JASKS
Jurnal Aplikasi Statistika & Komputasi Statistik

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 17, NO 1, JUNE 2025 P-ISSN 2086 – 4132 E-ISSN 2615 – 1367

POLITEKNIK STATISTIKA STIS
STIS

Jurnal Aplikasi Statistika & Komputasi Statistik

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

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

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

Editorial Foreword

Jurnal Aplikasi Statistika & Komputasi Statistik (JASKS) Volume 17 Number 1 June 2025 Edition has undergone transformations such as writing articles in English, establishing a journal logo, establishing the Politeknik Statistika STIS publisher logo, changing paper template designs, updating Author Guidelines, etc. The aim of this transformation is to improve the journal's performance and expand the reach of JASKS readers.

This issue consists of 7 articles contributed by 27 authors affiliated with various institutions from Indonesia and abroad. The contributing institutions include IAIN Palangka Raya, Universitas Palangka Raya, Politeknik Statistika STIS, BPS-Statistics from various regional offices, the National Team for the Acceleration of Poverty Reduction, Universitas Islam Jakarta, PKN STAN, the Center for Industrial, Services, and Trade Research, the National Research and Innovation Agency, University of Tsukuba (Japan), Jiangsu University (China), and Nanyang Technological University (Singapore). The research presented covers a diverse range of topics, including survival support vector machine analysis on separated couples during the COVID-19 outbreak, LLM implementation in survey interviews, quantile regression using constrained B-splines for modeling schooling and household expenditure, binary and traditional partial least squares structural equation modeling for poverty dimensions and social protection, geospatial analysis on hotel occupancy rates, partial proportional odds model application for household food insecurity in Papua, and the estimation of gross regional domestic product per capita using machine learning and geospatial data.

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, February 2025
Editor-in-Chief,

Setia Pramana

Jurnal Aplikasi Statistika & Komputasi Statistik

VOLUME 17, NO 1, JUNE 2025

Contents

DOI: https://doi.org/10.34123/jurnalasks.v17i1.739

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

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

Separated Couples during the COVID-19 Outbreak: A Survival Support Vector Machine Analysis

Muhammad Luthfi Setiarno Putera[1*], Rafik Patrajaya[2], Setiarno[3]

[1,2]IAIN Palangka Raya, Palangka Raya, Indonesia, [3]Universitas Palangka Raya, Palangka Raya, Indonesia
**Corresponding Author: E-mail address:* m.luthfi@iain-palangkaraya.ac.id

ARTICLE INFO

Article history:
Received 16 February, 2024
Revised 22 November, 2024
Accepted 10 December, 2024
Published 24 February, 2025

Keywords:
Cox Proportional Hazard; Feature Selection; Religious Court; Separation; Survival Support Vector Machine

Abstract

Introduction/Main Objectives: The separation between spouses has been rising noticeably in recent years in Palangka Raya, particularly during the COVID-19 outbreak. **Background Problems:** An analysis of time-to-event on those separations will be undertaken quantitatively using survival analysis by comparing the results yielded by Cox proportional hazards (PH) regression and non-parametric Survival Support Vector Machine (SUR-SVM). **Novelty:** This work suggests a feature selection method that looks for influencing elements related to the c-index by employing backward elimination. **Research Methods:** This study's data came from Indonesia's Supreme Court webpage, including a database of separation verdicts from the Palangka Raya Religious Court, spanning from April 2020 to March 2021. The response variables were the time-to-separation (marriage length until separation) (t) and the censored state of the occurrence (θ). **Finding/Results:** Based on SUR-SVM, the factors contributing the most to the separation are the absence of children, unsteady employment of appellants, and finance motive as the primary reason. In terms of concordance index and Akaike Information Criterion (AIC), the SUR-SVM outperformed the Cox proportional hazard model. These values of SUR-SVM were 59.24 and 1899.78, respectively. SUR-SVM correctly classified 59.24% of separations based on the chronological order of events.

1. Introduction

The COVID-19 outbreak spreading throughout the world has paralyzed various aspects of life, such as social, economic, health, and many more. As of November 2022, there are currently more than 6 million cases of COVID-19 in Indonesia since the emergence of the first victim in March 2020 [1]. One of the impacts felt by society as a result of the pandemic is couple’s hardship [2], [3]. Such problem arises due to various factors, such as negative emotions due to financial difficulties, home confinement, and even dissatisfaction with government policies [4]. Tensions and disputes which occur, notably in vulnerable couples, will be more prone to enduring conflicts and could point to divorce or marriage separation [5], [6]. Annual report of Religious Court of Palangka Raya showed that number of marriage separation there indicated an upturn since COVID-19 outbreak in 2020.

Analysis of time-to-event situations might be undertaken quantitatively using survival analysis, one of which is marriage separation. In Indonesia, separation for Muslim couples is generally administered by a Religious Court situated in the regency or city capital [5]. Such cases could reach thousands each year. Several survival studies on marriage separation were conducted by [7] and [8].

Both found that separation was motivated by variety factors, ranging from socio-economic and cultural ones [7], [8].

Survival analysis on marriage separation is commonly dealt with censored data. The censored data is a set of values whose occurrence status is unknown since the study ended before the failure occurred [9]. As [7] and [8] revealed about separation, their dataset were classified to the right-censored like most other events. Right censorship in separation occurs as the spouses return to living together eventually or separation is not noticed by the end of the research.

Separation cases could generally be modeled using Cox proportional hazards (PH) regression [10], [11]. Beyond its common use as a conventional method of survival analysis, this method has plenty shortcoming yet. As the assumption of proportional across time for the hazard of two individuals fails to be met, Cox PH regression is challenging to employ [12]. Likewise, if covariate dependencies occur, another model is needed in the form of a non-parametric model which does not require these assumptions [13]. Non-parametric methods have been used in many techniques, including Survival Support Vector Machine (SUR-SVM).

The other study by Abdel Sater discover that numerous social and economic factors drastically influence the survival time of a marriage. With respect to survival, having children decreases marriage survival and a high level of education or a high income increases marriage survival. Applications of a survival analysis model has great potential in social science settings. The model applied Support Vector Machine as well to identify the influencing factor to the duration of marriage. The difference lies on the absence of feature selection to seek for the factor that contributes more to the duration of marriage among separated spouses[8].

A large number of couples registering for marital separation amid the COVID-19 outbreak strengthens the need to analyze the characteristics of separated Muslim couples and the influencing factors behind it. Such motives would be explained by comparing the result of two models, Cox PH and SUR-SVM. This work also suggested a feature selection method that looks for influencing elements related to the c-index by employing backward elimination. By doing so, it will enable to determine the socio-economic factors motivating the separation among Muslim couples in early period of pandemic.

2. Material and Methods

2.1. Reference Review

SUR-SVM is a machine learning method which does not require proportional hazard assumption to be fulfilled [11]. Besides, SUR-SVM is able to be conducted on high-dimensional data [14]. Because separations were observed in numerous regions and were specifically linked to a number of socioeconomic factors, it is therefore appropriate for these matters. Plenty of study demonstrated the superiority of SUR-SVM over alternative methods. Studies applying several kinds of datasets discovered that SUR-SVM performed considerably better than Cox PH [15]. Furthermore, recent work on glioma dataset suggested that SUR-SVM topped Cox PH in terms of prediction [16]. However, our motives to conduct this study based on the less number of application of SUR-SVM in social matters or particularly law matters.

2.2. Kaplan-Meier Curve and Survival Model Assumption

To assess the survival rates of marital separation, one might utilize the Kaplan-Meier curve which is related to the use of log-rank test. Another common use of the log-rank test is to see whether survival curves for different categories in a variable differ from one another [17]. The following is the hypothesis

H_0 : Survival curve categories do not significantly differ from one another
H_1 : Survival curve categories differ significantly from one another

where test statistic is expressed as

$$\chi^2 = \sum_{l=1}^{L} \frac{\left(O_l - E_l\right)^2}{E_l} \tag{1}$$

where $O_l - E_l = \sum_{k=1}^{n}(m_{lk} - e_{lk})$ and $e_{lk} = \left(\frac{n_{lk}}{\sum_{l=1}^{L}\sum_{k=1}^{n} n_{lk}}\right)\left(\sum_{l=1}^{L}\sum_{k=1}^{n} m_{lk}\right)$.

Description :

O_l : number of case in the l-th category

E_l : expected number of cases in the l-th category

m_{lk} : number of cases in the l-th category bearing event at time t_k

n_{lk} : number of at-risk cases bearing an instantaneous event in the l-th category before time t_k

e_{lk} : expected value in the p-th category at time t_k

l : number of categories in a variable

If $\chi^2 >$ $\chi^2_{\alpha,(L-1)}$, then H$_0$ is rejected, indicating at least one difference in the survival curve for a variable [11].

