.59	991,790.79	991,460.83
Median	928,571.89	917,185.09	**906,978.11**	916,912.71	904,545.28	903,064.59
Q3	1,212,535.10	1,162,556.29	**1,137,248.19**	1,178,216.06	1,139,907.94	1,143,204.75
Max	4,306,173.42	2,400,269.65	**2,631,645.32**	1,977,508.37	2,551,022.92	2,341,868.74
Range	3,936,740.51	2,008,926.28	**2,234,818.05**	1,587,593.24	2,153,768.82	1,944,694.19
Std Dev	486,660.60	345,840.83	**356,331.37**	338,007.99	355,409.60	356,292.87

BPS-Statistics Indonesia uses a maximum threshold for the Relative Standard Error (RSE) value that is good to use at 25%. RSE values between 25 and 50 percent are prudent values to use with caution. Furthermore, the resultant estimate lacks reliability since the RSE value exceeds 50%. Figure 11 shows that the RSE from LST model has values that exceed 25%, there are Majalengka (35.17%), Sukahening (30.61%), and Cigugur (28.40%) Sub-Districts. Besides that, Sub-Districts Jatiwaras (25.30%) and Majalengka (25.25%) in PODES as auxiliary variables have RSE values greater than 25%. Majalengka Sub-District also has RSE maximum value in the SAE model with all remote sensing data (25.29%) and all variables dataset (25.03%) as auxiliary variables. The only model that doesn't have an RSE value of more than 25% is the one that uses NTL as an auxiliary variable.

The model containing all variables (10.39%) has the smallest average RSE value among all the proposed models. Followed by the model containing all remote sensing auxiliary variables (10.48%), and then the SAE model with NTL (10.50%) as an auxiliary variable. The minimum RSE value of all models is around 4.7% and the maximum value for all models is around 25%, except for the model with LST as an auxiliary variable (35%).

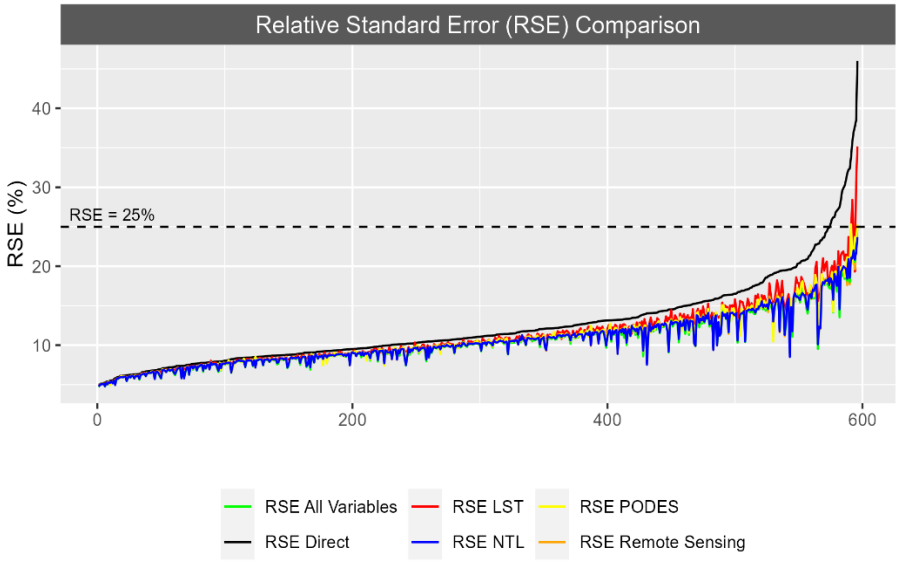


Figure 11. Relative Standard Error (RSE) Comparison

Table 12. Summary of Relative Standard Error (RSE) in All Models (%)

Indicator	Direct Estimation	Auxiliary Variables Used				
		PODES	**NTL**	LST	Remote Sensing	All Variables
Min	5.00	4.77	**4.78**	4.89	4.77	4.71
Q1	8.77	8.23	**8.19**	8.38	8.16	8.12
Mean	12.44	10.70	**10.50**	11.10	10.48	10.39
Median	11.06	10.20	**9.97**	10.41	9.95	9.89
Q3	14.43	12.36	**12.10**	12.88	12.09	11.94
Max	45.99	25.30	**23.68**	35.17	25.29	25.03
Total Sub-District with RSE > 25%	22	2	**0**	3	1	1

