JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 16, NO 1, JUNE 2024 P-ISSN 2086 – 4132 E-ISSN 2615 – 1367

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika & Komputasi Statistik

"Journal of Applications of Statistics & Statistical Computing (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

.

"Journal of Applications of Statistics & Statistical Computing (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, June 2024

Editor in Chief,

Rani Nooraeni

JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 16, NO 1, JUNE 2024

Contents

DOI: https://doi.org/10.34123/jurnalasks.v16i1.719

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

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 1 Juni Tahun 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

Modelling The Stunting Prevalence Rate in Indonesia Using Multi-Predictor Truncated Spline Nonparametric Regression

Pressylia Aluisina Putri Widyangga[1], Alda Fuadiyah Suryono[2] , Maria Setya Dewanti[3], Dewanti Ardi Kurniawan[4*]

[1,2,3,4]*Fakultas Sains dan Teknologi, Universitas Airlangga, Surabaya, Indonesia*

**Corresponding Author: E-mail address:* ardi-k@fst.unair.ac.id

ARTICLE INFO

Article history:
Received 08 January, 2024
Revised 15 May, 2024
Accepted 17 May, 2024
Published 30 June, 2024

Keywords:
Mother, Nutrition, SDGs, Spline, Stunting

Abstract

Introduction/Main Objectives: Stunting is the impaired growth and development that children experience from poor nutrition, repeated infection, and inadequate psychosocial stimulation. **Background Problems:** Based on data from the National Nutrition Status Survey (SSGI) in 2022, the prevalence of stunting in Indonesia was 21.6%, which is still above the WHO standard of below 20%. **Novelty:** This study was conducted with the aim of analysing the factors that influence the stunting prevalence rate in Indonesia using multi-predictor truncated spline nonparametric regression. **Research Methods:** The research data is secondary data taken from Health Statistics 2022 with response variables in the form of stunting prevalence. **Finding Result:** Based on the analysis, the best model to model the stunting prevalence rate is a multi-predictor truncated spline with three knots. In addition, it was found that four predictor variables which are the percentage of infants under 6 months old receiving exclusive breastfeeding, the average age of a mother's first pregnancy, the percentage of married women aged 15-49 using contraception, and the percentage of mothers who gave birth to a live child in the past two years and initiated early breastfeeding had a significant effect simultaneously and partially on the stunting prevalence rate in Indonesia.

1. Introduction

The health problem of stunted toddlers in Indonesia remains a serious issue that needs to be addressed by the government and society [1]. Based on data from the National Nutritional Status Survey in 2022 (SSGI), the prevalence of stunting in Indonesia is 21.6%. In 2022, Indonesia became the 10th highest contributor to stunting rates in the Southeast Asia region, based on data from the Asian Development Bank. Even though this figure has decreased from 24.4% in 2021, it is still above 20%, the standard rate according to the WHO. The prevalence of stunting in Indonesia has indeed decreased, but it is still far from the National Medium-Term Development Plan (RPJMN) 2024's target of a 14% reduction [2].

Stunting is a chronic malnutrition disease that occurs in children, often caused by long-term malnutrition. This condition has a serious impact on children's physical, cognitive, and developmental growth and may hinder their ability to achieve optimal productivity as adults [3]. According to World Health Organization (WHO) standards, toddlers are considered stunted if their height is less than minus two standard deviations compared to the average child's growth.

Stunting in toddlers is a chronic nutritional problem caused by many factors, including the mother's nutritional status during pregnancy, the family's socio-economic conditions, the child's health status, and child malnutrition. In such conditions, stunted toddlers will have difficulty achieving optimal physical and cognitive development [4]. Thus, special handling is needed in accordance with the targets of the Indonesian government [5]. Handling stunting is also an important effort in achieving the second point of the Sustainable Development Goals (SDGs), namely "End hunger, achieve food security and improved nutrition, and promote sustainable agriculture" [6].

Many studies on stunting in toddlers have been carried out, including the one using the Geographically Weighted Regression method and Multivariate Adaptive Regression Splines by Alif Yuanita [7]. Based on the study, it was concluded that the factors that influence the prevalence of stunting include the percentage of newborns receiving Early Breastfeeding Initiation (IMD), the percentage of infants receiving exclusive breastfeeding, the percentage of toddlers getting vitamin A, the percentage of toddlers having KMS or KIA books, the percentage of toddlers being weighed four times or more in the last six months, and the percentage of households that use clean water. Apart from that, another research was conducted by Marisa Rifada [8] regarding stunting modelling in toddlers using parametric and non-parametric ordinal logistic regression models [8]. Through a parametric ordinal logistic regression model approach, the study concluded that three variables, namely birth length, mother's height, and health services had a significant effect on stunting in toddlers, with an accuracy rate of 73.98%.

In "The Elements of Statistical Learning" by Hastie, Tibshirani, and Friedman (2009), nonparametric regression is described as a method for predicting the relationship between a dependent variable and an independent variable without assuming a specific functional form. It emphasises the flexibility of nonparametric regression methods that allow the data to determine the shape of the relationships rather than imposing a predetermined model structure [9]. Based on several previous studies that have been carried out, other research is needed to determine the factors that influence the prevalence of stunting in Indonesia based on data obtained from the publication of the Central Bureau of Statistics, namely Statistics Indonesia using a multi-predictor truncated spline nonparametric regression model [10]. This research was carried out using a non-parametric regression method because the data on the prevalence of stunting and the factors suspected to influence it did not form a particular pattern. The function that can be used in nonparametric regression in this research is spline. The spline function is part of a segmental polynomial, so it shows high flexibility and can adapt to the local characteristics of the data [11].

The use of nonparametric spline regression is very effective in handling data samples that show variations in behaviour within certain sub-intervals and has excellent generalisation capabilities in complex and complicated statistical models. [12]. Splines have an advantage in their ability to interpret data patterns according to their movement [13]. Moreover, this modelling uses the truncated concept to limit the response variable, namely the prevalence of stunting to a certain range appropriate to the context. This can help avoid bias that may arise due to outliers or irrelevant data [14].

Based on this description, research is needed to model the prevalence of stunting in Indonesia by taking data by province in 2022 [15]. It is hoped that the results of this modelling will provide a better understanding of the factors that contribute to stunting in Indonesia in 2022 as well as the grouping of provinces in Indonesia based on the best model knots. The modelling in this research is also expected to be a solution in handling stunting in order to achieve the second point of SDGs sustainable development goals, namely to end hunger, achieve food security, and improve the nutrition of Indonesian people.

2. Material and Methods

2.1. Type of Research

This research regarding modelling the prevalence of stunting in Indonesia utilises quantitative research methods. This method emphasises research on certain samples, analysis of quantitative or statistical data, and testing of predetermined hypotheses.

2.2. Location and Time Research

This research regarding modelling the prevalence of stunting in Indonesia was carried out in approximately three months from September to November 2023.

2.3. Data Collection Sources and Strategies

The data used in this research is secondary data on the prevalence of stunting in Indonesia in 2022 which comes from the website katadata.id and the data on factors that influence the prevalence of stunting which come from Health Statistics 2022 and the website bps.go.id.

2.4. Research Variables

The variables used in modelling the prevalence of stunting in Indonesia consist of response and predictor variables. The response variable used is the prevalence of stunting based on provinces in Indonesia in 2022. Meanwhile, the predictor variables used are contributing factors in the prevalence of stunting based on previous research obtained from the Health Statistics 2022. Details of the variables used in the research are aluented in Table 1.

Table 1. Research Variables

Variable	Description	Unit	Measure
Y	The Stunting Prevalence Rate in Indonesia	Percent	Ratio
X_1	Percentage of Infants Under 6 Months of Age Who Received Exclusive Breastfeeding	Percent	Ratio
X_2	Average Age of Mother's First Pregnancy	Year	Ratio
X_3	Percentage of Married Women Aged 15-49 Years Old Currently Using Contraception Methods	Percent	Ratio
X_4	Percentage of Mothers Who Gave Birth to Live-Born Children in the Past Two Years and Initiated Early Breastfeeding	Percent	Ratio

2.5. Operational Definition

Operational definitions of response variable and predictors used in this study are presented in Table 2.

Table 2. Operational Definition of Variable

Variable	Operational Definition
The Prevalence of Stunting	Children aged 0-59 months in the nutritional status category based on the body length index for age (PB/U) have a z-score of less than -2SD.
Percentage of Infants Under 6 Months of Age Who Received Exclusive Breastfeeding	The percentage of infants 0-6 months who are exclusively breastfed is calculated by accumulating the numerator (infants 0-6 months who are exclusively breastfed) and the denominator (the number of infants 0-6 months recorded in the breast-feeding registration register).
Average Age of Mother's First Pregnancy	The average age of a mother's first pregnancy at a certain age.
Percentage of Married Women Aged 15-49 Years Old Currently Using Contraception Methods	The proportion or percentage of the married female population of childbearing age (15-49 years) who use contraception methods.
Percentage of Mothers Who Gave Birth to Live-Born Children (ALH) in the Past Two Years and Initiated Early Breastfeeding	Early initiation of breastfeeding is the process of a baby breastfeeding immediately after birth, where the baby is allowed to look for its mother's nipple on its own.
Percentage of Children Aged 12-23 Months Who Received Complete Basic Immunisation	The complete basic immunisation that children aged 12-23 months must receive consists of Hepatitis B immunisation

Variable	Operational Definition
	four times, BCG once, Polio four times, DPT-HB three times, and Measles once.

2.6. Research Procedures

To model the prevalence rate of stunting in Indonesia and analyse the factors that influence it, it is necessary to carry out a data analysis process through the description and flowchart presented in Figure 1.

1. Look for data regarding the stunting prevalence rate by province in Indonesia in 2022 as well as the factors that influence it.
2. Tabulate raw data in which the response variable is the stunting prevalence rate in Indonesia in 2022 and predictor variables are the factors that affect the stunting prevalence rate.
3. Make a scatter plot to ensure that the data used does not form a particular functional relationship pattern.
4. Try several orders and knot points to find the model with the highest R^2 and minimum Generalised Cross Validation (GCV) value.
5. Choose the best model that has the highest R^2 value and minimum GCV.
6. Interpret the best model that has been obtained.
7. Create data visualisations using maps to map the prevalence of stunting in Indonesia based on the best model that has been obtained.
8. Draw conclusions.
9. Compile research reports and articles.

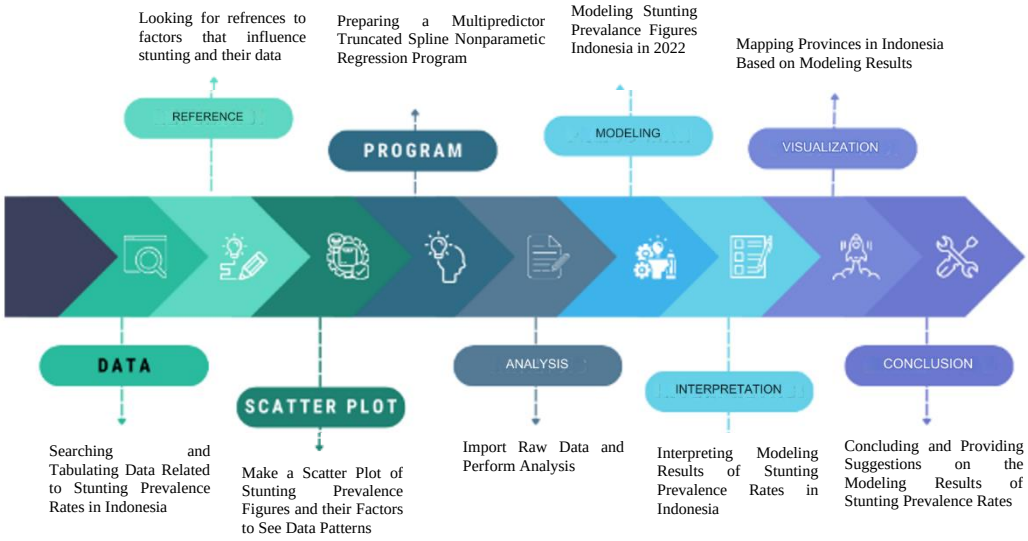


Figure. 1. Research Flow Diagram

3. Results and Discussion

The prevalence rate of stunting in Indonesia in 2022 is shown in Figure 2:

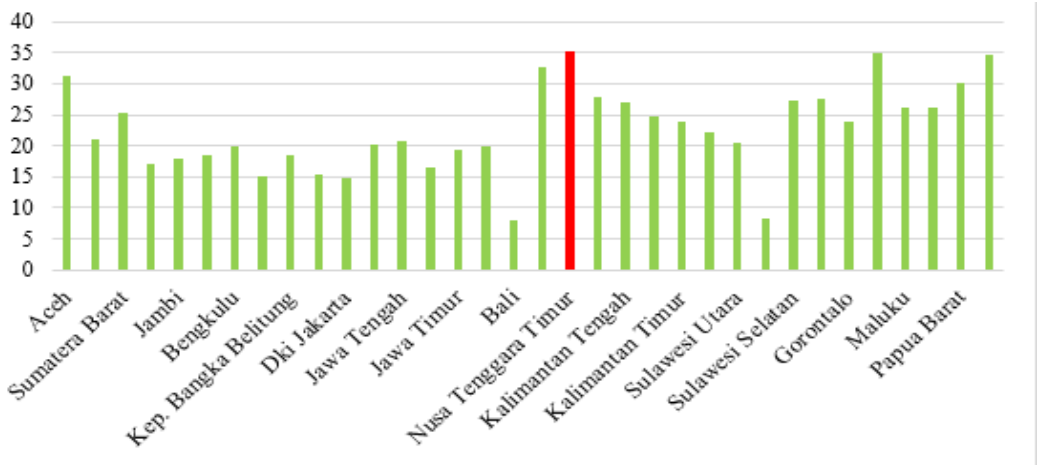


Figure. 2. Stunting Prevalence in Every Province of Indonesia

The province with the lowest stunting prevalence rate is Bali at 8% and the province with the highest stunting prevalence rate is East Nusa Tenggara at 35.3%.

3.1. Regression Model with One Knot Point

The truncated spline nonparametric regression model formed with one knot point and five predictors to model the stunting prevalence rate in Indonesia is as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)_+^1 + \hat{\beta}_3 x_2 + \hat{\beta}_4 (x_2 - K_2)_+^1 + \hat{\beta}_5 x_3 + \hat{\beta}_6 (x_3 - K_3)_+^1 + \hat{\beta}_7 x_4 + \hat{\beta}_8 (x_4 - K_4)_+^1 + \hat{\beta}_9 x_5 + \hat{\beta}_{10} (x_5 - K_5)_+^1 \quad (1)$$

Based on equation (1), the GCV and R^2 value obtained for each knot point are presented in Table 3.

Table 3. Measures of Goodness of Fit in Nonparametric Regression Model with One Knot

GCV	R^2	X_1	X_2	X_3	X_4	X_5
38.24	62.21	71.70	22.63	53.89	66.02	65.10
38.37	62.08	72.24	22.69	54.83	66.53	66.36
38.57	61.88	78.63	23.39	66.05	72.58	81.39
38.77	61.69	71.17	22.58	52.96	65.52	63.85
38.80	61.66	73.30	22.81	56.70	67.54	68.86

Based on Table 3, the minimum GCV value is 38.24 with knot points for each predictor variable respectively being 71.7033, 22.6345, 53.8935, 66.0227, 65.1033.

3.2. Equation of Two Knots

The truncated spline nonparametric regression model formed with two knot points and five predictor variables to model the stunting prevalence rate in Indonesia is as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)_+^1 + \hat{\beta}_3 (x_1 - K_2)_+^1 + \hat{\beta}_4 x_2 + \hat{\beta}_5 (x_2 - K_3)_+^1 + \hat{\beta}_6 (x_2 - K_4)_+^1 + \hat{\beta}_7 x_3 + \hat{\beta}_8 (x_3 - K_5)_+^1 + \hat{\beta}_9 (x_3 - K_6)_+^1 + \hat{\beta}_{10} x_4 + \hat{\beta}_{11} (x_4 - K_7)_+^1 + \hat{\beta}_{12} (x_4 - K_8)_+^1 + \hat{\beta}_{13} x_5 + \hat{\beta}_{14} (x_5 - K_9)_+^1 + \hat{\beta}_{15} (x_5 - K_{10})_+^1 \quad (2)$$

Based on equation (2), the GCV and R^2 value obtained for each knot point are presented in Table 4.

Table 4. Measures of Goodness of Fit in Nonparametric Regression Model with Two Knots

GCV	R^2	X_1	X_2	X_3	X_4	X_5
37.29	48.32	53.60	20.65	22.10	48.87	22.52
		79.69	23.51	67.92	73.59	83.89
38.24	62.21	53.60	20.65	22.10	48.87	22.52
		71.70	22.63	53.89	66.02	65.10
38.24	62.21	71.70	22.63	53.89	66.02	65.10
		79.69	23.51	67.92	73.59	83.89
38.37	62.08	53.60	20.65	22.10	48.87	22.52
		72.24	22.69	54.83	66.53	66.36
38.37	62.08	72.24	22.69	54.83	66.53	66.36
		79.69	23.51	67.92	73.59	83.89

Based on Table 4, the minimum GCV value obtained is 37.29 with knot points for each predictor variable being 53.6 and 79.69 for variable X_1, 20.65 and 23.51 for variable X_2, 22.1 and 67.92 for variable X_5.

3.3. Three Point Knot Equation

The truncated spline nonparametric regression model formed with three knot points and five predictors to model the stunting prevalence rate in Indonesia is as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 (x_1 - K_1)_+^1 + \hat{\beta}_3 (x_1 - K_2)_+^1 + \hat{\beta}_4 (x_1 - K_3)_+^1 + \hat{\beta}_5 x_2 + \hat{\beta}_6 (x_2 - K_4)_+^1 + \hat{\beta}_7 (x_2 - K_5)_+^1 + \hat{\beta}_8 (x_2 - K_6)_+^1 + \hat{\beta}_9 x_3 + \hat{\beta}_{10} (x_3 - K_7)_+^1 + \hat{\beta}_{11} (x_3 - K_8)_+^1 + \hat{\beta}_{12} (x_3 - K_9)_+^1 + \hat{\beta}_{13} x_4 + \hat{\beta}_{14} (x_4 - K_{10})_+^1 + \hat{\beta}_{15} (x_4 - K_{11})_+^1 + \hat{\beta}_{16} (x_4 - K_{12})_+^1 + \hat{\beta}_{17} x_5 + \hat{\beta}_{18} (x_5 - K_{13})_+^1 + \hat{\beta}_{19} (x_5 - K_{14})_+^1 + \hat{\beta}_{20} (x_5 - K_{15})_+^1 \quad (3)$$

Based on equation (3), the GCV value and the R^2 value obtained for each knot point are presented in Table 5.

Table 5. Measures of Goodness of Fit in Nonparametric Regression Model with Three Knot Points

GCV	R^2	X_1	X_2	X_3	X_4	X_5
36.04	88.62	61.59	21.53	36.13	56.44	41.31
		65.31	21.93	42.67	59.97	50.07
		78.63	23.39	66.05	72.58	81.39
36.19	88.57	68.51	22.28	48.28	63.00	57.59
		72.77	22.75	55.76	67.03	67.61
		73.83	22.87	57.63	68.04	70.11
36.26	86.72	64.78	21.88	41.74	59.46	48.82
		66.38	22.05	44.54	60.98	52.58
		66.91	22.11	45.48	61.48	53.83
37.26	88.24	67.98	22.23	47.35	62.49	56.34
		72.77	22.75	55.76	67.03	67.61
		73.83	22.87	57.63	68.04	70.11
37.42	86.30	65.31	21.93	42.67	59.97	50.07
		66.38	22.05	44.54	60.98	52.58
		66.91	22.11	45.48	61.48	53.83

Based on Table 5, the minimum GCV value is 36.04 and the R^2 is 88.62.

3.4. Selection of the Best Model

Next, it is necessary to compare the GCV values to obtain the best model for modelling the stunting prevalence rate in Indonesia which is presented in Table 6.

Table 6. GCV Values for Models with Multiple Knot Points

Number of Knot Points	GCV
One Knot Point	38.24
Two Knot Points	37.29
Three Knot Points	36.04

Based on Table 6, it was found that the best model for modelling the stunting prevalence rate in Indonesia is a model with three knot points. The best model equation for modelling the stunting prevalence rate in Indonesia is presented in the following equation.

$$\begin{aligned}\hat{y} = {} & 353.7738 + 1.03x_1 - 3.821(x_1 - 61.5867)_+^1 + 3.308(x_1 - 65.3139)_+^1 + 9.076(x_1 - \\ & 78.6251)_+^1 - 6.072x_2 + 33.404(x_2 - 21.5255)_+^1 - 16.632(x_2 - 21.9341)_+^1 - \\ & 285.91(x_2 - 23.3933)_+^1 - 2.989x_3 + 8.789(x_3 - 36.1265)_+^1 - 6.344(x_3 - \\ & 42.6722)_+^1 + 5.294(x_3 - 66.0498)_+^1 - 2.967x_4 + 6.432(x_4 - 56.4373)_+^1 - \\ & 4.184(x_4 - 59.9688)_+^1 + 61.464(x_4 - 72.581)_+^1 - 1.149x_5 + 3.963(x_5 - \\ & 41.3067)_+^1 - 2.877(x_5 - 50.0739)_+^1 - 34.55(x_5 - 81.3851)_+^1\end{aligned} \quad (4)$$

3.5. Simultaneous Parameter Testing

Simultaneous tests are used to determine whether the regression model parameters are significant or not. The following is a hypothesis from simultaneous parameter testing.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{20} = 0$

H_1: There is at least one $\beta_j \neq 0, j = 1, 2, \ldots, 20$

Table 7. GCV Values for Models with Multiple Knot Points

Source	df	*Sum Square*	*Mean Square*	F	*P-Value*
Regression	20	1395.2	69.8	5.06	0.002
Error	13	179.1	13.8		
Total	33	1574.3			

The p-value obtained was 0.002216 which was less than the significance level α (0.05). Thus, the decision is to reject H_0, meaning that there is at least one parameter that is not equal to zero or there is at least one predictor variable that has a significant effect on the response variable.

3.6. Partial Parameter Testing

Individual or partial significance testing of model parameters was carried out to find out which parameters have a significant effect. The hypothesis used in this test is as follows.

$H_0 : \beta_j = 0$

$H_1 : \beta_j \neq 0, j = 1, 2, \ldots, 20$

Table 8. Result of Parameter Testing

Variable	Parameter	*P-Value*	Significance
	β_0	0.025	Significant
X_1	β_1	0.128	Not Significant
	β_2	0.025	Significant
	β_3	0.017	Significant
	β_4	0.057	Not Significant
X_2	β_5	0.259	Not Significant
	β_6	0.071	Significant

Table 8. Result of Parameter Testing

Variable	Parameter	*P-Value*	Significance
	β_7	0.403	Not Significant
	β_8	0.014	Significant
X_3	β_9	0.003	Significant
	β_{10}	0.004	Significant
	β_{11}	0.008	Significant
	β_{12}	0.048	Significant
X_4	β_{13}	0.073	Not Significant
	β_{14}	0.017	Significant
	β_{15}	0.012	Significant
	β_{16}	0.011	Significant
X_5	β_{17}	0.015	Significant
	β_{18}	0.002	Significant
	β_{19}	0.004	Significant
	β_{20}	0.002	Significant

Using α=0.05, all variables have a significant effect on the model because the p-value is less than α.

3.7. Best Model Equation

If other variables except X_1 are held constant, the influence of the percentage of infants aged under 6 months of age who received exclusive breastfeeding (X_1) on the prevalence of stunting in Indonesia in 2022 is presented through the following equation.

$$\hat{y} = 353.7738 + 1.03x_1 - 3.821(x_1 - 61.5867)_+^1 + 3.308(x_1 - 65.3139)_+^1 + 9.076(x_1 - 78.6251)_+^1 \quad (5)$$

$$\hat{y} = \{353.7738 + 1.03x_1 \ , \ x_1 < 61.5867 \ 589.0966 - 2.79x_1 \ , \ 61.5867 \leq x_1 < 65.3139 \ 373.0382 + 0.52x_1 \ , \ 65.3139 \leq x_1 < 78.6251 \ -340.5632 + 9.59x_1 \ , \ x_1 \geq 78.6251$$

The mapping of provinces in Indonesia based on the percentage of infants under 6 months of age who received exclusive breastfeeding with the best knot points is as follows.

Figure. 3. The Mapping of Provinces based on X_1

Based on Figure 3, provinces that have a percentage of infants under 6 months of age who received exclusive breastfeeding below 61.6% consist of Maluku, North Sumatra, Central Kalimantan, and Gorontalo. Meanwhile, provinces that have a percentage of infants under 6 months of age who received exclusive breastfeeding above 78.6% consist of the provinces of North Kalimantan, Central Java and West Nusa Tenggara. The higher the percentage of infants under 6 months of age who are exclusively breastfed, the smaller the chance of the baby being stunted. On the other hand, the lower the percentage of infants under 6 months of age who are exclusively breastfed, the higher the stunting prevalence rate in that province.

If other variables except X_2 are held constant, the effect of the average age of a mother's first pregnancy (X_2) on the prevalence of stunting in Indonesia in 2022 is presented through the following equation.

$$\hat{y} = 353.7738 - 6.072x_2 + 33.404(x_2 - 21.5255)_+^1 - 16.632(x_2 - 21.9341)_+^1 - 285.91(x_2 - 23.3933)_+^1 \quad (6)$$

$$\hat{y} = \{353.7738 - 6.072x_2 \ , \ x_2 < 21.5255 \ -365.264 + 27.332x_2 \ , \ 21.5255 \leq x_2 < 21.9341 \ -0.4561 + 10.7x_2 \ , \ 21.9341 \leq x_2 \ll 23.3933 \ 6687.9224 - 275.21x_2 \ , \ x_2 \geq 23.3933$$

If an area has an average age of a mother's first pregnancy less than 21.5255, then a one-year increase in the average age of a mother's first pregnancy will cause a reduction in the prevalence of stunting by 6,072 units. If an area has an average age of a mother's first pregnancy in the interval 21.5255 to 21.9341, then a one-year increase in the average age of a mother's first pregnancy will cause an increase in the prevalence of stunting by 27,332 units. If an area has an average age of a mother's first pregnancy in the interval 21.9341 to 23.3933, then a one-year increase in the average age of a mother's first pregnancy will cause an increase in the prevalence of stunting by 10.7 units. If an area has an average age of a mother's first pregnancy of more than 23.3933, then a one-year increase in the average age of a mother's first pregnancy will cause a reduction in the prevalence of stunting by 275.21 units.

The mapping of provinces in Indonesia based on the average age of a mother's first pregnancy with the best knot points is presented in Figure 4.

Figure. 4. The Mapping of Provinces based on X_2

Based on Figure 4, the provinces that have an average age of a mother's first pregnancy for mothers under 21.5 years consist of the provinces of Central Kalimantan, Gorontalo, West Sulawesi, Jambi, South Kalimantan, Central Sulawesi, Bengkulu, Bangka Belitung Islands, West Kalimantan, West Nusa Tenggara, South Sumatra, West Java, Lampung, Southeast Sulawesi, North Maluku, Papua, East Java, Central Java and North Sulawesi. Meanwhile, the provinces that have an average age of first pregnancy for mothers above 23.4 years consist of the provinces of Jakarta and Riau Islands. Based on modelling results, the average gestational age of mothers who are not too old or still of childbearing age will reduce the prevalence rate of stunting in an area.

If other variables except X_3 are held constant, the effect of the percentage of married women aged 15-49 years who use contraception methods (X_3) on the prevalence of stunting in Indonesia in 2022 is presented through the following equation

$$\hat{y} = 353.7738 - 2.989x_3 + 8.789(x_3 - 36.1265)_+^1 - 6.344(x_3 - 42.6722)_+^1 + 5.294(x_3 - 66.0498)_+^1 \quad (7)$$
$$\hat{y} = \{353.7738 - 2.989x_3 \ , \ x_3 < 36.1265 \ 36.258 + 5.8x_3 \ , \ 36.1265 \leq x_3 < 42.6722 \ 306.9704 - 0.544x_3 \ , \ 42.6722 \leq x_3 < 66.0498 \ -42.6972 + 4.75x_3 \ , \ x_3 \geq 66.0498$$

The mapping of provinces in Indonesia based on the percentage of married women aged 15-49 who are currently using/wearing contraception methods with the best knot points is as follows.

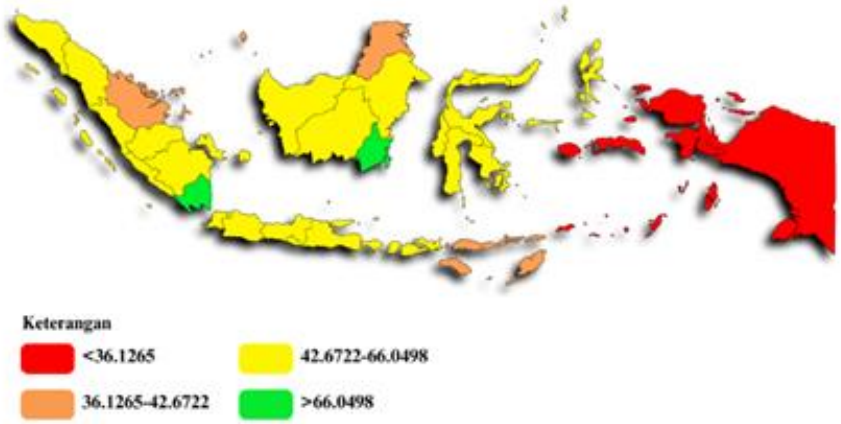


Figure. 5. The Mapping of Provinces based on X_3

Based on Figure 5, provinces that have a percentage of married women aged 15-49 years who are currently using/wearing contraception/birth control methods below 36.1% consist of the provinces of Papua, West Papua and Maluku. Meanwhile, provinces with a percentage of married women aged 15-49 years who are currently using/wearing contraception methods above 66.1% consist of Lampung and South Kalimantan provinces. The higher the percentage of women aged 15-49 years and married who are currently using/wearing contraception methods, the smaller the chance of stunting occurring in toddlers in an area. On the other hand, the lower the percentage of married women aged 15-49 years who are using/wearing contraception/birth control methods, the higher the stunting prevalence rate in an area.

If other variables except X_4 are held constant, the effect of the percentage of mothers who gave birth to live-born children in the past two years and initiated early breastfeeding (X_4) on the prevalence of stunting in Indonesia in 2022 is presented through the following equation.

$$\hat{y} = 353.7738 - 2.967x_4 + 6.432(x_4 - 56.4373)_+^1 - 4.184(x_4 - 59.9688)_+^1 + 61.464(x_4 - 72.581)_+^1 \quad (8)$$
$$\hat{y} = \{353.7738 - 2.967x_4 \ , \ x_4 < 56.4373 \ -9.231 + 3.465x_4 \ , \ 56.4373 \leq x_4 < 59.9688 \ 241.679 - 0.719x_4 \ , \ 59.9688 \leq x_4 < 72.581 \ -4219.44 + 60.745x_4 \ , \ x_4 \geq 72.581$$

The mapping of provinces in Indonesia based on the percentage of mothers who gave birth to live-born children (ALH) in the past two years and initiated early breastfeeding with the best knot points is as follows.

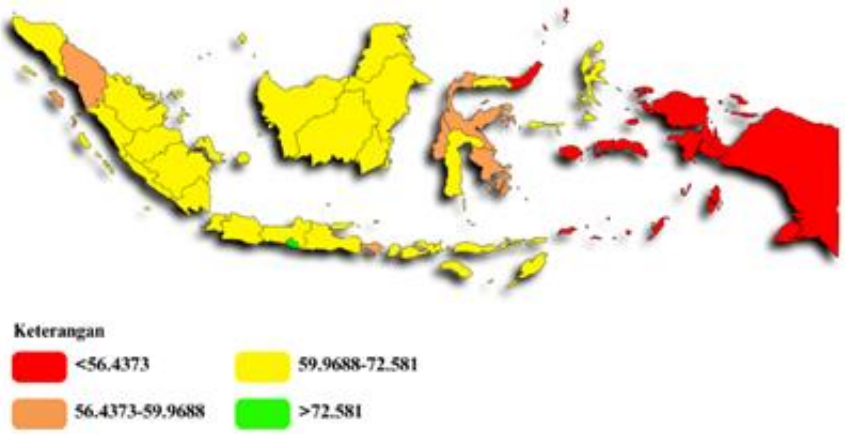


Figure. 6. The Mapping of Provinces based on X_4

Based on Figure 6, the provinces that have a percentage of mothers who gave birth to live-born children (ALH) in the last two years and initiated early breastfeeding below 56.4% consist of the provinces of Maluku, Papua, West Papua and North Sulawesi. Meanwhile, the province that has a percentage of mothers who gave birth to live-born children (ALH) in the last two years and initiated early breastfeeding above 72.6% is the province of Yogyakarta. The higher the percentage of mothers who gave birth to live-born children (ALH) and initiated early breastfeeding, the lower the risk of a child being stunted. On the other hand, the lower the percentage of mothers who gave birth to live-born children and initiated early breastfeeding, the higher the stunting prevalence rate in an area.