Besides testing on differences among categories, the Cox PH model should meet the proportional hazard assumption, suggesting that the hazard ratio is independent of time [18]. Such test is relied on Schoenfeld error calculated by

$$SR_{mk} = z_{mk} - E\left(z_{mk} \mid R\left(t_{(mk)}\right)\right) \tag{2}$$

and the conditional probability in Equation (2) is acquired from

$$E\left(z_{mk} \mid R\left(t_{(mk)}\right)\right) = \frac{\sum_{i \in R\left(t_{(mk)}\right)} z_{mk} \exp(\beta x_i)}{\sum_{i \in R\left(t_{(mk)}\right)} \exp(\beta x_i)} \tag{3}$$

where

SR_{mk} : Schoenfeld error of the m-th predictor for cases experiencing an event at time $t_{(k)}$

Z_{mk} : the value of the m-th predictor for cases experiencing an event at time $t_{(k)}$

Following that step is to generate a rank variable f_r corresponding to its survival time. A value of 1 is assigned to the scenario where the event occurs for the first time, and so forth. Next, examine the association between the Schoenfeld error and the ranking variable f_r using the hypothesis, as follows:

H$_0$: $\rho = 0$

H$_1$: $\rho \neq 0$

and below is the statistic test

$$t_{\text{test}} = \frac{corr_{f_r,SR_{mk}}\sqrt{n-2}}{\sqrt{1-\left(corr_{f_r,SR_{mk}}\right)^2}} \tag{4}$$

with $corr_{f_r,SR_{mk}} = \frac{\operatorname{cov}(f_r, SR_{mk})}{\sqrt{\operatorname{var}(f_r)\operatorname{var}(SR_{mk})}}$. The decision is to reject H$_0$ if $|t_{\text{test}}| > t_{(\alpha/2,n-2)}$. Since there is a clear association between the survival time rank variable and Schoenfeld error, the proportional hazard assumption is not valid [19].

2.3. Survival Function and Hazard Function

The survival function $S(t)$ represents the likelihood that an item will endure or avoid an occurrence or failure until a specific point in time. Given that T is the length of time till an event happens,

$$S(t)=P(T>t)=\int_t^\infty f(u)\,du=1-F(t) \tag{5}$$

the function $S(t)$ is calculated by applying equation (5). The hazard function $h(t)$ calculates the probability at which the event occurs during any given time point, which is given in Equation (6)

$$h(t)=-\frac{\partial \log S(t)}{\partial t} \tag{6}$$

As time goes on, the probability of occurrence of events will be higher [20]. Besides those two, there is cumulative hazard function. Such function could be written as

$$H(t)=\int_0^t h(u)\,du \text{ or } -H(t)=\ln S(t). \tag{7}$$

The function $H(t)$ could be interpreted as cumulative amount of hazard up to time t.

2.4. Cox Proportional Hazard Model

Cox Proportional Hazard model is a way to understand how predictors influence the survival function of the event, such as marriage separation. Let Z denotes the predictors. [20] declared that Cox regression can be expressed as

$$h(t|Z=z)=h_0(t)\exp\left(z^{\mathrm{T}}\beta\right) \tag{8}$$

with

$h(t|Z=z)$: hazard function

B : vector of coefficients of each predictor

Z : vector of predictor

$h_0(t)$: baseline hazard function

2.5. Survival Support Vector Machine

Survival Support Vector Machine (SUR-SVM) is machine learning model which employs a prognostic index, unlike Cox model which relies on hazard function. The prognostic could be expressed as the probability of the couple to reconcile due to mediation or others. Refer to [11], SUR-SVM has a utility function u

$$\mathbf{u}(\mathbf{x})=\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) \tag{9}$$

where $\mathbf{w}$ is vector of parameter and $\boldsymbol{\varphi}$ $(\mathbf{x})$ is the transformation of predictor x. Besides, SUR-SVM has an objective function, expressed in

$$\min_{\mathbf{w},\xi}\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+\frac{\gamma}{2}\sum_j\sum_{k,j<k}v_{jk}\xi_{jk};\gamma\geq 0 \tag{10}$$

and the constraint function is expressed in

$$\begin{gathered}\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_k)-\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_j)\geq 1-\xi_{jk};\forall j<k\\ \xi_{jk}\geq 0;\forall j<k.\end{gathered} \tag{11}$$

Indicator v_{ij} is a comparison between i-th case and j-th case which fulfills

$$v_{jk}=\begin{cases}1,\left(t_j<t_k,\delta_j=1\right)\\ 0,\left(t_j<t_k,\delta_j=0\right)\end{cases} \tag{12}$$

with ξ_{jk} in Equation (11) is the value of violations due to an error in ordering the occurence time-to-event [21].

2.6. *Feature Selection and Goodness of Survival Model*

Feature selection is a technique for identifying features that contain specific relevant predictors in order to produce a more favorable model. One of these techniques is backward elimination. The elimination is accomplished by modeling all predictors with Cox PH and SUR-SVM. After that, remove one of the least significant factors and regress the remaining predictors using both models. Continue the elimination process until every predictor gets an opportunity to be eliminated. The less contributing predictor could be determined by having a higher c-index [17]. The collection of predictors with the highest c-index is then used to re-model the Cox PH and SUR-SVM models. The concordance index, often known as the c-index, measures the order of the prognostic function and the observed survival time for both censored and uncensored data. A model with a higher c-index value represents a stronger survival model [11]. The Akaike Information Criterion (AIC) is another metric used to assess the goodness of a survival model. Model with lower AIC metric indicates better stability of survival model [22].

2.7. *Data and Variables*

This study's data came from Indonesia's Supreme Court webpage, including a database of separation verdicts from the Palangka Raya Religious Court which could be accessed in https://putusan3.mahkamahagung.go.id/pengadilan/profil/pengadilan/pa-palangkaraya.html. Spanning from April 2020 to March 2021, the time frame under observation was 1 year. There were 319 decisions regarding separation released within observed period. The analysis for regression aim took several variables into account. The response variables were the time-to-separation (marriage length until separate) (t) and censored state of the occurence (θ). Summary of variables is shown in Table 1.

Table 1. Variable summary

Variable	Explanation	Scale
Survival time (t)	Marriage length (time-to-separation), in years	Ratio
Status (θ)	State of the occurence 0: not denounced to separation (censored) 1: denounced to separation	Nominal
$z1$	Complainant's age at marriage	Ratio
$z2$	Complainant's level of school	Ordinal
$z3$	Complainant's job status	Nominal
$z4$	Appellant's age at marriage	Ratio
$z5$	Appellant's level of school	Ordinal
$z6$	Appellant's job status	Nominal
$z7$	Number of kids	Ratio
$z8$	Motive of separation	Nominal

Refer to Table 1, variable z_2 and z_5 comprise identical four categories, those are elementary school, junior high school, secondary school, and higher school. Variable z_3 and z_6 comprise identical four categories, including unskilled employees, skilled employees, semi-professionals, and working professionals. Variable z_8 consists of five categories, those are leaving duties, disputes, economic pressure, moral crises, and physical weaknesses.

2.8. *Analysis Method*

All categorical variables are converted to dummies using the number of categories (n_l) – 1 before being regressed. What is needed to do survival analysis on the separation dataset are explaining descriptive statistics, analyzing the curve of Kaplan-Meier and log-rank test, testing proportional hazard assumption, and analyzing survival and hazard curve. Afterwards, the data is being modeled using the Cox PH and Survival Support Vector Machine. Then, the step is selecting significant predictors from each survival method using feature selection. To seek for the better model, this work also provide goodness-of-fit based on c-index and AIC.

3. Results and Discussion

The number of separation cases filing by the Muslim couple in Palangka Raya Religious Court during April 2020 – March 2021 was 319 cases. Of all 319 occurrences, 64 were censored and 255 were uncensored. Marriage length (*t*) averages 10.26 years, with a median of exactly 9 years. The marriage lasted the longest 37 years, while the shortest was 0 years. A discrepancy between the mean and median revealed that the distribution of the marriage length of separated couples is asymmetric. The survival probability of the marriage is shown in Figure 1.

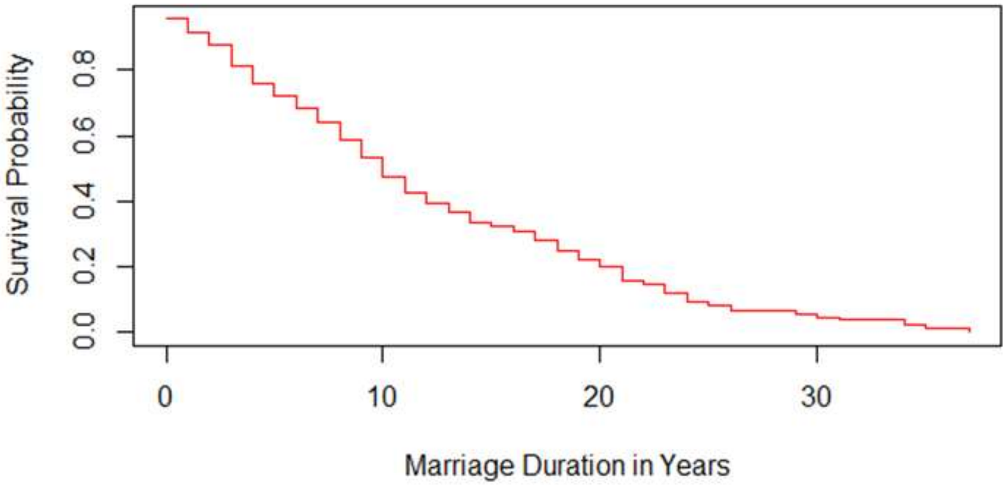


Figure 1. Survival probability of marriage of separated couples

Refer to Figure 1, the marriage length of separated couples indicated to decrease drastically at early phase of marriage within 0 – 5 years and 5 – 10 years. The survival probability was indicated steady after 25 years of marriage. It is coherent with the work of [23] noted that couple with longer marriage life was more likely to maintain their harmonious household life rather than newly-wed couple. The results of the log-rank test for each predictor are shown in Table 2, computed using Equation (1), where significant predictor is in bold.