Table 13. Normality Test in Error and Random Effect Area

Auxiliary Variables Model EBLUUP	Error	Random Effect Area
PODES	0.0314 (0.4939)	0.0179 (0.9909)
NTL	0.0287 (0.7085)	0.0259 (0.8196)
LST	0.0575 (0.0386)*	0.0339 (0.4992)
Remote Sensing	0.0288 (0.7045)	0.0214 (0.9486)
All Variables	0.0519 (0.0804)	0.0257 (0.8244)

* significant in 5%

In formula (2), the errors ($e_i \sim iid\ N(0, \sigma_e^2)$) and the random effect area ($v_i \sim iid\ N(0, \sigma_v^2)$) of EBLUP are distributed in a normal distribution [5]. Kolmogorov Smirnov is a normality test by comparing one distribution with another distribution, in this case the distribution of error (e_i) and random effect area (v_i) are compared with the normal distribution. The null hypothesis of the normality test is H_0: Error (e_i) or random effect area (v_i) is not normally distributed and the alternative hypothesis is H_a: Error (e_i) or random effect area (v_i) is normally distributed.

Table 13 shows that only the EBLUP model with LST as an auxiliary variable has errors (e_i) that are not spread in a normal distribution. Whereas all random effect areas are normally distributed.

Table 14. Model Evaluation

Auxiliary Variables Model EBLUP	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)
PODES	16,721	16,747
NTL	**16,637**	**16,651**
LST	16,842	16,855
Remote Sensing	16,627	16,645
All Variables	16,591	16,635

Based on Table 14, the model with all variables used has the lowest value of AIC and BIC. This means that the estimation with all variables is better than the others. However, modelling with all variables does not fulfill the principle of parsimony. The model using NTL and all remote sensing variables as auxiliary variable has better AIC and BIC values than the PODES model. Hence, this provides an opportunity that the remote sensing data can be used as auxiliary variables in the estimation expenditure per capita using EBLUP.

Based on the description above, the estimation of expenditure per capita using the EBLUP model with Nighttime Light Intensity (NTL) is the best model. The estimated value using EBLUP with NTL tends to have similar characteristics and represent to the true value. The RSE values is relatively lower than direct estimation and especially in this model all RSE values produced has been in category that can be used for the government.

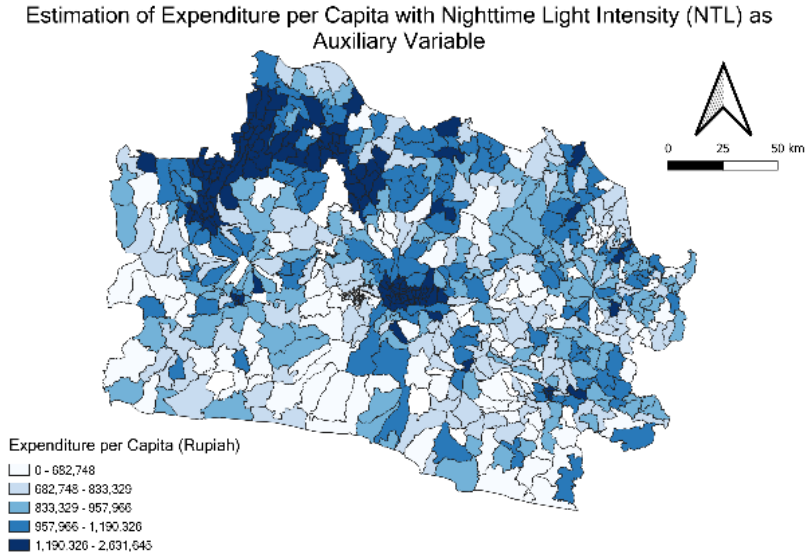


Figure 12. Estimation Expenditure per Capita Best EBLUP model with NTL as auxiliary variables

Figure 12 shows the results best model of estimation expenditure per capita, Empirical Best Linear Unbiased Prediction (EBLUP) with Nighttime Light Intensity (NTL) as an auxiliary variable. The Norwest and the central regions of West Java province have relatively high expenditure per capita. On the other hand, the Southwest region tends to have a low value of expenditure per capita. Therefore, based on the above analysis, estimation using Empirical Best Linear Unbiased Prediction (EBLUP) with Nighttime Light Intensity (NTL) tends to be similar to direct estimation and has the advantage of relatively lower RSE values than direct estimation.