If other variables except X_5 are held constant, the effect of the percentage of children aged 12-23 months who receive complete basic immunisation (X_5) on the prevalence of stunting in Indonesia in 2022 is presented through the following equation.

$$\hat{y} = 353.7738 - 1.149x_5 + 3.963(x_5 - 41.3067)_+^1 - 2.877(x_5 - 50.0739)_+^1 - 34.55(x_5 - 81.3851)_+^1 \quad (9)$$

$$\hat{y} = \{353.7738 - 1.149x_5 \ , \ x_5 < 41.3067 \ 190.075 + 2.814x_5 \ , \ 41.3067 \leq x_5 < 50.0739 \ 334.138 - 0.063x_5 \ , \ 50.0739 \leq x_5 < 81.3851 \ 3145.993 - 34.613x_5 \ , \ x_5 \geq 81.3851$$

The mapping of provinces in Indonesia based on the percentage of children aged 12-23 months who receive complete basic immunisation with the best knot points is as follows.

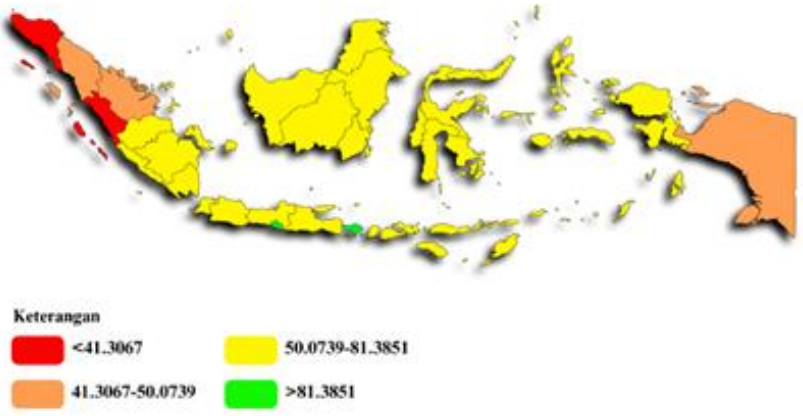


Figure. 7. The Mapping of Provinces based on X_5

Based on Figure 7, provinces that have a percentage of children aged 12-23 months who receive complete basic immunisation below 41.3% are Aceh and West Sumatra. Meanwhile, the provinces that have a percentage of children aged 12-23 months who receive complete basic immunisation above 81.4% consist of the provinces of Bali and Yogyakarta. The higher the percentage of children who receive complete basic immunisation, the smaller the chance of children suffering from stunting. Conversely, the lower the percentage of children who receive complete basic immunisation, the higher the risk of the child experiencing stunting.

3.8. Solutions to Reduce Stunting Prevalence

The problem of the high prevalence of stunting can be overcome with several solutions. First, provide education or training regarding stunting to increase public awareness. Other solutions include providing access to adequate health services, encouraging the government to implement policies that support stunting management, empowering women, regularly conducting research on stunting in Indonesia, and finally carrying out evaluations to improve in the future.

4. Conclusions

The best model for the prevalence of stunting in Indonesia using non-parametric truncated spline multi-predictor regression is modelling with three knot points because it has the minimum Generalised Cross Validation (GCV) value and the largest coefficient of determination compared to models with one and two knot points. Based on the results of hypothesis testing, at a significance level of 5%, it was concluded that the variables, which are the percentage of infants under 6 months of age who were exclusively breastfed, the average age of the mother's first pregnancy, the percentage of married women aged 15-49 years who are currently using/wearing contraception/birth control methods, the percentage of mothers who gave birth to Live-Born Children (ALH) in the last two years and initiated early breastfeeding, and the percentage of children aged 12-23 months who received complete basic immunisation had a significant simultaneous and partial effect on the prevalence of stunting in Indonesia.

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 Universitas Airlangga for the support and resources they have provided. We would also like to thank Badan Pusat Statistik (BPS) to publication data and datasets, which have greatly enriched our research results. The support and cooperation of all parties has been invaluable to the success of 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.

References

[1] N. A. N. Djide, 'Hubungan Intervensi Spesifik Dari Indikator Program Indonesia Sehat Dengan Pendekatan Keluarga (Pis-Pk) Dengan Prevalensi Stunting Di 10 Desa Lokus Program Pencegahan Stunting di Kab. Banggai Tahun 2018-2019', *Jurnal Kesehatan Masyarakat*, vol. 12, no. 5, pp. 121–231, 2021.

[2] L. Makripudin, D. A. Roswandi, and F. T. Tazir, 'Kebijakan dan Strategi Percepatan Penurunan Stunting Di Indonesia'. Jakarta: Pusat Pendidikan dan, 2021.

[3] K. R. Indonesia, 'Standar antropometri penilaian status gizi anak', *Departemen Bina Gizi, Jakarta*, 2010.

[4] Kementerian Kesehatan Republik Indonesia (Kemenkes RI), 'Buletin Jendela Data dan Informasi Kesehatan: Situasi Balita Pendek (Stunting) di Indonesia', *Kementerian Kesehatan RI*, p. 20, 2018.

[5] Badan Perencanaan Pembangunan Nasional (Bappenas), 'Tentang SDGs'. [Online]. Available: https://www.bappenas.go.id/id/berita-dan-siaran-pers/tentang-sdgs/.

[6] Badan Perencanaan Pembangunan Nasional (Bappenas), 'Tujuan SDGs Poin Kedua "Tanpa Kelaparan."' [Online]. Available: https://www.bappenas.go.id/id/berita-dan-siaran-pers/tujuan-sdgs-poin-kedua-tanpa-kelaparan/.

[7] A. Y. K. Kartini and L. N. Ummah, 'Pemodelan Kejadian Balita Stunting di Kabupaten Bojonegoro dengan Metode Geographically Weighted Regression dan Multivariate Adaptive Regression Splines', *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, vol. 15, no. 1, 2022.

[8] M. Rifada, N. Chamidah, R. A. Ningrum, and L. Muniroh, 'Stunting determinants among toddlers in Probolinggo district of Indonesia using parametric and nonparametric ordinal logistic regression models', *Commun. Math. Biol. Neurosci.*, vol. 2023, p. Article-ID, 2023.

[9] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[10] Irodah, 'Hubungan Berat Badan Lahir Dan Pemberian ASI Eksklusif Dengan Kejadian Stunting Pada Balita Usia 12-59 Bulan Di Puskesmas Pegandon Kabupaten Kendal', Universitas Muhammadiyah Semarang, 2018.

[11] D. Widiantini and Others, 'Estimasi Model Regresi Nonparametrik Multivariat Berdasarkan Estimator Polinomial Lokal Orde Dua', Universitas Airlangga, 2012.

[12] I. N. Budiantara, 'Spline dalam Regresi Nonparametrik dan Semiparametrik: Sebuah Pemodelan Statistika Masa Kini dan Masa Mendatang', *Pidato Pengukuhan untuk Jabatan Guru Besar dalam Bidang Ilmu Matematika Statistika dan Probabilitas, pada Jurusan Statistika, Fakultas MIPA*, 2009.

[13] W. W. Chin and Others, 'The partial least squares approach to structural equation modeling', *Modern methods for business research*, vol. 295, no. 2, pp. 295–336, 1998.

[14] H. Muryaninggar, 'Pemodelan Angka Putus Sekolah Usia SMA di Indonesia dengan Pendekatan Regresi Nonparametrik Spline', Institut Teknologi Sepuluh Nopember, 2017.

[15] A. D. N. Yadika, K. N. Berawi, and S. H. Nasution, 'Pengaruh stunting terhadap perkembangan kognitif dan prestasi belajar', *Jurnal Majority*, vol. 8, no. 2, pp. 273–282, 2019.

DOI: https://doi.org/10.34123/jurnalasks.v16i1.666

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

Digital Literacy in Mediating the Influence of Education, Demography, and Employment on Poverty

Satria Liswanda[1*], Rini Oktavia[2] , Rahma Zuhra[3]

[2,3] STEM Research Center, Universitas Syiah Kuala, 23111, Banda Aceh, Indonesia, [1] Master of Mathematics Study Program, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, 23111, Banda Aceh, Indonesia, [1,2,3] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Syiah Kuala, 23111, Banda Aceh, Indonesia

**Corresponding Author: E-mail address:* satria23math@gmail.com

ARTICLE INFO

Article history:
Received 10 December, 2023
Revised 4 May, 2024
Accepted 7 May, 2024
Published 30 June, 2024

Keywords:
Digital Literacy, Education, Demography, Employment, Poverty, SEM-PLS

Abstract

Introduction/Main Objectives: This study investigated the influence of education, demography, and employment on poverty with digital literacy as a mediating variable. **Background Problems:** interrelationship has not been investigated before, either in Indonesia or in other countries, even though some studies have indicated the importance of providing digital literacy training to eradicate poverty in poor communities. **Novelty:** using the Structural Equation Modeling (SEM) Partial Least Square (PLS) method in analyzing data and modeling relationships between Education, demographic, and employment factors and poverty with digital literacy as a mediating variable that acts as a link. **Research Methods:** A structural Equation Modeling (SEM) with the Partial Least Square (PLS) method was applied. **Finding/Results:** Significant indicators found are four indicators of education and digital literacy variables, two indicators of demographic and employment variables, and three indicators of poverty variables. It was found that education and employment variables had a significant influence on poverty with a negative influence. We found that no variable has a significant effect on digital literacy and there is no significant effect of digital literacy on Poverty.

1. Introduction

Poverty according to the Badan Pusat Statistik (BPS) is a condition where a person has an income below the basic needs of the general public [1]. Efforts to reduce poverty are a United Nations program in 2015 Sustainable Development Goals (SDGs) as a global invitation are in line with efforts to protect the earth and ensure that by 2030, all humans can live in peace and prosperity. No Poverty is the first point in te SDGs which means the global community collaboratively agrees to eradicate poverty on earth. Poverty can be attributed to several other Global Goal factors such as the quality of education, demography, employment, and the use of digitalization [2].

According to Lawrence Cremin in Klimczuk, education is efforts made to convey or acquire knowledge, values, attitudes, and skills. Demography is the study of human populations and changes in their quantity related to migration, births, and deaths [3]. Employment is a social interaction between

employers and workers, where workers provide certain services and receive predetermined and negotiable wages [4]. According to Schallmo and Williams [5], digitalization is the use of digital technology and data to generate revenue, improve business, change business processes, and create an ecosystem for digital business. Digital literacy is a skill needed to do something by utilizing communication and access to information using digital technology such as the Internet, social media, and mobile devices.

Previous research using SEM-PLS was conducted by Nurhafifah et al. [6], to see the relationship between economic, educational, and health factors to poverty in Indonesia in districts/cities that have a percentage of poor people above the percentage of poor people in Indonesia. The results of the study stated that the heterogeneity of poverty variables can be explained by economic, educational, and health variables. The increase of digitalization in all aspects of our lives made digital literacy one of the important life skills needed to have a better life. Advancing the study done by Nurhafifah et al., the authors were interested in using the Structural Equation Modeling (SEM) Partial Least Square (PLS) method in analyzing data and modeling relationships between Education, demographic, and employment factors and poverty with digital literacy as a mediating variable that acts as a link [6]. This interrelationship has not been investigated before, either in Indonesia or in other countries, even though some studies have indicated the importance of providing digital literacy training to eradicate poverty in poor communities [7].

SEM is a statistical approach to testing hypotheses about the relationship between observed variables (indicators) and latent variables [8]. PLS is an SEM method used to estimate path relationships between latent variables and between latent variables and their indicators in complex problems [9]. PLS data analysis techniques consist of parameter estimation and model evaluation, parameter estimation is the first step of SEM-PLS data analysis that will determine the next data analysis. Parameter estimation aims to produce values that relate between indicators and their latent variables and the relationship between latent variables. So for the next stage of analysis, which indicators can still be used or not will be evaluated.

2. Material & Methods

2.1. Partial Least Square- Structural Equation Modeling (PLS-SEM)

Structural Equation Modeling (SEM) is a multivariate statistical technique used to model complex relationships between directly observable variables (indicators) and non-directly observable variables (latent variables). SEM analysis techniques involve the solution of systems of linear equations, regression, factor analysis, path analysis, and growth curve modeling [10]. There are two variables in SEM analysis, namely variables that can be observed directly or commonly called indicators and variables that cannot be observed directly called latent variables. There are two types of latent variables, namely endogenous latent variables as response variables with notation η (eta) and exogenous latent variables as explanatory variables with notation ξ (xi).

There are two types of SEM, namely the covariant-based SEM (CB-SEM) and the variance-based SEM (SEM-PLS). CB-SEM is appropriate for theoretical tests and obtaining explanations based on test results with a series of complex analyses, CB-SEM also requires a relatively large number of samples for accurate results. While SEM-PLS aims to test the existence of relationships or predictive influences between variables, SEM-PLS can use samples that are not large. The data and objectives of this study that fit these criteria are using SEM-PLS [11].

Partial Least Square is an SEM data analysis technique to simulate the connection between response variables and other explanatory variables. A simple interpretation is given to show an easy-to-apply method of forming predictive equations. PLS-SEM consists of two models, including a measurement model and a structural model. The measurement model is a model of the relationship between latent variables and indicators, this model is analyzed to see the validity and reliability of each indicator used. The structural model is a model of relationships between latent variables, this model is to see the relationships in the model and test hypotheses on prediction models [12].

Indicators in SEM-PLS data analysis are built through two models, including the reflective model and the formative model. The reflective model describes indicators that are affected by latent variables,

the formative model is an indicator model that affects latent variables. The structural models (inner models) are designed to model the relationship between latent variables, in the form of:

$$\eta_j = \sum \gamma_i \xi_i + \zeta, \quad (1)$$

where:

I = the index of exogenous latent variables

γ_i = the connecting pathway coefficient of the endogenous (η) with exogenous (ξ)

ζ = the measurement error rate.

Measurement models (outer models) describe the relationship between latent variables and their indicators. In this study, the authors used a reflective-type measurement model with the following conditions:

$$X = \lambda_X \xi + \delta_X, \quad (2)$$

$$Y = \lambda_Y \eta + \varepsilon_Y, \quad (3)$$

X is the indicator of exogenous latent variables (ξ). Y is the indicator of endogenous latent variables (η). ξ is the exogenous latent variable. η is the endogenous latent variable. λ_X and λ_Y are a loading matrix, associating latent variables with indicators. δ_X ε_Y is the measurement error rate.

That model specifications do not yet describe latent variable values therefore weight relations must be defined. One of the characteristics of SEM-PLS analysis is estimating the value of the latent variable score. Weights are described with the symbol w_{jk}, so the estimated latent variable score can be written as follows:

$$\xi_j = \sum_{k=1}^{K_j} w_{jk} X_{jk}. \quad (4)$$

The PLS algorithm is a link between simple and multiple regression using an estimation approach ordinary least square [13].

2.2. PLS Algorithm Stage 1

The first stage aims to generate weights to calculate the score of latent variables. In calculating weight, we used iteration techniques based on the built model, namely structural and measurement models. At this stage, it depends heavily on the relationship between the score of the latent variable in the structural model and the indicator linked to the score of the latent variable. The estimation of the parameters of the measurement model is formed through equation (4). While the estimation of structural model parameters is formed through the formula:

$$z_j \propto \sum_{i=1, i \neq j}^{J} e_{ji} \xi_i, \quad (5)$$

z_j is the symbol of the latent variable to be reestimated. Symbol $\propto$ means left-hand variables represent right-hand variables. The weight of the structural model e_{ji} can be estimated using a factoring scheme, this scheme takes into account the direction of the sign and the strength of the path on the structural model. The factor scheme is defined as follows [14]:

$$e_{ji} = \{cor(\xi_i, \xi_j)\ 0\ , related\ \xi_i\ and\ \xi_j\ \ others. \quad (6)$$

The next step at this stage is to update the measurement model after estimating the approximate value of the weight value. Updating the weights of the measurement model can use the reflective indicator model:

$$X_{jk} = \lambda_{jk} \xi_j + \delta_{jk}. \quad (7)$$

The renewal weight value for an exogenous latent variable is defined as follows:

$$w_{jk} = \left(z_j' z_j\right)^{-1} z_j' X_{jk}, \tag{8}$$

Next is the convergence check at the iteration stage, the convergence is checked by comparing the weights of the new values at each current step and the previous step with the following criteria:

$$\left|w_{jk}^{s} - w_{jk}^{s-1}\right| < 10^{-5}. \tag{9}$$

2.3. PLS Algorithm Stage 2

The second stage is to calculate the estimation of the path coefficient and loading factor in structural models and measurement models. Path coefficients in structural models are estimated using OLS (Ordinary Least Square) such as multiple linear regression analysis by looking at the relationship between ξ_j and ξ_i.

$$\xi_j = \sum_{i=1}^{I_j} \gamma_{ji} \xi_i, \tag{10}$$

$$\gamma_{ji} = (\xi_i' \xi_i)^{-1} \xi_i' \xi_j, \tag{11}$$

In reflective measurement models, loading factors are estimated such as multiple linear regression of relationships described as follows:

$$X_{jk} = \lambda_{jk} \xi_j, \tag{12}$$

$$\lambda_{jk} = \left(\xi_j' \xi_j\right)^{-1} \xi_j' X_{jk}. \tag{13}$$

2.4. PLS Algorithm Stage 3

The last stage is to estimate location parameters. There are two estimated location parameters γ_{0j} (structural model constant) and λ_{0jk} (reflective measurement model constant). The specification of an equation containing a constant is defined as a linear regression as follows:

$$E(\xi_i) = \gamma_{0j} + \sum_{i}^{I_j} \gamma_{ji} \xi_i, \tag{14}$$

$$E\left(\xi_j\right) = \lambda_{0jk} + \lambda_{jk} \xi_j. \tag{15}$$

The location parameter takes into account the mean of the latent variables and indicators. The mean for the estimation of latent variables is defined as follows:

$$\widehat{m_j} = \sum_{k=1}^{K_j} w_{jk} X_{jk}, \tag{16}$$

$$\widehat{\xi_j} = \xi_j + \widehat{m_j}. \tag{17}$$

Based on the form of the equation above γ_{0j} and λ_{0jk} can be defined as follows [15]:

$$\gamma_{0j} = \widehat{m_j} - \sum_{i}^{I_j} \gamma_{ji} \widehat{m_i}, \tag{18}$$

$$\lambda_{0jk} = X_{jk} - \lambda_{jk} \widehat{m_j}. \tag{19}$$

Model evaluation is done to see indicators and variables that can work well in model analysis. Model evaluation is two, namely measurement model evaluation and structural model evaluation.

2.5. Evaluation of Measurement Models on Reflective Indicators

2.5.1. Convergent Validity

Convergent Validity is indicated by the value of the loading factor (λ). This value describes the relationship between the indicator and its latent variable. A loading factor value above 0.7 is said to be an ideal value that can work well in model analysis and is said to be significant as an indicator that measures latent variables. A loading factor value below 0.7 can be eliminated from the model [16].

2.5.2. Composite Reliability

Composite Reliability is part of the indicator that measures the relationship between latent variables. A latent variable is said to be reliable if the value of Composite Reliability is more than 0.7. Composite Reliability can be defined by the following formula [16].

$$CR = \frac{\left(\sum_{k=1}^{K_j} \lambda_{jk}\right)^2}{\left(\sum_{k=1}^{K_j} \lambda_{jk}\right)^2 + \sum_{k=1}^{K_j}(1-\lambda_{jk}^2)}. \quad (20)$$

2.6. Structural Model Evaluation

Evaluation of the structural model is carried out by looking at the estimated value of the path coefficient which describes the strength of the relationship between latent variables, R^2 which indicates the magnitude of variability of endogenous latent variables described by exogenous latent variables, and Q^2 which can be used for the predictive ability of the model. If the value of Q^2 gets closer to 1, then the model has good predictions [9]. The value R^2 and Q^2 can be described as follows:

$$R^2 = \sum_{k=1,j=1}^{K,J} \gamma_{jk}\, cor(\xi_k, \eta_j), \quad (21)$$

$$Q^2 = 1 - (1 - R_1^2)(1 - R_2^2) \ldots (1 - R_n^2). \quad (22)$$

The bootstrap sampling method involves resampling from the original sample. By resampling or repeating sampling, the bootstrap method is used to determine the average value of derivatives from skewed data [17]. The bootstrap resample approach employs hypothesis testing. The following are the suggested hypotheses.

The statistical hypothesis of structural models, the influence between latent variables is:

$$H_{0i}: \gamma_i = 0;\ H_{1i}: \gamma_i \neq 0$$

The statistical hypothesis of the measurement model is:

$$H_{0i}: \lambda_i = 0;\ H_{1i}: \lambda_i \neq 0.$$

Digital literacy is the ability to use digital technology to obtain, manage, understand, integrate, communicate, evaluate, and produce information for employment, decent work, and entrepreneurship safely and ethically. This includes skills referred to by various terms such as media literacy, information literacy, computer literacy, and ICT literacy. Digitalization in Indonesia will impact revenues of up to 150 billion US dollars by 2025 and create 3.7 million new jobs. This potential, among others, is evidenced by the increasing number of emerging start-up technology companies [18].

Education is the act, practice, or application of discipline to the intellect or a process of character training, the process of education is crucial to human development [19]. The ability to access employment, resources, and skills that enable one to not just survive but also thrive is one of the reasons why education is frequently referred to as the great equalizer. Because of this, having access to a good education is considered a known antidote to poverty. Numerous other problems that might make

individuals, families, and even entire communities vulnerable to the cycle of poverty can be resolved with education [20].

Demography is the scientific study of human population growth and development, with a focus on migration, marriage, health, and living arrangements as well as factors such as fertility, mortality, and migration [21]. Over the previous few decades, poverty has substantially decreased worldwide. This shift was accompanied in many emerging nations by quick advancements in demographic outcomes, such as declining child mortality and fertility [22].

Employment is a contract of a worker will perform to an employer. In return, the worker receives a salary or wage with some negotiable terms for both parties [4]. Poverty can be anticipated if a person has a job. In all European countries, poverty will increase for the unemployed and the Great Recession brings unfavorable social repercussions due to widespread unemployment [23].

Poverty is the state or situation of an individual or society when there is a lack of means of livelihood. Individuals or communities living in poverty experience insufficient access to adequate housing, clean water, healthy food, and medical care. Poverty is not only caused by income earned, poverty can also be caused by other factors such as education, employment, and so on [24].

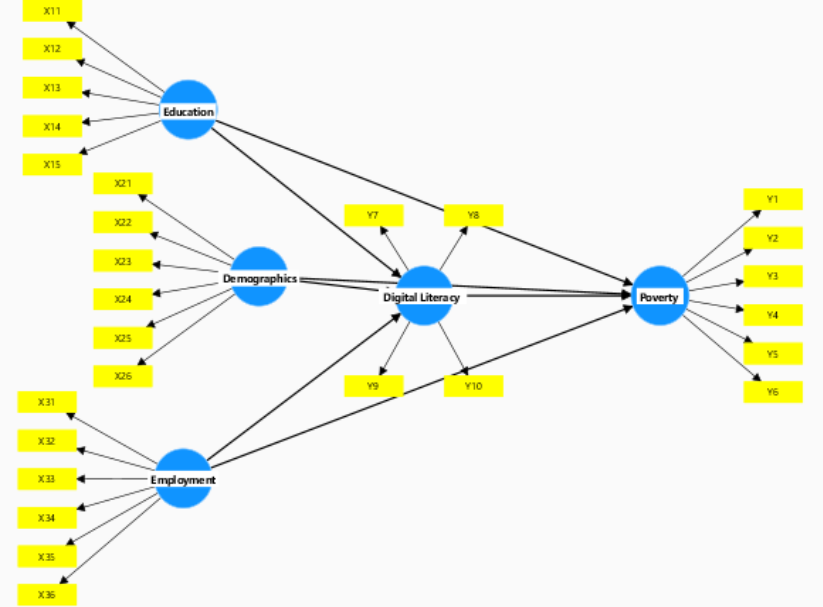


Figure. 1. Path Chart of Latent Variable and Indicator.

Table 1. Path Chart of Latent Variable and Indikator

Latent Variables	Indicators	
Education (ξ_1)	X11	Average Years of Schooling for Residents Aged 15 Years and Over
	X12	Availability of high schools
	X13	College Availability
	X14	The number of high school students
	X15	The number of college students
Demographics (ξ_2)	X21	Life expectancy
	X22	Population growth rate
	X23	Area
	X24	Population
	X25	Human Development Index (HDI)
	X26	Population density
Employment (ξ_3)	X31	Percentage of employed against the labor force Open unemployment rate
	X32	Labor force participation rate
	X33	Registered job seekers
	X34	Registered vacancies

Table 1. Path Chart of Latent Variable and Indikator

Latent Variables	Indicators	
	X35	Workforce fulfilment
	X36	
Poverty (η_1)	Y1	Provincial Minimum Wage (UMP)
	Y2	Percentage of poor people
	Y3	Poverty severity
	Y4	Depth of poverty
	Y5	Poverty line
	Y6	Provincial per capita expenditure
Digital Literacy (η_2)	Y7	Information and data literacy
	Y8	Communication and Collaboration
	Y9	Security in the use of digital technology
	Y10	Ability to use technology

2.7. Analysis Method

To see the influence between factors in this study, the author applies data analysis techniques using Structural Equation Modeling (SEM) with the Partial Least Square (PLS) method. This research uses secondary data obtained from several sources, including the Badan Pusat Statistik (BPS) and the Ministry of Communication and Information Technology of Indonesia (Kominfo). The data used are data related to education, demography, employment, digital literacy, and poverty which consists of 27 indicators and five latent variables. This data is taken from the data in 34 provinces in Indonesia in 2020. The five latent variables are education, demography, employment, digital literacy, and poverty. An explanation of latent variables and their indicators can be seen in Table 1.

The data analysis techniques in this study are as follows: (1) Descriptive statistical analysis, (2) Designing models, (3) Creating a path chart, (4) Performing a path diagram conversion to the equation, (5) Estimating parameters, (6) Evaluate the model, (7) Conducting hypothesis testing, and (8) Draw the conclusions.

3. Results

Descriptive statistics are used to describe in general the variables in the study. The results of descriptive statistics on the indicators in this study can be seen in Table 2. The value of the parameter coefficient/weight (w_{jk}) of the measurement model λ and the parameters of the structural model γ are obtained in Table 3. Evaluation of measurement models on reflective indicators includes the value of validity and reliability of each indicator against its latent variables.

This study consisted of 34 observations and 5 variables, validity tests were analyzed through degrees of freedom ($df = 34 - 5 = 29$) so that the t-table value for the significance level of 10% with two-tailed and df 29 was 1.7. Validity is a value that describes the relationship between a reflective indicator and its latent variable. The evaluation is by looking at the value of the loading factor (λ), if the value of the loading factor ($\lambda \geq 0.5$) then the indicator is declared valid. However, if the value of the loading factor ($\lambda < 0.5$) then the indicator is invalid and must be eliminated from the model. In Table 3 there are still a loading factor value ($\lambda < 0.5$) that is on the indicator X_{11}, X_{21}, X_{22}, X_{23} $X_{24}, X_{32}, X_{34}, X_{35}, X_{36}, Y_2, Y_3$, dan Y_4. The loading factor value ($\lambda < 0.5$) indicates that the indicator is invalid and must be eliminated in the next analysis. So that at the next stage of analysis the indicator $X_{11}, X_{21}, X_{22}, X_{23}, X_{24}, X_{32}, X_{34}, X_{35}, X_{36}, Y_2, Y_3$, dan Y_4 is no longer used. The models in Figures 2 and 3 are over-identified because the number of parameters of 27 is smaller than the number of data of 34. The following Figures 2 and 3 are path diagrams of the SEM-PLS model before and after eliminating invalid indicators.

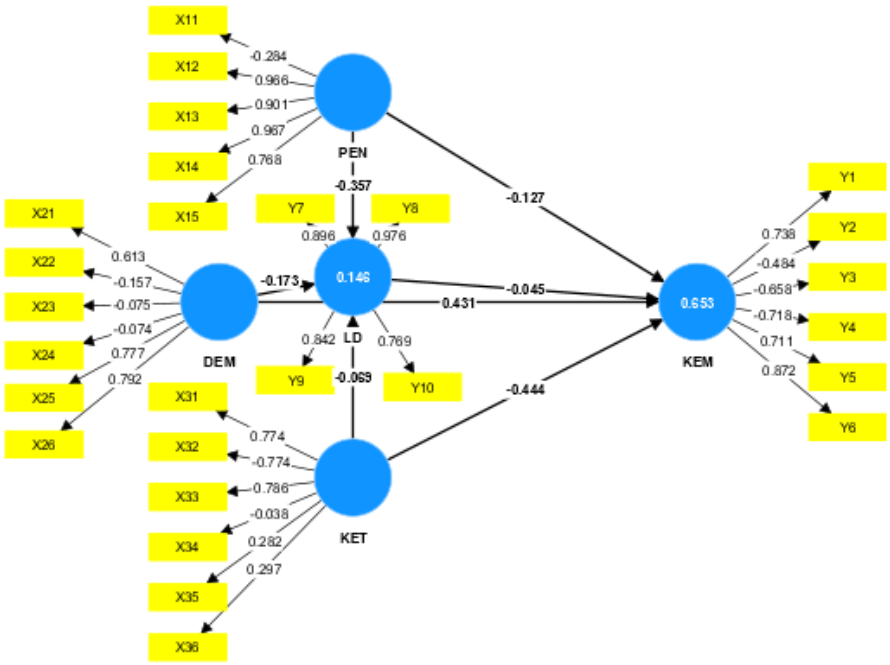


Figure. 2. Path Chart and Loading Factor Value Before Elimination of Invalid Indicators.

Table 2. Descriptive Statistics

Variable	Min	Max	Median	IQR
Poverty (η_1)				
Y_1	1704608	4276350	2678863	639170.75
Y_2	4.45	26.80	9.140	6.715
Y_3	0.09	0.72	0.280	0.2375
Y_4	0.43	2.85	1.150	0.8325
Y_5	356967	723478	504445	152280.3
Y_6	794361	1140075	1251783.76	336501
Digital Literacy (η_2)				
Y_7	2.68	3.77	3.250	0.305
Y_8	3	3.97	3.420	0.3975
Y_9	3.30	4.38	3.670	0.315
Y_{10}	3.08	4.55	3.740	0.445
Education (ξ_1)				
X_{11}	6.96	11.17	9.2	1.2675
X_{12}	107.00	5884	589	805
X_{13}	8	389	54	78.5
X_{14}	29169	2047024	167443	252433
X_{15}	11834	1308214	94760	170810
Demographics (ξ_2)				
X_{21}	65.14	75.03	70	3.0625
X_{22}	0.58	4.13	1.4	0.7075
X_{23}	664.01	319036.05	42012.890	56906.45
X_{24}	701.80	48274.20	1.590	6551.375
X_{25}	60.44	80.77	71.450	3.11
X_{26}	9	15907	104	220.5
Employment (ξ_3)				
X_{31}	89.05	96.68	94.490	2.295

Table 2. Descriptive Statistics

Variable	Min	Max	Median	IQR
X_{32}	3.32	10.95	5.630	2.295
X_{33}	63.40	74.32	68.670	4.9625
X_{34}	16765	1310894	154007	233129.5
X_{35}	8384	595499	53106	81916.25
X_{36}	6987	449249	44258	74374.5

Based on Figure 3, all loading factor values are greater or equal to 0.5 ($\lambda \geq 0.5$) for each indicator of each latent variable: education, demography, employment, digital literacy, and poverty. Based on these provisions, all indicators used are good and valid so that they can be used in measuring latent variables.

Table 3. Model Parameter Estimation

Loading Factor Measurement Model	
$\lambda_{X11} = -0.284$	$\lambda_{X34} = -0.038$
$\lambda_{X12} = 0.966$	$\lambda_{X35} = 0.282$
$\lambda_{X13} = 0.901$	$\lambda_{X36} = 0.297$
$\lambda_{X14} = 0.967$	$\lambda_{Y1} = 0.738$
$\lambda_{X15} = 0.768$	$\lambda_{Y2} = -0.484$
$\lambda_{X21} = 0.613$	$\lambda_{Y3} = -0.658$
$\lambda_{X22} = -0.157$	$\lambda_{Y4} = -0.718$
$\lambda_{X23} = -0.075$	$\lambda_{Y5} = 0.711$
$\lambda_{X24} = -0.074$	$\lambda_{Y6} = 0.872$
$\lambda_{X25} = 0.777$	$\lambda_{Y7} = 0.896$
$\lambda_{X26} = 0.792$	$\lambda_{Y8} = 0.976$
$\lambda_{X31} = 0.774$	$\lambda_{Y9} = 0.842$
$\lambda_{X32} = -0.774$	$\lambda_{Y10} = 0.769$
$\lambda_{X33} = 0.786$	

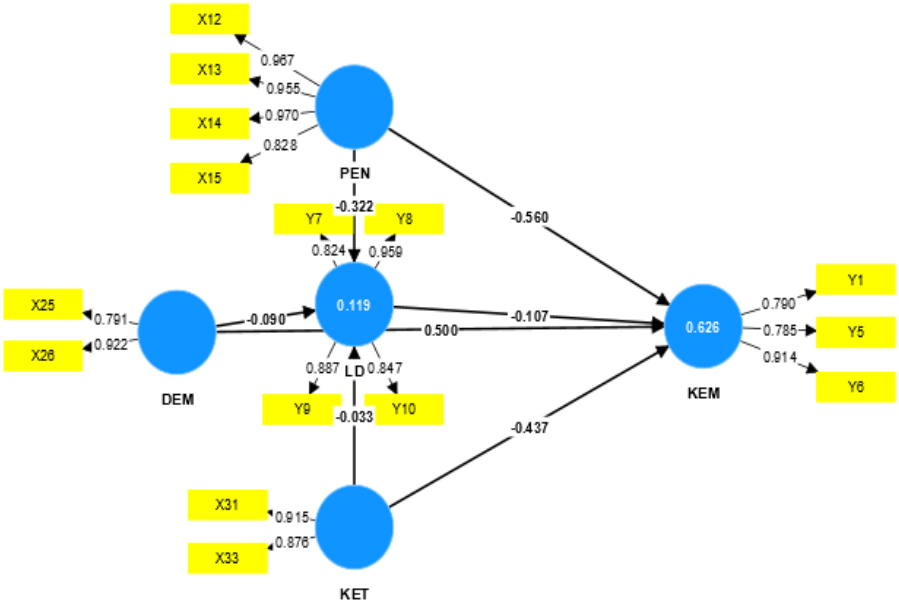


Figure. 3. Path Chart and Loading Factor Value After Elimination of Invalid Indicators.