Table 2. Summary of log-rank test

Variable	Log-rank Value	d.f	p-value
*z*1	19.1	1	0.001
*z*2	14.8	3	0.002
*z*3	12.0	3	0.007
*z*4	3.2	1	0.070
*z*5	4.0	3	0.300
*z*6	0.8	3	0.900
*z*7	65.7	1	0.001
*z*8	3.4	4	0.500

At α = 5%, Table 2 showed that four significant predictors, e.g. complainant's age at marriage (*z*1), complainant's education (*z*2), complainant's job status (*z*3), and number of kids (*z*7). It implied those four variables have different survival curves between groups (categories). Henceforth, complainant's age at marriage and number of kids could cause significant differences to the survival probability of marriage. It is relevant to the work of [11] which declared that some categories many socio-economic determinant have were able to differ the survival curve between categories.

There were eight variables to identify whether they all had a substantial impact on how long the problematic couples survive. The test result for the proportional hazard assumption, which was determined using Equations (2), (3), and (4), is shown in Table 3. The significant predictor at α = 5% is in bold.

Table 3. Summary of proportional hazard assumption test

Variable	Correlation (ρ)	Chi-square	p-value
*z*1	0.048	0.507	0.476
***z*2**	**-0.176**	**7.471**	**0.006**
***z*3**	**0.145**	**5.496**	**0.019**
*z*4	-0.007	0.014	0.904
*z*5	-0.020	0.103	0.748
*z*6	-0.022	0.136	0.712
***z*7**	**0.374**	**40.273**	**0.001**
*z*8	-0.003	0.002	0.961

Table 3 displays complainant's education (*z*2), complainant's job status (*z*3), and number of kids (*z*7), all in bold, with test results that contradict H_0 of Equation (4). As a result, the proportional assumption is not met. It shows a strong association between the Schoenfeld error and the survival time ranking. In particular, SUR-SVM, a substitute means which dispenses with the proportional hazard assumption, is required because all of the variables involved are necessary to represent survival time. Having been calculated by using Equations (5) and (6), Figure 2 displays the cumulative survival function and the cumulative hazard function.

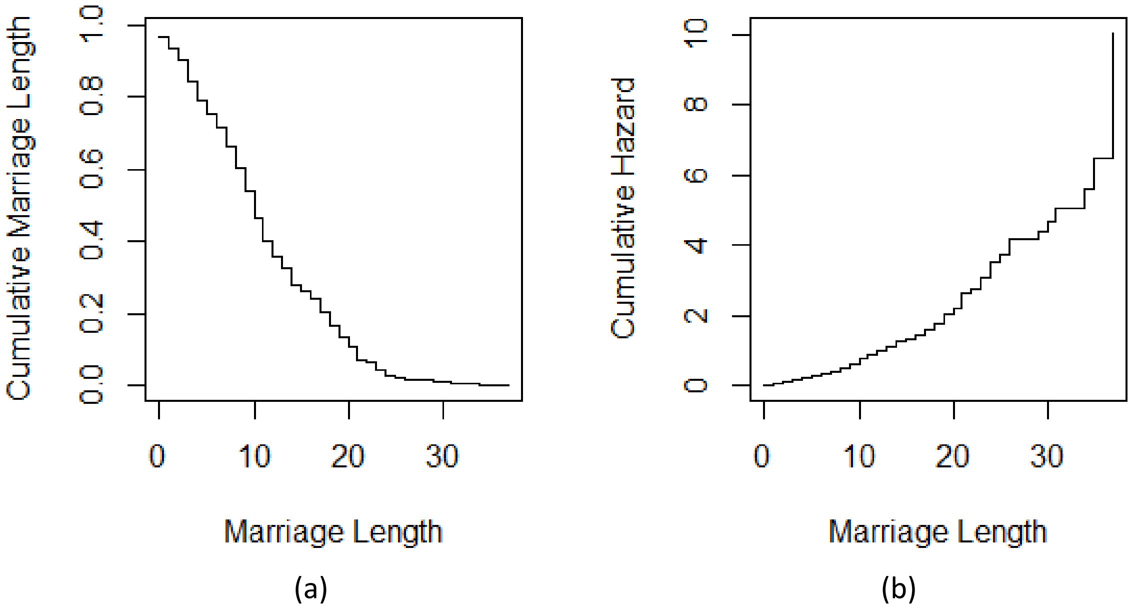


Figure 2. Cumulative survival and cumulative hazard of marriage length

Figure 2(a) shows that spouses with 0 - 10 years of marriage are more likely to be separate than they with longer marriages. The likelihood that a separated spouse will survive drops off significantly between the ages of 0 and 10 years, beyond that, it appears to be more stable. Figure 2(b) shows a trend that rises from left to right, similar to a staircase. That increase suggested that the likelihood of a pair divorcing increased with the length of the marriage. It is similar to the study of [24] which revealed that prolonged disputes and negative emotions within household life might contribute to elevated separation rate. Table 4 shows parameter estimation from the Cox proportional hazard model applied to separated couples' marital length data. The numbers which come after the dot (.) symbol in a variable, for instance *z*3.1, *z*3.2, etc., correspond to the category in a categorical variable.

Table 4. Cox model parameter estimates

Variable	Coefficient (β)	Hazard Ratio	p-value	Variable	Coefficient (β)	Hazard Ratio	p-value
*z*1	0.001	1.00	0.93	*z*5.3 (higher school)	-0.051	0.95	0.85
*z*2.1 (junior high school)	0.131	1.14	0.58	*z*6.1 (unskilled employees)	0.313	1.37	0.27
***z*2.2 (secondary school)**	**0.667**	**1.95**	**0.01**	*z*6.2 (skilled employees)	0.436	1.55	0.12
***z*2.3 (higher school)**	**0.830**	**2.29**	**0.01**	*z*6.3 (semi-professionals)	0.107	1.11	0.72
*z*3.1 (unskilled employees)	0.106	1.11	0.71	***z*7***	**-0.643**	**0.52**	**0.01**
*z*3.2 (skilled employees)	0.196	1.22	0.47	*z*8.1 (leaving duties)	-0.468	0.63	0.66
*z*3.3 (semi-professionals)	0.449	1.57	0.11	*z*8.2 (disputes)	0.331	1.39	0.75
*z*4	0.025	1.02	0.05	*z*8.3 (economy pressures)	0.136	1.14	0.89
*z*5.1 (junior high school)	0.187	1.21	0.40	*z*8.4 (moral crisis)	-0.112	0.89	0.91
*z*5.2 (secondary school)	0.023	1.02	0.91				
Likelihood ratio test		131.7		d.f. = 19	p-value = <0,01		

As shown in Table 4, the likelihood ratio test yields a test statistic of 131.7 and a p-value less than 0.01 in order to assess the importance of the parameters jointly. The choice to reject H$_0$ in cases where at least one variable significantly affects the separation rate was suggested by the p-value. According Table 4, the partial test on parameter found that three terms representing two variables were significant to the separation rate, shown in bold. These variables are the complainant's level of schooling, which includes secondary school and higher school, as well as the number of kids. The best Cox model based on Equation (7) is

$$h(t|Z=z)=h_0(t)\exp(0.667z_{2.2}+0.830z_{2.3}-0.643z_7)$$

A couple with more kids may have a lower hazard ratio, as indicated by the negative sign (-) for the number of kids. Table 4's hazard ratio, particularly for the major predictors, can be understood as a gauge of how these factors affect the rate of separation. For example, the number of kids hazard ratio (represented by the * symbol) is 0.52. Increasing by one kid would presumably result in a 0.52-fold decrease in the separation rate. Therefore, it is possible that couples who have more kids may have longer marriages.

Conversely, the hazard ratio is roughly 2.29 for categorical factors such as the complainant's degree from a higher school. Such figure implies that complainants who graduated higher school have separation rate 2.29 times higher than they who graduated elementary school as the reference. Hence, it can be said that the complainant’s level of school contributed statistically to the separation. The c-index of the SUR-SVM model was 58.83, and the c-index of the Cox model was 23.22, according to an algorithm that used the Kernel Radial Basis Function. After feature selection, the c-index of the SUR-SVM and Cox-PH models is shown in Table 5.