4. Conclusion

The problem of the high Relative Standard Error (RSE) value per capita expenditure at the sub-district level in West Java can be overcome by using the Small Area Estimation (SAE) model. The proposed model can reduce the RSE value of the estimate. The best model obtained is the Empirical Best Linear Unbiased Prediction (EBLUP) model using Nighttime Light Intensity (NTL) as an auxiliary variable. Therefore, there is potential for remote sensing data to be utilized as an auxiliary in the SAE

model. This remote sensing data has the advantage of data that is relatively quick in updating and cheap compared to conventional census data.

The government may consider that remote sensing data can be utilized as auxiliary variables in estimating expenditure per capita in particularly and estimating other economics indicators in general using the Small Area Estimation (SAE) model. Other studies may consider exploring other remote sensing data in developing auxiliary variables in the Small Area Estimation (SAE) model.

Ethics approval

Not required.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable feedback.

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

Muhamad Feriyanto: Conceptualization, Data Collection, Formal Analysis, Writing - Original Draft, Visualization**. Arie Wahyu Wijayanto**: Methodology, Writing - Review & Editing, Supervision**. Ika Yuni Wulansari**: Writing - Review & Editing, Supervision. **Novia Budi Parwanto**: Writing - Review & Editing, Supervision.

References

[1] (BPS) Badan Pusat Statistik, "Konsumsi dan Pengeluaran," Konsumsi dan Pengeluaran [Consumption and Expenditure]. Accessed: Nov. 30, 2023. [Online]. Available: https://www.bps.go.id/subject/5/konsumsi-dan-pengeluaran.html

[2] (BPS) Badan Pusat Statistik, "Ekonomi Indonesia 2018 Tumbuh 5,17 Persen," Ekonomi Indonesia 2018 Tumbuh 5,17 Persen [Indonesia's 2018 Economy grew 5.17 Percent]. Accessed: Nov. 30, 2023. [Online]. Available: https://www.bps.go.id/pressrelease/2019/02/06/1619/ekonomi-indonesia-2018-tumbuh-5-17-persen.html

[3] (BPS) Badan Pusat Statistik, "Statistik Indonesia 2018," Statistik Indonesia 2018 [Statistics Indonesia 2018]. Accessed: Nov. 30, 2023. [Online]. Available: https://www.bps.go.id/publication/2018/07/03/5a963c1ea9b0fed6497d0845/statistik-indonesia-2018.html

[4] Badan Pusat Statistik, *PEDOMAN KEPALA BPS PROVINSI, STATISTISI AHLI MADYA/KOORDINATOR FUNGSI STATISTIK SOSIAL BPS PROVINSI, DAN KEPALA BPS KABUPATEN/KOTA SUSENAS MARET 2022* [*GUIDELINES FOR THE HEAD OF THE*

PROVINCIAL BPS, INTERMEDIATE EXPERT STATISTICIAN/COORDINATOR OF THE SOCIAL STATISTICS FUNCTION OF THE PROVINCIAL BPS, AND THE HEAD OF THE DISTRICT / CITY BPS SUSENAS MARCH 2022]. Badan Pusat Statistik, 2022.

[5] J. N. K. Rao and I. Molina, *Small area estimation*, Second edition. in Wiley series in survey methodology. Hoboken, New Jersey: John Wiley & Sons, Inc, 2015.

[6] D. T. Lindgren, *Land use planning and remote sensing*. in Remote sensing of earth resources and environment, no. 2. Dordrecht ; Boston : Hingham, MA, USA: M. Nijhoff Publishers ; Distributors for the U.S. and Canada, Kluwer Academic Pub, 1985.

[7] S. Hutasavi and D. Chen, "Estimating District-Level Electricity Consumption Using Remotely Sensed Data in Eastern Economic Corridor, Thailand," *Remote Sensing*, vol. 13, no. 22, p. 4654, Nov. 2021, doi: 10.3390/rs13224654.

[8] S. R. Putri, A. W. Wijayanto, and A. D. Sakti, "Developing Relative Spatial Poverty Index Using Integrated Remote Sensing and Geospatial Big Data Approach: A Case Study of East Java, Indonesia," *IJGI*, vol. 11, no. 5, p. 275, Apr. 2022, doi: 10.3390/ijgi11050275.