Reliability indicates the level of consistency of the data. A variable is said to be reliable if it has a composite reliability value greater than 0.7. The composite reliability value of each latent variable can be seen in Table 4.

Table 4. Composite Reliability of Latent Variable

Variable	Composite Reliability
Education	0.963
Demographics	0.848
Employment	0.890
Digital Literacy	0.932
Poverty	0.870

Based on Table 4 the composite reliability value of each latent variable is more than 0.7. This means that all indicators of the measured latent variable are declared reliable. Based on these criteria, it can be concluded that the measurement model is good because it has met the validity test and reliability test.

Table 5. Values of R-square, Q-square, and F-square

Evaluation Size	Value
R_1^2	0.626
R_2^2	0.119
Q^2	0.671
Variable	F-square
Education → Digital Literacy	0.099
Demographics → Digital Literacy	0.007
Employment → Digital Literacy	0.001
Education → Poverty	0.644
Demographics → Poverty	0.489
Employment → Poverty	0.354
Digital Literacy → Poverty	0.027

Structural model evaluation can be done to see the relationship between latent variables and their effects by looking at the value of the estimated result coefficient and the level of significance. Measures for the evaluation of structural models can use values of R-square and Q-square.

Table 5 indicates the value of $R_1^2 = 0.626$ for the variables of poverty, meaning that the poverty variable that can be explained by education, demographics, employment, and digital literacy variables is 62.6 percent. The remaining 37.4 percent was explained by other variables not mentioned in the model. The digital literacy variable value $R_2^2 = 0.119$ means that the 11.9 percent digital literacy variable can be explained by education, demographic, and employment variables. The remaining 88.1 percent was explained by other variables not included in the study. The resulting value of Q^2 is 0.671, meaning that the model has a good prediction because the value of Q^2 is close to 1.

Evaluation of the model is also seen based on discriminant validity, this value is seen based on three categories: cross-loading, HTMT, and Fornell Larcker. These values can be seen in Tables 6, 7, and 8.

Table 6 shows valid values for all variables because the indicator values on the value-bound variables are greater than the values of other indicators. Table 7 is valid because all values are less than 0.9. Table 8 is valid because the value of the relationship to the same latent variable is greater than the other values.

Bootstrapping is used to determine the standard deviation/standard error in determining the significance of statistical values without relying on any assumptions [25]. This bootstrapping technique involves the main sample to be resampled. This method aims to find statistics in the distribution of data [17]. The bootstrapping procedure is performed using 500 resampling at a t-table value of 1.65 (2-tailed) and a significant level of 0.1. The hypotheses used are:

$H_0: \lambda_i = 0$
$H_1: \lambda_i \neq 0.$

The results of t-statistical testing for measurement models can be seen in Table 9. In the Table 9 shows a good measurement model for each of the latent variables obtained. This is indicated by a t-statistic value greater than 1.65 (2-tailed) at a significant level of 0.1 or with a p-value less than 0.1. The hypothesis used for the measurement model is H_{0i}: $\lambda_i = 0$ and H_{1i}: $\lambda_i \neq 0$. The model shows that $\lambda_i \neq 0$ meaning the hypothesis H_{0i} is rejected and we agree with the alternative hypothesis H_{1i}, meaning that there is an influence between the latent variable and the indicator. Based on this hypothesis, we can conclude that each latent variable has a relationship with its indicators and has path coefficient values that are all positive. The smallest contribution was indicator Y5 with a coefficient to the latent variable of poverty is 0.785; the largest contribution was indicator X14 with a path coefficient to the latent variable of education is 0.970. The path coefficient values for each bootstrap resampling in Table 10 show the same value as the value of the actual path coefficient.

Table 6. Values of Cross-Loading

Variable	ξ_2	η_2	ξ_1	ξ_3	η_1
Poverty (η_1)					
Y_1	0.323	-0.067	-0.397	-0.367	**0.790**
Y_5	0.231	0.036	-0.283	-0.233	**0.785**
Y_6	0.707	-0.082	0.015	-0.533	**0.914**
Digital Literacy (η_2)					
Y_7	-0.0938	**0.824**	-0.414	-0.020	0.109
Y_8	-0.090	**0.959**	-0.323	0.065	-0.011
Y_9	-0.195	**0.887**	-0.219	0.154	-0.175
Y_{10}	-0.209	**0.847**	-0.224	0.256	-0.135
Education (ξ_1)					
X_{12}	0.123	-0.309	**0.967**	-0.269	-0.334
X_{13}	0.462	-0.336	**0.955**	-0.410	-0.085
X_{14}	0.154	-0.241	**0.970**	-0.277	-0.317
X_{15}	0.442	-0.386	**0.828**	-0.485	0.010
Demographics (ξ_2)					
X_{25}	**0.791**	-0.034	0.236	-0.456	0.389
X_{26}	**0.922**	-0.219	0.273	-0.423	0.565
Employment (ξ_3)					
X_{31}	-0.559	0.098	-0.496	**0.915**	-0.484
X_{33}	-0.317	0.145	-0.146	**0.876**	-0.389

Table 7. Values of HTMT

Variable	ξ_2	η_2	ξ_1	ξ_3	η_1
Demographics (ξ_2)					
Digital Literacy (η_2)	0.219				
Education (ξ_1)	0.398	0.369			
Employment (ξ_3)	0.706	0.178	0.437		
Poverty (η_1)	0.668	0.146	0.376	0.583	

Table 8. Values of Fornell Larcker

Variable	ξ_2	η_2	ξ_1	ξ_3	η_1
Demographics (ξ_2)	0.859				
Digital Literacy (η_2)	-0.169	0.881			
Education (ξ_1)	0.297	-0.337	0.932		
Employment (ξ_3)	-0.500	0.133	-0.375	0.896	
Poverty (η_1)	0.571	-0.061	-0.212	-0.491	0.832

Table 9. Results of T-Statistical Values of Loading Measurement Model

	Loading Factor	*T-Statistic*	*P-value*
Education			
X_{12}	0.967	8.281	0.000
X_{13}	0.955	6.683	0.000
X_{14}	0.970	8.282	0.000
X_{15}	0.828	4.632	0.000
Demography			
X_{25}	0.791	3.185	0.002
X_{26}	0.922	3.684	0.000
Employment			
X_{31}	0.915	4.100	0.000
X_{33}	0.876	4.714	0.000
Poverty			
Y_1	0.790	6.085	0.000
Y_5	0.785	5.711	0.000
Y_6	0.914	3.367	0.001
Digital Literacy			
Y_7	0.824	6.006	0.000
Y_8	0.959	9.689	0.000
Y_9	0.887	7.349	0.000
Y_{10}	0.847	6.436	0.000

Table 10. Bootstrap Resampling Estimation Results of Path Coefficient Value

Variable	Path Coefficient	Bootstrap Resampling (Path Coefficient)		
		100	500	1000
Education → Digital Literacy	-0.322	-0.322	-0.322	-0.322
Demographics → Digital Literacy	-0.090	-0.090	-0.090	-0.090
Employment → Digital Literacy	-0.033	-0.033	-0.033	-0.033
Education → Poverty	-0.560	-0.560	-0.560	-0.560
Demographics → Poverty	0.500	0.500	0.500	0.500
Employment → Poverty	-0.437	-0.437	-0.437	-0.437
Digital Literacy → Poverty	-0.107	-0.107	-0.107	-0.107

Mathematically the structural model of PLS analysis can be written as follows:

$$\eta_1 = -0.560\,\xi_1 + 0.500\,\xi_2 - 0.437\,\xi_3 - 0.107\,\eta_2 + \zeta_1 \quad (23)$$

$$\eta_2 = -0.322\,\xi_1 - 0.090\,\xi_2 - 0.033\,\xi_3 + \zeta_2 \quad (24)$$

Based on Table 11, the t-statistics value at 500 resampling has a greater value on most of the relationships between variables compared to other resampling, so 500 resampling is best used in subsequent analyses. The results of bootstrapping t-statistics resampling testing using 500 resamplings are shown in Table 12.

Table 11. Bootstrap Resampling Estimation Results T-Statistic

Variable	Resampling Bootstrap (T-Statistics)		
	100	500	1000
Education → Digital Literacy	1.216	1.235	1.225
Demographics → Digital Literacy	0.381	0.425	0.428
Employment → Digital Literacy	0.128	0.128	0.124
Education → Poverty	2.805	3.700	3.404
Demographics → Poverty	1.013	1.056	1.016
Employment → Poverty	3.169	2.888	2.666
Digital Literacy → Poverty	1.053	1.028	0.986

Table 12. Value of Loading Factor, T-Statistic, and P-Value used Bootstrap

Variable	*Loading Factor*	*T-Statistic*	*P-value*
Education → Digital Literacy	-0.322	1.235	0.218
Demographics → Digital Literacy	-0.090	0.425	0.671
Employment → Digital Literacy	-0.033	0.128	0.898
Education → Poverty	-0.560	3.700	0.000*
Demographics → Poverty	0.500	1.056	0.292
Employment → Poverty	-0.437	2.888	0.004*
Digital Literacy → Poverty	-0.107	1.028	0.305

Based on Table 11, the t-statistics value at 500 resampling has a greater value on most of the relationships between variables compared to other resampling, so 500 resampling is best used in subsequent analyses. The results of bootstrapping t-statistics resampling testing using 500 resamplings are shown in Table 12.

The results of bootstrapping t-statistics resampling testing using 500 resamplings are shown in Table 12. Based on Table 12 the causal relationships between latent variables are described as follows:

H_1: $\gamma_{11} \neq 0$. Education (ξ_1) affects digital literacy (η_2). The t-statistics value of 1.235 is smaller than the t-table of 1.65 and the p-value of 0.218 is greater than 0.1 (not significant), meaning that education has an influence on digital literacy with a negative influence of -0.322 but not significant.

H_2: $\gamma_{21} \neq 0$. Demographics (ξ_2) affect digital literacy (η_2). The t-statistics value of 0.425 is smaller than the t-table of 1.65 and the p-value of 0.671 is greater than 0.1 (not significant), meaning that demographics influence digital literacy with a negative influence of -0.090 but not significant.

H_3: $\gamma_{31} \neq 0$. Employment (ξ_3) affects digital literacy (η_2). The t-statistics value of 0.128 is smaller than the t-table of 1.65 and the p-value of 0.898 is greater than 0.1 (not significant), meaning that employment has an influence on digital literacy with a negative influence of -0.033 but not significant.

H_4: $\gamma_{12} \neq 0$. Education (ξ_1) affects poverty (η_1). The t-statistics value of 3.700 is greater than the t-table of 1.65 and the p-value of 0.000 is less than 0.1 (significant), meaning that education has an influence on poverty with a negative influence of -0.560 and significant.

H_5: $\gamma_{22} \neq 0$. Demographics (ξ_2) affect poverty (η_1). The t-statistics value of 1.056 is smaller than the t-table 1.65 and the p-value of 0.292 is greater than 0.1 (not significant), meaning that demographics influence poverty with a positive influence of 0.500 but not significant.

H_6: $\gamma_{32} \neq 0$. Employment (ξ_3) affects poverty (η_1). The t-statistics value of 2.888 is greater than the t-table of 1.65 and the p-value of 0.004 is smaller than 0.1 (significant), meaning that employment has an influence on poverty with a negative influence of -0.437 and significant.

H_7: $\gamma_4 \neq 0$. Digital literacy (η_2) affects poverty (η_1). The t-statistics value of 1.028 is smaller than the t-table of 1.65 and the p-value of 0.305 is greater than 0.1 (not significant), meaning that digital literacy has an influence on poverty with a negative influence of -0.107 but not significant.

Poverty (η_1) is influenced by education (ξ_1) with a loading factor of -0.560, and employment (ξ_3) with a loading factor of -0.437. That is, if education increases by one unit while assuming permanent employment, then poverty decreases by 0.560. In addition, if employment increases by one unit assuming permanent education, then poverty decreases by 0.437. No variable has a significant effect on digital literacy and there is no significant effect of digital literacy in mediating the Influence of Education, Demography, and Employment on Poverty in Indonesia.

The results of this research are similar to those conducted by Zhou, et al., according to this research, digital literacy can reduce poverty by increasing the performance and scale of entrepreneurship. Digital literacy is no less important than employment [26]. This study has an important impact on understanding and improving the governance framework for long-term poverty alleviation through digital literacy.

4. Conclusions

The poverty modeling in Indonesia with education, demographics, and employment as factors as well as digital literacy as mediating variables using the PLS approach obtained results that for the measurement model there are indicators that have met the criteria of validity and reliability. The indicators include four indicators of education variables (ξ_1) namely the availability of high school, college availability, the number of students at the high school, and the number of College-level students; two indicators of demographic variables (ξ_2) namely the Human Development Index (HDI) and population density; two indicators of employment variables (ξ_3) namely the percentage of employment to the labor force and the labor force participation rate; three variable indicators of poverty (η_1) namely the Provincial Minimum Wage (UMP), poverty line, and provincial per capita expenditure; and four variable indicators of digital literacy (η_2), namely information and data literacy, communication and collaboration, security in the use of digital technology, and the ability to use technology.

Based on the factoring scheme, the variation in the poverty model (η_1) that can be explained by education, demographics, employment, and digital literacy variables is 62.6 percent, the remaining 37.4 percent is explained by other variables not mentioned in the model. While the variation in the digital literacy model (η_2) that can be explained by education, demographic, and employment variables is only 11.9 percent, the remaining 88.1 percent is explained by other variables not mentioned in the model. The formed structural model is as follows:

$$\eta_1 = -0.560\,\xi_1 + 0.500\,\xi_2 - 0.437\,\xi_3 - 0.107\,\eta_2 + \zeta_1 \quad (25)$$

$$\eta_2 = -0.322\,\xi_1 - 0.090\,\xi_2 - 0.033\,\xi_3 + \zeta_2 \quad (26)$$

The recommendation that can be given based on the results of this research analysis is: that the Indonesian government needs to consider factors found to be significant in reducing poverty. Improving all factors of education and employment will effectively reduce poverty because these variables and poverty variables have an inverse and significant relationship.

Ethics approval

The research has met research ethical rules so that it can be carried out without harm to the research subjects.

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.

References

[1] Central Agency of Statistics (BPS), 'Badan Pusat Statistik', 2023. [Online]. Available: https://www.bps.go.id/subject/23/kemiskinan-dan-ketimpangan.html#subjekViewTab1.

[2] United Nations Development Programme (UNDP), 'Sustainable Development Goals', 2023. [Online]. Available: https://www.undp.org/sustainable-development-goals.
[3] A. Klimczuk, 'Introductory Chapter: Demographic Analysis', Demographic Analysis: Selected Concepts, Tools, and Applications, p. 3, 2021.
[4] S. M. Heathfield, 'Why Human Resources Management Is So Important', The Balance Careers, 2020.
[5] D. R. A. Schallmo and C. A. Williams, 'Digital Transformation Now! Guiding the Successful', 2018.
[6] Nurhafifah, S. Mahdi, and R. Oktavia, Analisis Faktor yang Saling Memengaruhi Kemiskinan di Indonesia Menggunakan Metode Structural Equation Modeling-Partial Least Square (SEM-PLS). Banda Aceh: Universitas Syiah Kuala, 2023.
[7] R. Parianom, I. Arrafi, and D. Desmintari, 'The Impact of Digital Technology Literacy and LifeSkills On Poverty Reduction', Devotion: Journal of Research and Community Service, vol. 3, no. 12, pp. 2002–2007, 2022.
[8] R. H. Hoyle, 'The structural equation modeling approach: Basic concepts and fundamental issues', 1995.
[9] M. Sarstedt, C. M. Ringle, and J. F. Hair, 'Partial least squares structural equation modeling', in Handbook of market research, Springer, 2021, pp. 587–632.
[10] C. M. Stein, N. J. Morris, and N. L. Nock, 'Structural equation modeling', Statistical human genetics: Methods and protocols, pp. 495–512, 2012.
[11] R. Solling Hamid and S. M Anwar, 'Structural Equation Modeling (SEM) Berbasis Varian'. PT Inkubator Penulis Indonesia, 2019.
[12] P. H. Garthwaite, 'An interpretation of partial least squares', Journal of the American Statistical Association, vol. 89, no. 425, pp. 122–127, 1994.
[13] M. Tenenhaus, V. E. Vinzi, Y.-M. Chatelin, and C. Lauro, 'PLS path modeling', Computational statistics & data analysis, vol. 48, no. 1, pp. 159–205, 2005.
[14] G. S. Trujillo, 'Pathmox approach: Segmentation trees in partial least squares path modeling', Universitat Politècnica de Catalunya (UPC), 2009.
[15] A. Anekawati, B. W. Otok, and Others, 'Structural equation modelling with three schemes estimation of score factors on partial least square (Case study: the quality of education level SMA/MA in Sumenep Regency)', in Journal of Physics: Conference Series, 2017, vol. 855, p. 012006.
[16] W. W. Chin and Others, 'The partial least squares approach to structural equation modeling', Modern methods for business research, vol. 295, no. 2, pp. 295–336, 1998.
[17] F. P. A. P. Rachman, R. Goejantoro, and M. N. Hayati, 'Penentuan Jumlah Replikasi Bootstrap Menggunakan Metode Pretest Pada Independent Sampel T Test', EKSPONENSIAL, vol. 9, no. 1, pp. 35–40, 2018.
[18] A. Das, M. Chowdhury, and S. Seaborn, 'ICT diffusion, financial development and economic growth: new evidence from low and lower middle-income countries', Journal of the Knowledge Economy, vol. 9, pp. 928–947, 2018.
[19] P. O. Adesemowo and O. A. Sotonade, 'Basic of education: The meaning and scope of education', Olabisi Onabanjo University, 2022.
[20] O. Giovetti, 'How does education affect poverty? It can help end it. Concern Worldwide'. 2022.
[21] N. J. Spence, 'Family Demography', in Encyclopedia of Gerontology and Population Aging, Springer, 2022, pp. 1789–1792.
[22] F.-B. Wietzke, 'Poverty, inequality, and fertility: the contribution of demographic change to global poverty reduction', Population and Development Review, vol. 46, no. 1, pp. 65–99, 2020.
[23] M. Vaalavuo and O. Sirniö, 'Jobs against poverty: a fixed-effects analysis on the link between gaining employment and exiting poverty in Europe', European Societies, vol. 24, no. 4, pp. 431–462, 2022.
[24] J. Chen, 'What's Poverty? Meaning, Causes, and How to Measure'. 2023.
[25] J. F. Hair Jr, G. T. M. Hult, C. M. Ringle, M. Sarstedt, N. P. Danks, and S. Ray, Partial least squares structural equation modeling (PLS-SEM) using R: A workbook. Springer Nature, 2021.

[26] D. Zhou, F. Zha, W. Qiu, and X. Zhang, ‘Does digital literacy reduce the risk of returning to poverty? Evidence from China’, Telecommunications Policy, vol. 48, no. 6, p. 102768, 2024.

DOI: https://doi.org/10.34123/jurnalasks.v16i1.598

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

Geographically Weighted Poisson Regression for Modeling The Number of Maternal Deaths in Papua Province

Toha Saifudin[1*], Nur Rahmah Miftakhul Jannah[2], Risky Wahyuningsih[3], Gaos Tipki Alpandi[4]

[1,2,3,4] *Departemen Matematika, Fakultas Sains dan Teknologi, Universitas Airlangga, Surabaya, Indonesia*
**Corresponding Author: E-mail address: toha.saifudin@fst.unair.ac.id*

ARTICLE INFO

Article history:
Received 25 June, 2023
Revised 2 May, 2024
Accepted 7 May, 2024
Published 30 June, 2024

Keywords:
Geographically Weighted Poisson Regression (GWPR); Maternal Mortality Rate (MMR); Papua; Sustainable Development Goals (SDGs)

Abstract

Introduction/Main Objectives: Maternal Mortality Rate (MMR) in Indonesia is one of the main focuses in achieving the third Sustainable Development Goals (SDGs) in 2030. **Background Problems:** The Central Statistics Agency states that the MMR in Papua Province is the highest, reaching 565. **Novelty:** Given the diverse geographical conditions of each district/city in Papua Province, an analysis was carried out. **Research Methods:** Using the Geographically Weighted Poisson Regression (GWPR) method with the response variable being maternal mortality rates and variables predictors of health, social, and environmental factors. **Finding/Results:** Fixed Gaussian kernel GWPR is the best model with an AIC value of 27.6. Variable significantly influencing MMR include the percentage of households with access to adequate sanitation, the number of recipients of food assistance programs, and the number of doctors.

1. Introduction

One of the indicators of societal welfare in a country is a low maternal mortality rate (MMR). However, so far, maternal mortality cases in Indonesia remain relatively high compared to neighboring countries. According to data from the Indonesian Health Commission in 2022, the maternal mortality rate was around 183 per 100,000 live births, which is significantly higher compared to Malaysia, with an MMR of 20 per 100,000 live births. The data shows that the number of maternal deaths in 2023 increased to 4,129 compared to 4,005 in 2022. The Indonesian Ministry of Health recorded that the Maternal Mortality Rate (MMR) is still around 305 per 100,000 live births, not yet reaching the target of 183 per 100,000 live births by 2024. This certainly poses a barrier to achieving the third SDG, which is good health and well-being.

Papua is one of the provinces in Indonesia that faces complex health issues. According to a publication by the Ministry Health in 2022, Papua has the highest maternal mortality rate in Indonesia [1]. The 2021 report from the Indonesian Ministry of Health indicated that Papua's maternal mortality ratio was 305 per 100,000 live births, significantly higher than the national average of 102 per 100,000 live births. On the other hand, Papua also has heterogeneous geographical and socio-economic conditions, meaning not all areas in Papua face the same health problems. Therefore, a statistical

modeling approach that accounts for spatial variability in factors affecting maternal mortality in Papua is needed.

Previous research on maternal mortality risk factors was conducted by Rachmah et al. in 2014 using Poisson regression [2]. The results of this study indicated that the Poisson regression analysis encountered overdispersion, and no measures were taken to address the overdispersion issue. Additionally, the significant influence of predictor variables obtained was not relevant to the existing facts because the predictor variables used lacked variability.

The novelty of this research lies in the use of Geographically Weighted Regression (GWR) in modeling the number of maternal deaths. GWR is used in modeling maternal mortality rates to allow for spatially varying regression parameter estimates [3]. However, the assumption of a Gaussian distribution for the response variable in GWR does not always hold in practice, leading to the development of research aims to contribute to achieving the third SDG, which is to improve health and well-being by reducing maternal mortality rates and increasing access to quality healthcare services in Papua Province.

2. Material and Methods

2.1 Maternal Mortality

The Maternal Mortality Ratio (MMR) is the number of women who die from causes related to pregnancy or its management (excluding accidents, suicide, or incidental cases) during pregnancy, childbirth, and the postpartum period (42 days after childbirth) per 100,000 live births.

Maternal mortality, as defined by the International Classification of Diseases 11th (ICD-11), is the death of a woman while pregnant or within 42 days of termination of pregnancy, regardless of the duration and location of the pregnancy, from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes [4].

2.1.1 Factors Influencing Maternal Mortality

Factors contributing to maternal mortality can broadly be categorized into direct and indirect causes. Direct causes of maternal mortality are factors related to complications of pregnancy, childbirth, and the postpartum period, such as hemorrhage, preeclampsia/eclampsia, infection, obstructed labor, and abortion. Indirect causes of maternal mortality include factors that exacerbate the condition of pregnant women, such as the four "too's" (too young, too old, too frequent childbirth, and too close spacing between births).

Additionally, maternal mortality is influenced by pregnant women suffering from infectious diseases such as malaria, HIV/AIDS, tuberculosis, and syphilis; non-communicable diseases such as hypertension, diabetes mellitus, heart disease, mental disorders, and malnutrition [5].

Medical factors such as Hemoglobin (Hb) levels are significant in increasing the risk of maternal mortality. Anemia in pregnant women increases the relative risk of maternal mortality by 15.3 times compared to non-anemic pregnant women. A delay in decision-making also increases the risk of maternal mortality by 50.8 times compared to women who do not experience referral delays [6].

Health service factors, such as delays in medical treatment, can increase the risk of maternal mortality. Referral hospitals may face shortages of blood supplies and medical procedures may be delayed due to the absence of specialists.

2.2 Data and Data Source

The data used is secondary data, specifically the Health Profile Data of Papua Province. This data was obtained from publications by the Papua Provincial Health Office in 2021, with the unit of observation being the regencies/cities in Papua Province.

2.3 Research Variables

The research variables used are defined as follows:

1. The number of maternal deaths by age group in Papua Province in 2021 (Y)

2. The number of families actively participating in the Family Planning (Keluarga Berencana/KB) program in Papua Province in 2021 (X_1)
3. The percentage of households with access to adequate sanitation in Papua Province in 2021 (X_2)
4. The realization of the number of recipients of food Social Assistance (Bantuan Sosial/Bansos) in Papua Province in 2021 (X_3)
5. The number of health workers (doctors) in Papua Province in 2021 (X_4).

2.4 Multicollinearity

Multicollinearity in a regression model can be determined using the Variance Inflation Factor (VIF), which is expressed in the following equation:

$$VIF_j = \frac{1}{1-R_j^2} \tag{1}$$

Where $R_j^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{X}_{ij}-\bar{X}_{ij})^2}{\sum_{i=1}^{n}(X_{ij}-\bar{X}_{ij})^2}$. SSR (Sum SSR (Sum of Squares Regression) represents the variation caused by the relationship between predictor variables. SST (Sum of Squares Total) is a measure of the variation in the values of X from their mean. X_{ij} is the value of X_j in the i[th] observation, $\bar{X}_j$ is the mean value of X_j and $\hat{X}_j$ is the predicted value of X_j in the i[th] observations [7].

2.5 Poisson Regression Model

Poisson regression is used to analyze discrete data, where the response in the data follows a Poisson distribution with parameter μ. The Poisson distribution is then used to model an event that is relatively rare or uncommon to occur within a certain unit of observation.

The Poisson regression model can be expressed as follows:

$$E(y_i) = \mu(x_i, \beta) = exp\left(x_i^T \beta\right) \tag{2}$$

With $x_i^T = (x_{1i}, x_{2i}, \dots, x_{ki})$ and $\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$. Function $\mu(x_i, \beta)$ In the Poisson regression model, it is a function of x_i as the predictor variable and β as the regression parameter to be estimated [8].

One of the methods for estimating Poisson regression parameters is Maximum Likelihood Estimation (MLE). Finding the likelihood equation from the probability mass function of the Poisson distribution is:

$$\begin{aligned} lnL(\beta) =& ln\ \frac{exp\left(-\sum_{i=1}^{n} exp\ \left(x_i^T\beta\right)\ \right)\left(exp\left(\sum_{i=1}^{n} y_i x_i^T \beta\right)\right)}{\prod_{i=1}^{n} y_i!} \\ &= -\sum_{i=1}^{n} exp\left(x_i^T\beta\right)\ +\ \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} ln\left(y_i\right) \end{aligned} \tag{3}$$

The equation is derived concerning β^T, which is in vector form, as it involves multiple parameters in this case.

$$\frac{\partial ln\ L(\beta)}{\partial \beta^T} = -\sum_{i=1}^{n} x_i\ exp\left(x_i^T\beta\right)\ +\ \sum_{i=1}^{n} y_i x_i \tag{4}$$

The equation is equated to zero and then solved using the Newton-Raphson iteration. The iteration stops when the parameter estimates converge [9].

2.6 Geographically Weighted Poisson Regression Model (GWPR)

Geographically Weighted Poisson Regression (GWPR) is a local form of Poisson regression that produces locally parameterized model estimates for each point or observation location assuming Poisson-distributed data [10][11]. The GWPR model can be written as:

$$E(y_i) = \mu(x_i, \beta(u_i, v_i)) = exp\left(x_j^T \beta(u_j, v_j)\right), for\ i = 1,2,\dots,n \quad (5)$$

With $x_i^T = (x_{1i}, x_{2i}, \dots, x_{ki})$ and $\beta = (\beta_1, \beta_2, \dots, \beta_k)^T$. The function $\mu(x_i, \beta(u_i, v_i))$ in the Poisson regression model is a function of x_i as the predictor variable and β as the regression parameter to be estimated.

The GWPR model parameters can be estimated using the MLE method, resulting in the equation:

$$lnL(\beta) = -\sum_{i=1}^{n} exp\ (x_i^T \beta)\ +\ \sum_{i=1}^{n} y_i x_i^T \beta - \sum_{i=1}^{n} ln\ (y_i)! \quad (6)$$

Spatial factors, such as geographic location, act as weighting factors in the GWPR model. These factors have different values for each region, indicating the local nature of the model. Therefore, the weighting is incorporated into the log-likelihood form of the GWPR model, thus obtained:

$$ln\ L^*(\beta(u_i, v_i)) = \sum_{j=1}^{n} \left(y_j x_j^T \beta(u_i, v_i) - exp\left(x_j^T \beta(u_i, v_i)\right) - ln\ y_j!\right) w_{ij}(u_i, v_i)\ (2.8) \quad (7)$$

Then, the estimation of the parameter $\beta(u_i, v_i)$ is obtained by differentiating the equation, leading to:

$$\frac{\partial ln L^*(\beta(u_i, v_i))}{\partial \beta^T(u_j, v_j)} = \sum_{j=1}^{n} \left(y_j x_j - x_j exp\left(x_j^T \beta(u_j, v_j)\right)\right) w_{ij}(u_i, v_i) \quad (8)$$

The equation is equated to zero and then solved using the Newton-Raphson iteration. The iteration stops when the parameter estimates converge [12].

2.7 Partial Parameter Significance Testing

Model parameter testing is conducted by testing parameters individually. This test aims to determine which parameters significantly affect the response variable with the hypothesis:

$H_0 : \beta_k(u_i, v_i) = 0 ; i = 1,2,\dots,n ; k = 1,2,\dots,p$

$H_1 : \beta_k(u_i, v_i) \neq 0$

In the hypothesis test above, the following test statistics can be used:

$$t = \frac{\beta_k(u_i, v_i)}{se(\beta_k(u_i, v_i))} \quad (9)$$

The value of $se(\beta_k(u_i, v_i))$ or the standard error of $\beta_k(u_i, v_i)$ is obtained by taking the square root of $var(\beta_k(u_i, v_i))$

$$se(\beta_k(u_i, v_i)) = \sqrt{var(\beta_k(u_i, v_i))} \quad (10)$$

The testing criterion is to reject H_0 if $|t_{count}| > t_{\frac{\alpha}{2};n-(p+1)}$

2.8 Model Goodness-of-Fit Test

Model goodness-of-fit is performed to compare the global goodness-of-fit with the model considering spatial elements. One method used to select the model is Akaike's Information Criterion (AIC), which can be defined as follows:

$$AIC(h) = D(h) + 2K(h) \quad (11)$$

Where $D(h)$ is the deviance value of the model with bandwidth h, and $K(h)$ is the number of parameters in the bandwidth model.

2.9 Analysis Procedures

The steps of data analysis conducted are as follows:

1. Describing the Maternal Mortality Rate (MMR) in Papua Province and the various influencing variables.

2. To understand the description of regencies/cities in Papua Province based on the research variables, thematic maps of Papua Province are used to describe the dependent variable (Y) and the predictor variables from a regional perspective.
3. Analyzing the Poisson distribution assumption on the dependent variable.
4. Conducting a Multicollinearity test with the criterion that if the Variance Inflation Factor (VIF) value exceeds 10, multicollinearity is indicated.
5. Conducting Poisson regression modeling.
   a. Estimating the parameters of the Poisson regression model.
   b. Conducting a simultaneous test of Poisson regression parameters using the Maximum Likelihood Ratio Test (MLRT).
   c. Conducting a partial test of Poisson regression parameters with the criterion $|Z_{count}| > Z_{\alpha/2}$.
6. Analyzing the GWPR model as follows:
   a. Calculating the Euclidean distance between observation locations based on geographic positions. The Euclidean distance between location i at coordinates (u_i, v_i) toward location j at coordinates (uj, vj).
   b. Sorting the Euclidean distances of all locations relative to location i to obtain the nearest neighbor order of location i.
   c. Determining the optimal bandwidth.
   d. Calculating the weighting matrix using the kernel weighting function.
   e. Estimating the parameters of the GWPR model.
   f. Testing the significance of the GWPR regression model parameters partially.
   g. Calculating the AIC value of the GWPR model.
   h. Testing the model fit between the Poisson regression model and the GWPR model using the deviance values of each model and calculating the $F_{count.}$
7. Determining the best model between Poisson Regression, GWPR-Fixed Bisquare, and GWPR-Fixed Gaussian based on the smallest AIC value. AIC is chosen as the criterion for the best model selection as it measures the quality of the model by balancing model fit and data complexity [13].
8. Concluding to achieve the research objectives

3. Result and Discussion

3.1 Data Description

Before conducting further analysis, a simple analysis with descriptive statistics was performed. Descriptive statistics methods are used to get a general overview of the data. used. The following are the results of the descriptive statistics as shown in Table 1.