Table 5. Eliminated variables' contribution to all models

Eliminated Variable	C-index of Cox PH	C-index of SUR-SVM	Eliminated Variable	C-index of Cox PH	C-index of SUR-SVM
$z1$	23.23	57.56	$z5$	23.25	58.73
$z2$	24.31	58.63	$z6$	22.62	58.30
$z3$	22.73	58.65	$z7$	22.48	56.76
$z4$	23.64	57.66	$z8$	23.48	58.00

As seen in Table 5, the number of children ($z7$) had the largest decrease in the SUR-SVM c-index. Table 5 produces a difference of 2.07 between 58.83 and 56.76. It is thought to be the primary factor affecting the order in which the prognostic index and survival time are related. Besides number of kids, others contributing to the c-index of SUR-SVM are complainant's age at marriage ($z1$), appellant's age at marriage ($z4$), reason of separation ($z8$), and appellant's job status ($z6$). Other predictors were not considered since they had slight margin, not more than 0.5.

Additionally, Table 5 demonstrates that the number of kids had the most notable decline in the Cox model c-index. Since there is notable difference between the final c-index and the overall c-index, the number of kids has the most impact on the length of marriage. Afterwards, the selected features for SUR-SVM model are complainant's age at marriage ($z1$), appellant's age at marriage ($z4$), appellant's job status ($z6$), number of kids ($z7$), and reason of separation ($z8$), while the selected feature for Cox-PH model is complainant's job status ($z3$), appellant's job status ($z6$), and number of kids ($z7$). Table 6 shows an overview of both survival models' performance after applying the selected features.

Table 6. Performance metric of survival model

Predictors Selected	**Indices**	**Cox-PH**	**SUR-SVM**
All	C-index	23.22	58.83
	AIC	2361.13	2318.22
Feature selection	C-index	25.41	59.24
	AIC	2004.35	1899.78

According to Table 6, SUR-SVM's post-feature selection c-index is 0.41 higher than SUR-SVM's c-index across all predictors, which generated results between 59.24 and 58.83. In other side, the Cox model's c-index after feature selection is 2.19 times higher than the Cox model's c-index across all predictors. As evidenced by the greater c-index and lower AIC of SUR-SVM, it suggests that SUR-SVM is superior than the Cox model. Thus, 59.24% time-to-separation and a suitable prognostic sequence could be obtained using the enhanced SUR-SVM model. It suggested that SUR-SVM correctly classified 59.24% of separation situations as occurring in the correct order.

4. Conclusion

SUR-SVM underperforms semi-parametric Cox proportional hazard model in terms of quantifying the relationship between the predictors and survival duration of dissolved Muslim couples. The result demonstrated that the feature selection applied in both models was effective in optimizing the survival model by eliminating less contributed predictors. Various socio-economic factors substantially influenced the duration of marriage among separated Muslim couples in Palangka Raya, including the number of kids, the appellant's employment, and the reason for separation. Future studies could compare the length of marriages befor COVID-19 and its aftermath by applying machine learning techniques. The challenge is lied on the exploring best machine learning method along with best feature selection technique.

Ethics approval

The ethical guidelines were followed for conducting this investigation. Every individual participant in the study gave their informed permission.

Acknowledgments

We would like to thank the Religious Court of Palangka Raya for their invaluable support in this study. Our research findings have been substantially enhanced by the Religious Court of Palangka Raya's publication statistics and databases. All stakeholders' participation and assistance have been crucial to the accomplishment of this study.

Competing interests

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

Funding

This study received no external funding.

Underlying data

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

Credit Authorships

Muhammad Luthfi Setiarno Putera: Methodology, Software Data Processing and Analysis, Data Visualization, Writing. **Rafik Patrajaya:** Writing, Layout and Editing. **Setiarno:** Conceptualization, Supervision, Validation.

References

[1] M. A. Widiawaty, K. C. Lam, M. Dede, and N. H. Asnawi, “Spatial differentiation and determinants of COVID-19 in Indonesia,” *BMC Public Health*, vol. 22, no. 1, p. 1030, May 2022, doi: 10.1186/s12889-022-13316-4.

[2] P. R. Pietromonaco and N. C. Overall, “Implications of social isolation, separation, and loss during the COVID-19 pandemic for couples’ relationships,” *Current Opinion in Psychology*, vol. 43, pp. 189–194, Feb. 2022, doi: 10.1016/j.copsyc.2021.07.014.

[3] H. Prime, M. Wade, and D. T. Browne, “Risk and resilience in family well-being during the COVID-19 pandemic,” *American Psychologist*, vol. 75, no. 5, pp. 631–643, 2020, doi: 10.1037/amp0000660.

[4] M. D. Puspitasari and M. Gayatri, “COVID-19 and Marital Dissolution in West Java, Indonesia,” *The Family Journal*, no. 10664807221124246, 2022, doi: https://doi.org/10.1177/10664807221124246.

[5] I. Rais, “The impact of COVID-19 pandemic on divorce rates among Indonesian Muslim societies,” *Indonesian Journal of Islam and Muslim Societies*, vol. 11, no. 2, 2021, doi: 10.18326/ijims.v11i2.271-297.

[6] A. Thadathil and S. Sriram, “Divorce, Families and Adolescents in India: A Review of Research,” *Journal of Divorce & Remarriage*, vol. 61, no. 1, pp. 1–21, 2020, doi: https://doi.org/10.1080/10502556.2019.1586226.

[7] S. Norouzi *et al.*, “Marriage survival in new married couples: A competing risks survival analysis,” *PLoS ONE*, vol. 17, no. 8, p. e0272908, 2022, doi: https://doi.org/10.1371/journal.pone.0272908.

[8] R. Abdel-Sater, “Marriage Dissolution in the United States: A Survival Analysis Approach,” *Journal of Divorce & Remarriage*, vol. 63, no. 4, pp. 235–261, 2022, doi: https://doi.org/10.1080/10502556.2022.2042788.

[9] M. Nemati, J. Ansary, and N. Nemati, “Machine-Learning Approaches in COVID-19 Survival Analysis and Discharge-Time Likelihood Prediction Using Clinical Data,” *Patterns*, vol. 1, no. 5, p. 100074, Aug. 2020, doi: 10.1016/j.patter.2020.100074.

[10] V. Berg, A. Miettinen, M. Jokela, and A. Rotkirch, “Shorter birth intervals between siblings are associated with increased risk of parental divorce,” *PLOS ONE*, vol. 15, no. 1, p. e0228237, Jan. 2020, doi: 10.1371/journal.pone.0228237.

[11] M. L. S. Putera and S. Setiarno, “Machine learning survival analysis on couple time-to-divorce data,” *Desimal: Jurnal Matematika*, vol.5, no. 3, Art. no. 3, Dec. 2022, doi: 10.24042/djm.v5i3.13742.

[12] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. K. Suykens, “Support vector methods for survival analysis: a comparison between ranking and regression approaches,” *Artificial Intelligence in Medicine*, vol. 53, no. 2, pp. 107–118, Oct. 2011, doi: 10.1016/j.artmed.2011.06.006.

[13] Md. S. Satu, K. Howlader, M. P. Hosen, N. Chowdhury, and M. A. Moni, “Identifying the Stability of Couple Relationship Applying Different Machine Learning Techniques,” in *2020 11th International Conference on Electrical and Computer Engineering (ICECE)*, Dec. 2020, pp. 246–249. doi: 10.1109/ICECE51571.2020.9393131.

[14] S. Goli, H. Mahjub, J. Faradmal, and A.-R. Soltanian, “Performance Evaluation of Support Vector Regression Models for Survival Analysis: A Simulation Study,” *ijacsa*, vol. 7, no. 6, 2016, doi: 10.14569/IJACSA.2016.070650.

[15] S. Lee and H. Lim, “Review of statistical methods for survival analysis using genomic data,” *Genomics Inform*, vol. 17, no. 4, p. e41, Dec. 2019, doi: 10.5808/GI.2019.17.4.e41.

[16] R. Zhao, Y. Zhuge, K. Camphausen, and A. V. Krauze, “Machine learning based survival prediction in Glioma using large-scale registry data,” *Health Informatics Journal*, vol. 28, no. 4, p. 14604582221135427, 2022, doi: https://doi.org/10.1177/14604582221135427.

[17] P. Wang, Y. Li, and C. K. Reddy, “Machine Learning for Survival Analysis: A Survey,” *ACM Comput. Surv.*, vol. 51, no. 6, p. 110:1-110:36, Feb. 2019, doi: 10.1145/3214306.

[18] Z. Zhang, J. Reinikainen, K. A. Adeleke, M. E. Pieterse, and C. G. M. Groothuis-Oudshoorn, “Time-varying covariates and coefficients in Cox regression models,” *Ann Transl Med*, vol. 6, no. 7, p. 121, Apr. 2018, doi: 10.21037/atm.2018.02.12.

[19] A. Sheng and S. K. Ghosh, “Effects of Proportional Hazard Assumption on Variable Selection Methods for Censored Data,” *Statistics in Biopharmaceutical Research*, vol. 12, no. 2, pp. 199–209, Apr. 2020, doi: 10.1080/19466315.2019.1694578.

[20] D. G. Kleinbaum and M. Klein, “Kaplan-Meier Survival Curves and the Log-Rank Test,” in *Survival Analysis : A Self-Learning Text, Third Edition*, in Statistics for Biology and Health. , Springer International Publishing, 2012, pp. 55–96. Accessed: Jan. 27, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4419-6646-9_2

[21] D. D. Prastyo, H. A. Khoiri, S. W. Purnami, Suhartono, S.-F. Fam, and N. Suhermi, “Survival Support Vector Machines: A Simulation Study and Its Health-Related Application,” in *Supervised and Unsupervised Learning for Data Science*, M. W. Berry, A. Mohamed, and B. W. Yap, Eds., in Unsupervised and Semi-Supervised Learning. , Cham: Springer International Publishing, 2020, pp. 85–100. doi: 10.1007/978-3-030-22475-2_5.