[9] S. Shi, Y. Ye, and R. Xiao, "Evaluation of Food Security Based on Remote Sensing Data—Taking Egypt as an Example," *Remote Sensing*, vol. 14, no. 12, p. 2876, Jun. 2022, doi: 10.3390/rs14122876.

[10] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, "Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping," *AAAI*, vol. 30, no. 1, Mar. 2016, doi: 10.1609/aaai.v30i1.9906.

[11] J. Yin, Y. Qiu, and B. Zhang, "Identification of Poverty Areas by Remote Sensing and Machine Learning: A Case Study in Guizhou, Southwest China," *IJGI*, vol. 10, no. 1, p. 11, Dec. 2020, doi: 10.3390/ijgi10010011.

[12] S. R. Putri, A. W. Wijayanto, and S. Pramana, "Multi-source satellite imagery and point of interest data for poverty mapping in East Java, Indonesia: Machine learning and deep learning approaches," *Remote Sensing Applications: Society and Environment*, vol. 29, p. 100889, Jan. 2023, doi: 10.1016/j.rsase.2022.100889.

[13] S. Ahmed, "Assessment of urban heat islands and impact of climate change on socioeconomic over Suez Governorate using remote sensing and GIS techniques," *The Egyptian Journal of Remote Sensing and Space Science*, vol. 21, no. 1, pp. 15–25, Apr. 2018, doi: 10.1016/j.ejrs.2017.08.001.

[14] 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 Sensing*, vol. 13, no. 4, p. 766, Feb. 2021, doi: 10.3390/rs13040766.

[15] Y. Sun, S. Wang, X. Zhang, T. O. Chan, and W. Wu, "Estimating local-scale domestic electricity energy consumption using demographic, nighttime light imagery and Twitter data," *Energy*, vol. 226, p. 120351, Jul. 2021, doi: 10.1016/j.energy.2021.120351.

[16] L. Lin, L. Di, C. Zhang, L. Guo, and Y. Di, "Remote Sensing of Urban Poverty and Gentrification," *Remote Sensing*, vol. 13, no. 20, p. 4022, Oct. 2021, doi: 10.3390/rs13204022.

[17] G. E. Battese, R. M. Harter, and W. A. Fuller, "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, vol. 83, no. 401, pp. 28–36, Mar. 1988, doi: 10.1080/01621459.1988.10478561.

[18] R. Singh, D. P. Semwal, A. Rai, and R. S. Chhikara, "Small area estimation of crop yield using remote sensing satellite data," *International Journal of Remote Sensing*, vol. 23, no. 1, pp. 49–56, Jan. 2002, doi: 10.1080/01431160010014756.

[19] G. F. Santoso and S. Muchlisoh, "Estimasi Produktivitas Padi Level Kecamatan di Kabupaten Tulungagung Menggunakan Geoadditive SAE" [Sub-district Level Rice Productivity Estimation in Tulungagung Regency Using Geoadditive SAE], *asks*, vol. 12, no. 3, pp. 23–36, Mar. 2022, doi: 10.34123/jurnalasks.v14i1.385.

[20] N. T. Longford, M. G. Pittau, R. Zelli, and R. Massari, "Poverty and inequality in European regions," *Journal of Applied Statistics*, vol. 39, no. 7, pp. 1557–1576, Jul. 2012, doi: 10.1080/02664763.2012.661705.

[21] T. Siswantining, "GEOINFORMATIKA PADA KASUS AREA KECIL DAN PENERAPANNYA UNTUK MENDETEKSI KANTONG- KANTONG KEMISKINAN DI JEMBER" [GEOINFORMATICS IN THE CASE OF SMALL AREAS AND ITS

APPLICATION TO DETECT POCKETS OF POVERTY IN JEMBER], *Institute Pertanian Bogor Pasca Sarjana Statistika*, 2013.

[22] N. Istiana, "SMALL AREA ESTIMATION WITH EXCESS ZERO (CASE STUDY: INFANT MORTALITY RATE IN JAVA ISLAND)," *asks*, vol. 13, no. 1, pp. 25–34, Sep. 2021, doi: 10.34123/jurnalasks.v13i1.270.