Table 1. Descriptive Statistics Results

Variable	Mean	Std.Dev	Min	Max
$\boldsymbol{Y}$	5.4	3.089	1	10
$\boldsymbol{X_1}$	9264	12280	76	50876
$\boldsymbol{X_2}$	53.39	28.34	2.67	85.8
$\boldsymbol{X_3}$	13011	10036	2327	42948
$\boldsymbol{X_4}$	76.3	98.4	5	393

In Table 1, it can be seen that the average number of maternal deaths *(Y)* in 15 regencies/cities in Papua Province is 5.4. This figure means that there are at least 5 maternal deaths per 100,000 live births in each regency/city in Papua Province. However, this result only represents a portion of the regions in Papua Province, as other regions did not report these cases to the local authorities. In Figure 1, it is shown that the highest number of maternal deaths reaches 10 in Merauke City and Biak Numfor Regency. Meanwhile, the lowest number of maternal deaths is 1, found in Puncak Jaya Regency and Supiori Regency. Furthermore, the number of families actively participating in the Family Planning (Keluarga Berencana/ KB) program (X_1), the number of food Social Assistance (Bantuan Sosial/Bansos) recipients (X_3), and the number of health workers (doctors) (X_4) are 138,959, 195,161, and 1,145,

respectively. The average percentage of households with access to adequate sanitation (X_2) is 53.39%. The descriptive statistics results above generally show that the health, environmental, and social sectors in Papua Province are relatively low.

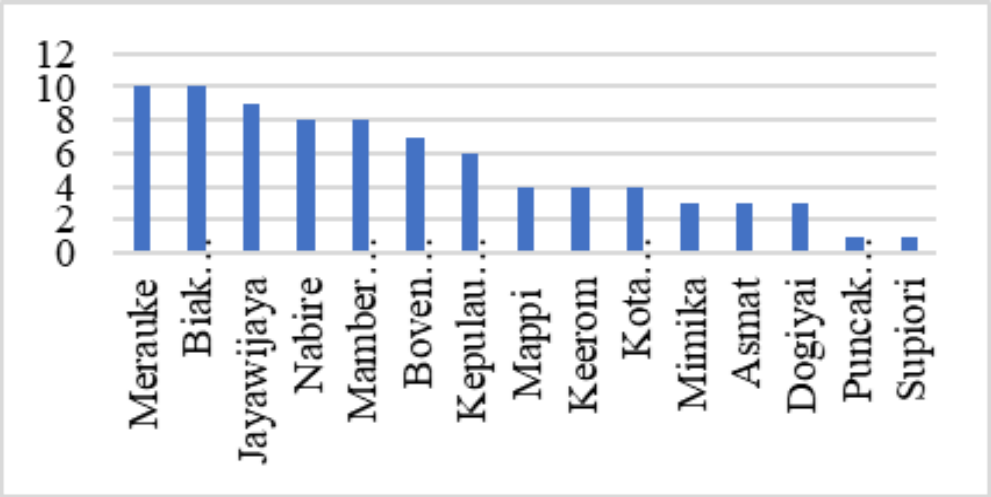


Figure 1. The Distribution of Maternal Mortality Rates in Papua Province 2021

3.2 Poisson Distribution Assumption of the Dependent Variable

The GWPR (Geographically Weighted Poisson Regression) modeling assumes that the dependent variable (Y_i) must follow a Poisson distribution. Therefore, a test was conducted on the variable Y, which represents the number of maternal deaths in Papua Province for each district/city. The following are the results of the Poisson distribution test.

Table 2. Results of the Poisson Distribution Test

DF	*Chi-Square*	*P-Value*
2	4.92993	0.085

Based on Table 2, a *p-value* of 0.085 was obtained, leading to the decision to accept H_0. It can be concluded that the maternal mortality data per district/city in Papua Province follows a Poisson distribution. Therefore, the variable Y_i meets the assumption and can proceed to the multicollinearity assumption test.

3.3 Multicollinearity Test

One of the criteria that must be met in conducting Poisson regression analysis is the detection of multicollinearity. The presence of multicollinearity can be determined based on the Variance Inflation Factor (VIF) values. A VIF value for any predictor variable exceeding 10 indicates the presence of multicollinearity. The following are the results of the multicollinearity test.

Table 3. VIF Values of Predictor Variables

Variables	VIF
$\boldsymbol{X_1}$	3.329252
$\boldsymbol{X_2}$	1.610570
$\boldsymbol{X_3}$	4.312278
$\boldsymbol{X_4}$	1.158255

Based on Table 3, it is concluded that none of the predictor variables indicate the presence of multicollinearity, as all VIF values are less than 10. Therefore, the analysis can proceed with Poisson regression and GWPR modeling.

3.4 Poisson Regression Modeling

The following are the estimated parameter values for the Poisson regression model.

Table 4. Parameter Estimates for the Poisson Regression Model

Parameter		Estimate	Z-count
$\boldsymbol{\beta_0}$		0.833145	2.246126
$\boldsymbol{\beta_1}$		0.000028	1.556444
$\boldsymbol{\beta_2}$		0.010092	1.861479
$\boldsymbol{\beta_3}$		0.000027	2.607824
$\boldsymbol{\beta_4}$		-0.004527	-1.745916
Deviance	18.834		
AIC	28.833900		

Table 4 shows that the AIC and deviance values for the Poisson regression model are 28.83 and 18.834, respectively. After obtaining the estimated parameters of the Poisson regression model, simultaneous and partial tests were conducted as follows:

1. Simultaneous Test of Poisson Regression Parameters

In the simultaneous test, the deviance value of the model is 18.834, with a critical region of $D > \chi^2(10; 0.1)$ or $D > 15.987$. Therefore, we reject H_0 and conclude that there is at least one $\beta_j \neq 0$, indicating that at least one parameter significantly affects the Poisson regression model.

2. Partial Test of Poisson Regression Parameters

The partial test uses the criterion $|Z_{count}| > Z_{\alpha/2}$. With $\alpha = 10\%$, the critical region is $|Z_{count}| > 1.645$. It can be concluded that the parameters significantly affecting the Poisson regression model are β_2, β_3, and β_4 as they have $|Z_{count}| = 1.645$ or *p-value* < 0.1. Generally, the Poisson regression model for maternal mortality in Papua Province can be written as follows:

$$\hat{\mu} = exp\ (0.833145 + 0.000028X_1 + 0.010092X_2 + 0.000027X_3 - 0.004527X_4)$$

3.5 GWPR Modelling

Parameter testing for the GWPR (Geographically Weighted Poisson Regression) model is conducted to determine the significance of parameter β with the following hypotheses:

$H_0: \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \beta_3(u_i, v_i) = \beta_4(u_i, v_i) = 0$

H_1: At least one parameter $\beta_p(u_i, v_i) \neq 0, p = 1,2,3,4$

Based on Table 5, the deviance value of the GWPR model is 15.512. With a significance level of 10%, the critical region for rejecting H_0 is $D > \chi^2(8.407; 0.1)$ or $D > 13.902$. Therefore, we reject H_0 indicating that at least one predictor variable significantly affects the GWPR model. The significance of the GWPR model parameters is then tested partially to identify predictors that significantly affect each district/city with the following hypotheses:

$H_0: \beta_p(u_i, v_i) = 0, i = 1,2, \dots, 15, p = 1,2,3,4$

$H_1: \beta_p(u_i, v_i) \neq 0$

Table 5. Parameter Test of GWPR Model in Merauke City with Fixed Bisquare Kernel

Parameter		*Estimate*	*Z-count*
$\boldsymbol{\beta_0}$		0.979567	2.521377
$\boldsymbol{\beta_1}$		0.000016	0.909077
$\boldsymbol{\beta_2}$		0.011295	0.006098
$\boldsymbol{\beta_3}$		0.000022	1.992033
$\boldsymbol{\beta_4}$		-0.003585	-1.396888
Deviance	15.512		
AIC	27.583728		

A predictor variable is considered to significantly affect the model if it has $|Z| > Z_{\alpha/2}$ or $|Z| > 1.645$. Based on the results, the percentage of households with access to proper sanitation (X_2), the number of food aid recipients (X_3), and the number of healthcare workers (X_4) significantly affect maternal mortality in Papua Province. In contrast, the number of families actively participating in the family planning program (X_1) does not significantly affect the model. The significant variables in the GWPR model with the Fixed Bisquare kernel for each district/city in Papua Province are grouped as follows:

1. The variables $X_1, X_2, X_3,\ and\ X_4$ are significant for the following districts/cities: Jayawijaya, Yapen Island, Biak Numfor, Puncak Jaya, Mimika, Supiori, Mamberamo Tengah, and Dogiyai.
2. The variables $X_2, X_3, and\ X_4$ are significant only for Asmat District.
3. The variables $X_2\ and\ X_3$ are significant for the following districts/cities: Merauke, Boven Digoel, and Mappi.
4. The variables $X_3\ and\ X_4$ are significant for the following districts/cities: Nabire, Keerom, and Jayapura City.

Below is a visualization of the grouping of significant variables for each district/city in Papua Province in the GWPR model with the Fixed Bisquare kernel.

Figure 2. Mapping of Significant Variables in the GWPR Model with Fixed Bisquare Kernel

Partial parameter testing was conducted using the first research location (u_1, v_1), which is Merauke City, as shown in Table 5. In this table, it is evident that the only significant variable is parameter β_3 with $|Z| > Z_{\alpha/2}$. The GWPR model for maternal mortality in Merauke City can be formulated as follows:

$$\hat{\mu} = exp\ (0.979567 + 0.000016X_1 + 0.011295X_2 + 0.000022X_3 - 0.003585X_4)$$

Based on the GWPR model for Merauke City, it can be interpreted that each increase in the number of food aid recipients (X_3) will increase maternal mortality (Y) in Merauke City by a factor of $exp\ exp\ (0.000022) = 1.000022$, assuming other variables remain constant. Additionally, each increase in the number of healthcare workers (X_4) will decrease maternal mortality in Merauke City by a factor of $exp\ exp\ (0.003585) = 1.00359$ assuming other variables remain constant. The same

interpretation method applies to the percentage of households with access to proper sanitation and the number of families actively participating in the family planning program.

3.6 Determining the Best Model

The criterion for determining the best model is the AIC value. The smaller the AIC value, the better the model is considered to be. The following are the AIC values for the Poisson regression model, GWPR with Fixed Bisquare kernel, and GWPR with Fixed Gaussian kernel.

Table 6. AIC Values

Regression Model	AIC
Poisson Regression	28.8
GWPR – Fixed Bisquare	27.6
GWPR – Fixed Gaussian	30.6

The AIC value for the Poisson regression model is 28.83, the AIC value for the GWPR model with Fixed Bisquare kernel is 27.58, and the AIC value for the GWPR model with Fixed Gaussian kernel is 30.64. From these results, it can be concluded that the GWPR model with a Fixed Bisquare kernel is the best model based on the smallest AIC value among the three models.

The GWPR model with Fixed Bisquare kernel for maternal mortality in 14 districts/cities in Papua Province can be seen in Table 7.

Table 7. GWPR Model for Maternal Mortality in Papua Province

District/City	GWPR Model
Merauke	$\hat{\mu} = exp(0.979567 + 0.000016X_1 + 0.011295X_2 + 0.000022X_3 + -0.003585X_4$
Jayawijaya	$\hat{\mu} = exp(0.903864 + 0.000031X_1 + 0.009243X_2 + 0.000024X_3 + -0.00492X_4)$
Nabire	$\hat{\mu} = exp(0.97635 + 0.00003X_1 + 0.008222X_2 + 0.000024X_3 + -0.004838X_4)$
Yapen Island	$\hat{\mu} = exp(0.695837 + 0.000047X_1 + 0.010478X_2 + 0.000028X_3 + -0.006554X_4$
Biak Numfor	$\hat{\mu} = exp(0.636394 + 0.00005X_1 + 0.010973X_2 + 0.000029X_3 + -0.006929X_4)$
Puncak Jaya	$\hat{\mu} = exp(0.849427 + 0.000035X_1 + 0.009494X_2 + 0.000025X_3 + -0.005358X_4$
Mimika	$\hat{\mu} = exp(0.8062 + 0.000037X_1 + 0.010116X_2 + 0.000026X_3 + -0.00545X_4)$
Boven Digoel	$\hat{\mu} = exp(1.011314 + 0.000021X_1 + 0.009252X_2 + 0.000022X_3 + -0.003898X_4$
Mappi	$\hat{\mu} = exp(0.965983 + 0.000023X_1 + 0.009879X_2 + 0.000023X_3 + -0.004121X_4$
Asmat	$\hat{\mu} = exp(0.907459 + 0.000029X_1 + 0.009657X_2 + 0.000024X_3 + -0.004691X_4$
Keerom	$\hat{\mu} = exp(0.988211 + 0.000028X_1 + 0.00838X_2 + 0.000023X_3 + -0.004572X_4)$
Supiori	$\hat{\mu} = exp(0.590476 + 0.000052X_1 + 0.011448X_2 + 0.00003X_3 + -0.007097X_4)$
Mamberamo Tengah	$\hat{\mu} = exp(0.906912 + 0.000032X_1 + 0.00909X_2 + 0.000024X_3 + -0.004984X_4)$
Dogiyai	$\hat{\mu} = exp(0.736044 + 0.000042X_1 + 0.010528X_2 + 0.000027X_3 + -0.005965X_4$

The high maternal mortality rate in Papua Province in 2021 indicates that the achievement of sustainable development in the region is still not optimal. One of Indonesia's development targets by 2030 is to reduce the maternal mortality ratio to less than 70 per 100,000 live births. BPS also mentioned that the high maternal mortality rate in Papua is influenced by various factors including health status, education, economy, socio-cultural aspects, and healthcare services in the region [14]. Therefore, this issue should be a concern for the government to further enhance equity in various aspects, especially in Eastern Indonesia such as Papua, so that maternal mortality issues and others can be reduced.

4. Conclusion

The average maternal mortality rate in each district/city in Papua Province was 5.4 in 2021, with the highest mortality of 10 cases occurring in Merauke City and Biak Numfor District, and the lowest mortality of 1 case in Puncak Jaya District and Supiori District. The best kernel function chosen was Fixed Bisquare with an AIC value of 27.58. Based on the analysis results, it was found that the percentage of access to proper sanitation, the number of food aid recipients, and the number of healthcare workers significantly affected the maternal mortality rate with an error rate of 10%.

Ethics approval

Not required.

Acknowledgments

The author extends heartfelt thanks to Dr. Toha Saifudin, S.Si., M.Si., for his invaluable support and encouragement. As the lecturer of the Spatial Data Analysis course at the Faculty of Science and Technology, Universitas Airlangga, his guidance was crucial to the completion of this research.

Competing interests

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

Funding

This study received no external funding.

Underlying data

The data used was obtained from publications by the Papua Provincial Health Office in 2021, with the unit of observation being the regencies/cities in Papua Provinces. Data can be accessed via the following link: https://dinkes.papua.go.id/informasi-publik/informasi-berkala/.

References

[1] Kementerian Kesehatan RI, *Profil Kesehatan Indonesia Tahun 2021*. Jakarta: Kementerian Kesehatan RI, 2022.

[2] N. F. Rachmah and P. Purhadi, 'Pemodelan Jumlah Kematian Ibu dan Jumlah Kematian Bayi di Provinsi Jawa Timur Menggunakan Bivariate Poisson Regression', *Jurnal Sains dan Seni ITS*, vol. 3, no. 2, pp. D194–D199, 2014.

[3] A. S. Fotheringham, C. Brunsdon, and M. E. Charlton, 'Geographically weighted regression', *The Sage handbook of spatial analysis*, vol. 1, pp. 243–254, 2009.

[4] World Health Organization, 'Publications: Trends in Maternal Mortality 2000 to 2020: Estimates by WHO, UNICEF, UNFPA, World Bank Group and UNDESA/Population Division', Feb-2023. [Online]. Available: https://www.who.int/.

[5] Kementerian Kesehatan Republik Indonesia, 'Peraturan Menteri Kesehatan Republik Indonesia Nomor 97 Tahun 2014 tentang Pelayanan Kesehatan Masa Sebelum Hamil, Masa Hamil, Persalinan, dan Masa Sesudah Melahirkan, Penyelenggaraan Pelayanan Kontrasepsi, Serta Pelayanan Kesehatan Seksual'. 2014.

[6] S. Juharni, I. K. T. Widarsa, and D. N. Wirawan, 'Faktor risiko kematian ibu sebagai akibat komplikasi kehamilan, persalinan dan nifas di Kabupaten Bima tahun 2011â€"2012', *Public health and Preventive medicine Archive*, vol. 1, no. 2, pp. 96–102, 2013.

[7] S. W. Tyas, L. A. Puspitasari, and Others, 'Geographically weighted generalized poisson regression model with the best kernel function in the case of the number of postpartum maternal mortality in east java', *MethodsX*, vol. 10, p. 102002, 2023.

[8] D. Darnah, 'Menentukan Model Terbaik dalam Regresi Poisson dengan Menggunakan Koefisien Determinasi', *Jurnal Matematika, Statistika dan Komputasi*, vol. 6, no. 2, pp. 59–71, 2010.

[9] D. Noviani, R. Wasono, and I. M. Nur, 'Geographically Weighted Poisson Regression (GWPR) untuk Pemodelan Jumlah Penderita Kusta di Jawa Tengah', *Jurnal Statistika Universitas Muhammadiyah Semarang*, vol. 2, no. 2, 2014.

[10] F. Ling *et al.*, 'Geographically weighted Poisson regression for exploring spatial variations in maternal mortality in Papua Province, Indonesia', *International Journal of Environmental Research and Public Health*, vol. 15, no. 9, p. 1845, 2018.

[11] C. Brunsdon, A. S. Fotheringham, and M. E. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. John Wiley & Sons, 2002.

[12] I. N. Septiani, 'Pemodelan Jumlah Kematian Bayi dengan Pendekatan Geographically Weighted Poisson Regression (GWPR) Studi Kasus di Provinsi Kalimantan Barat', *Bimaster: Buletin Ilmiah Matematika, Statistika dan Terapannya*, vol. 10, no. 1, 2021.

[13] T. W. Arnold, 'Uninformative parameters and model selection using Akaike's Information Criterion', *The Journal of Wildlife Management*, vol. 74, no. 6, pp. 1175–1178, 2010.

[14] E. F. Santika, 'Papua Jadi Provinsi dengan Angka Kematian Ibu Terbanyak 2020-2022', 2023. [Online]. Available: https://databoks.katadata.co.id/.Kementerian Kesehatan RI, *Profil Kesehatan Indonesia Tahun 2021*. Jakarta: Kementerian Kesehatan RI, 2022.

DOI: https://doi.org/10.34123/jurnalasks.v16i1.587

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

Application of Geographically Weighted Logistic Regression in Modeling the Human Development Index in East Java

Toha Saifudin[1], Leni Sartika Panjaitan[2], Sabrina Falasifah[3*], Yan Dwi Pracoko[4]

[1,2,3,4]*Airlangga University, Surabaya, Indonesia*
**Corresponding Author: E-mail address: sabrina.falasifah-2020@fst.unair.ac.id*

ARTICLE INFO

Article history:
Received 13 June, 2023
Revised 03 May, 2024
Accepted 07 May, 2024
Published 30 June, 2024

Keywords:
AIC; HDI; GWLR Method; Fixed Gaussian Kernel; Logistic Regression

Abstract

Introduction/Main Objectives: pinpoint the factors influencing HDI, taking into consideration location and spatial factors. **Background Problems:** The Human Development Index (HDI) in East Java often fails to reflect actual conditions accurately, as disparities exist among districts and cities, with some falling below government expectations. **Novelty**: GWLR extends logistics regression by incorporating spatial factors, allowing for the identification of regional differences and influential factors affecting HDI based on actual data. **Research Methods:** To address this issue, the Geographically Weighted Logistic Regression (GWLR) method is employed. The independent variables used are Expected Years of Schooling (X_1), Open Unemployment Rate (X_2), and Morbidity Rate (X_3) in 2021, while dependent variable is the Human Development Index (Y). **Finding/Results:** The study reveals that GWLR provides a superior model compared to Ordinary Logistic Regression, indicated by a lower Akaike Information Criterion (AIC) of 28.72. Additionally, the GWLR model with Fixed Gaussian Kernel weights outperforms other weighting methods. At 90% confidence level, the significant variables influencing HDI are expected years of schooling (X_1) and the open unemployment rate (X_2). Given the relatively low HDI in Indonesia, the East Java Government should focus on improving these key areas to enhance HDI across districts and cities in the region.

1. Introduction

The Sustainable Development Goals (SDGs) consist of 17 goals that must be attained [1]. Within the realm of sustainable development, human development is paramount. The SDGs address human development, specifically the third goal (ensuring healthy lives and promoting well-being), the fourth goal (ensuring inclusive and equitable quality education), and the seventh goal (promoting sustained, inclusive, and sustainable economic growth).

The United Nations Development Program (UNDP) defines human development as a continuous process that involves making choices that contribute to a long and healthy life, acquiring knowledge, and living a decent life with access to essential resources. The Human Development Index (HDI) serves as a metric to assess a community's level of development in terms of health, education, income, and other related factors. HDI achievements are categorized into four groups: very high (with values of 80

and above), high (with values between 70 and below 80), medium (with values between 60 and below 70), and low (with values below 60) [2].

HDI in East Java varies significantly across districts/cities due to distinct location characteristics and differing development priorities. Despite the annual increase in HDI, the growth does not entirely mirror the true state of HDI in the region. Disparities exist among districts/cities, with some falling below government expectations.

Geographically Weighted Logistic Regression (GWLR) is a variant of logistic regression that incorporates spatial or location factors when predicting a categorical dependent variable [3]. By incorporating a weighting function, the model takes into account the geographical location of the observed data. The local nature of GWLR allows it to address the issue of spatial heterogeneity, making it a valuable tool in overcoming this challenge.

Indah Manfaati Nur and M. Al Haris [4] conducted a study on the factors affecting HDI in Central Java using the Geographically Weighted Logistic Regression (GWLR) method. The research revealed that the influencing factors included the number of health facilities, literacy rate, morbidity rate, and open unemployment rate. Additionally, Lili et al. [5] conducted a similar study on the factors influencing HDI in Kalimantan, also utilizing the Geographically Weighted Logistic Regression (GWLR) method. Their findings indicated that the influencing factors consisted of the percentage of open unemployment, the percentage of the population with university degrees, the percentage of the poor, and the number of health workers, including doctors, midwives, nurses, and pharmacists.

Based on this explanation, a study was conducted in East Java to model HDI using the Geographically Weighted Logistic Regression (GWLR) method. This modeling aimed to pinpoint the factors influencing HDI, taking into consideration location and spatial factors. The distinct models for each location enable researchers to gain a more specific insight into the challenges faced in different areas. The goal of this research is to assist the East Java Government in enhancing these influencing factors to elevate HDI levels in every district/city in the province and work towards achieving the third, fourth, and eighth Sustainable Development Goals (SDGs).

2. Material and Methods

2.1 Human Development Index (HDI)

The Human Development Index (HDI) serves as a holistic indicator that measures human development by considering three fundamental aspects of well-being: health, education, and living standards. Established by the United Nations Development Program (UNDP), this index utilizes a range of metrics to assess development achievements, including average years of schooling, life expectancy, school enrollment rates, and per capita income levels [6]. To gauge the level of human development in different countries and regions worldwide, the commonly used metric is the Human Development Index (HDI). This index employs a three-dimensional geometric average, with each dimension being standardized on a scale of 0 to 1. A higher HDI value indicates a higher level of human development. Unlike more conservative measures such as income and economic growth, the HDI acknowledges that factors like healthcare, education, access to social services, and others significantly impact human well-being.

Development goes beyond the mere enhancement of per capita income; it also encompasses various other facets of society. Relying solely on Gross Domestic Product (GDP) growth is inadequate when addressing human development. It is imperative to consider additional factors such as societal challenges, shifts in people's perceptions and behaviors, and more [7-8]. The quality of Human Resources (HR) stands out as a critical factor for successful development. Human resources, as a development target, serve as a driving force for progress that impacts the success of development initiatives. With proficient human resources, training programs could be effectively carried out. Competent human resources play a crucial role in a country's development [9-10-11].

In 2021, East Java recorded an HDI of 71.4, signifying a substantial level of human development in the region. The HDI value for East Java is calculated based on data obtained from diverse sources, including BPS, the Ministry of Health, the Ministry of Education and Culture, and other relevant institutions. These sources supply data on critical indicators such as life expectancy at birth, years of schooling, and per capita income. The 2021 East Java HDI report underscores the significant progress made by the province across several domains. For instance, life expectancy at birth in East Java has increased from 69.8 years in 2015 to 70.9 years in 2021. Moreover, the literacy rate for individuals aged 15 and above has shown an upward trend, rising from 96.2% in 2015 to 98.2% in the period of 2015-2021 [12].

Several factors impact the attainment of the HDI growth goal in East Java [13-14].

1. Expected Years of Schooling

   BPS report highlights that the Expected Years of Schooling (HLS) represents the projected length of education (in years) that children of a specific age group are expected to undergo in the future [12]. The significance of expected schooling duration in shaping educational development at each level is duly acknowledged by considering the expected duration of schooling for individual children. This underscores the commitment to providing quality education to all children [15].

2. Open Unemployment Rate

   The open unemployment rate specifically denotes the proportion of individuals who are not currently engaged in any form of employment. Unemployment could be attributed to various factors, and one of the significant causes is a decline in economic growth. Additionally, the decline in industrial development and the substitution of human labor with advanced technology directly contribute to unemployment [16].

3. Pain Rate

   The pain rate, as stipulated in Health Law no. 28, refers to the numerical or proportional representation of individuals afflicted with a particular illness within a specific community over a designated timeframe [17]. Morbidity statistics serve as an accurate reflection of the prevailing circumstances, as they exhibit a strong correlation with factors such as local poverty levels, living conditions, the standard of potable water, and the quality of healthcare services [18].

2.2 Data Source

This research makes use of secondary data obtained from BPS East Java [12]. The focus of this study comprises all districts and cities within East Java in 2021, encompassing a total of 38 districts/cities.

2.2.1 Research Variable

The research encompasses two categories of variables: response variables and predictor variables. The specifics are outlined as follows:

1. Response Variables

   The dependent variable (Y) represents the Human Development Index (HDI) of all districts/cities in East Java in 2021. This variable is measured on a nominal scale. According to the classification by BPS (2014), HDI is considered low if it is below 60, medium if it is $60 \leq \text{IPM} < 70$, high if it ranges from $70 \leq \text{IPM} < 80$, and very high if it is $\text{IPM} \geq 80$. For this research, binary responses are utilized, allowing the classification based on the BPS criteria as follows:

   $0 = \text{Low if the HDI number} < 70$

   $1 = \text{High if the HDI number} \geq 70$

2. Predictor Variables

   The data used as predictors corresponds to the year 2021. Table 1 provides an inventory of several predictor variables that are significant for this specific research.

Table 1. Research predictor variables

Var	Variable Name	Data Type
X_1	Expected Years of Schooling	Continuous
X_2	Open Unemployment Rate	Continuous
X_3	Pain Rate	Continuous

2.3 Binary Logistic Regression Analysis

Logistic regression is a statistical technique commonly used to determine the association between a categorical dependent variable and one or more independent variables [19-20]. Unlike linear regression, logistic regression utilizes a binary dummy variable for the dependent variable, eliminating the need for normality, heteroscedasticity, or autocorrelation assumption tests [21]. However, it cannot handle multicollinearity, which occurs when predictor variables counteract each other. Despite this limitation, logistic regression offers the advantage of Y estimating a wide range of values for the

dependent variable, potentially extending beyond the conventional 0 to 1 range. In the case of a dependent variable encompassing two categories with values of 0 and 1, binary logistic regression is employed. When utilizing a model, the response variable adheres to a Bernoulli distribution.

$$f(y_i) = \pi_i(x_i)^{y_i}(1-\pi_i(x_i))^{1-y_i} \quad (1)$$

In the context provided, π_i signifies the probability assigned to the i event, while y_i refers to the ith random variable that could take on the values of 0 or 1.

The multivariable logistic regression model is created by representing the value of E(Y=1|X) as π(x), leading to the following equation:

$$\pi(x_i) = \frac{exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k)}{1+exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_kx_k)} \quad (2)$$

where k denotes the number of predictor variables

Equation (2) could be represented as equation (3) in the following manner:

$$\pi(x) = \frac{exp(g(x))}{1+exp(g(x))} \quad (3)$$

To facilitate the estimation of regression parameters, the logistic regression model could be converted into the π(x) equation (3). This transformation leads to the logit form of logistic regression, resulting in the following equation:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k \quad (4)$$

Estimating unknown parameters in logistic regression involves the utilization of the *Maximum Likelihood Estimation* (MLE) technique. This method focuses on maximizing the likelihood function to estimate the β parameter, necessitating that the data follows a specific distribution. In binary logistic regression, every observation satisfies a *Bernoulli distribution*, allowing for the derivation of a *likelihood* function. The following section presents a systematic description of the *likelihood* function for a binary logistic regression model.

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad (5)$$

To simplify calculations, the likelihood function is maximized as a log function so that from equation (6) the *log-likelihood* function is obtained as follows:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}[y_i\, ln\,\pi(x_i) + (1-y_i)\, ln(1-\pi(x_i)) \quad (6)$$

Obtaining estimates for the parameter vector β involves maximizing the *likelihood* function. This is done by equating the first derivative of the likelihood function for each parameter to zero. $\left(\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0\right)$, It follows from equation (7) that:

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta_j}} = \sum_{i=1}^{N} X_{ij}\,(y_i - \pi(x_i)) = 0 \quad (7)$$

given j=0,1,2,…,k, it follows that X_{i0}=1. Due to the non-linear nature of the derivative of $\ell(\boldsymbol{\beta})$ concerning $\boldsymbol{\beta_j}$ for j=0,1,2,…,k, the *Newton-Raphson* method is utilized to calculate the estimate of $\boldsymbol{\beta}$ denoted as $\widehat{\boldsymbol{\beta}}$.

2.4 Testing Spatial Assumptions

Spatial influences in data are determined through various testing methods, including the examination of spatial dependence and spatial heterogeneity. The *Breusch-Pagan* test method is

employed to assess spatial heterogeneity, whereas the *Morans'I* test method is utilized to evaluate spatial dependence. In the case of partial testing using the *Breusch-Pagan* test, the hypothesis is as follows:
$H_0 = \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2 = \sigma$
$H_1 =$ At the very least, there is one $\sigma_i^2 \neq \sigma$

The test statistics applied in the *Breusch-Pagan* test are.

$$BP = \left(\frac{1}{2}\right) \boldsymbol{f}^T \boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{f} \tag{8}$$

The vector components of $\boldsymbol{f}$ are given by $f_i = \left(\frac{e_i}{\sigma^2} - 1\right)$, where $e_i = y_i - \hat{y}$ and σ^2 represents the variance of y. On the other hand, $\boldsymbol{Z}$ is an $n \times p$ matrix that includes standardized normal vectors for each observation. The critical region, denoted as H_0, is rejected when the BP test statistic value exceeds $BP > \chi^2_{(p)}$, suggesting spatial heterogeneity in the data.

In the context of spatial dependency testing using the *Morans'I* test method, the null hypothesis H_0 could be formulated as follows:
H_0 : The data does not exhibit any spatial dependency.
H_1 : The data exhibits spatial dependency.

The *Morans'I* test employs the following test statistics for conducting the analysis.

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}} \tag{9}$$

where is the value $E(I) = -\frac{1}{n-1}$ and $var(I) = \frac{(n^2S_1 - nS_2 + 3W^2)}{W^2(n^2-1)} - [E(I)]^2$. Spatial dependency in the data is inferred when the test statistic value $|Z| = Z_{\frac{\alpha}{2}}$, leading to the rejection of the critical region H_0.

2.5 Geographically Weighted Logistic Regression (GWLR) Model

Geographically Weighted Logistic Regression (GWLR) is a regression analysis technique that incorporates spatially varying coefficients into a logistic regression model [22]. Unlike traditional logistic regression models where coefficients are assigned to all observations, GWLR considers location dependence of coefficients to explain spatial heterogeneity in the data [23]. The GWLR method incorporates the geographic location of areas by using a weighting function, where each observation is assigned a weight (w_{ij}). Thus, the resulting model from equation (2) could be expressed as follows:

$$\pi(x_i) = \frac{exp(\sum_{k=0}^{p} \beta_k(u_i, v_i)x_{ik})}{1 + exp(\sum_{k=0}^{p} \beta_k(u_i, v_i)x_{ik})} \tag{10}$$

where the regression coefficient $\beta_k(u_i, v_i)$ denotes the impact of the predictor variable x_{ik} at a particular location (u_i, v_i).