[22] K. Burke, M. C. Jones, and A. Noufaily, “A Flexible Parametric Modelling Framework for Survival Analysis,” *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 69, no. 2, pp. 429–457, 2020, doi: https://doi.org/10.1111/rssc.12398.

[23] P. Chi, Q. Wu, H. Cao, N. Zhou, and X. Lin, “Relationship-oriented values and marital and life satisfaction among Chinese couples,” *Journal of Social and Personal Relationships*, vol. 37, no. 8–9, pp. 2578–2596, Aug. 2020, doi: 10.1177/0265407520928588.

[24] D. A. Widyastari, P. Isarabhakdi, P. Vapattanawong, and M. Völker, “Marital Dissolution in Postmodern Java, Indonesia: Does Early Marriage Increase the Likelihood to Divorce?,” *Journal of Divorce & Remarriage*, vol. 61, no. 8, pp. 556–573, Nov. 2020, doi: 10.1080/10502556.2020.1799308.

DOI: https://doi.org/10.34123/jurnalasks.v17i1.792

POLITEKNIK STATISTIKA STIS

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

Jurnal Aplikasi Statistika &
Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

Early Study of LLM Implementation in Survey Interviews

Lailatul Hasanah [1*], Budi Yuniarto[2]

[1,2]Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address: 222011364@stis.ac.id*

ARTICLE INFO

Article history:
Received 30 August, 2024
Revised 12 December, 2024
Accepted 1 February, 2025
Published 24 February, 2025

Keywords:
ChatGPT; Interview; LLM; Official Statistics; Survey

Abstract

Introduction/Main Objectives: This research aims to conduct a preliminary study into the use of LLMs for extracting information to fill out questionnaires in survey interviews. **Background Problems:** BPS-Statistics Indonesia used paper-based questionnaires for interviews and is recently utilizing the Computer Assisted Personal Interviewing (CAPI) method. However, the CAPI method has some drawbacks. Enumerators must input data into the device, which can be burdensome and prone to errors. **Novelty:** This study uses a large language model (LLM) to extract information from survey interviews. **Research Methods:** This study utilizes a text-to-speech application to translate interview results into text. Translation accuracy is measured by the Word Error Rate (WER). Then the text was extracted using the ChatGPT 3.5 Turbo model. GPT-3.5 Turbo is part of the GPT family of algorithms developed by OpenAI. **Finding/Results:** The extraction results are formatted into a JSON file, which is intended to be used for automatic filling into the database and then evaluated using precision, recall, and F1-score. Based on research conducted by utilizing the Speech Recognition API by Google and the ChatGPT 3.5 Turbo model, an average WER of 10% was obtained in speech recognition and an average accuracy of 76.16% in automatic data extraction.

1. Introduction

Large Language Model (LLM) is a relatively new technology, which utilizes artificial intelligence. LLMs are a class of Artificial Intelligence (AI) that can understand, interpret, and generate texts. LLMs are capable of understanding and generating texts at a level indistinguishable from humans [1]. Large Language Models (LLMs) are deep learning models trained on vast amounts of data, enabling them to understand and generate natural language [2]. Recent studies have demonstrated that LLMs have achieved significant success in various natural language tasks, including automatic summarization (creating a condensed version of a text), machine translation (automatically translating text from one language to another), and question answering (developing automated systems that respond to questions based on a given text) [3].

Statistical organizations can use the advantages of an LLM. They could be helpful in automating several duties within a statistical organization because of their exceptional comprehension of textual data, ability to summarize vast amounts of information, and ability to provide responses that like those of a human [1]. Badan Pusat Statistik (BPS-Statistics Indonesia) is the national statistical office of Indonesia. BPS plays a crucial role in providing accurate and reliable statistical information for both the public and the government [4]. To fulfill this role, BPS collects data through censuses and surveys, employing interviewers to gather information from respondents. Initially, BPS used paper-based questionnaires for interviews. However, with technological advancements, BPS began utilizing the

Computer Assisted Personal Interviewing (CAPI) method [5]. Interviews are now conducted using Android-based devices, allowing respondents' answers to be directly stored in a database and automatically uploaded to a central server.

The implementation of CAPI has proven to have numerous advantages. Using Android devices for CAPI minimizes the costs associated with printing paper questionnaires. Additionally, this system enables the direct transmission of respondents' answers to the database, expediting the data collection process. The validation features in the CAPI application also help prevent entry errors by enumerators during interviews [6]. However, the CAPI method has some drawbacks. Enumerators must input data into the device, which can be burdensome and prone to errors [7]. Concurrent interviewing, where the same questions are asked to different household members, can increase the workload on enumerators, leading to physical fatigue [6]. Furthermore, enumerators must multitask—conducting interviews, listening to respondents, and entering data—which can result in typing errors and incomplete documentation of respondents' information, as enumerators often only type key points [8].

With the advancement of technology, human-computer interaction has evolved from keyboard input to voice input [9]. Collecting data through voice input for survey interviews is a promising method compared to using paper questionnaires for interviews. For responses, the enumerator only needs to press the record button to capture the answer [10]. Collecting voice data also facilitates in-depth interviews, resulting in richer and more comprehensive information [7].

The recorded results must be converted into text. Therefore, using an automatic speech recognition application, the recordings are transcribed into text. However, the output of the automatic speech recognition process is still in the form of unstructured text. To fill out the interview questionnaire, an information extraction process is required. This process involves several natural language processing methods, such as categorizing information by identifying parts of the text that contain words matching the fields in the questionnaire.

Based on the research conducted by [12], speech recognition, commonly referred to as Speech-to-Text, can be achieved through various processing techniques. Speech-to-Text application by dividing the voice processing process into two stages: feature extraction and feature matching. She employed the Mel Frequency Cepstral Coefficients (MFCC) method for feature extraction, while the Dynamic Time Warping (DTW) method was used for feature matching. After conducting experiments on 217 data samples, she achieved an accuracy rate of 95.85%. Although Dinata et al.'s research demonstrated high accuracy, their experiments were limited to words and sentences containing only five words. This limitation makes their approach unsuitable for voice conversion in interview processes.

Indonesian Speech-to-Text on self-recorded voices and voices from YouTube, with durations ranging from a minimum of 17 seconds to a maximum of 2 minutes [13]. The Speech-to-Text process in this study involved several stages, including feature extraction, voice detection, and language detection. At the feature extraction stage, noise correction is performed to isolate the desired sound from background noise using the FastICA algorithm. For the voice and language detection stages, this study employs the speech recognition library module in Python, utilizing the Google Speech Recognition API. Experiments combining the FastICA and Google Speech Recognition algorithms achieved an accuracy of 94.75%. Buana's research demonstrates that the use of the speech recognition library module for extensive word sets yields high accuracy.

The effectiveness of the speech recognition library module is further supported by research conducted by Adnan et al. [14]. This study compares the accuracy of the Speech-to-Text process for audio recorded in real time versus audio files. The experiments revealed that the accuracy of the Speech-to-Text process for real-time audio recordings was slightly lower than for audio files, with respective accuracies of 93% and 97%. To obtain category information, the LLM can be utilized. Recent studies have demonstrated that LLMs have achieved significant success in various natural language tasks, including automatic summarization (producing a condensed version of a text), machine translation (automatically translating text from one language to another), and question answering (developing automated systems that respond to questions based on a given text) [11].

Gartlehner et al. [15] conducted a study on data extraction, experimenting with extracting data elements from published research. The experiment involved 10 English-language controlled trial publications, each containing 16 data elements. Utilizing the Large Language Model Claude 2, the study achieved an overall accuracy of 96.3%, with high test-retest reliability (replication 1: 96.9%; replication 2: 95.0%). In contrast to the controlled trial sample used by Gartlehner et al., Zou et al. [16] conducted research on data extraction from ESG reports published by companies in 12 industries listed on the Hong Kong Stock Exchange in 2022. This study utilized Large Language Models (LLM), specifically ChatGPT-4, combined with the Retrieval Augmented Generation (RAG) technique. The trial, which included a sample of 166 companies, achieved an accuracy rate of 76.9% in extracting structured data.

The currently popular Large Language Models include: BERT developed by Google, T5 developed by Google, and GPT developed by OpenAI. Although they all implement Transformer, these three models have different architectures. Therefore, this study aims to conduct a preliminary study into the use of LLMs for extracting information to fill out questionnaires in survey interviews.

2. Material and Methods

2.1. *Collection and Preparation*

As an early study on the application of LLM in survey interviews, this research focuses on the voice conversion system from interview recordings, using sample questions from the 2020 Population Census. The interviews conducted in this study used the Indonesian language. All respondents are students of Politeknik Statistika STIS. Here is the profile of the respondents of this study.