[23] S. Guha and H. Chandra, "Measuring disaggregate level food insecurity via multivariate small area modelling: evidence from rural districts of Uttar Pradesh, India," *Food Sec.*, vol. 13, no. 3, pp. 597–615, Jun. 2021, doi: 10.1007/s12571-021-01143-1.

[24] Md. J. Hossain, S. Das, H. Chandra, and M. A. Islam, "Disaggregate level estimates and spatial mapping of food insecurity in Bangladesh by linking survey and census data," *PLoS ONE*, vol. 15, no. 4, p. e0230906, Apr. 2020, doi: 10.1371/journal.pone.0230906.

[25] P. A. Kaban, B. I. Nasution, R. E. Caraka, and R. Kurniawan, "Implementing night light data as auxiliary variable of small area estimation," *Communications in Statistics - Theory and Methods*, pp. 1–18, May 2022, doi: 10.1080/03610926.2022.2077963.

[26] A. Ubaidillah, K. A. Notodiputro, A. Kurnia, and I. W. Mangku, "Multivariate Fay-Herriot models for small area estimation with application to household consumption per capita expenditure in Indonesia," *Journal of Applied Statistics*, vol. 46, no. 15, pp. 2845–2861, Nov. 2019, doi: 10.1080/02664763.2019.1615420.

[27] A. Asian Development Bank, "How Nighttime Lights Help Us Study Development Indicators | Asian Development Blog," How Nighttime Lights Help Us Study Development Indicators. Accessed: Nov. 04, 2024. [Online]. Available: https://blogs.adb.org/blog/how-nighttime-lights-help-us-study-development-indicators

[28] R. Beyer, J. Yao, and Y. Hu, "Measuring Quarterly Economic Growth from Outer Space," *IMF Working Papers*, vol. 2022, no. 109, p. 1, Jun. 2022, doi: 10.5089/9798400211553.001.

[29] S. Yu *et al.*, "Land Surface Temperature Changes in Different Urbanization Increments in China since 2000," *Land*, vol. 13, no. 4, p. 417, Mar. 2024, doi: 10.3390/land13040417.

[30] M. H. Kutner, Ed., *Applied linear statistical models*, 5th ed. in The McGraw-Hill/Irwin series operations and decision sciences. Boston: McGraw-Hill Irwin, 2005.

[31] Q. Zhang and K. C. Seto, "Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data," *Remote Sensing of Environment*, vol. 115, no. 9, pp. 2320–2329, Sep. 2011, doi: 10.1016/j.rse.2011.04.032.

[32] S. P. Subash, R. R. Kumar, and K. S. Aditya, "Satellite data and machine learning tools for predicting poverty in rural India," *Agri. Econ. Rese. Revi.*, vol. 31, no. 2, p. 231, 2018, doi: 10.5958/0974-0279.2018.00040.X.

[33] Md. Bodruddoza Mia, Rahul Bhattacharya, and ASM Woobaidullah, "Correlation and Monitoring of Land Surface Temperature, Urban Heat Island with Land use-land cover of Dhaka City using Satellite imageries," *International Journal of Research in Geography*, vol. 3, no. 4, 2017, doi: 10.20431/2454-8685.0304002.

DOI: https://doi.org/10.34123/jurnalasks.v15i2.442

POLITEKNIK STATISTIKA STIS

Available online at https://jurnal.stis.ac.id/

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 1, June 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

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, left alignment))

First Author Name[1*], Second Author Name[2] , Third Author Name[3] (TNR font, size 13 pt, exactly spacing at at 15 pt and 8 pt spacing for the next heading.)

[1]First Affiliation, City, Country, [2,3]Author Affiliation, City, Country [2]Second 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://ieee-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 1.12 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].

a  b 

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.

$$\rho = \frac{\vec{E}}{J_C(T=\text{const.})\cdot\left(P\cdot\left(\frac{\vec{E}}{E_C}\right)^m+(1-P)\right)} \tag{1}$$

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

Zhang San: Conceptualization, Methodology, Software. **Priya Singh**: Data curation, Writing- Original draft preparation. **Wang Wu:** Visualization, Investigation. **Jan Jansen:** Supervision. **Ajay Kumar:** Software, Validation. **Sun Qi:** Writing- Reviewing and Editing. (Example)

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.