In the case of Geographically Weighted Logistic Regression (GWLR), the logit form is given by:

$$\big(\pi(x_i)\big) = ln\left(\frac{\pi(x_i)}{1-(x_i)}\right) = \boldsymbol{x_i}\boldsymbol{\beta}(u_i, v_i) \tag{11}$$

where $0 < \pi(\boldsymbol{x_i}) < 1$

The response variable (Y_i) follows a *Bernoulli distribution* with the probability function given by:

$$Pr(Y = y_i) = \pi^{y_i}\big(1 - \pi(x_i)\big)^{1-y_i} \tag{12}$$

where value $y_i = 0.1$

Utilizing the *Maximum Likelihood Estimator* (MLE) technique, the parameters of the GWLR model are estimated. The *log-likelihood* function of the logistic spatial model is derived by maximizing the likelihood function in log form:

$$\ln L\left(\beta(u_i, v_i)\right) = \ln\left(\prod_{i=1}^{n}\left[\frac{(\exp(x_i^T\boldsymbol{\beta}))^{y_i}}{1+\exp(x_i^T\boldsymbol{\beta})}\right]\right) \quad (13)$$

2.6 Odds Ratio

The odds ratio value is determined by comparing the probability of success to the probability of failure. The equation used to estimate the odds ratio value is:

$$Odd = \frac{\pi(x)}{1-\pi(x)} = exp(\boldsymbol{X}_i\boldsymbol{\beta}) \quad (14)$$

where $i = 1,2,\ldots,n$ and $0 < \pi(x) < 1$

2.7 Selection of Best Bandwidth and Model

The radius of the circle, known as *bandwidth*, assumes that the points within this radius continue to exert influence. In GWLR modeling, the role of *bandwidth* is of utmost importance as it directly affects the accuracy of the model by adjusting the variance and bias in the data.

The Akaike Information Criterion (AIC) method is the preferred approach for model selection. When dealing with a large sample size in the context of the GWLR method, the AIC formula is utilized:

$$AIC = 2k - 2\ln(\hat{L}) \quad (15)$$

The calculation for AIC in the context of small sample sizes through the GWLR method is outlined in the equation:

$$AICc = AIC + \frac{2k^2-2k}{n-k-1} \quad (16)$$

The model that demonstrates the lowest AIC value is regarded as the most favorable model, where *k* denotes the number of parameters and *n* represents the number of samples.

2.8 GWLR Model Fit Test

To assess the level of geographic influence, hypothesis testing for the GWLR model involves conducting similarity tests and simultaneous tests between logistic regression models. A similarity test was carried out between the GWLR model and the logistic regression model to determine the significant level of influence. The hypothesis being tested is as follows:

H_0: $\beta_k(u_i, v_i) = \beta_k\ ; k = 1,2,\ldots,p$ (The GWLR and logistic regression models exhibit no substantial disparity)

H_1: At the very least, there is one $\beta_k(u_i, v_i) \neq \beta_k$ (The GWLR and logistic regression models display significant disparities)

The first stage of the Maximum Likelihood Ratio Test (MLRT) method involves deriving the test statistics by determining the set of parameters

within the population $(\Omega) = \{\beta_0(u_i, v_i),\ \beta_1(u_i, v_i), \ldots, \beta_k(u_i, v_i)\}$. Then, the likelihood function is constructed as shown below:

$$L(\Omega) = \prod_{i=1}^{n}\pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i} \quad (17)$$

$$L(\widehat{\Omega}) = \max_{\Omega} L(\Omega) = \prod_{i=1}^{n}\pi(x_i)^{y_{1j}}(1-\pi(x_i))^{y_{0i}} \quad (18)$$

Begin by establishing the parameters set under $H_0(\omega), \omega = \{\beta_0(u_i, v_i)\}$. Proceed to formulate both the likelihood function and the maximum likelihood function.

$$L(\omega) = \prod_{i=1}^{n}\pi(x_i)^{y_{1i}}(1-\pi(x_i))^{y_{0i}} \quad (19)$$

with

$$L(\widehat{\omega}) = \max_{\omega} L(\omega)\ \prod_{i=1}^{n}\left\{\left(\frac{n_{1i}}{n}\right)^{y_{1i}}\left(\frac{n_{0i}}{n}\right)^{y_{0i}}\right\} \quad (20)$$

the ratio equation of the maximum likelihood function under H_0 and the maximum likelihood function under the population could be expressed as follows, where n_{1i} represents the number of i observations included in category 1, n_{0i} represents the number of i observations included in category 0, and n represents the total number of observations.

$$\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} = \frac{\prod_{i=1}^{n}\left\{\left(\frac{n_{1i}}{n}\right)^{y_{1i}}\left(\frac{n_{0i}}{n}\right)^{y_{0i}}\right\}}{\prod_{i=1}^{n}\hat{\pi}(x_i)^{y_{1i}}(1-\hat{\pi}(x_i))^{y_{0i}}} \tag{21}$$

Equation (21) provides insight into what could be derived from the GWLR model deviation:

$$D(\hat{\beta}^*) = -2\ln\Lambda = 2\left[\ln L(\hat{\Omega}) - \ln L(\hat{\omega})\right] \tag{22}$$

For instance, if $D(\hat{\beta})$ represents the deviance of the logistic regression model with db_1 degrees of freedom, and $D(\hat{\beta}^*)$ represents the deviance of the GWLR model with db_2 degrees of freedom, then the test statistics used to examine the correlation between these two models are:

$$F_{hit} = \frac{D(\hat{\beta})/db_1}{D(\hat{\beta}^*)/db_2} \tag{23}$$

The F distribution with db_1 and db_2 degrees of freedom is attained by the test statistic in equation (23). To assess the GWLR model parameters collectively, the testing criteria involve rejecting H_0 when the $F_{hit} > F_{\alpha,db_1,db_2}$. Evaluate the GWLR model parameters concurrently by considering the following hypotheses:

$H_0: \beta_1(u_i, v_i) = \beta_2(u_i, v_i) = \cdots = \beta_p(u_i, v_i) = 0\ ; i = 1,2,\ldots,n$
$H_1:$ At the very least, there is one $\beta_k(u_i, v_i) \neq 0\ ; k = 1,2,\ldots,p$

According to the MLRT model, the test statistics used for simultaneous testing are as follows:

$$G^2 = 2\left[ln\, L(\hat{\Omega}) - ln\, L(\hat{\omega})\right] \tag{24}$$

The test statistic G^2 in equation (24) approximates the chi-square distribution with degrees of freedom $v = n - k - 1$. The criteria for the test involve rejecting H_0 if the value of $G^2 > \chi^2_{\alpha,v}$.

2.9 Partial Test of the GWLR Model

GWLR model parameter testing determines which parameters have the greatest impact on the model. The partial test hypothesis for the β_k parameter is as follows:

$H_0: \beta_k(u_i, v_i) = 0\ ;\ i = 1, 2, \ldots, n;\ k = 1, 2, \ldots, p$
$H_1: \beta_k(u_i, v_i) \neq 0$

The *Wald* test could be utilized to obtain test statistics for this test, as demonstrated below:

$$Z_{hit} = \frac{\hat{\beta}_k(u_i,v_i)}{Se\left(\hat{\beta}_k(u_i,v_i)\right)}\ ; k = 1,2,\ldots p \tag{25}$$

The test statistic presented in equation (25) provides an approximation of the standard normal distribution. The purpose of the test is to reject the null hypothesis, H_0 if the absolute value $|Z_{hit}| > Z_{\frac{\alpha}{2}}$.

2.10 GWLR Model Classification Accuracy

The classification accuracy of the GWLR model could be assessed by computing the Apparent Error Rate (APPER) value, which represents the likelihood of error in object classification. The APPER value could be calculated using the following formula:

$$APPER = \frac{n_{12}+n_{21}}{n_{11}+n_{12}+n_{21}+n_{22}} \times 100 \tag{26}$$

where,

n_{11} = Number of observed errors that fall into the error category based on the prediction results
n_{12} = Number of observed errors that fall into the success category based on the prediction results

n_{21} = Number of observed successful events that fall into the error category based on the prediction results

n_{22} = Number of observed success events that fall into the success category based on the predicted results

2.11 Research Procedure

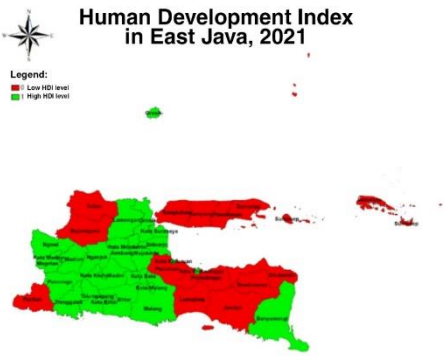
The GWLR method allows for the analysis procedure to be conducted in the following manner:

1. Based on a thematic map generated with ArcGIS software, the map illustrates the factors that impact the Human Development Index of districts/cities in East Java.
2. The process of modeling and estimating Human Development Index data utilizing the Geographically Weighted Logistic Regression (GWLR) approach could be carried out using GWR4 software through the subsequent steps:
    a. Validate the fundamental assumptions of spatial regression, specifically examining spatial heterogeneity through the Breusch-Pagan technique.
    b. Verify the core assumptions of spatial regression, particularly focusing on spatial dependence using Moran's I approach.
    c. Establish the optimal bandwidth (h) by utilizing the CV method.
    d. Compute the weighting matrix using kernel functions such as Fixed Gaussian, Fixed Bi-square, and Adaptive Gaussian. Subsequently, identify the most suitable kernel function by comparing the AIC values.
    e. Evaluate the appropriateness of the GWLR model using logistic regression.
    f. Conduct partial testing on the GWLR parameters and identify the variables that impact the HDI in each district/city.
    g. Develop a GWLR model specific to each district/city within East Java Province.
    h. Estimate the Odd Ratio value of predictor variables that exhibit a significant impact.
    i. Contrast the Logistic Regression model with the GWLR Model.
3. Examine and explain the different factors that play a crucial role in determining the Human Development Index in East Java Province and assess the precision of the categorization.
4. Formulate conclusions

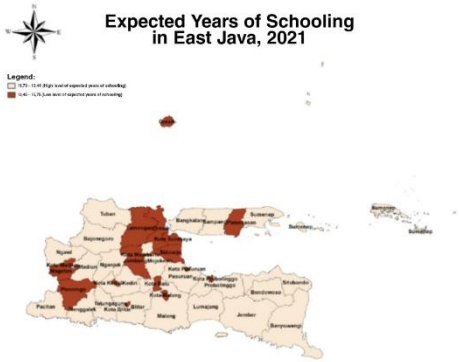
3. Result and discussion

The obtained data is utilized to classify dependent and independent variables based on districts/cities in East Java. Thematic maps are employed for this purpose. The subsequent classification represents the research variables categorized according to districts/cities in East Java.

a



b



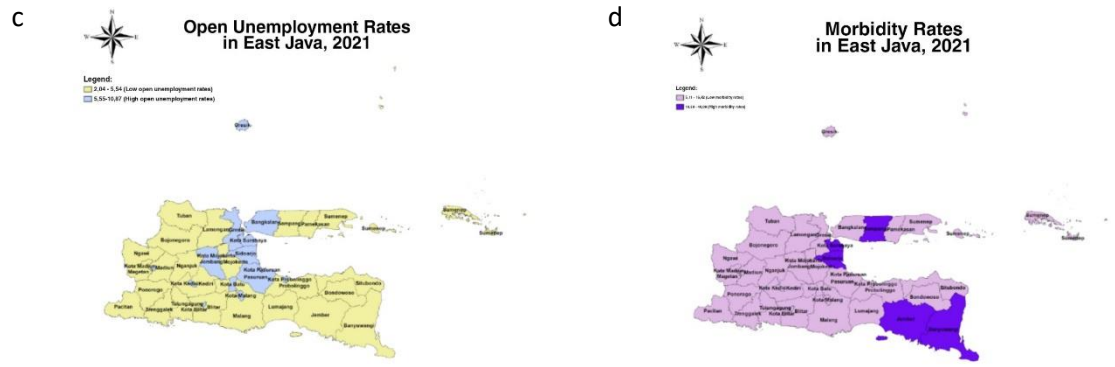


Figure. 1. (a) Thematic Map of Human Development Index in East Java; (b) Thematic Map of Expected Years of Schooling in East Java; (c) Thematic Map of Open Unemployment Rates in East Jawa; (d) Thematic Map of Morbidity Rates in East Java.

Before estimating model parameters using spatial regression analysis, the data is subjected to a spatial assumption test. This test is instrumental in assessing whether the data conforms to the fundamental assumptions of spatial regression. A significant level of 10% was applied in this study. In spatial analysis, there could be substantial variations between neighboring locations. By employing a higher significance level, such as 10%, the GWLR model could better capture these discrepancies and offer more precise estimates for each location. Furthermore, given the social context of this research, a 10% significance level is deemed appropriate. The GWLR method analysis entails meeting several assumptions, including the spatial dependency test using Moran's I test and the spatial heterogeneity test using the Breusch-Pagan test.

Table 3. Spatial Dependency Test Output

Observed	0.1809297
Expected	-0.02702703
SD	0.0516518
P-Value	0.0000056702

According to the calculation of Moran's I test, a probability value of 0.000056702 is obtained, which is less than the significant level $\alpha = 10\%$. Therefore, it could be concluded that the null hypothesis H_0 is rejected, indicating the presence of spatial dependency in the data. These findings provide evidence that the data has successfully passed the GWLR spatial assumption test, confirming the existence of spatial dependency in the dataset.

Table 4. Spatial Heterogeneity Test Output

BP	9.6154
DF	3
P-Value	0.02213

The Breusch-Pagan test conducted using R software yields a probability value of 0.02213, which is below the significant level of ($\alpha = 10\%$). Therefore, rejecting the null hypothesis suggests the existence of spatial heterogeneity in the data. This outcome aligns with the GWLR spatial assumption test, indicating spatial heterogeneity in the dataset.

Table 5. AIC Value of Each Kernel Weight

Kernel Function Weighting	Bandwith	AIC
Adaptive Bi-Square	-	-
Adaptive Gaussian	-	-
Fixed Bi-Square	1.819	28.777
Fixed Gaussian	0.659	28.732

Based on the information provided in the table, it is evident that the Kernel Fixed Gaussian weighting function demonstrates the lowest AIC value among the various weighting functions, specifically 28.723, Therefore, the Kernel Fixed Gaussian weighting function is employed to estimate the optimal model in this study. Subsequently, a model suitability test was conducted to assess whether GWLR modeling is more suitable compared to logistic regression modeling.

Table 6. GWLR Model Fit Test

Model	Dev	db	Dev/db	F
Logistic Regression	28.686	34	0.844	2.398
GWLR	10.103	28.690	0.352	

The analysis reveals that the calculated $F = 2.398$. By considering a significance level of α=10% (0.1), the value is determined to be $F_{(0.1;34;28.690)} = 1.603$. The critical region in this study dictates rejecting H_0 when $F > 1.603$, resulting in the rejection of H_0. Consequently, a notable difference exists between the logistic regression model and GWLR, indicating the superiority of the GWLR model for modeling purposes.

Following the completion of the analysis, an HDI model was derived for every district/city within East Java Province utilizing the GWLR method. For instance, a specific area in Blitar Regency was selected, resulting in a model equation as presented below:

$$g(x) = -67.957 + 4.007X_1 - 0.059X_2 - 0.169X_3 \quad (27)$$

In Blitar Regency, the partial test results revealed a significant effect of X_1 (Expected Years of Schooling). The Odd Ratio calculation was then used to determine the influence of the predictor variables in the model. Notably, only the predictor variables that showed significance in Blitar Regency, such as Expected Years of Schooling, were included in this calculation.

$$Odd\ Ratio = exp(\beta_1) = exp(4.007) = 10.892 \quad (28)$$

In the context of the Odd Ratio calculation, it has been established that a 1 unit increase in Expected Years of Schooling, assuming the other X values remain constant, leads to 10.892 times increase in the Odd value. Moving forward, let's assess the efficiency of the logistic regression model and the GWLR model in the HDI case in East Java.

Table 7. Model Goodness Test

Model	AIC
Logistic Regression	36.686398
GWLR	28.723450

The AIC value in the GWLR model is smaller than that in the logistic regression model, indicating that the GWLR model is more suitable for analyzing HDI data in East Java.

Upon further analysis of the estimated results of the HDI distribution, significant factors influencing the distribution are identified. This allows for a comparison of HDI distribution before and after estimation, shedding light on the factors influencing districts/cities in East Java. The presentation of HDI Index results in graphs and thematic maps provide a visual representation of the findings.

Estimation results for each district/city in East Java were compared by inputting predictor variables into the equation model. For instance, Bojonegoro Regency was selected, showing an anticipated years of schooling of 12.68, an open unemployment rate of 4.82, and a morbidity rate of 11.02, then it is obtained.

$$\pi(X_j) = \frac{exp(\sum_{k=0}^{p} \beta_k(u_i,v_i)x_{ik})}{1+exp(\sum_{k=0}^{p} \beta_k(u_i,v_i)x_{ik})}$$

$$\pi(X_j) = \frac{exp\ (-67.96+7.66(12.68)- \ -0.36(4.82)+0.002(11.02)}{1+exp\ (\ -0.36(4.82)+0.002\ (11.02)-0.36(4.82)\ +0.002(11.02)}$$

$$\pi(X_j) = \frac{\exp(27.47)}{1+exp(27.47)}$$

$$\pi(X_j) = 1 \quad (29)$$

Manual calculations are employed in every district/city to allow for a comparison between the estimated results and actual observations, as illustrated in the graph below.

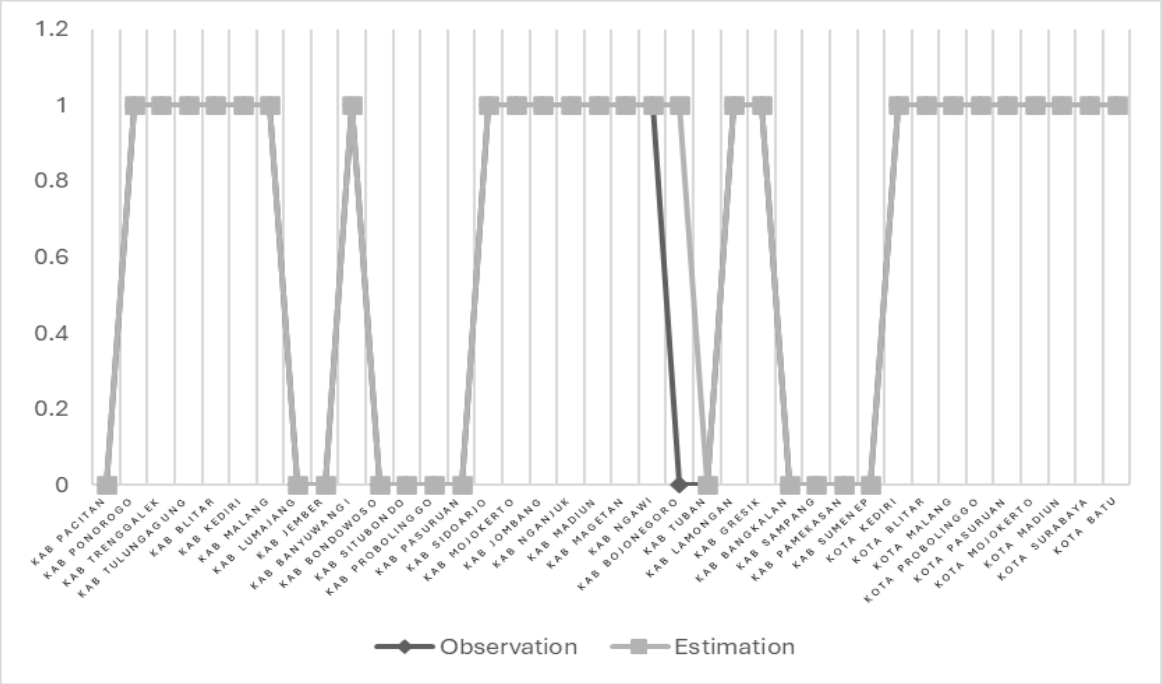


Figure. 5. Graph of Comparison of Classification Results with Preliminary Data on the Human Development Index in East Java

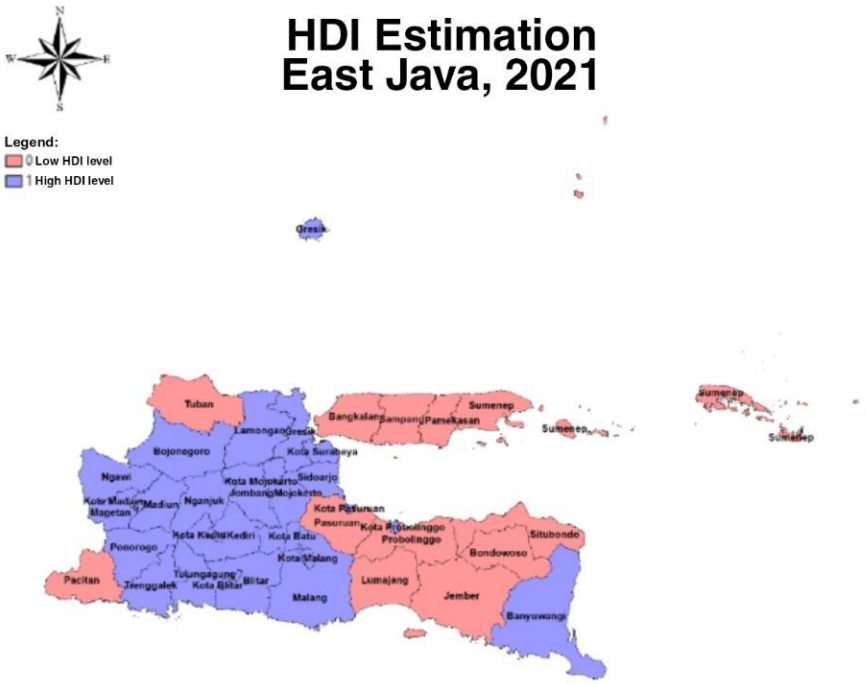


Figure. 6. Thematic Map of Human Development Index Estimation Results in East Java

After examining the projected outcomes of the HDI distribution and the factors that impact it, the HDI results in East Java could be categorized using the GWLR method.

Table 8. Classification of Human Development Index Results in East Java GWLR Model

Observation	Estimation	
	Low (0)	High (1)
Low (0)	12	1
High (1)	0	25

The estimation results successfully classified 12 regions with low HDI in the appropriate category, but there was one misclassification from low to high

HDI, specifically Bojonegoro Regency. Conversely, the estimation accurately categorized 25 regions with high HDI in the high HDI category, without any incorrect classifications in the low HDI category. As a next step, a thematic map will be developed based on the influential factors affecting HDI in East Java.

Table 9. Variables that have a significant influence on HDI

Significant Variable	Regency/City
X_1	Blitar Regency, Kediri Regency, Malang Regency, Probolinggo Regency, Pasuruan Regency, Sidoarjo Regency, Mojokerto Regency, Jombang Regency, Lamongan Regency, Gresik Regency, Bangkalan Regency, Kediri City, Blitar City, Malang City, Pasuruan City, Mojokerto City, Surabaya City, Batu City.
X_2	Lumajang Regency
X_3	Nothing

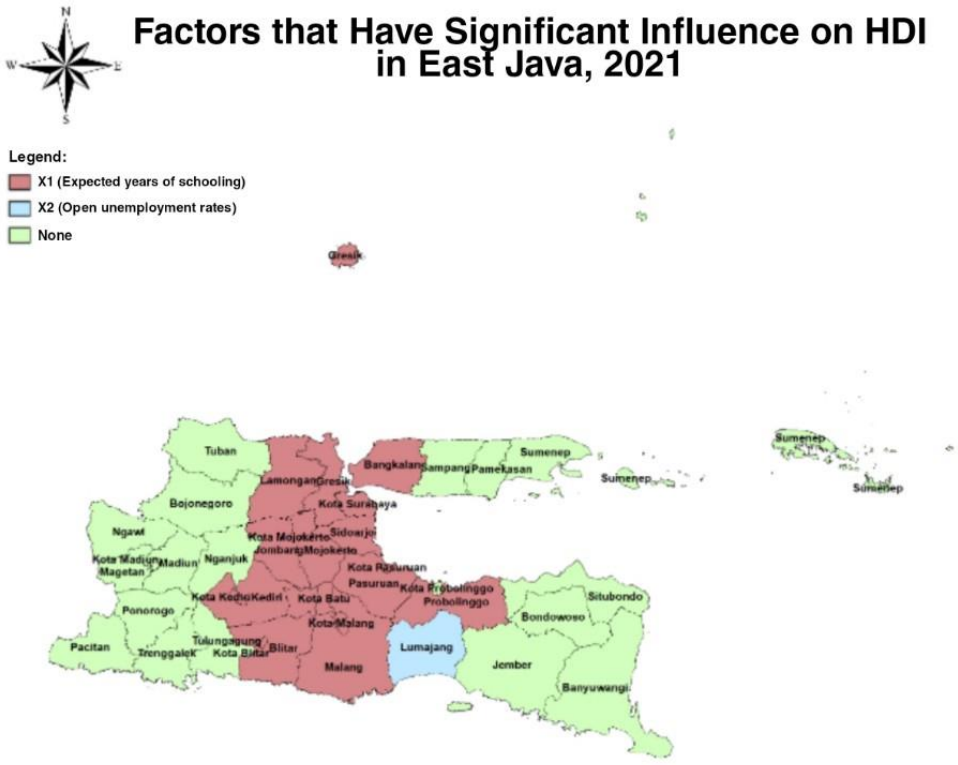


Figure 4. Thematic Map of Factors that Significantly Influence the Estimation of HDI in East Java

The provided diagram illustrates the distribution of predictor variables that impact the Human Development Index (HDI) in districts and cities located in East Java. The two influential predictor variables are X_1, which represents the expected years of schooling, and X_2, which represents the open unemployment rate. Out of all the districts and cities in East Java, 18 of them, accounting for 47.36% of

the total, have their HDI value solely influenced by variable X_1, indicating the expected years of schooling. Conversely, there is only one district or city, making up 2.63% of the total, where the HDI value is solely influenced by variable X_2, representing the open unemployment rate. The figure also demonstrates a tendency for neighboring locations to exhibit similar conditions, with the same predictor variables significantly influencing each other.

The obtained analysis results indicate that the significant variables differ among districts and cities, and in some cases, certain variables may not hold significance at all. This discrepancy could be attributed to the existence of spatial heterogeneity, which reinforces the argument favoring the use of spatial models for effectively describing and analyzing complex phenomena across various regions. Therefore, it is crucial to identify predictor variables that significantly influence the Human Development Index (HDI) in each district and city within East Java Province, to enhance the overall HDI.

4. Conclusion

The conclusions that have been derived after conducting analysis and discussion are outlined below:

The year 2021 will see 13 districts/cities in East Java categorized as low HDI, accounting for 34% of the total. Sampang Regency holds the lowest HDI score at 62,8, with Surabaya City leading at 82,31. The analysis suggests that the Fixed Gaussian Kernel function weighting is the optimal model for HDI.

The HDI varies across districts and cities, and the variables that have a significant impact on it differ as well. However, two variables that at least have some influence on the HDI are the expected years of schooling (X_1) and the open unemployment rate (X_2). According to the GWLR model, the results indicate that districts or cities with higher expected years of schooling tend to have a higher HDI. Conversely, districts or cities with higher levels of open unemployment tend to have a lower HDI.

Based on the discussion, the two districts with the lowest HDI are Sampang Regency and Bangkalan Regency. In Sampang Regency, the HDI is not influenced by the three predictors that were studied. On the other hand, the HDI for Bangkalan Regency is influenced by the Expected Years of Schooling variable (X_1), as indicated by the equation $f(x) = -67.957 + 3{,}866X_1 + 0.964X_2 - 0.065X_1$. The higher the expected years of schooling, the higher the HDI value.

The outcomes and deliberations suggest that special attention should be directed towards regencies/cities with low HDI values, particularly in Sampang Regency and Bangkalan Regency. In Bangkalan Regency, the government should focus on outreach and support to enhance education, particularly the duration of schooling. It is crucial for the people of East Java to actively engage in and endorse the government's initiatives to achieve the SDGs 2030, with the overarching aim of advancing society in Indonesia, especially in East Java. Furthermore, readers should investigate other variables apart from the ones analyzed in this research to facilitate the prompt improvement of districts/cities with low HDI figures, such as Sampang Regency, where the factors contributing to the low HDI remain unknown.

Ethics approval

Not Required

Acknowledgments

We gratefully acknowledge the support and contributions of all those involved in this research. Special thanks to Badan Pusat Statistik (BPS) for providing the essential data that made this study possible Our thanks also go to the reviewers and proofreaders for their meticulous efforts and invaluable feedback, which greatly enhanced 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.

References

[1] UNDP, "Sustainable Development Goals", undp.org. https://www.undp.org/sustainable-development-goals (Accessed May. 5, 2023).
[2] BPS, *New Method Human Development Index*. Jakarta: Central Statistics Agency, 2014.
[3] D. L. Fika, K. Dadan, and N. B. Naomi, "Estimasi Parameter Model Geographically Weighted Logistic Regression", *Buletin Ilmiah Math. Stat. dan Terapannya (Bimaster)*, vol. 9, no. 1, pp. 159-164, 2020.
[4] I. M. Nur and M. Al Haris, "Geographically Weighted Logistic Regression (GWLR) with Adaptive Gaussian Weighting Function in Human Development Index (HDI) in The Province of Central Java", in *Journal of Physics: Conference Series*, 2021, vol. 1776, p. 012048.
[5] W. Lili, Y. Desi, and N. H. Memi, "Pemodelan Faktor-Faktor yang Berpengaruh Tehadap Indeks Pembangunan Manusia (IPM) di Kalimantan dengan Geographically Weighted Logistic Regression (GWLR)", *Jurnal Eksponensial*, vol. 9, no. 1, pp. 67-74, 2018.
[6] A. Siti, 'Peran Badan Usaha Milik Desa (Bumdes) Terhadap Kesejahteraan Masyarakat di Desa Wanasaba Lauk Kecamatan Wanasaba Kabupten Lombok Timur', Universitas_Muhammadiyah_Mataram, 2023.
[7] I. A. Juliannisa, M. B. N. Ariani, and T. Siswantini, 'Efforts to Increase HDI from an Educational Side in the Johar Baru Community, Central Jakarta', *IKRA-ITH ABDIMAS, DKI Jakarta*, 2023.
[8] H. Lubis, 'Analysis of the Effect of Minimum Wage, GRDP, HDI, and Population on Poverty Levels in Districts/Cities of North Sumatra Province 2015-2019', 2023.
[9] A. Tyas and Ikhsani, 'Natural Resources & Human Resources for Indonesia's Economic Development'. 2015.
[10] A. J. Mahya and W. Widowati, 'Analysis of the Influence of Expected Years of Schooling, Average Years of Schooling, and Per Capita Expenditures on the Human Development Index', *Prismatics: Journal of Mathematics Education and Research*, vol. 3, no. 2, 2021.
[11] L. Widyastuti, D. Yuniarti, and M. N. Hayati, 'Pemodelan Faktor-Faktor yang Berpengaruh Terhadap Indeks Pembangunan Manusia (IPM) di Kalimantan dengan Geographically Weighted Logistic Regression (GWLR)', *Jurnal Eksponensial*, vol. 9, no. 1, pp. 67–74, 2018.
[12] BPS East Java, *East Java in numbers 2021*. Surabaya: BPS East Java, 2021.
[13] M. N. Faritz and A. Soejoto, 'Pengaruh pertumbuhan ekonomi dan rata-rata lama sekolah terhadap kemiskinan di Provinsi Jawa Tengah', *Jurnal Pendidikan Ekonomi (JUPE)*, vol. 8, no. 1, pp. 15–21, 2020.
[14] Á. S. Batista, *Logistic Regression: An Introduction to Statistical Model with an Example of Revolving Credit*. Lisbon: Createspace Independent Publishing Platform, 2014.
[15] E. N. Manurung and F. Hutabarat, 'The Influence of Expected Years of Schooling, Average Years of Schooling, and Expenditures per Capita on the Human Development Index', *Scientific Journal of Management Accounting*, vol. 4, no. 2, pp. 121–129, 2021.
[16] E. Permadi and E. Chrystanto, 'Analysis of the Influence of Population, Gross Regional Domestic Product (GRDP), and Regency/City Minimum Wage on the Open Unemployment Rate in Regency/City in East Java Province 2012-2018', *OECONOMICUS Journal of Economics*, vol. 5, no. 2, pp. 86–95, Jun. 2021.
[17] D. S. Amru and E. D. Sihaloho, 'The influence of per capita expenditure and health expenditure on morbidity rates in districts/cities throughout Java', *Asian Business and Economics Scientific Journal*, vol. 14, no. 1, pp. 14–25, 2020.
[18] S. Kardjati, *Aspects of Health and Nutrition for Children Under Five*, First. Jakarta: Indonesian Obor Foundation, 1985.