1. Respondent 1: female, a student from Kudus, Central Java which has 4 family members.
2. Respondent 2: female, a student from Palembang, South Sumatra which has 4 family members.
3. Respondent 3: female, a student from Banyuwangi, East Java which has 3 family members.
4. Respondent 4: female, a student from Makassar, South Sulawesi which has 2 family members.
5. Respondent 5: male, a student from Bantul, DI Yogyakarta which has 4 family members.
6. Respondent 6: male, a student from Agam, West Sumatra which has 3 family members.

The questions used in this interview can be seen in Table 1. In one-way interviews, respondents directly speak to answer written questions, while in two-way interviews, the enumerator asks each question, and the respondent provides the answers. The questions asked relate to 11 fields whose information will be extracted as shown in table 1.

Table 1. List of questions

Fields (English / *Indonesian*)	**Type**
Province (*Provinsi*)	Text
Regency/City (*Kabupaten/kota*)	Text
Religion (*Agama*)	Text
Head of household name (*Nama kepala rumah tangga*)	Text
Email (*Email*)	Text
Name of respondent (*Nama responden*)	Text
Cellphone number (*Nomor HP*)	Number
Number of household member (*Jumlah anggota rumah tangga*)	Number
ID (*NIK*)	Number
Address (*Alamat*)	Text
Business classification (*Lapangan usaha*)	Text

Six respondents were interviewed across several different scenarios. In addition, the results of interviews conducted in one direction during the day will be given additional noise using Gaussian noise from the Python library. This repetition with different scenarios on the same respondents aims to see the consistency of speech recognition.

The scenarios are as follows:

1. One-way interviews in the morning using external recording.
2. One-way interviews in the afternoon using external recording.
3. One-way interviews in the evening using external recording.
4. One-way interviews with respondents adding regional accents using external recording.
5. One-way interviews with randomized questions using external recording.
6. One-way interviews using a recorder from web-based application.
7. Two-way interviews using external recording.
8. Two-way interviews with randomized questions using external recording.
9. One-way interviews during the day with the addition of Gaussian noise from the python library.

For the purposes of case number 6, a simple web-based application will also be developed using python, javascript, and HTML programming.

2.2. Audio Translation with Speech Recognition

The collected recordings from scenarios 1-5 and 7-9 above are saved in MP3 format. Then the recorded audio format is converted from MP3 to WAV. This is necessary because the Automatic Speech Recognition process requires the library to accept voice input only in WAV and FLAC formats. The conversion of audio format from MP3 to WAV is performed using the AudioFileClip module in the *moviepy.editor* library.

The AudioFileClip module can convert all types of audio supported by ffmpeg. The advantage of using the moviepy library for converting sound formats is that it does not require a separate *ffmpeg* installation. Things to consider when converting sound with *moviepy* are the metadata of the audio. Audio metadata contains, among others: duration, format, bitrate, audio channel, and audio frequency. The recorder provided by the cellphone usually already has complete metadata. However, recordings in web-based applications still use the default recorder from the browser application, such as Mozilla Firefox, Google Chrome, and so on. Recording with this browser application produces audio with incomplete metadata. Therefore, metadata is refined using the *ffmpeg* library.

The translation process using the SpeechRecognition library begins with inputting audio by entering the file directory. The audio file will be recorded or read using the recognizer available in the SpeechRecognition library function. The recorded file is detected using Google Speech Recognition with the Indonesian language filter. After the detection process is complete, the output will be in the form of a text string which is stored in the input_text variable. The audio that is already in WAV format is then used as input in the speech recognition process. Several popular libraries can be used for Automatic Speech Recognition, including SpeechRecognition, PyAudio, and Librosa [17]. This study will use the SpeechRecognition library which utilizes the Google Speech Recognition API and translates into text.

2.3. Evaluation of Automatic Speech Recognition

In the next step, translation accuracy will be measured by the *Word Error Rate* (WER). WER is the percentage of words, which are to be inserted, deleted or replaced in the translation in order to obtain the sentence of reference [18]. WER is the most popular metric for Automatic Speech Recognition evaluation, which measures the percentage of word errors (Substitution (S), Insertion (I), Deletion (D)) against the number of words processed (N). WER can be written as the following equation 1.

$$WER = \frac{(S + D + I)}{N} \times 100\% \tag{1}$$

2.4. Model Fine Tuning

In the next process, the text is used as input into the fine-tuned Chat-GPT3.5 Turbo model. This study chooses the OpenAI and ChatGPT-3.5 Turbo tools as models in this study. This is because this model does not require a computer with high computer capabilities and easy free trial access. GPT3.5 Turbo is part of the GPT family of algorithms developed by OpenAI [19]. GPT is an autoregressive model that uses an attention mechanism to predict the next token in a sequence based on previous tokens. This process is conducted to obtain data extraction results from the entered text input.

The process of fine-tuning involves several stages, including:

1. Preparing and Uploading Training Data:

   Training data must be stored in a JSONL file with the following format.

   {"messages": [{"role": "system", "content": " *The contents of the task/prompt/command you want to run* "}, {"role": "user", "content": " Contents of Speech Recognition text results "}, {"role": "assistant", "content": " Desired output result "}]}

   It is important to note that a minimum of 10 training data samples is required for fine-tuning.
2. Uploading Training Data:

Once the training data has been created, upload it to the OpenAI API web by selecting the model used, in this case GPT-3.5 Turbo-0125.

3. Confirmation:

   If the upload is successful, the name of the fine-tuned model used in this study will be displayed.

2.5. *Data Extraction Using Fine-Tuned Model*

The next step is to extract information from translated text using the fine-tuned model above. To use the Chat-GPT3.5 Turbo Fine Tuned model in Python, the openai library is required. OpenAI provides an API that gives users worldwide access to the LLM model so that developers can build interactions with OpenAI applications. In performing information extraction using ChatGPT, this study uses Indonesian language prompt tuning with several commands as follows.

- *Tugas dari model adalah untuk melakukan pengekstrak data dari variabel input_text dan data yang diperlukan, antara lain: Provinsi, Kabupaten/Kota, Nama Kepala Keluarga, Email, Jumlah Anggota Keluarga, Nomor Handphone, dan Alamat Rumah*
- *Untuk data Nama, NIK, Agama, dan Deskripsi Pekerjaan harus diekstrak secara berulang sesuai dengan jumlah anggota keluarga.*
- *Pengekstrakan Provinsi harus sesuai dengan penulisan Provinsi, Kabupaten/Kota, dan Agama yang ada di Badan Pusat Statistik dan apabila provinsi tidak disebutkan dalam audio rekaman, maka bisa didekati dengan melihat Kabupaten/Kota yang disebutkan dalam audio rekaman.*
- *Pengekstrakan jumlah Anggota Keluarga dan NIK memiliki data berupa numerik.*
- *Pengekstrakan Alamat Rumah harus serinci mungkin.*
- *Pengekstrakan NIK harus memiliki 16 digit angka.*
- *Pengekstrakan Deskripsi Pekerjaan harus sedetail mungkin dan berisi kegiatan yang dilakukan, bahan yang digunakan, output yang dihasilkan, dan tempat bekerja.*

2.6. *Formatting into JSON Form*

The extraction results are formatted into a JSON file. Formatting into JSON format is intended so that it can be used for automatic filling into the database. Subsequently, the precision, recall, and accuracy of the model in recognizing answer entities according to the questions are calculated. Additionally, a matching process is performed to assess the alignment of the extracted text with real answer.

2.7. *Evaluation of Extraction Results*

Evaluation of information extraction results using fine-tuned models for each question in questionnaire is done by calculating precision, recall and F1-score of each respondent's answer. Precision, recall and F1-score are calculated using the following formula [20].

$$Precision: P = \frac{TP}{TP + FP} \tag{2}$$

$$Recall: R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 - score = 2\,\frac{P * R}{P + R} \tag{4}$$

Meanwhile, to evaluate the accuracy of the answers recognized by ChatGPT compared to the actual answers, the accuracy of the answers will be evaluated.

$$Accuracy\ = \frac{correct\ answer\ recognized}{number\ of\ questions} \times 100\% \tag{5}$$

3. Result and Discussion

Based on the results of Automatic Speech Recognition (ASR), it is evident that less common words often result in translation errors, such as “NIK” being interpreted as “nikah.” Additionally, most symbols are translated phonetically, for example, “@” being rendered as “ath.” Translation errors are also prevalent in the pronunciation of names and email addresses, which often differ from the intended spelling.

This issue arises because names and emails are created by individuals, allowing for variations in spelling despite identical pronunciation. Consequently, there is no standardized spelling for names and email addresses in ASR. Examples of incorrect spellings from the Automatic Speech Recognition results include Indriani instead of Indriyani, Sumiati instead of Sumiyati, and Zainab instead of Zaenab, among others. Repeated mention of digits also often experiences translation errors. This is because the same number but mentioned repeatedly will be read as a single digit number. For example; 0001 becomes 001. Two examples of the speech recognition results for the fourth respondent during the afternoon interview can be seen in Table 2.

Table 2. Two examples of speech recognition results

Respondent, case	**Result of *Speech Recognition***
Respondent 1, One-way interviews in the afternoon using external recording	**Saya tinggal di kabupaten Kudus provinsi Jawa Tengah nama kepala keluarga Ibu Jariyah alamat rumah di desa damaran nomor 26 RT 3 RW 1 Kecamatan Kota kabupaten Kudus nomor handphone 0815 7561 1774 untuk jumlah anggota rumah tangga ada 4 email 222011275 ~~etis~~ @stis.ac.id Kemudian untuk kepala anggota rumah tangga untuk anggota Rumah Tangga pertama itu Edwin Brian Rahmanto Nik 3319020 905080003 Kemudian untuk agama Islam dan dia masih bersekolah di SMK Muhammadiyah 1 Kudus jurusan teknik sepeda motor Kemudian untuk nama ibu ~~Sumiati~~ Sumiyati ~~nikah~~ nik 331 9025 ~~10445001~~ 104450001 kemudian agama Islam dan merupakan pensiunan di Rumah Sakit Islam Kudus Rumah Sakit Islam Kudus Kemudian untuk nama Ibu Jariyah agama Islam merupakan guru di Sekolah Dasar Muhammadiyah 1 Kudus Nik 3319024 60672 ~~001~~ 0001 Kemudian untuk nama untuk agama Islam kemudian untuk nama Karina Cindy Rahmanto merupakan mahasiswa di ~~Polsek~~ Polstat Stis beragama Islam nikah 331 902540302 0002**
Respondent 2, One-way interviews in the afternoon using external recording	**Provinsi Sumatera Selatan Kota Palembang nama kepala keluarga Bapak Amrin Joni email Niken Yuliana 853 @ gmail.com nomor HP 0895 0310 3779 jumlah anggota rumah tangga 4 alamat rumah jalan bungaran Nomor 61 RT 09 kota Palembang nama anggota rumah tangga pertama Bapak Amrin Joni agama Islam Nik 1671 09151062 ~~005~~ 0005 pekerjaan membuka usaha di depan rumah anggota rumah tangga kedua nama ibu ~~Zainab~~ Zaenab agama Islam Nik 1671 0941 0165 0008 lapang Pekerjaan ibu rumah tangga anggota rumah tangga ketiga nama Elisa Anggraini agama Islam ~~nikah~~ nik 1671 0943 0299 0007 rumah tangga keempat Niken Yuliana nama nama Niken Yuliana agama Islam ~~nikah~~ nik 1671 ~~09.58~~ 0958 0701 ~~005~~ 0005 pekerjaan mahasiswa di politeknik statistika Stis pekerjaan anggota rumah tangga ketiga sebagai mahasiswa di Universitas Negeri Malang**

Notes: Words in green, red, and blue indicate correct, incorrect, and corrective words, respectively.

Based on the Word Error Rate (WER) values provided, the one-way recording condition yields a smaller WER value than the two-way recording. When considering the recording method, audio files from external recorders exhibit a smaller WER value compared to recordings obtained directly from the web-based application. This difference can be attributed to the superior sound quality of external recorder files. Additionally, the presence of noise significantly impacts the results of Automatic Speech Recognition (ASR) translation. Specifically, audio files with noise tend to have a higher WER value than those without noise.

Table 3. Word error rate of speech reconition result

Case	Respondent					
	1	2	3	4	5	6
1	6%	5%	9%	6%	-	7%
2	7%	5%	7%	5%	11%	11%
3	11%	7%	11%	6%	-	5%
4	5%	7%	26%	6%	11%	8%
5	7%	9%	7%	6%	8%	11%
6	14%	13%	9%	8%	7%	7%
7	12%	12%	18%	11%	13%	11%
8	10%	7%	15%	7%	-	15%
9	10%	7%	11%	6%	18%	15%
Average : 10%						

Overall, the results of Speech Recognition in translating voice into text produce good values as can be seen in Table 3. This can be seen from the average Word Error Rate of 10%. This means that Speech Recognition has succeeded in correctly translating 90% of words. The next step is to extract information from the text that has been translated with ASR. The extracted information corresponds to the questions in the questionnaire. An example of the information extraction results on one of the respondents can be seen in Table 4.

Table 4. Example of information extraction result

Respondent, cases	**Results of information extraction using ChatGPT3.5 Turbo**
Respondent 1, One-way interviews in the afternoon using external recording	"provinsi": "Jawa Tengah", "kabupaten_kota": "Kudus", "nama_kepala_keluarga": "Jariyah", "email": "imel222011275@attis.com", "nomor_hp": "081575611774", "jumlah_anggota_keluarga": 4, "alamat_rumah": "Desa Damaran Nomor 26, RT 3, RW 1, Kecamatan Kota, Kabupaten Kudus, Jawa Tengah", "anggota_keluarga": "nama": "Jariyah", "nik": 3319024606720001, "agama": "Islam", "deskripsi_pekerjaan": "Guru di SD Muhammadiyah 1 Kudus" "nama": "Edwin Brian Rahmanto", "nik": 3319020905080003, "agama": "Islam", "deskripsi_pekerjaan": "Sedang bersekolah di SMK Muhammadiyah 1 Kudus Kelas 1 SMK"

	"nama": "Karina Cindy Rahmanto", "nik": 3319025403020002, "agama": "Islam", "deskripsi_pekerjaan": "Sedang berkuliah di Politeknik Statistika (STIS) jurusan komputasi statistik" "nama": "Sumiati", "nik": 3319025104450001, "agama": "Islam", "deskripsi_pekerjaan": "Pensiunan perawat di RS Rumah Sakit Islam Kudus"

Then the results of this information extraction stored in JSON format as can be seen in Table 5. Formatting into JSON is done in this study by adding the symbols { and , to the output results of data extraction. An example is as follows.

Table 5. Extracted information from survey interviews in JSON format

"provinsi": "Jawa Tengah", "kabupaten_kota": "Kudus", "nama_kepala_keluarga": "Jariyah", "email": "imel222011275@attis.com", "nomor_hp": "081575611774", "jumlah_anggota_keluarga": 4, "alamat_rumah": "Desa Damaran Nomor 26, RT 3, RW 1, Kecamatan Kota, Kabupaten Kudus, Jawa Tengah", "anggota_keluarga": "nama": "Jariyah", "nik": 3319024606720001, "agama": "Islam", "deskripsi_pekerjaan": "Guru di SD Muhammadiyah 1 Kudus" "nama": "Edwin Brian Rahmanto", "nik": 3319020905080003, "agama": "Islam", "deskripsi_pekerjaan": "Sedang bersekolah di SMK Muhammadiyah 1 Kudus Kelas 1 SMK"	{ "provinsi": "Jawa Tengah", "kabupaten_kota": "Kudus", "nama_kepala_keluarga": "Jariyah", "email": "imel222011275@attis.com", "nomor_hp": "081575611774", "jumlah_anggota_keluarga": 4, "alamat_rumah": "Desa Damaran Nomor 26, RT 3, RW 1, Kecamatan Kota, Kabupaten Kudus, Jawa Tengah", "Anggota_keluarga": [{ "nama": "Jariyah", "nik": 3319024606720001, "agama": "Islam", "deskripsi_pekerjaan": "Guru di SD Muhammadiyah 1 Kudus" }, { "nama": "Edwin Brian Rahmanto", "nik": 3319020905080003, "agama": "Islam", "deskripsi_pekerjaan": "Sedang bersekolah di SMK Muhammadiyah 1 Kudus Kelas 1 SMK" }]}

The next step involves classifying the respondents' answers (information extraction results) according to their questions using a fine-tuned ChatGPT model. As shown in Figure 1, the fine-tuned ChatGPT-3.5 model successfully classified the answers from the Speech Recognition text according to the appropriate question entity. This is evidenced by the precision, recall, and F-1 Score, all of which reached 100%.