[19] F. Febrianti and H. Helma, 'Binary Logistic Regression Analysis on Factors that Influence the Willingness of the Nagari Paninauan Community to be Vaccinated with COVID-19', *UNP Journal of Mathematics*, vol. 8, no. 1, pp. 36–44, 2023.

[20] I. Arofah and S. Rohimah, 'Analisis Jalur Untuk Pengaruh Angka Harapan Hidup, Harapan Lama Sekolah, Rata-Rata Lama Sekolah Terhadap Indeks Pembangunan Manusia Melalui Pengeluaran Riil Per Kapita Di Provinsi Nusa Tenggara Timur', *Jurnal Saintika Unpam: Jurnal Sains Dan Matematika Unpam*, vol. 2, no. 1, p. 76, 2019.

[21] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[22] P. H. M. Alburquerque, F. A. S. Medina, and A. R. Silva, 'Geographically Weighted Logistic Regression Applied to Scoring Model', in *XL ANDAP Congress*, 2016.

[23] F. D. Lestari, D. Kusnandar, and N. N. Debataraja, 'Estimasi Parameter Model Geographically Weighted Logistic Regression', *Mathematics Scientific Bulletin. Stat. and Applications (Bimaster)*, vol. 9, no. 1, pp. 159–164, 2020.

DOI: https://doi.org/ 10.34123/jurnalasks.v16i1.576

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

Geographically Weighted Lasso Method in Modeling The Gross Regional Domestic Product of The Bali-Nusa Tenggara Region

Hairunnisa[1], Mustika Hadijati[2*], Nurul Fitriyani[3]

[1] Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Mataram, Mataram, 83125, Indonesia, [2,3] Department of Statistics, Faculty of Mathematics and Natural Sciences, University of Mataram, Mataram, 83125, Indonesia
**Corresponding Author: E-mail address:* mustika.hadijati@unram.ac.id

ARTICLE INFO

Article history:
Received 10 June, 2023
Revised 29 April, 2024
Accepted 10 May, 2024
Published 30 June, 2024

Keywords:
Gross Regional Domestic Product; Geographically Weighted Lasso; Geographically Weighted Regression, Multicollinearity; Spatial Heterogeneity.

Abstract

Introduction: Statistics Indonesia announced that economic growth in 2020 is still in the negative zone, and the group of provinces in the Bali-Nusa Tenggara region has the most negligible impact on economic growth. The value of Gross Regional Domestic Product (GRDP) measures Indonesia's economic growth. GRDP is the total added value all regional business units generate at a particular time. **Background Problems:** This research aims to apply and interpret the results of the Geographically Weighted Lasso (GWL) method for GRDP in the Bali-Nusa Tenggara region. **Novelty:** Preliminary analysis in this study shows that the data used has the effect of spatial heterogeneity, which is a requirement for modelling using the GWR method. In addition, there is a multicollinearity problem between independent variables. Therefore, the GWL method is used to solve the problem of spatial heterogeneity and multicollinearity in modelling the GRDP of the Bali-Nusa Tenggara Region. **Research Methods:** The GWL method further develops the Geographically Weighted Regression (GWR) approach by adding the Least Absolute Shrinkage and Selection Operator (LASSO) method. The GWL method simultaneously selects insignificant variables by reducing the value of the regression coefficient to zero using the LASSO method. The data used has the effect of spatial heterogeneity and multicollinearity, a prerequisite for modelling with the GWL method. **Results:** Based on the analysis conducted, there are 41 different GRDP models for each district/city in the Bali-Nusa Tenggara region. The resulting GWL model provides a coefficient of determination of 95.84 % so that the resulting model can be used and is considered valid.

1. Introduction

In 2020, the Indonesian economy decreased by 2.07 % compared to 2019. Statistics Indonesia indicated that the Indonesian economy in the third quartile of 2020 was -3.49 %, meaning that economic growth in the third quartile of 2020 was still in the negative zone. When viewed spatially, all provincial economies in all islands in Indonesia also experienced negative growth. The most profound economic contraction was in the group of Bali-Nusa Tenggara (Bali-Nusra) provinces, with economic growth of -

6.80%. Based on this, the contribution of the economic growth of the Bali-Nusra region to the Indonesian economy is only 2.92%. The primary condition for the continued economic development of an area is the condition of high economic growth. According to Statistics Indonesia, a region's overall economic growth rate can be measured using Gross Regional Domestic Product (GRDP) at constant prices. GRDP preparation can be done through the Production, Expenditure, and Income approaches [1].

In a previous study, the factors used were the value of GRDP through the Production and Human Resources (HR) Approach, which consisted of 13 independent variables [2]. Based on this description, this study used factors that can affect GRDP using independent variables that have been studied previously by adding other factors, namely through the Expenditure Approach that is adjusted to the factors that affect GRDP in the Bali-Nusra region. A regression analysis model is needed to modelling the relationship between the GRDP of an area and the factors that influence it. The regression analysis used in this study is spatial regression because the data used is data that contains information on regional or location (spatial) related to the latitude and longitude coordinates of an area.

Geographically Weighted Regression (GWR) is a non-stationary method that models spatially varying relationships with coefficients in the GWR method, namely the function of spatial location. The GWR method aims to explore spatial heterogeneity in data relationships [3]. Several studies on the application of the regression model to spatial data have been carried out by previous researchers, including modelling the poverty rate in Central Java using the GWR model [4], the human development index model in West Nusa Tenggara using the Geographically Weighted Ridge Regression (GWRR) Method [5], the poverty model in the West Nusa Tenggara using the Geographically Weighted Logistic Regression (GWLR) model [6], and modelling the number of infant mortality in the East Lombok using Geographically Weighted Poisson Regression [7].

The drawback of the GWR method is that it has not been able to overcome cases of multicollinearity. Multicollinearity indicates a situation where there is a strong correlation between the independent variables [8]. The impact that can be caused by multicollinearity is the variance of the regression coefficient to be large. The magnitude of the variance can cause several problems, including the standard error, and the resulting interval will be large or wide. If the standard error is too large, the estimate of β will likely be insignificant. Therefore, other methods are needed to overcome the multicollinearity case [9].

The method to overcome the case of multicollinearity in the spatial model is the GWL method. The GWL method is an evolution of the GWR method, adding the LASSO method to the modeling. The objective of the GWL method is to overcome the problems of spatial heterogeneity and multicollinearity present in the least squares method [10]. The GWL method overcame the problem of spatial heterogeneity and multicollinearity in spatial data with a case study of food insecurity in Tanah Laut Regency [11]. This research results show that the GWL method performs better than the GWR method. Another study in the case of GRDP in West Java, using the GWL method with weighting the Fixed Exponential Kernel kernel function, was able to overcome the problems of spatial heterogeneity and multicollinearity. The result is that the model obtained is feasible to use or can be said to be valid [2].

Preliminary analysis in this study shows that the data used has the effect of spatial heterogeneity, which is a requirement for modeling using the GWR method. In addition, there is a multicollinearity problem between independent variables. Therefore, the GWL method is used to solve the problem of spatial heterogeneity and multicollinearity in modeling the GRDP of the Bali-Nusa Tenggara Region.

2. Material and Methods

This research aims to determine the GRDP model for the Bali-Nusra region and determine the factors that significantly influence the GRDP in the Bali-Nusra region using the GWL method. The method used is spatial regression analysis. The steps in this research are divided into the following stages:

2.1. Preparations

The preparations carried out in this research include setting the research topic, collecting as much information as possible from the literature related to the research, and then collecting the data used in the study. In this study, p represents the number of independent variables, n indicates the number of observations, Y_i is the value of the dependent variable on i-th observation, β_0 is the constant, β_j represents the regression coefficient of independent variable X_j, X_{ij} indicates the value of j-th

independent variable on i-th observation, and ε_i represents the error value which is assumed to be identical, independent, and normally distributed with zero means and σ^2 variance.

2.2. Performing multicollinearity checking

Multicollinearity is a condition where there is a high correlation between independent variables. The multicollinearity test aims to identify cases of multicollinearity in the independent variables using the VIF test criteria. If the VIF value is greater than 10, then multicollinearity occurs, whereas if the VIF value range from $1 \leq VIF \leq 10$, then multicollinearity does not occur [9]. The VIF value can be obtained using Formula (1):

$$VIF_j = \frac{1}{1-R_j^2}; j = 1,2,\dots,p \quad (1)$$

where R_j^2 is the coefficient of determination of a multiple regression model with the j-th variable (X_j) as the model's dependent variable and the other independent variables as its explanatory variables [11].

2.3. Performing spatial heterogeneity test

The spatial heterogeneity test used the Breusch-Pagan test formula. Tests with Breusch-Pagan were carried out to identify the data's spatial heterogeneity, i.e., there was a diversity of observational data between locations. If there is spatial heterogeneity in the model, then the requirements for testing using the GWR method are met [12]. The Breusch-Pagan test formula is shown in Formula (2).

$$BP = \left(\frac{1}{2}\right) \boldsymbol{f^T Z (Z^T Z)^{-1} Z^T f} \quad (2)$$

with,

$$f_i = \left(\frac{{e_i}^2}{\sigma^2} - 1\right),\ Z = \frac{x_i - \underline{x_i}}{\sigma_{x_i}},\ i = 1,2,\dots,n.$$

2.4. Performing GWR modeling

a. Calculating the Euclidean distance between the i-th observation location and the j-th observation location, which is defined in Formula (3).

$$d_{ij} = \sqrt{\left(u_i - u_j\right)^2 + \left(v_i - v_j\right)^2} \quad (3)$$

with u latitude and v longitude coordinates [13].

b. Estimating the value of bandwidth (b) with a kernel function that minimizes the value of cross-validation (CV). The mathematical equation of the CV value, according to [14], is in Formula (4).

$$CV = \sum_{i=1}^{n} [y_i - \hat{y}_{\neq i}(b)]^2 \quad (4)$$

where $\hat{y}_{\neq i}$ is the estimated value of y_i at the i-th observation location with a certain bandwidth removed from the estimation process [15].

The kernel function used in this research is the fixed exponential kernel function which is defined in Formula (5):

$$w_{ij} = exp\left(\frac{-d_{ij}}{b}\right) \quad (5)$$

where d_{ij} is the Euclidean distance, and b is the bandwidth [13].

c. Forming a weighting matrix for each observation location using the bandwidth value that has been obtained previously. The weighting matrix used to estimate the parameters at each observation location is expressed in a diagonal matrix, with its elements being the kernel function of each point of observation location. The form of the weighting matrix is shown in Formula (6):

$$\boldsymbol{W}(u_i, v_i) = \begin{bmatrix} {w_{(u_i,v_i)}}^{i,1} & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & {w_{(u_i,v_i)}}^{i,n} \end{bmatrix} \quad (6)$$

with ${w_{(u_i,v_i)}}^{in}$ is the value of the kernel function for the data at the n-th point in the model test at the i-th observation location.

d. Estimating GWR parameters using the WLS method by adding a weighting element in the estimate. The form of parameter estimation of the GWR model for each observation location is shown in Formula (7):

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = (\boldsymbol{X^T W X})^{-1}\boldsymbol{X^T W y} \quad (7)$$

e. Detecting cases of local multicollinearity in the GWR model. It can be seen from the local VIF value greater than 10 to determine the presence of local multicollinearity in the GWR model.

2.5. *Performing GWL modeling.*

The GWL method is a concept from the LASSO method applied in GWR modeling to overcome the problem of spatial heterogeneity and local multicollinearity. The parameter estimation in the GWL model uses the WLS method by adding a Lagrange multiplier function (λ) to the estimate. The value of λ is an additional component in the LASSO regression that controls the magnitude of the shrinkage of the regression parameter value by producing a solution where the number of parameters is zero. The form of parameter estimation of the GWL model for each observation location is shown in Formula (8):

$$\widehat{\boldsymbol{\beta}}(u_i, v_i) = \left(\boldsymbol{X^T W X} + \boldsymbol{\lambda I}\right)^{-1}\boldsymbol{X^T W Y} \quad (8)$$

with λ referred to as a tuning parameter that is useful for controlling the amount of shrinkage in the value of the regression parameter [8].

2.6. *Testing the sustainability of the model.*

Testing the suitability of the model using the coefficient of determination (R^2) with the equation in Formula (9):

$$R^2 = 1 - \frac{SSE}{SST} \quad (9)$$

SSE is the sum of squared errors obtained by the formula $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and SST is the sum of the squared total obtained by the formula $\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2$. The value of the coefficient of determination ranges from $0 \leq R^2 \leq 1$. That is if the value of R^2 obtained is close to 1, it can be said that the independent variable has a powerful influence on the dependent variable; in other words, the model used is good at explaining the effect of the variable [16].

3. Results And Discussion

It can be seen by looking at the VIF value to determine whether multicollinearity occurs. If the VIF value is > 10, then multicollinearity occurs, whereas if the VIF value is around $1 \leq VIF \leq 10$, there is no multicollinearity. The results of the multicollinearity test are presented in Table 1:

Table 1. Rice coefficient for various climatic conditions

Variables	VIF Value
X_1	30.7318
X_2	11.5388
X_3	4.4546
X_4	3.6861
X_5	4.5036

Variables	VIF Value
X_6	17.5488
X_7	2.9251
X_8	1.7150
X_9	13.2390
X_{10}	60.7279
X_{11}	90.6596
X_{12}	10.7505
X_{13}	33.1903
X_{14}	73.5313
X_{15}	10.8764
X_{16}	62.3569
X_{17}	65.2403

Table 1 shows that the VIF values that are less than 10 are X_3, X_4, X_5, X_7 dan X_8, while the other independent variables have a VIF value of more than 10. Therefore, it can be said that there is multicollinearity between the independent variables, so it needs to be overcome by the GWL method.

3.1. Spatial Heterogenity

A spatial heterogeneity test was conducted to identify whether or not there was a spatial effect on the data. If there is spatial heterogeneity in the regression model, then the requirements for completing the test using the GWR method are met to continue the research. The spatial heterogeneity test was carried out using the Breusch-Pagan (BP) test with the following hypothesis:

$H_0 : \sigma_1^2 = \sigma^2$ (no spatial heterogenity)

H_1 : at least one $\sigma_i^2 \neq \sigma^2$ (there is spatial heterogeneity)

3.2. Geographically Weighted Regression (GWR) Modeling

Based on the results of the tests carried out, the BP value = 43.73, and the Chi-Square table value $\chi^2_{(p,\alpha)}$ (with $p = 17$ and α=0.05) =27.59. Because $BP > \chi^2_{(17,0.05)}$ then H_0 is rejected, meaning that there is spatial heterogeneity in the regression model. Therefore, the heterogeneity test is met, so the research can be continued to model the GRDP data for the Bali-Nusra region using the GWR spatial model.

The GWR model obtained is 41 models from 41 regencies/ cities in the Bali-Nusra region, so the parameter estimates from the model are very large. Therefore, the following is an example of a GRDP model in the Bali-Nusra region using the GWR method, one of which is the GRDP model for West Lombok Regency with model in Formula (10):

$$\hat{y}_1 = -3325.2670 + 0.2765x_1 - 0.0055x_2 + 43.1414x_3 + 66.2337x_4 - 5.8356x_5 - 0.1076x_6 + 0.0020x_7 + 0.9999x_8 + 0.3629x_9 - 15.9082x_{10} + 0.9629x_{11} + 0.3102x_{12} + 1.2276x_{13} + 1.3702x_{14} - 4.4293x_{15} + 1.9280x_{16} - 0.3204x_{17} \quad (10)$$

3.3. Local Multicollinearity

The next step is to identify local multicollinearity. It can be seen from the local VIF value greater than 10 to determine the presence of local multicollinearity in the GWR model. A summary of the total local VIF values for each observation location is presented in Table 2.

Table 2. Summary of local VIF values for all observation sites.

Independent Variable	VIF > 10
X_1	41
X_2	34
X_3	0

Table 2. Summary of local VIF values for all observation sites.

Independent Variable	VIF > 10
X_4	0
X_5	0
X_6	41
X_7	0
X_8	0
X_9	41
X_{10}	41
X_{11}	41
X_{12}	41
X_{13}	41
X_{14}	41
X_{15}	25
X_{16}	41
X_{17}	41

Based on the results of the local multicollinearity test, 12 independent variables have a VIF value greater than 10, with ten independent variables in which all observation locations have a VIF value greater than 10 and 2 other independent variables, respectively 34 and 25 observation locations have a VIF value greater than 10. greater than 10. The VIF value generated from the GWR model analysis is greater than that generated in the OLS method due to adding a weighting matrix to the parameter estimation. In addition, the results of local multicollinearity (with the GWR method) are similar to the results of global multicollinearity (with the OLS method), namely 12 independent variables have a VIF value greater than ten, and only five independent variables have a VIF value of less than 10. The results of the VIF values indicate a local multicollinearity problem in the GWR model. Therefore, the GWL method was applied to solve this problem.

3.4. Geographically Weighted Lasso (GWL) Modeling

The GWL method is a concept from the LASSO method applied in GWR modeling to overcome the problem of spatial heterogeneity and local multicollinearity. The regression coefficient generated by the GWL method will be depreciated to zero using the LASSO approach. Therefore, a coefficient that is zero does not affect the model; in other words, the GWL method selects variables that are not significant. Based on the parameter estimation results, 41 different models were obtained for each observation location. In other words, there were only significant variables in the obtained model. As an example of the results of the GRDP model for the Bali-Nusra Region using the GWL method, the following models are given:

1. GWL model for West Lombok, as an example for West Nusa Tenggara Province:

$$\hat{Y}_1 = 0{,}0090X_6 + 0{,}2565X_{11} + 0{,}1678X_{14} \quad (11)$$

Based on the model obtained, shown in Formula (11), it can be seen that the variables that significantly influence the value of GRDP in West Lombok Regency are per capita expenditure (X_6), construction (X_{11}), and household consumption expenditure (X_{14}).

2. GWL model for Jembrana, as an example for Bali Province:

$$\hat{Y}_{11} = 0{,}1495X_6 + 0{,}0802X_{13} + 0{,}3910X_{14} + 0{,}1886X_{17} \quad (12)$$

Based on the model obtained, shown in Formula (12), it can be seen that the variables that significantly influence the value of GRDP in Jembrana are per capita expenditure (X_6), provision of accommodation and food and drink (X_{13}), household consumption expenditure (X_{14}), and gross fixed capital formation (X_{17}).

3. GWL model for Alor, as an example for East Nusa Tenggara Province:

$$\hat{Y}_{20} = -0{,}0085 + 0{,}0064X_4 + 0{,}0334X_6 + 0{,}0137X_7 + 0{,}2527X_8 + 0{,}0097X_9 + 0{,}0080X_{10} + 0{,}0256X_{11} + 0{,}2755X_{13} + 0{,}6010X_{14} + 0{,}0687X_{16} \quad (13)$$

Based on the model obtained, shown in Formula (13), it can be seen that the variables that significantly influence the value of GRDP in Jembrana are old-school expectations (X_4), expenditure per capita (X_6), agriculture, forestry, and fisheries (X_7), mining and quarrying (X_8), processing industry (X_9), water procurement, waste management, waste, and recycling (X_{10}), construction (X_{11}), provision of accommodation and food and drink (X_{13}), household consumption expenditure (X_{14}), and government consumption expenditure (X_{16}).

After all the tests were carried out, 41 different models were obtained using the GWL method for each observation location, with the significant variables for each district/ city given in Table 3.

Table 3. Significant variables for each district/city

No	District/ City	Significant Variables
1	West Lombok	X_6, X_{11}, X_{14}
2	Central Lombok	$X_6, X_8, X_{13}, X_{14}, X_{17}$
3	East Lombok	$X_6, X_8, X_{10}, X_{13}, X_{14}, X_{17}$
4	Sumbawa	$X_6, X_{11}, X_{13}, X_{14}, X_{17}$
5	Dompu	$X_6, X_8, X_{10}, X_{13}, X_{14}, X_{17}$
6	Bima	$X_6, X_8, X_{10}, X_{11}, X_{13}, X_{14}, X_{17}$
7	West Sumbawa	X_{11}
8	North Lombok	$X_2, X_3, X_4, X_5, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
9	Mataram City	$X_6, X_8, X_{13}, X_{14}, X_{17}$
10	Bima City	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
11	Jembrana	$X_6, X_{13}, X_{14}, X_{17}$
12	Tabanan	$X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
13	Badung	$X_6, X_8, X_{10}, X_{13}, X_{14}, X_{17}$
14	Gianyar	$X_6, X_8, X_{13}, X_{14}, X_{17}$
15	Klungkung	$X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{13}, X_{14}, X_{15}, X_{16}$
16	Bangli	$X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}$
17	Karangasem	X_6, X_{11}, X_{14}
18	Buleleng	$X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}$
19	Denpasar City	$X_3, X_6, X_7, X_8, X_{10}, X_{13}, X_{14}, X_{16}$
20	Alor	$X_4, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
21	Belu	$X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
22	Ende	$X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
23	East Flores	$X_6, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}$
24	Kupang	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
25	Lembata	$X_1, X_2, X_4, X_6, X_7, X_8, X_9, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}$
26	Malaka	$X_6, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
27	Manggarai	$X_6, X_7, X_8, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
28	West Manggarai	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
29	East Manggarai	$X_6, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
30	Nagekeo	$X_2, X_3, X_4, X_7, X_8, X_9, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
31	Ngada	$X_6, X_8, X_{10}, X_{11}, X_{13}, X_{14}, X_{17}$
32	Rote Ndao	$X_1, X_2, X_4, X_6, X_7, X_8, X_9, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$

No	District/ City	Significant Variables
33	Sabu Raijua	$X_1, X_2, X_4, X_6, X_7, X_8, X_9, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
34	Sikka	$X_2, X_3, X_4, X_5, X_7, X_8, X_9, X_{11}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
35	West Sumba	$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}$
36	Southwest Sumba	$X_2, X_3, X_5, X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
37	Sumba Tengah	$X_6, X_8, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
38	East Sumba	$X_6, X_8, X_{10}, X_{11}, X_{13}, X_{14}, X_{17}$
39	Southern Central Timor	$X_6, X_8, X_{10}, X_{11}, X_{13}, X_{14}$
40	Northern Central Timor	$X_6, X_7, X_8, X_9, X_{10}, X_{11}, X_{13}, X_{14}, X_{16}$
41	Kupang City	$X_6, X_{10}, X_{11}, X_{14}$

3.5. Model Fit Test

The goodness of fit test is used to determine the stability of the regression model obtained in representing the observational data, and it can be seen from the value of the coefficient of determination (R^2) obtained. The R^2 value generated from the GWL model is 0.9584, meaning that the independent variable in the model obtained using the GWL method is able to explain the diversity of the GRDP value at 41 observation locations in the Bali-Nusra region of 95.84%, with 4.16% explained by other variables not included in the model. Based on the results obtained, it can be said that the model obtained is feasible to use or is considered valid because it has a high R^2 value.

4. Conclusions

Based on the analysis that has been carried out, the GRDP model for the Bali-Nusra region based on the GWL method was obtained from as many as 41 different models, which can be declared suitable for use and considered valid based on the value of the determination coefficient obtained. Based on the model obtained for each district/ city, it can be seen that the variables that significantly influence the value of GRDP are different from each other. Therefore, it is necessary to provide different policies related to Gross Regional Domestic Product (GRDP), depending on the characteristics of the location.

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

Data obtained via the Statistics Indonesia website https://bps.go.id/.

References

[1] BPS, "Statistics Indonesia, https://bps.go.id/ (Accessed in 2022)."

[2] M. Kasyfurrahman, "Application of the Geographically Weighted Lasso Method in the Case of the Gross Regional Domestic Product of West Java ," Dissertation, Indonesian University of Education, 2020.

[3] B. Lu, M. Charlton, P. Harris, and A. Fotheringham, "Geographically Weighted Regression with A Non-Euclidean Distance Metric: A Case Study using Hedonic House Price Data," *International Journal of Geographical Information Science*, vol. 28, no. 4, pp. 660–681, 2014.

[4] M. Agustina, R. Wasono, and Darsyah MY, "Pemodelan Geographically Weighted Regression (GWR) pada Tingkat Kemiskinan di Provinsi Jawa Tengah (translate: Geographically Weighted Regression (GWR) Modeling at the Poverty Level in Central Java Province)," *Statistika*, vol. 3, no. 2, pp. 67–74, 2015.

[5] M. Prajanati, L. Harsyiah, and N. Fitriyani, "Model of Human Development Index in West Nusa Tenggara Province using Geographically Weighted Ridge Regression Method (GWRR)," in *Proceeding of The 3rd International Conference on Natural Sciences, Mathematics, Applications, Research, and Technology*, Bali, Indonesia, 2022.

[6] I. Sumarni, N. Fitriyani, and Z. Baskara, "Modeling of Factors Affecting Poverty in West Nusa Tenggara Province in 2020 With Geographically Weighted Logistic Regression," in *Mathematics National Conference (Konferensi Nasional Matematika) XXI (not published)*, Mataram, Indonesia, 2022.

[7] B. Justitiaski, N. Fitriyani, and S. Bahri, "Modeling the Number of Infant Mortality in East Lombok using Geographically Weighted Poisson Regression," *Eigen Mathematics Journal*, vol. 5, no. 2, pp. 100–108, 2022.

[8] A. Ramadhan, H. Pramoedyo, and R. Fitriani, "Perbedaan Metode Geographically Weighted Lasso (GWL)-Lokal dan Geographically Weighted Lasso (GWL) Global dalam Mengatasi Kasus Multikolinieritas pada Model Geographically Weighted Regression (GWR) (translate: Differences in Geographically Weighted Lasso (GWL)-Local and Global Geographically Weighted Lasso (GWL) Methods in Overcoming Multicollinearity Cases in Geographically Weighted Regression (GWR) Models)," *Jurnal Mahasiswa Statistik*, vol. 1, no. 2, pp. 93–96, 2013.

[9] R. Kurniawan and B. Yuniarto, *Analisis Regresi Dasar dan Penerapannya dengan R (translate: Basic Regression Analysis and Its Application with R)*. Jakarta: PT. Karisma Putra Utama, 2016.

[10] A. Setiyorini, J. Suprijadi, and B. Handoko, "Implementations of Geographically Weighted Lasso in Spatial Data with Multicollinearity (Case Study: Poverty Modeling of Java Island)," in *AIP Conference Proceedings, 1827(1):020003*, AIP Publishing LLC, 2017.

[11] J. Kim, "Multicollinearity and Misleading Statistical Results," *Korean J Anesthesiol*, vol. 72, no. 6, pp. 558–569, 2019.

[12] D. Desriwendi, A. Hoyyi, and Wuryandari, "Pemodelan Geographically Weighted Logistic Regression (GWLR) dengan Fungsi Pembobot Fixed Gaussian Kernel dan Adaptive Gaussian Kernel; Studi Kasus: Laju Pertumbuhan Penduduk Provinsi Jawa Tengah (translate: Geographically Weighted Logistic Regression (GWLR) Modeling with Weighted Functions of Fixed Gaussian Kernel and Adaptive Gaussian Kernel; Case Study: Population Growth Rate of Central Java Province)," *Jurnal Gaussian*, vol. 4, no. 2, pp. 193–204, 2015.

[13] R. Pamungkas, H. Yasin, and R. Rahmawati, "Perbandingan Model GWR dengan Fixed dan Adaptive Bandwidth untuk Persentase Penduduk Miskin di Jawa Tengah (translate: Comparison of the GWR Model with Fixed and Adaptive Bandwidth for the Percentage of Poor People in Central Java)," *Jurnal Gaussian*, vol. 5, no. 3, pp. 535–544, 2016.

[14] A. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. West Sussex: John Wiley & Sons, 2002.

[15] M. Nadya, W. Rahayu, and V. Santi, "Analisis Geographically Weighted Regression (GWR) pada Kasus Pneumonia Balita di Provinsi Jawa Barat (translate: Geographically Weighted Regression (GWR) Analysis in Toddler Pneumonia Cases in West Java Province)," *Jurnal Statistika dan Aplikasinya*, vol. 1, no. 1, pp. 23–32, 2017.

[16] A. Widarjono, *Ekonometrika Teori dan Aplikasinya (translate: Econometrics Theory and Its Applications)*. Yogjakarta: Ekonisia, 2005.

DOI: https://doi.org/10.34123/jurnalasks.v16i1.459

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 16 No. 1, June 2024
e- ISSN: 2615-1367, p-ISSN:2086-4132

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

Modeling Multi-Output Back-Propagation DNN for Forecasting Indonesian Export-Import

Rengganis Woro Maharsi [1*], Wisnowan Hendy Saputra[2], Nila Ayu Nur Roosyidah[3], Dedy Dwi Prastyo[4], Santi Puteri Rahayu[5]

[1]BPS- Statistics Indonesia, Tanggamus Regency, Lampung, [2,4,5]Departement of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya 60111, Indonesia, [2]BPS-Statistics Indonesia, Poso Regency, Central Sulawesi
**Corresponding Author: E-mail address: [1]rengganis@bps.go.id*

ARTICLE INFO

Article history:
Received 11 December, 2022
Revised 27 December, 2023
Accepted 13 June, 2024
Published 30 June, 2024

Keywords:
Back-propagation, Deep Neural Network, Export-Import, Forecasting, Multi-output

Abstract

Introduction/Main Objectives: International trade through the mechanisms of exports and imports plays a significant role in the Indonesian economy, making the timely availability of export and import value data crucial. **Background Problems:** Export and import values are influenced by inflation and exchange rate factors. **Novelty:** This study identifies two categories of variables, namely output (export value and import value) and input (inflation rate and the exchange rate of the Rupiah against the US Dollar). **Research Methods:** the research approach utilizes a Multi-output Deep Neural Network (DNN) with a Back-propagation algorithm to model the input-output relationship. The method can provide forecasting results for two or more bivariate or multivariate output variables. **Finding/Results:** The modeling analysis results indicate that the optimal model network structure is DNN (3.4). This model successfully predicts output 1 (export value) and output 2 (import value) with Mean Absolute Percentage Error (MAPE) rates of 13.76% and 13.63%, respectively. Additionally, the forecasting results show predicted export and import values for November to be US$ 16,208.13 billion and US$ 15,105.33 billion, respectively. These findings offer important insights into the direction of Indonesia's international trade movement, which can serve as a basis for future economic decision-making.

1. Introduction

Globalization economy resulted in a significant increase in international trade activity, particularly in the value of exports and imports. The growth in the value of exports and imports, which was triggered by rising commodity prices and recovering global demand, is a major concern for export and import industry players. In Indonesia, the impressive performance of the trade balance has succeeded in keeping the current account deficit low, namely below 1 percent of Gross Domestic Product (GDP) in 2020 and Semester I-2021 [1].

The increase in Indonesian exports to the international market has encouraged growth in the domestic production of goods and services, which indicates the need for greater production input factors, including labor. This phenomenon also triggers an increase in demand for labor, which in turn leads to full employment and efficiency in economic growth.

In the context of imports, research by Rofiyandi [2] shows that the right imports can provide high efficiency in the production process by providing raw materials at cheaper prices and sophisticated industrial equipment that can increase output. However, it is important to remember that excessive imports, especially of consumer goods, can hurt the economy and cause a balance of payments deficit.

International trade is a key element in supporting a country's economy, with accurate and timely data regarding exports and imports being the key for the government to formulate appropriate fiscal and monetary policies [3]. Therefore, developing a prediction model that can predict these two components accurately at run time, based on relevant indicators, is important (Forecasting).

In the context of the influence of external factors, such as currency exchange rates and inflation [4], on exports and imports, previous research shows that changes in currency exchange rates can directly influence domestic prices of goods and services, with a significant impact on the volume of exports and imports. Likewise with inflation, which can affect exports and imports simultaneously by causing an increase in the prices of goods and services, which ultimately has an impact on the volume of international trade.

In dealing with the complexity of export and import data, as well as the limitations of classical assumptions in regression analysis, the use of Machine Learning methods, especially Multi-output Back-propagation DNN [5], becomes relevant. This method was chosen because of its ability to predict data at a running time based on independent variable information, without having to rely on certain assumptions which are often difficult to fulfill in an economic context.Thus, this research attempts to fill the gap in the literature by exploring the potential of Machine Learning methods in predicting the value of Indonesian exports and imports.

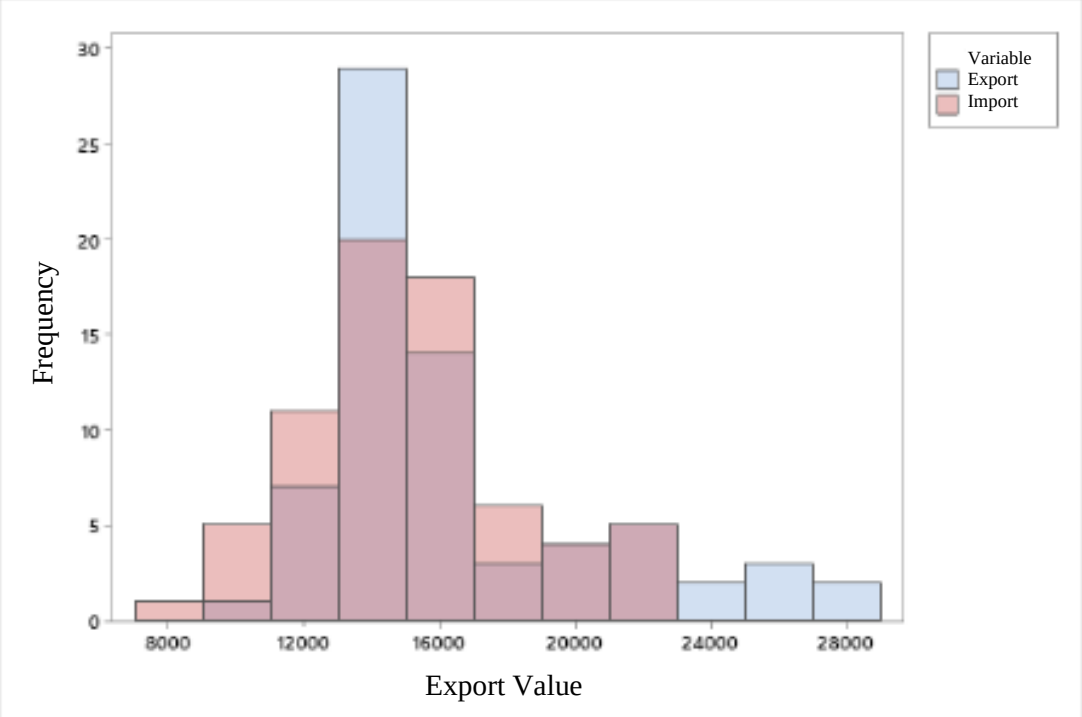


Figure. 1. Histogram of Indonesian Export and Import Values, January 2017 to October 2022

2. Material and Methods

2.1. Model Prediction

The modeling introduced to overcome prediction problems is mostly a linear model. The linear model has absolute requirements that must be met, namely the existence of a linear relationship. Most methods are only designed to solve problems in the case of linear predictions.

Some real cases certainly cannot demand certain conditions from the resulting data and are relatively freer. Therefore, in some cases, non-linear conditions are often found. Most of the literature directs the decomposition method to separate linear and non-linear aspects so that they can be analyzed separately.

Most of the time series data in various cases are non-linear conditions. The decomposition process is considered to be able to reduce or even eliminate important information from non-linear time series data. Yokota et al. [6] suggested using non-linear methods to overcome problems in these conditions. Apart from that, Levenberg [7] confirmed that non-linear methods were developed to solve problems in non-linear cases.

As time goes by, many modern methods have emerged that focus on non-linear conditions [8]. The Machine Learning (ML) method is a non-linear modeling method that utilizes computational capabilities as an optimization tool [9]. Artificial Neural Network (ANN) or Neural Network (NN) is an ML that is adapted from the functioning of nerves in humans. NN was first introduced by McCulloch and Pitts in 1943 [10]. The idea of this method is to predict output values based on several input values that have been given from several observations. In the case of time series, observations can be interpreted as a time sequence with a certain length.

In general, many types of NN methods are developed according to the objectives and background of the existing problem [11-12-13]. The Deep Neural Network (DNN) method was developed to obtain accurate output predictions with increasingly complex computational algorithms. Apart from that, Mishra and Passos [14] explained that ordinary NNs cannot be used in multivariate cases and explained that there is a need to develop Multi-output NNs.

Modeling that is focused on predicting time series data certainly requires evaluation. The Mean Absolute Percentage Error (MAPE) is the best performance measure used on non-negative time series data. MAPE has a clearer range of values and interpretation boundaries.

2.2. Linearity

The theory of linearity in time series data explains that time series data can be approximated by a straight line or has a relatively straight pattern. In some cases, especially for time series data, linear conditions are very difficult to recognize. Therefore, many tests have been developed which are used to validate the linear condition of data.

Teräsvirta [15] is one of the researchers who developed the linearity test. This test was first introduced in 1994 and was then called the Terasvirta test [15]. The idea of the Terasvirta test was specifically adapted based on the NN model. According to Prabowo et al. [16], the Terasvirta test is the most accurate linearity test because it has the highest sensitivity based on other tests.

2.3. Artificial Neural Network

Artificial Neural Network (ANN) is a science developed by Warren McCulloch and Walter Pitts in 1943 [17]. ANN is an information processing pattern initiated by the biological nervous system, such as information processing in the human brain. The essence of this idea is the structure contained in information processing systems which consists of interconnected process elements (neurons) that then work simultaneously to solve certain problems.

Not far from the human network, the ANN structure consists of an input layer, a hidden layer, and an output layer. Information (α) will be received by the input layer using a certain arrival weight (w). After that, the weights will be added to the hidden layer. Then the results of the sum will be compared with the threshold value. If the value passes the threshold, it will be passed to the output layer, whereas if the value does not pass the threshold, it will not be passed to the output layer. So that maximum output is obtained. The following is the network structure of ANN.

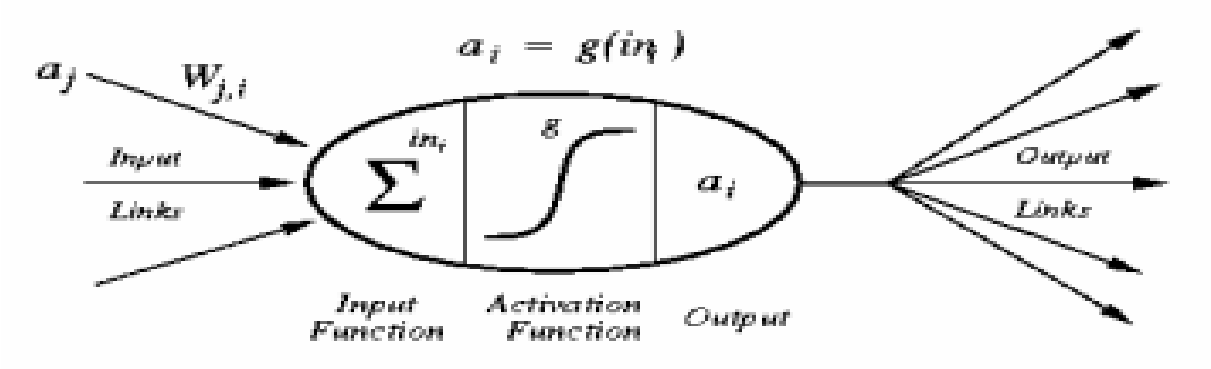


Figure. 2. Structure Network on ANN

The weights in the ANN are estimated using several one of these algorithms is Back-propagation. This algorithm was first introduced by Paul Werbos in 1974 [17]. This algorithm allows repeated weight estimation based on the weights that have been obtained up to a certain error threshold or a certain number of iterations.

2.4. Deep Neural Network

Deep Neural Network (DNN) is a development of the ANN method which involves a large number of hidden layers. In 2011, Deep NN began combining convolutional layers with max-pooling layers whose output was then passed to several fully connected layers followed by an output layer. This condition is called Convolutional Neural Networks (CNN) [17]. In general, estimation process weights in DNN are relatively the same as ANN, but with more time long as well as complexity more computing complicated. Following is one of the form structure networks on DNN.

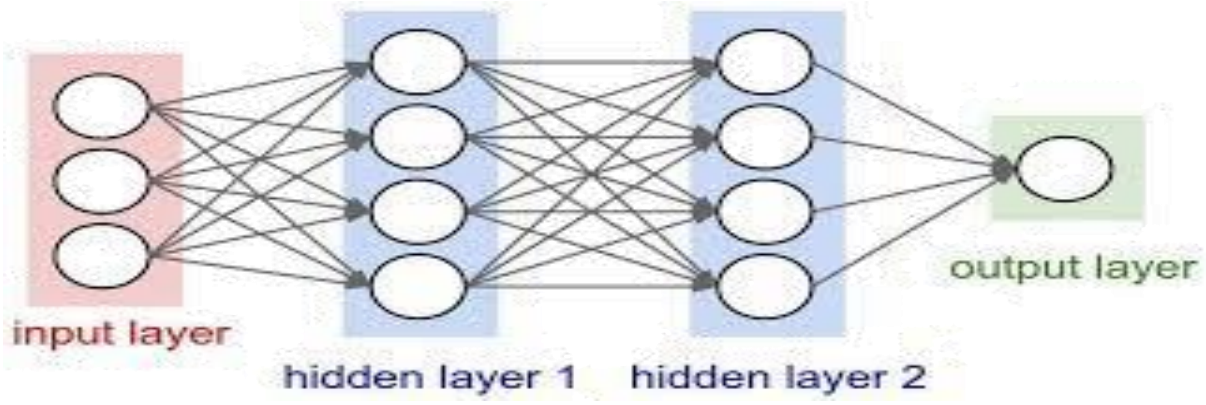


Figure. 3. One form of DNN network structure

2.5. Multi-output Neural Network

ANN is generally used for modeling data with output as much One. Multi-output NN is the development of the ANN used specifically on the type of multivariate model, meaning own total output more from One. As well as with the method Multivariate Ordinary, the output of the NN is modeled in a way direct (multivariate) to predict each value its output based on the input entered. The following is one form of Multi-output NN network structure.

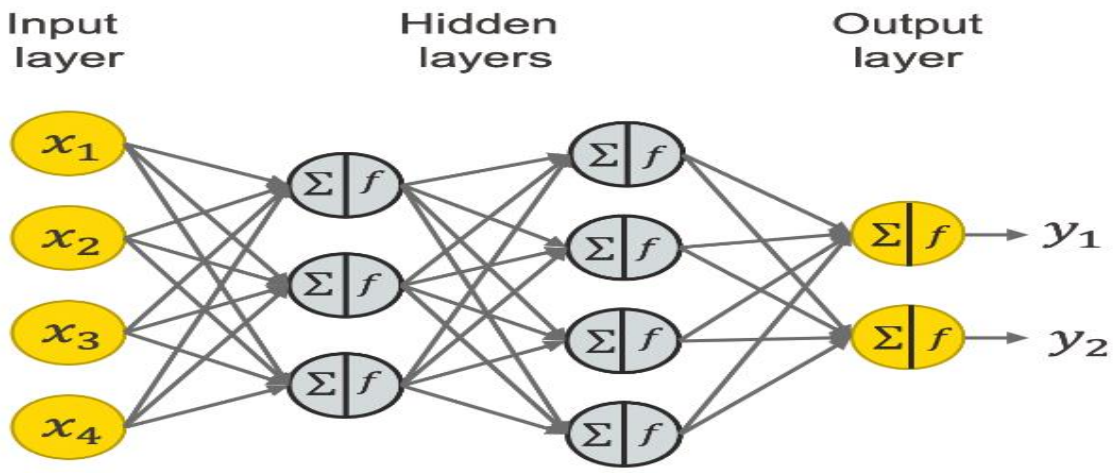


Figure. 4. One form of *multi-output* NN network structure

2.6. Prediction Performance Measures

Evaluating the performance of the time series data modeling method is carried out based on error calculations from the actual data. A prediction performance measure using Mean Absolute Percentage Error (MAPE). Counting level error based on MAPE following a formula like the following.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \times 100\% \quad (1)$$

where Y_t is the actual data, $\hat{Y}_t$ is the length of the data

Variations of MAPE values have different meanings. If the MAPE value is smaller than 10% then the predictive model's ability is very good [14]. If the MAPE value is between 10% - 20% then the prediction model's ability is good. If the MAPE value is in the range of 20% - 50% then the prediction model's ability is feasible. If the MAPE value ranges more than 50% then the forecasting ability of the model is poor.

2.7. Analysis Method

This research uses secondary data obtained from the Statistics Indonesia (BPS) and Yahoo Finance. The variables used in this research are divided into 2 categories, namely output which consists of export value (Y_1) and import value (Y_2), and input which consists of inflation (X_1) and the Rupiah exchange rate against the United States Dollar (X_2).

Variables of export value, import value, and inflation obtained from BPS accessed through www.bps.go.id. Meanwhile, exchange rate data is obtained via the Yahoo Finance website with the recorded exchange rate being the value in the closing period. The data in this study is monthly data. The research period starts from January 2017 to October 2022, so the number of observations is 70.

Table 1 below is the form of the data structure in this research.

Table 1. Research Data Structure

Month (t)	$Y_{1,t}$	$Y_{2,t}$	$X_{1,t}$	$X_{2,t}$
1	$Y_{1,1}$	$Y_{2,1}$	$X_{1,1}$	$X_{2,1}$
2	$Y_{1,2}$	$Y_{2,2}$	$X_{1,2}$	$X_{2,2}$
⋮	⋮	⋮	⋮	⋮
70	$Y_{1,70}$	$Y_{2,70}$	$X_{1,70}$	$X_{2,70}$

Data was analyzed computationally using the R language with the general stages stated in the following flowchart:

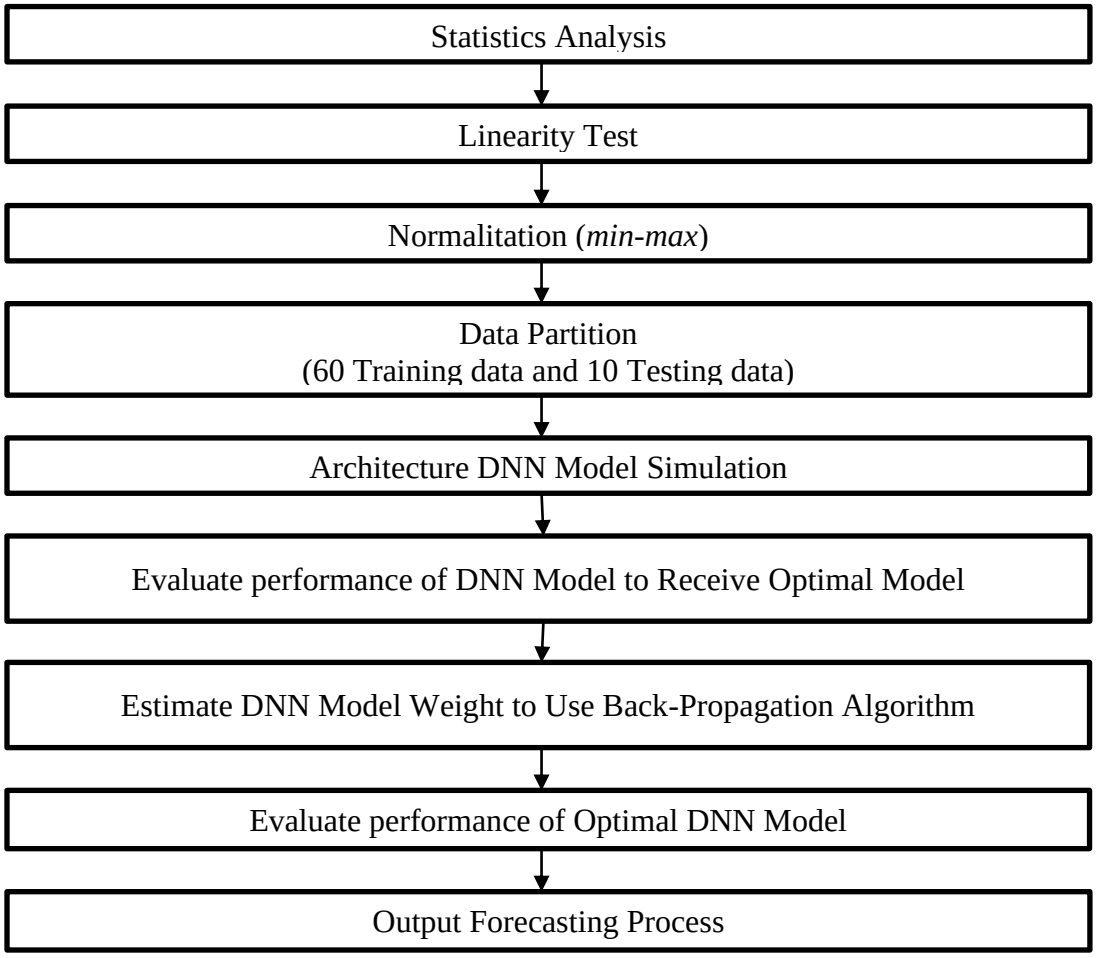


Figure. 5. *Flowchart of* general research stages.

3. Result and Discussion

3.1. Statistics Descriptive

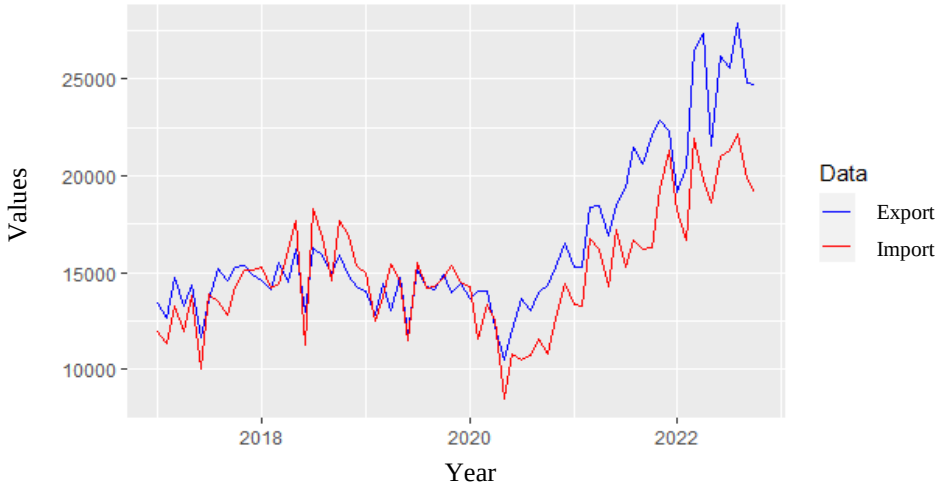


Figure. 6. Development of Indonesian Export and Import Values, January 2017 to October 2022

In general, the development of Indonesia's export and import values, as shown in Figure 5, has almost the same pattern. The majority of Indonesia's imports are capital goods as well as raw materials and industrial auxiliary materials [18]. Capital goods (capital) act as input factors together with other production factors, such as labor (labor) and natural resource factors (land) as raw materials and intermediate materials.

Some raw materials and industrial auxiliary materials cannot be obtained domestically, so imports are required. Apart from that, capital goods also play an important role in building facilities and infrastructure to support industrial activities, such as the construction of transportation infrastructure which supports mobility and distribution of output. Because optimizing input factors in the production process is closely related to the amount of output obtained, the increase in imports is correlated or in line with the increase in exports in Indonesia.

Judging from the movement, Figure 1 shows that both export and import values showed a pattern that tended to be constant from January 2017 to December 2019, then experienced the lowest decline during the research period, starting from January to March 2020 due to the impact of the spread of Covid-19 virus infections. which influences the continuity of production of goods and services in several business sectors in the world. However, starting from April 2020 the movement of both tended to increase even though they experienced a decline several times. The value of exports exceeded the value of imports until the end of the research period, which indicates that during that period the foreign trade balance experienced a surplus.

3.2. Linearity Test

Based on Figure 4, it can be seen that the export and import data relatively have a non-linear pattern. Based on this information, it is of course suspected that the relationship between input and output is not linear. To validate this assumption, a linearity test was carried out using the Terasvirta test. The test was carried out with the following hypothesis.

H_0: Data has a linear relationship

H_1: The data does not have a linear relationship

The results of the tests that have been carried out are obtained according to Table 2. as follows.

Table 2. Terasvirta Test Results

Inputs	Outputs	P-value
X_1 And X_2	Y_1	0.01210
	Y_2	0.05667

Based on Table 2, it is obtained information that with the level significance of the first 5% of output own non-linear relationship with the input used. However, for second output can conclude a relatively linear relationship with the input used. Since there are outputs that have nonlinear relationships, then the condition that machine learning modeling will be better if used than a linear model has been valid.

3.3. Simulation Structure Network

The process of simulating the structure of the network is carried out to obtain a combination of the number of neurons for each hidden layer that is optimal using normalized training data. The amount of hidden layer used is 2 with the putative optimal neuron amount starting from 2 to 4 (twice the number of inputs).

The simulation results obtained are presented in Table 3.

Table 3. Simulation Results Structure Network

Hidden 1	Hidden 2	MAPE Y_1	MAPE Y_2	Number of MAPE
3	4	0.1028	0.1275	0.2303*
2	2	0.1031	0.1274	0.2304
2	4	0.1025	0.1280	0.2306
3	2	0.1036	0.1329	0.2365
3	3	0.1037	0.1330	0.2367
2	3	0.1023	0.1361	0.2383
4	4	0.1041	0.1347	0.2388
4	3	0.1035	0.1357	0.2392
4	2	0.1046	0.1361	0.2407

Table 3 provides information that the most optimal NN network structure based on the MAPE value is NN(3,4). This network structure can predict normalized values Y_1 with an error rate of 0.1028% and normalized values Y_2 with an error rate of 0.1275%. Based on this error level, we can infer that the model has a very excellent prediction capacity for the actual output normalized value.

3.4. Model Performance Comparison

The model performance comparison is intended to see the model performance on both data compositions, namely training and testing. The performance results on the training data obtained the same value as the previous stage. The following are the results of the model performance comparison presented in Table 4.

Table 4. Model Performance Comparison

Data	MAPE(Y_1)	MAPE(Y_2)
Training	0.1028	0.1275
Testing	19.0417	13.1136

Based on Table 4, it can be seen that the MAPE in the testing data is very different from the training data. This is due to the limited amount of research data used for testing data. However, this large MAPE value is still considered good because it is below 20%. Thus, it can be concluded that the NN structure has the ability to predict the normalization of outputs on training and testing data well.

3.5. Final Model Neuron Weight Estimation

After obtaining the most optimal network structure, this structure is then applied to all data (training + testing) and then the weights are estimated for each neuron. The estimation was carried out based on the NN(3,4) architecture with the Back-propagation algorithm with a maximum number of interactions determined to be 1 million times. The weight estimation results obtained are presented in Table 5., Table 6., and Table 7. sequentially for hidden layer 1, hidden layer 2, and the following output layers.

Table 5. Weights in Hidden Layer 1

	Neuron 1	Neuron 2	Neuron 3
Biased	1.602	0.736	0.886
X_1	-1.036	8.553	7.070
X_2	-5.034	-4.871	-2.309

Table 6. Weights in Hidden Layer 2

	Neuron 1	Neuron 2	Neuron 3	Neuron 4
Biased	1.487	1.633	-2.354	1.072
N_1	2.567	-3.861	-0.235	1.631
N_2	2.144	2.559	0.194	0.835
N_3	-3.528	-4.053	0.465	1.242

Table 7. Weights in the Output Layer

	Y_1	Y_2
Biased	-1.231	-1.265
N_1	3.101	2.623
N_2	4.265	3.581
N_3	-0.702	0.969
N_4	-1.546	-1.089

* N_i: neuron *output* in *hidden layer 1*

Estimating the neuron weights of the final model according to Table 5., Table 6., and Table 7. requires 704201 iterations to reach convergence. Thus, the form of the final model network structure obtained is illustrated in Figure 6 below.

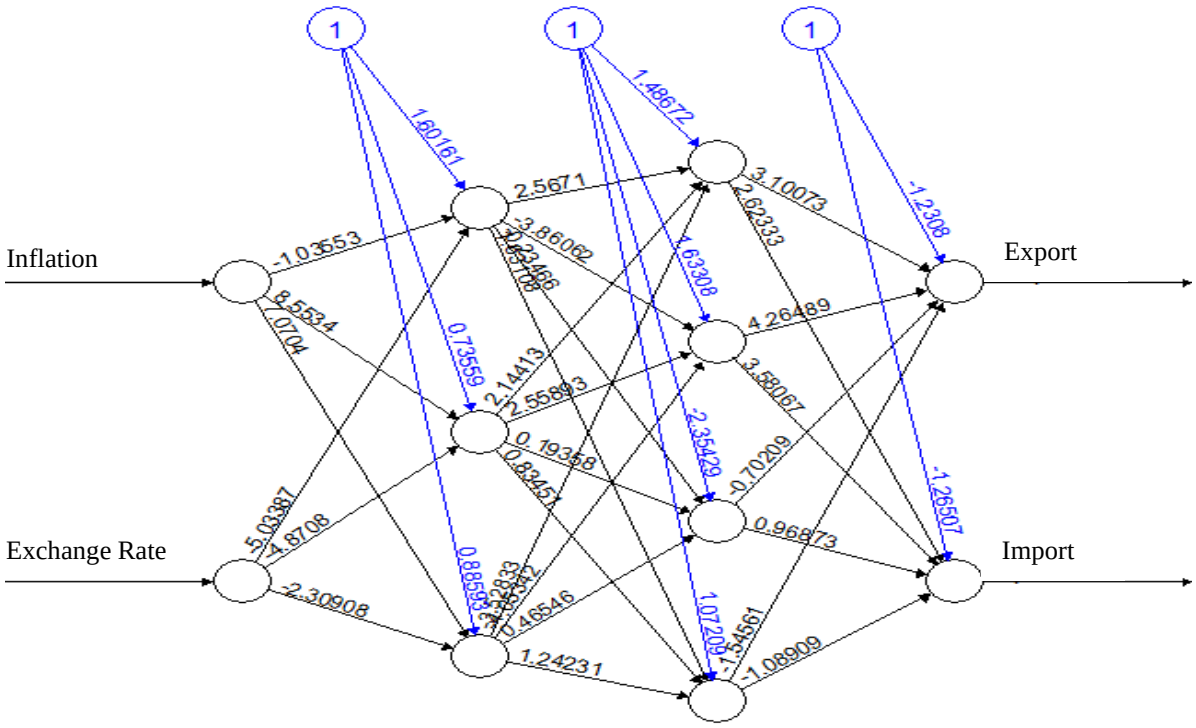


Figure. 7. Final Model Network Structure

Mathematically, the final model equation is written as:

$$\hat{y}_{t,m} = f^{(o)}\left(\sum_{k=1}^{K} v_{k,m}^{(o)} f^{(h_1)}\left(\sum_{j=1}^{J} v_{j,k,m}^{(h_2)} f^{(h_1)}\left(\sum_{i=1}^{p} v_{i,j,m}^{(h_1)} x_{i,t,m} + b_{j,m}^{(h_1)}\right) + b_{k,m}^{(h_2)}\right) + b_{m}^{(o)}\right) \quad (2)$$

where $m = 1,2$ shows the index for the first (export) and second (import) output. The weight and bias (b) parameters (v) for each output refers to the model architecture as in Figure 6.

3.6. Late Model Performance

Before carrying out the forecasting process, the model obtained must first go through a performance-checking stage to see the model's ability to predict actual data. Performance checking is done by calculating the prediction error rate using MAPE and by visually comparing actual and predicted data. The results of calculating the MAPE model obtained are presented in Table 8.

Table 8. Final Model MAPE

Outputs	MAPE
Y_1	13.76172
Y_2	13.62980

Based on Table 8, the MAPE of both outputs obtained values that are smaller than 20%. Thus, it can be concluded that the final model obtained can predict the actual data well.

After getting an interpretation based on calculating the error rate, visualization is then carried out to compare the actual output and the predictions. The visualization for each output 1 and 2 is presented in Figures 7 and 8 below.

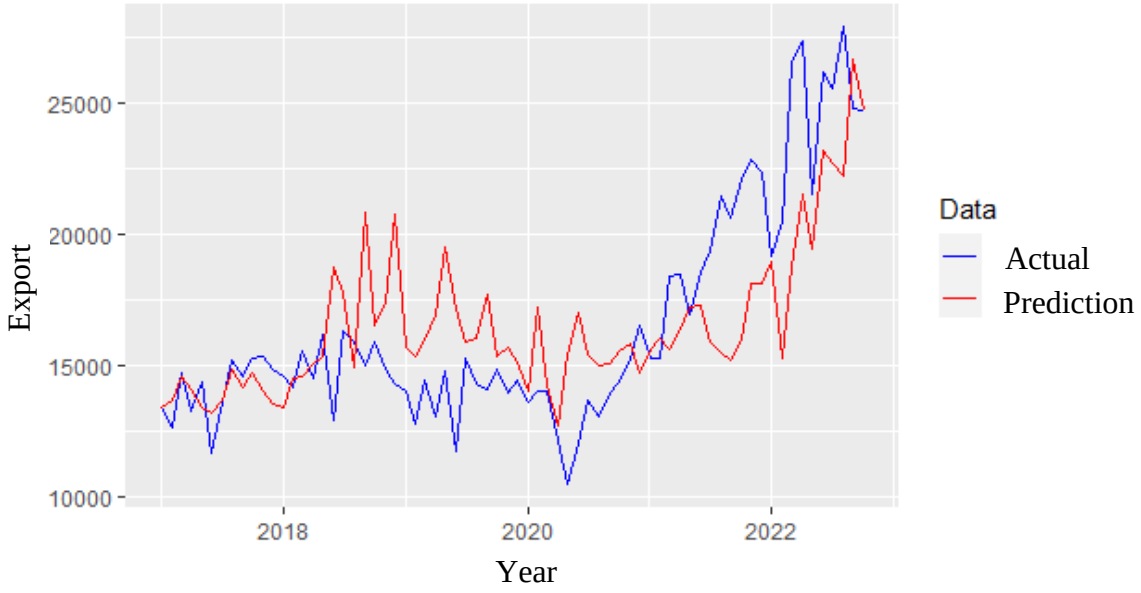


Figure. 8. Comparison of Actual and Predictions for Output 1

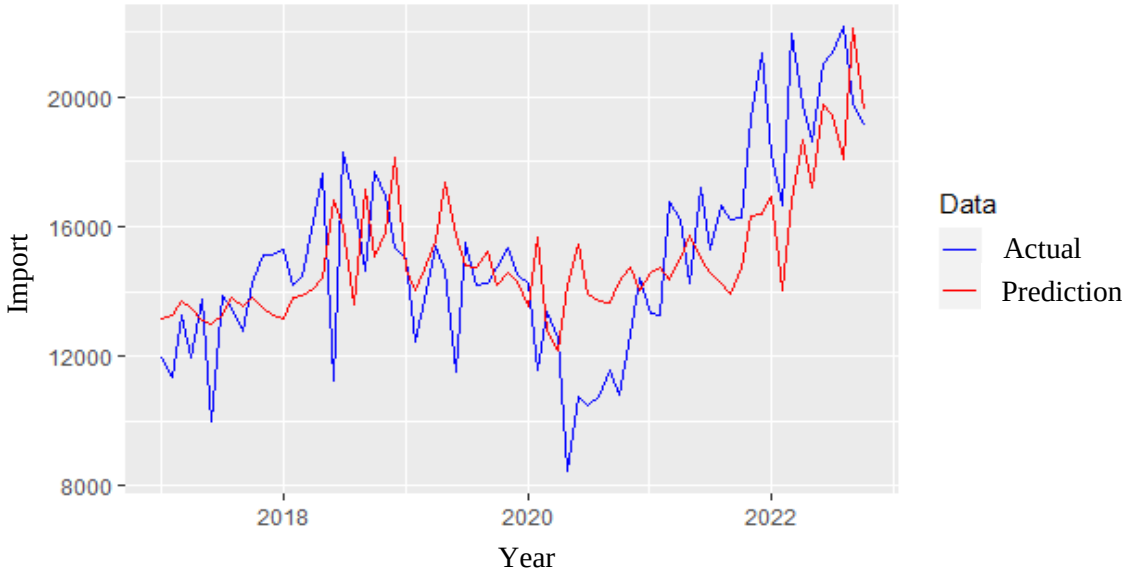


Figure. 9. Comparison of Actual and Predictions for Output 2

Based on Figure 8 and Figure 9, it can be seen that the model predictions are relatively able to follow the volatility pattern (pattern of ups and downs in data) which is quite fluctuating from the actual output.

Thus, it can be concluded that based on the level of prediction error and visualization the model has been able to predict actual data well.

3.7. Forecasting Model

After validating the final model structure based on Back-propagation DNN, forecasting is then carried out to obtain the current value of the output. The input variables used are the normalized values of inflation (X_1) and the exchange rate (X_2), the data of which are stated in Table 9 below.

Table 9. Forecasting Process Input

Inputs	Normalized Value
X_1	0.6181896
X_2	0.2500000

Based on the input listed in Table 9, a prediction (forecasting) is obtained for the output presented in Table 10 below.

Table 10. Forecasting Process Input

Outputs	Forecasting Value (Million US$)
Y_1	16208.13
Y_2	15105.33

3.8. Discussion

Forecasting statistical data that is multivariate (in this case called multi-output) generally has the assumption that the influence of predictors (in this case called input) must be linear, but in the case of real data, this assumption is difficult to fulfill. Therefore, modeling is carried out using DNN so that it can ignore linearity assumptions. DNN modeling cannot be done to interpret the influence of input on output, but rather to obtain the most accurate prediction values. In this research, modeling begins with

linearity testing to ensure the urgency of using the DNN model. After that, the DNN modeling simulation process continues with a limit of the number of hidden layers of 2 and the number of nodes in each hidden layer of 2 to 4. Validation of the model used is by calculating the MAPE accuracy value for training and testing data, both of which must have a value below 20% so it can be categorized as a good prediction. After obtaining a model with these conditions, the model is applied to the full data (a combination of training and testing) and then used to forecast 1 data observation in the future which is stated in Table 10.

4. Conclusion

The study tested the potential of using Machine Learning methods, in particular Multi-Output Back-propagation DNN, to predict Indonesian exports and imports. The results of the research show that the model developed was able to predict export and import values with relatively low error rates, proving the effectiveness of this method in dealing with the complexity of multi-output data. This research provides potential applications in economic analysis, i.e. more accurate and responsive modeling to market conditions can be an important tool in strategic decision-making. The limitations of this research are the limitations on the quality and quantity of available data for further research, it is better to expand the scope of other machine learning methods or develop more complex models to improve the accuracy of predictions. This research provides a basis for a better understanding of Indonesia's international trade dynamics and the impact of economic policy on it.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

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

Funding

This study received no external funding.

Underlying data

This research uses secondary data obtained from BPS- Statistics Indonesia and Yahoo Finance.

References

[1] C. M. for Economic Affairs, ‘Maintaining Economic Growth and Controlling the Covid-19 Pandemic is Proof of the Accuracy of Government Policies and Programs’. 2021.

[2] M. Y. Rofiyandi, ‘Imports are not always bad, get to know the meaning of imports and their benefits’. 2022.

[3] N. P. Okenna and B. Adesanya, 'International trade and the economies of developing countries', *American International Journal of Multidisciplinary Scientific Research*, vol. 6, no. 2, pp. 31–39, 2020.
[4] B. Setyorani, 'Pengaruh nilai tukar terhadap ekspor dan jumlah uang beredar di indonesia', in *FORUM EKONOMI: Jurnal Ekonomi, Manajemen dan Akuntansi*, 2018, vol. 20, pp. 1–11.
[5] A. Smith and B. Johnson, 'Multi-output Deep Neural Networks for Economic Forecasting: A Review', *Journal of Economic Forecasting*, vol. 10, no. 3, pp. 45–62, 2021.
[6] T. Yokota, M. Gen, and Y.X. Li, "Genetic algorithm for non-linear mixed integer programming problems and its applications," Computers & Industrial Engineering, vol. 30, no. 4, pp. 905-917, 1996.
[7] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," Quarterly of Applied Mathematics, vol. 2, no. 2, pp. 164-168, 1944.
[8] K. D. Hartomo, J. A. Lopo, and H. D. Purnomo, 'Enhancing Multi-Output Time Series Forecasting with Encoder-Decoder Networks', *Journal of Information Systems Engineering & Business Intelligence*, vol. 9, no. 2, 2023.
[9] W. Samek and K.-R. Müller, 'Towards explainable artificial intelligence', *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 5–22, 2019.
[10] I. B. Darmawan, M. Maimunah, and R. N. Whidiasih, 'Identifikasi Warna Kerabang Telur Ayam Ras Menggunakan Jaringan Syaraf Tiruan', *PIKSEL: Penelitian Ilmu Komputer Sistem Embedded and Logic*, vol. 6, no. 2, pp. 189–200, 2018.
[11] C. Dai, 'A method of forecasting trade export volume based on back-propagation neural network', *Neural Computing and Applications*, vol. 35, no. 12, pp. 8775–8784, 2023
[12] R. Garcia and L. Santos, 'Forecasting Trade Flows Using Artificial Neural Networks: A Comparative Study', *Journal of Business and Economic Forecasting*, vol. 8, no. 3, pp. 112–125, 2020.
[13] M.-L. Shen, C.-F. Lee, H.-H. Liu, P.-Y. Chang, and C.-H. Yang, 'Effective multinational trade forecasting using LSTM recurrent neural network', *Expert Systems with Applications*, vol. 182, p. 115199, 2021.
[14] P. Mishra and D. Passos, 'Multi-output 1-dimensional convolutional neural networks for simultaneous prediction of different traits of fruit based on near-infrared spectroscopy', *Postharvest Biology and Technology*, vol. 183, p. 111741, 2022.
[15] T. Teräsvirta, "Testing linearity and modelling nonlinear time series," Kybernetika, vol. 30, no. 3, pp. 319-330, 1994.
[16] H. Prabowo, S. Suhartono, and D. D. Prastyo, 'The Performance of Ramsey Test, White Test and Terasvirta Test in Detecting Nonlinearity', *Inferensi*, vol. 3, no. 1, pp. 1–12, 2020.
[17] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, 'Convolutional neural networks: an overview and application in radiology', *Insights into imaging*, vol. 9, pp. 611–629, 2018.
[18] BPS, *Indonesia's Gross Domestic Product by Expenditure, 2017-2021*. Jakarta: BPS, 2022

DOI: https://doi.org/10.34123/jurnalasks.v16i1.542

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

Performance Study of Prediction Intervals with Random Forest for Poverty Data Analysis

Nina Valentika[1], Khairil Anwar Notodiputro[2], Bagus Sartono[3*]

[1]Universitas Pamulang, South Tangerang, Indonesia, [2,3]IPB University, Bogor, Indonesia
**Corresponding Author: E-mail address:* bagusco@apps.ipb.ac.id

ARTICLE INFO

Article history:
Received 18 September, 2023
Revised 9 April, 2024
Accepted 9 April, 2024
Published 30 June, 2024

Keywords:
LM; SPI; Quant; HDR; CHDR.

Abstract

Introduction/Main Objectives: Determine the prediction interval with for analyzing poverty data at the Regency/City level in Indonesia. **Background Problems:** Poverty will be a topic in various discussion and debates in the future. **Novelty:** This study's methods for constructed prediction intervals are LM, Quant, SPI, HDR, and CHDR. This method can improve the prediction interval performance with Random Forests. **Research Methods:** The method for building forests and obtaining BOP in this study is CART with the LS splitting rule. **Finding/Results:** The results of this study are that the best method for one replication is HDR with 500 trees. The best method for 100 repetitions is LM. Based on hypothesis testing, there is sufficient evidence to say no difference between the LM, SPI, Quant, HDR, and CHDR methods for 100 replications at a 5% significance level.

1. Introduction

The goal of predictive modeling in the concept of building a model is to predict unknown responses from observations given the covariates. Prediction models in their simplest form aim to provide point predictions for new observations. However, point predictions do not contain information about their precision that could tell how close to the actual response the prediction is expected to be, which is often important in decision-making contexts. Therefore, although point prediction is often the primary goal of predictive analysis, assessing its reliability is equally important, and this can be achieved with prediction intervals. A prediction interval consists of a series of probability values for an actual response with an associated confidence level, usually 90% or 95%. Given that shorter prediction intervals are more informative, developing predictive models that can produce shorter prediction intervals along with point predictions is critical in assessing and measuring prediction error. In real-world applications, knowing the error of predictions other than point predictions can increase the practical value of those predictions. The classic and most commonly used approach to construct prediction intervals is the parametric approach. However, its main weakness is that its validity and performance depend heavily on the assumed functional relationship between covariates and responses [1]. Roy & Larocque [1] have reviewed a new method that improves the performance of prediction intervals with Random Forests. The two aspects explored by Roy & Larocque [1] are the method used to construct the forest and the method used to construct the prediction interval. Four methods for building forests, three from the Classification And Regression Tree (CART) paradigm and the transformation forest method. The

method to build a forest and get Bag of Observations for prediction (BOP) which has been studied by Roy & Larocque [1] is CART with splitting rules, namely Least-Squares (LS), splitting L1, and Shortest Prediction Interval (SPI). The prediction interval is constructed using the BOP, which is the set of nearest-neighbor observations [1]. Methods for constructing prediction intervals have been reviewed by Roy & Larocque [1] are the Classical method (LM), Quantile, Shortest Prediction Interval (SPI), Highest Density Region (HDR), and Contiguous HDR (CHDR).

The LM is calculated based on an intercept-only linear model using the BOP as a sample and produces a symmetric prediction interval around the prediction point. Similar to the Quantile Regression Forest (QRF) method, the quantile method is based on BOP quantiles. SPI corresponds to the shortest interval among the intervals containing the least $(1-\alpha)100\%$ number of observations in the BOP. As an alternative to SPI, HDR is the smallest region in the BOP, with the desired $(1-\alpha)$ coverage. Note that HDR is not necessarily one interval. If the distribution is multimodal, it can be formed with several intervals. CHDR is a way to obtain a single prediction interval from an HDR interval by constructing an interval with the minimum and maximum boundaries of the HDR interval [1].

Alakus et al. [2] created a Package RFpredInterval that implements 16 methods for constructing prediction intervals with Random Forests and Boosted Forests. Alakus et al. [2] also carried out a simulation regarding the splitting rule method least-squares (LS) and prediction interval methods, namely: LM, Quant, SPI, HDR, and CHDR using the Ranger package in building Random Forest. However, there is another package for building Random Forest that is available in the RFpredInterval package, namely randomForestSRC. This research wants to examine the prediction interval for Random Forest with the random ForestSRC package. This research creates a 95% prediction interval using variations studied by Roy & Larocque [1] the method used in building the forest, namely CART with splitting rules. LSand methods for building prediction intervals are LM, Quant, SPI, HDR, and CHDR. This study also applied Out-Of-Bag (OOB) calibration and the acceptable coverage range was set to [0.945, 0.955].

Poverty is one of the problems that exists in developing countries, such as Indonesia. The Central Bureau of Statistics of Indonesia (BPS) uses the concept of the ability to meet basic needs to measure poverty. In this approach, poverty is understood as an economic inability to meet basic food and non-food needs as measured by expenditure [3].

Various factors that influence the Percentage of Poor Population (PPP) are Gross Regional Domestic Product (GRDP), Life Expectancy Rate (LER), Mean Years of Schooling (MYS), Expected Years of Schooling (EYS), and Real Per Capita Expenditure (PPK). There are 5 main characteristics, namely area of residence, gender, education level, number of household members, and work status of the head of the household, which have the potential to cause household poverty in Central Java [4]. According to [5], a region that has a high GDP means the region has a good economy. The opposite applies. The economy in question is an economy that can support people's lives so that poverty does not arise. In the economic field, development performance in achieving prosperity is measured based on Gross Domestic Product (GDP) and its growth rate [6]. The Health Dimension is measured by the life expectancy indicator [7]. Life Expectancy is a tool for evaluating the government's performance in improving the welfare of the population in general, and improving health status in particular [8]. According to Anggadini [9], the higher the life expectancy, the higher the quality of public health. In the Circle of Poverty Theory, the quality of public health is reflected in the increase in the life expectancy rate (LER). Increasing community productivity can encourage economic growth thereby reducing the poverty rate, namely the higher the life expectancy, the lower the poverty rate. StudyPramesti & Bendesa [10], found that there is an influence on poverty where increasing education will reduce poverty. Indonesia is a developing country and has a large population. The problem of poverty in Indonesia cannot be avoided (Aulele et al. [11]). Poverty will be a topic in various discussions and debates in the future [12]. Based on the description above, the prediction interval from PPP become an interesting topic for study. Thus, this research aims to determine the prediction interval with *Random Forest* for analyzing poverty data at the Regency/City level in Indonesia.

2. Material and Methods

The data used is secondary data that comes from the Central Statistics Agency (BPS). Study This using data from 514 districts /cities in Indonesia in 2021. Table 1 presents variables used in the study. The software used in this study is R.

Table 1. Variables Study

Role of Variable	Variable	Unit	Symbol
Response	PPP	Percent	Y
Explainer	GRDP	Billion Rupiah	X_1
	LER	Year	X_2
	MYS	Year	X_3
	EYS	Year	X_4
	PPK	Thousand Rupiah	X_5

The steps taken in this study are

1. Exploration and description.
2. Create prediction intervals for one repetition and 100 repetition.
   a. Divide training data and test data. Amount trees used are {200,500,1000,5000}.
      i. Divide training data and test data. Amount trees used are {200,500,1000,5000}.
      ii. For 100 repetitions, 70% training data and 30% test data (notated 70:30); 80% training data and 20% test data (notated 80:20); as well as 90% training data and 10% test data (notated 90:19).
   b. Deternain the prediction interval for all methods. The regression model used in Random Forest is

$$Y = \sum_{i=1}^{5} X_i + \varepsilon. \quad (1)$$

   c. Rule to create prediction interval in Random Forest is as following:
      i. Build forest and get BOP. Method used in study This is method Least Square (LS).
      ii. Calculating Prediction Intervals using BOP. Method used in study This are LM, Quant, SPI, HDR, and CHDR. Function from packages RFpredInterval to use in study is rfpi(). Packages for Random Forest used in this study is randomForestRSC The type bootstrap used moment by root active is Sampling With Replacement (SWR).
      iii. Prediction Interval Calibration with the Use of OOB i=Information Desired coverage arranged to 95% for all methods. For all methods, calibration-based validation cross done as procedure calibration main, but also checked OOB calibration. OOB calibration for finding α_w. In the second procedure calibration, the range of possible coverage accepted is arranged to [0.945, 0.955].
3. Compare results For one repetition and 100 repetitions.
   a. For one repetition, determine the amount of tree best seen from the mark of the smallest Root Mean Absolute Error (MAE). Determine the method best seen from the mean of the length of the smallest prediction interval. Then, create a plot for all the methods that own a prediction interval for every observation, that is all methods except the HDR method.
   b. For 100 repetitions, the method compare to based on mean coverage, average length of prediction interval and percentage enhancement prediction interval length. Percentage enhancement prediction interval length for method I calculated as $100\ x\ (ml_i - ml^*)/ml^*$, where ml_i is the mean length of the prediction interval of the method $i$andml^* is mean length of prediction interval shortest from all method. Smaller value of this size shows better performance [2].

3. Results and Discussion

3.1. Data Description

Table 2. Data Description

	Y	X_1	X_2	X_3	X_4	X_5
Min	2.38	1,087	55.43	1,420	3.87	3976
Q_1	7.15	5,963	67.39	7,510	12.42	8574
Q_2	10.46	13,643	69.97	8,305	12.93	10196
Mean	12.27	37,222	69.66	8,437	13.02	10325
Q_3	14.89	30,895	72.04	9,338	13.65	11719
Max	41.66	861,000	77.73	12,830	17.80	23888

Based on Table 2, it is found that the smallest percentage of poor people for districts/cities in Indonesia in 2021 is 2.38%, namely Sawah Lunto City. The largest percentage of poor people for districts/cities in Indonesia in 2021 is 41.66%, namely Intan Jaya Regency. Correlations between variables are presented in Table 3.

Table 3. Correlation beetween variables

	Y	X_1	X_2	X_3	X_4	X_5
Y	1.00	-0.08	-0.54	-0.54	-0.43	-0.64
X_1	-0.08	1.00	0.21	0.17	0.09	0.34
X_2	-0.54	0.21	1.00	0.42	0.37	0.57
X_3	-0.54	0.17	0.42	1.00	0.78	0.67
X_4	-0.43	0.09	0.37	0.78	1.00	0.52
X_5	-0.64	0.34	0.57	0.67	0.52	1.00

Based on Table 3, it is found that GRDP, LER, MYS, EYS, and PPK have a negative relationship with PPP.

3.2. Prediction Interval for One Repeat

This section presents the prediction intervals in one replication with the number of trees {200, 500, 1000, 5000} and the 70% rule for training data and 30% for test data. A comparison of all methods for prediction intervals in one replication is presented in Table 4.

Table 4. Comparison of all methods for prediction intervals in one replication

Number of Trees (MAE) (RMSE)	Method	Mean length Prediction Interval	*Coverage Levels* (in %)	α_w (in %)
200	L.M	16.6	92.9	5
(3,607)	SPI	17.4	94.2	2
(4,788)	Quant	17.5	94.2	3
	HDR	16.7	94.2	4
	CHDR	16.8	94.2	4
500	L.M	18.2	97.4	5
(3,219)	SPI	18.2	97.4	2.5
(4,325)	Quant	19.2	98.1	3
	HDR	17.1	98.1	4
	CHDR	17.1	97.4	5

1000	L.M	16.3	95.5	5
(3,343)	SPI	16.1	94.2	3
(4,435)	Quant	17.6	94.2	3
	HDR	16.9	96.8	3
	CHDR	15.7	94.8	4.5
5000	L.M	17.3	96.1	5
(3,657)	SPI	20.0	97.4	1
(4.8)	Quant	18.4	94.2	3
	HDR	19.4	97.4	0.4
	CHDR	18.6	95.5	2.8

Based on Table 4, it is found that the number of trees that have the smallest MAE and RMSE is 500. The method that has the smallest mean of prediction interval length for one repetition is HDR with 500 trees. Thus, the best method for a single replicate is HDR with 500 trees. Plots for methods that have only one prediction interval for each observation (all methods except the HDR method) are presented in Figures 1 to Figure 4. HDR allows multiple prediction intervals for one observation [2].

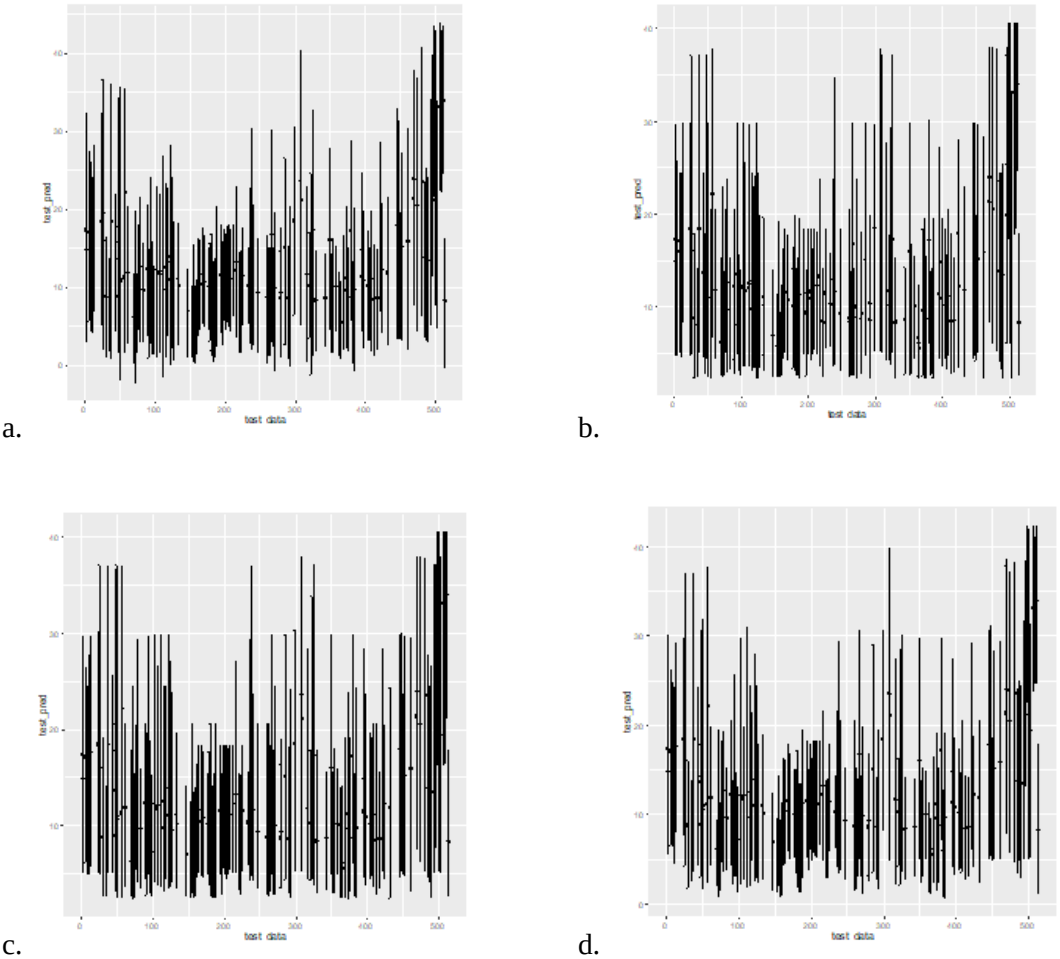


Figure. 1. (a) LS-LM method; (b) LS-SPI method; (c) LS-Quant method; (d) LS-CHDR method

Based on Figures 1, it is found that the response data in the test data is mostly within the prediction interval.

3.3. Prediction Interval for 100 Repetition

The measure used in this research to evaluate the following performance by Roy & Larocque [1] is by using mean coverage and average prediction interval length. Table 5 presents the mean level of coverage for each method from 100 replications.

Table 5. Mean level of coverage for each method from 100 replication

Number of trees (*Split*)*)	*Mean coverage* (in%)				
	L.M	SPI	Quant	HDR	CHDR
200(a)	95.2	95.9	95.3	95.6	95.5
200(b)	95.3	95.9	95.6	95.5	95.4
200(c)	95.4	95.4	95.1	95.8	95.7
500(a)	94.8	95.1	94.8	95.4	95.0
500(b)	95.0	95.3	95.2	95.2	95.2
500(c)	94.8	94.9	94.6	95.5	95.3
1000(a)	94.7	95.3	95.0	95.2	95.0
1000(b)	95.2	95.0	94.9	95.4	95.3
1000(c)	95.5	95.2	94.8	95.3	95.2
5000(a)	95.1	95.3	95.3	95.0	94.9
5000(b)	95.0	95.2	95.2	95.4	95.2
5000(c)	94.4	94.8	94.7	95.1	94.6

*)

(a) 70:30

(b) 80:20

(c) 90:10

Based on Table 5, most of the LM methods have a mean coverage difference with the desired coverage (95%) being the smallest compared to other methods. The LM method has better accuracy than other methods based on coverage. Thus, the LM method is indicated to be the best method based on mean coverage. Table 6 presents the average length of the prediction interval from 100 repetitions.

Table 6. Average length of prediction interval from 100 repetition

Number of trees (*Split)*)*	Average Prediction interval length				
	L.M	SPI	Quant	HDR	CHDR
200(a)	17.0	18.2	18.1	17.1	17.1
200(b)	16.8	17.8	17.9	17.0	17.1
200(c)	16.9	17.5	17.7	17.1	17.2
500(a)	16.7	17.4	17.6	16.8	16.8
500(b)	16.7	17.2	17.6	16.5	16.6
500(c)	16.6	16.9	17.3	16.3	16.5
1000(a)	16.7	17.3	17.5	16.9	16.7
1000(b)	16.5	17.0	17.3	16.6	16.4
1000(c)	16.4	16.6	17.0	16.1	16.3
5000(a)	16.7	17.2	17.5	18.4	17.0
5000(b)	16.5	16.9	17.3	18.3	17.0
5000(c)	16.4	16.5	17.0	17.9	16.7

*)

(a) 70:30

(b) 80:20

(c) 90:10

Based on Table 6, it is found that the LM method mostly has the smallest interval length compared to other methods. Thus, there is an indication that based on the interval length, the LM method has better accuracy compared to other methods. For all methods, for the most part, the average prediction interval length decreases as the sample size decreases. This result is different from Alakus et al. [2]. Based on Table 4, the number of trees is the best is 500. Figure 5, Figure 6, and Figure 7 illustrate mean coverage for all models with use amount tree 500 out of 100 repetitions.

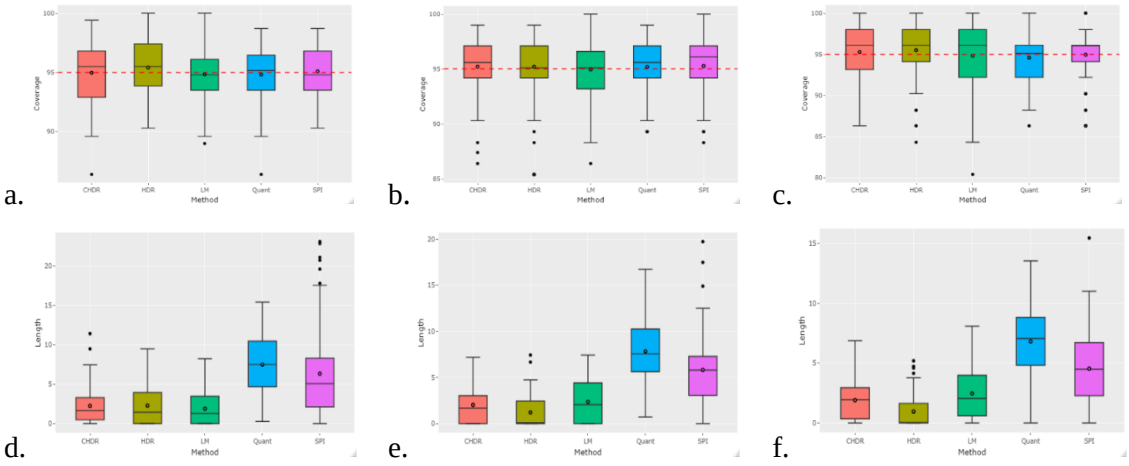


Figure. 2. (a) Mean coverage for all models from 100 replications for amount tree 500 split 70:30; (b) Mean coverage for all models from 100 replications For amount tree 500 split 80:20; (c) Mean coverage for all models from 100 replications For amount tree 500 split 90:10; (d) Mean coverage for all models from 100 replications For amount tree 500 split 90:10; (e) Percentage enhancement length of prediction interval from 100 repetitions For amount tree 500 split 80:20; (f) Percentage enhancement length of prediction interval from 100 repetitions For amount tree 500 split 90:10.

The red dotted line in Figure 2.a, Figure 2.b, and Figure 2.c is the desired level, namely 95%. The white circle is the average of the percentage increase in the length of the prediction interval from 100 replicates. Based on Table 4, the number of trees is the best is 500. Based on Figure 2.a, Figure 2.b, and Figure 2.c, it can be seen that all methods provide mean coverage that is close to the desired level. Figure 2.d, Figure 2.e, and Figure 2.f illustrate the percentage increase in the length of the prediction interval using several trees of 500 from 100 replications.

Based on Figure 2.d, Figure 2.e and Figure 2.f, the average percentage increase in the length of the smallest prediction interval in the 500 tree scenario in the 70:30 split is LM, the 80:20 split and the 90:10 split is HDR. The smaller the percentage increase, the better the method (Alakus et al., [2]. As a result, it is indicated that The best method based on the percentage increase in the length of the prediction interval is LM for split 70:30, HDR for split 80:20, and split 90:10.

Hypothesis testing is carried out with the following hypothesis:

H0: $\mu_{LM} = \mu_{SPI} = \mu_{Quant} = \mu_{HDR} = \mu_{CHDR}$(There is no difference between LM, SPI, Quant, HDR, and CHDR methods)

H1: $\mu_{LM} \neq \mu_{SPI} \neq \mu_{Quant} \neq \mu_{HDR} \neq \mu_{CHDR}$(There are differences between LM, SPI, Quant, HDR, and CHDR methods)

Table 7. Hypothesis testing results from mean coverage for 100 repetitions for each number of trees and split

Number of Trees (Split)*)	Fcount	F criteria
200(a)	1,802	2,390
200(b)	0.730	2,390
200(c)	0.711	2,390
500(a)	1,188	2,390
500(b)	0.245	2,390
500(c)	1,257	2,390
1000(a)	1,277	2,390
1000(b)	1,049	2,390
1000(c)	0.493	2,390
5000(a)	0.763	2,390
5000(b)	0.315	2,390
5000(c)	0.729	2,390
200(a)	1,802	2,390
200(b)	0.730	2,390
200(c)	0.711	2,390
500(a)	1,188	2,390
500(b)	0.245	2,390
500(c)	1,257	2,390
1000(a)	1,277	2,390
1000(b)	1,049	2,390
1000(c)	0.493	2,390
5000(a)	0.763	2,390
5000(b)	0.315	2,390
5000(c)	0.729	2,390

*)

(a) 70:30

(b) 80:20

(c) 90:10

Because Fcount is smaller than Fcriteria , then No reject H_0. So, based on Table 7, there is sufficient evidence to say that there is no difference between the LM, SPI, Quant, HDR, and CHDR methods for 100 repetitions at a 5% significance level

4. Conclusion

This research concludes that the best method for one replication is HDR with 500 trees. LM is the best method based on coverage, interval length, and percentage increase in prediction interval length for 100 repetitions. Based on hypothesis testing, there is sufficient evidence to say that there is no difference between the LM, SPI, Quant, HDR, and CHDR methods for 100 repetitions at a 5% significance level.

Ethics approval

Not required.

Acknowledgments

-

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.

References

[1] M.-H. Roy and D. Larocque, 'Prediction intervals with random forests', *Statistical Methods in Medical Research*, vol. 29, no. 1, pp. 205–229, 2020.
[2] C. Alakus, D. Larocque, and A. Labbe, 'RFpredInterval: An R Package for Prediction Intervals with Random Forests and Boosted Forests', *arXiv preprint arXiv:2106. 08217*, 2021.
[3] N. Ellah, 'Analysis of the Influence of Factors that Influence Poverty in East Java', *Student Scientific Journal Feb*, vol. 4, no. 1, 2016.
[4] L. Sugiyono and S. Ningsih, 'Analysis of Potential Causes of Poor Households in Central Java', *Journal of Applications of Statistics & Statistical Computing*, vol. 10, no. 2, p. 25, 2018.
[5] R. K. Damanik and S. A. Sidauruk, 'The influence of population and GRDP on poverty in North Sumatra Province', *Journal of Darma Agung*, vol. 28, no. 3, pp. 358–368, 2020.
[6] A. Rinaldi, 'Structural equation model to analyze household welfare indicators', *Decimal: A Journal of Mathematics*, vol. 2, no. 3, pp. 281–288, 2019.
[7] A. S. Wicaksono and A. M. Yolanda, 'Grouping Regencies/Cities in East Nusa Tenggara Province Based on Human Development Index Indicators Using K-Medoids Clustering', *Journal of Applied Statistics*, vol. 1, no. 1, pp. 79–90, 2021.
[8] S. P. Sinaga, A. Wanto, and S. Solikhun, 'Implementasi Jaringan Syaraf Tiruan Resilient Backpropagation dalam Memprediksi Angka Harapan Hidup Masyarakat Sumatera Utara', *Jurnal Infomedia: Teknik Informatika, Multimedia, dan Jaringan*, vol. 4, no. 2, pp. 81–88, 2019.
[9] F. Anggadini, 'Analysis of the effect of life expectancy, literacy rate, open unemployment rate and gross regional domestic income per capita on poverty in districts/cities in Central Sulawesi Province in 2010-2013', *Catalogis E-Journal*, vol. 3, no. 7, pp. 40–49, 2015.
[10] N. L. Pramesti and I. K. G. Bendesa, 'The Influence of Socio-Economic Factors on Poverty in Bali Province', *EP Unud E-Journal*, vol. 7, no. 9, pp. 1887–1917, 2018.
[11] S. N. Aulele, V. Y. I. Ilwaru, E. R. Wuritimur, and M. Y. Matdoan, 'Analysis of the Number of Poor People in Maluku Province Using a Spatial Regression Approach', *Journal of Applications of Statistics & Statistical Computing*, vol. 13, no. 2, pp. 23–34, 2021.
[12] C. Suryawati, 'Understanding Poverty Multidimensionally', *Journal of Health Service Management*, vol. 8, no. 03, pp. 585–597, 2005.

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.

3. 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.

2.12 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.

2.13 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

2.14 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

2.15 Section headings

Section headings should be left justified, bold, with the first letter capitalized and numbered consecutively, starting with the Introduction. Section headings use 14 pt Times New Roman and exactly spacing at 12 pt with before and after spacing in 12 pt, left alignment and special hanging indentation at 0.63 cm. Sub-section headings should be in capital and lower-case italic letters, numbered 1.1, 1.2, etc, exactly spacing at 12 pt with before and after spacing in 12 pt, left alignment with 0.12 cm left indentation and special hanging indentation at 0.63 cm, with second and subsequent lines indented. All headings should have a minimum of three text lines after them before a page or column break. Ensure the text area is not blank except for the last page. Both section heading and sub-section headings are in dark blue color with the code #034F84 (R: 3 G: 79 B: 132).

2.16 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.

2.17 Footnotes

Footnotes should be avoided if possible.

3 Illustrations

All figures should be numbered with Arabic numerals (1,2,3,….). Every figure should have a caption. All photographs, schemas, graphs and diagrams are to be referred to as figures. Line drawings should be good quality scans or true electronic output. Low-quality scans are not acceptable. Figures must be embedded into the text and not supplied separately. In MS word input the figures must be properly coded. Preferred format of figures are PNG, JPEG, GIF etc. Lettering and symbols should be clearly defined either in the caption or in a legend provided as part of the figure. Figures should be placed at the top or bottom of a page wherever possible, as close as possible to the first reference to them in the paper. Please ensure that all the figures are of 300 DPI resolutions as this will facilitate good output. Figures should be embedded and not supplied separately.

The figure number and caption should be typed below the illustration in 11 pt and left justified [***Note:*** one-line captions of length less than column width (or full typesetting width or oblong) centered].

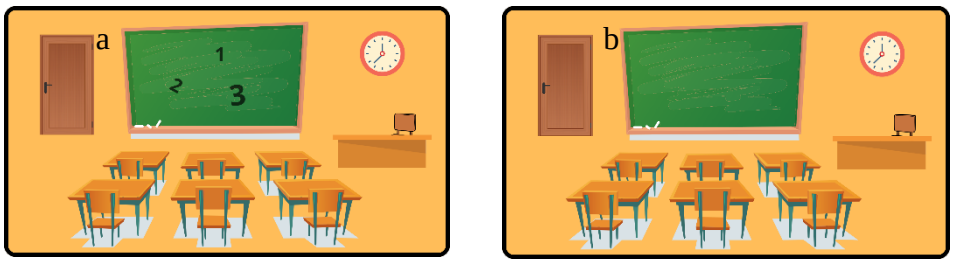


Figure. 1. (a) first picture; (b) second picture.

4 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.”

References

[24] W.K. Chen, *Linear Networks and Systems*. Belmont, CA: Wadsworth Press, 2003.
[25] R. Hayes, G. Pisano, and S. Wheelwright, *Operations, Strategy, and Technical Knowledge*. Hoboken, NJ: Wiley, 2007.
[26] 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.