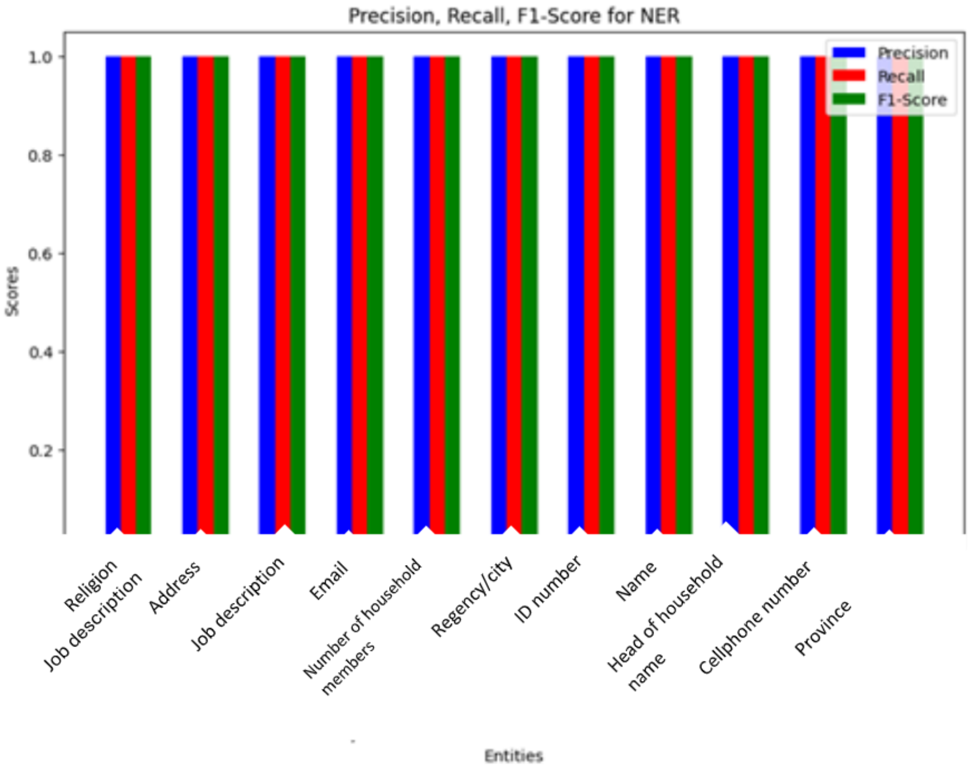


Figure 1. Precision, recall, dan f1-score

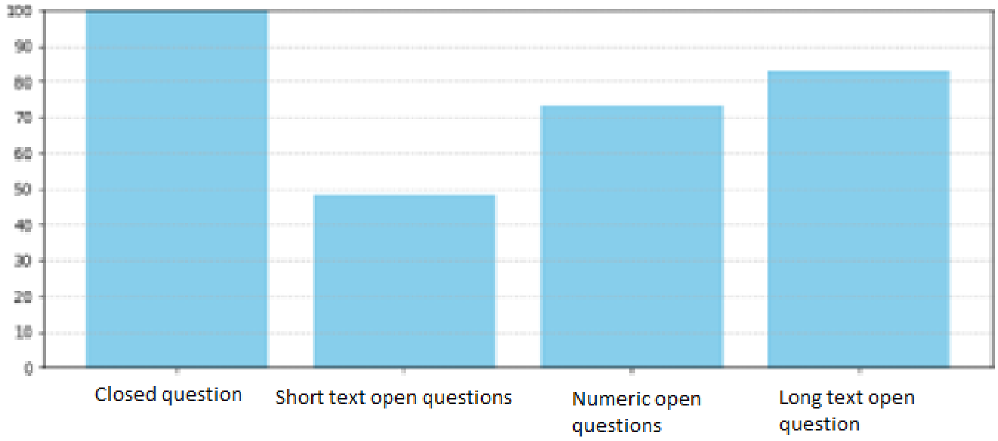


Figure 2. Percentage of correct answer accuracy by question type

The next step is to evaluate the accuracy of filling in the questions on the questionnaire based on the extraction results compared to the expected answers. To evaluate the results of information extraction, the questions in the questionnaire are grouped into several categories. The first category includes closed questions, covering province (provinsi), regency/city (kabupaten/kota), and religion (agama). The second category consists of short-text open questions, including head of household name (nama kepala rumah tangga), name (nama), and email. The third category encompasses numerical open questions, such as cellphone numbers (nomor HP), number of household members (jumlah anggota rumah tangga), and ID number (NIK). Finally, the fourth category includes long-text open questions, covering home addresses (alamat) and job descriptions (deskripsi pekerjaan). Figure 2 illustrates that the fine-tuned ChatGPT-3.5 model can perfectly answer closed-ended questions with clear standards, achieving a 100% success rate. However, the model is less effective at providing correct answers for open-ended questions without clear standards, especially in short-text open questions. This discrepancy is due to errors in the Speech Recognition translation, which result in incorrect extracted answers. Additionally, the most frequent errors occur with NIK digits that contain repeated numbers. But nevertheless, for all open-ended questions groups, the fine-tuned ChatGPT-3.5 model has produced satisfactory answers, achieving an accuracy rate of 83.01%.

4. Conclusion

Based on the results and discussion in the previous section, here are some conclusions that can be drawn: (i) ASR Performance: The Google Speech Recognition API effectively translates voice into text with an average Word Error Rate (WER) of 10%. Errors are more common with names, email addresses, and NIK (National Identification Numbers) due to their non-standardized spellings and the way numbers are read. (ii) Recording Conditions: One-way recordings and recordings with unscrambled questions yield lower WER values compared to direct web recordings and recordings with added noise. (iii) Data Extraction with ChatGPT-3.5 Turbo: The fine-tuned model successfully classifies answers based on question entities. However, the accuracy is affected by errors in the ASR output, particularly with names, emails, and NIKs. These findings highlight the importance of optimizing recording conditions and addressing specific challenges in ASR to improve overall performance. Additionally, refining the input data for models like ChatGPT can enhance the accuracy of automatic data extraction. This study is preliminary, so these things will be improved in subsequent research.

Then suggestions based on the research results are as follows: (i) Automatic Speech Recognition (ASR): Google Speech Recognition is not sufficiently effective for translating voices containing people’s names and long digits. It is better suited for translating voices containing common words. It is recommended to use ASR for one-way recordings to achieve better accuracy. (ii) Large Language Model (LLM) API: Most LLM APIs are paid services. For data with large tokens, it is advisable to use open-source applications that provide LLMs, such as lmstudio. Ensure that the computer used has high capabilities to handle the processing requirements.

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 thank all respondents who were willing to participate in this study.

Competing interests

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

Funding

This study received no external funding.

Underlying data

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

Credit Authorship

Lailatul Hasanah: Methodology, Software, Data Curation, Writing- Original Draft Preparation, Visualization. **Budi Yuniarto:** Conceptualization, Methodology, Supervision, Reviewing and Editing.

References

[1] UNECE HLG-MOS, *Large Language Models for Official Statistics*, United Nations Economic Commission for Europe – High Level Group on Modernisation of Official Statistics, Dec. 2023. https://unece.org/.

[2] R. Peng, K. Liu, P. Yang, Z. Yuan, and S. Li, "Embedding-based retrieval with LLM for effective agriculture information extracting from unstructured data," *arXiv preprint*, arXiv:2308.03107 [cs.AI], 2023.

[3] J. Gao, H. Zhao, C. Yu, and R. Xu, "Exploring the feasibility of ChatGPT for event extraction," *arXiv preprint*, arXiv:2303.03836, 2023.

[4] Pemerintah Republik Indonesia, *Undang-Undang Republik Indonesia Nomor 16 Tahun 1997 tentang Statistik [Law of the Republic of Indonesia Number 16 of 1997 concerning Statistics]*, 1997

[5] BPS, *Pemanfaatan Aplikasi CAPI (Berbasis Android) dalam Pendataan Updating PODES 2020 [Utilization of CAPI (Android-based) Application in Data Collection for Updating PODES 2020]*, 2020. https://bukittinggikota.bps.go.id/news/2020/07/15/41/pemanfaatan-aplikasi-capi--berbasis-android--dalam-pendataan-updating-podes-2020.html

[6] T. Takdir, "Analisis Kinerja, Kualitas Data, dan Usability pada Penggunaan CAPI untuk Kegiatan Sensus/Survey," *Jurnal Aplikasi Statistika & Komputasi Statistik*, vol. 10, no. 1, pp. 9–26, 2018.

[7] J. K. Höhne, K. Gavras, and J. Claassen, "Typing or Speaking? Comparing Text and Voice Answers to Open Questions on Sensitive Topics in Smartphone Surveys," *Social Science Computer Review*, vol. 08944393231160961, 2022.

[8] W. Wicara, *Meningkatkan Produktivitas dengan Menggunakan STT [Enhancing Productivity by Using STT]*, 2023. https://widyawicara.com/meningkatkan-produktivitas-dengan-menggunakan-speech-to-text/

[9] Y. Feng, "Intelligent speech recognition algorithm in multimedia visual interaction via BiLSTM and attention mechanism," *Neural Computing and Applications*, vol. 36, no. 5, pp. 2371–2383, 2024.

[10] T. Lenzner and J. K. Höhne, "Who is willing to use audio and voice inputs in smartphone surveys, and why?," *International Journal of Market Research*, vol. 64, no. 5, pp. 594–610, 2022.

[11] J. Gao, H. Zhao, C. Yu, and R. Xu, "Exploring the feasibility of chatgpt for event extraction," *arXiv preprint*, arXiv:2303.03836, 2023.

[12] C. Dinata, D. Puspitaningrum, and E. Erna, "Implementasi Teknik Dynamic Time Warping (DTW) pada Aplikasi Speech To Text [Implementation of Dynamic Time Warping (DTW) Technique in Speech-to-Text Applications]," *Jurnal Teknik Informatika*, vol. 10, no. 1, pp. 49–58, 2017, doi: 10.15408/jti.v10i1.6816.

[13] I. K. S. Buana, "Implementasi Aplikasi Speech to Text untuk Memudahkan Wartawan Mencatat Wawancara dengan Python [Implementation of Speech-to-Text Application to Facilitate Journalists in Recording Interviews using Python]," *Jurnal Sistem Dan Informatika (JSI)*, vol. 14, no. 2, pp. 135–142, 2020, doi: 10.30864/jsi.v14i2.293.

[14] F. Adnan and I. Amelia, "Implementasi Voice Recognition Berbasis Machine Learning [Implementation of Machine Learning-Based Voice Recognition]," *Edu Elektrika Journal*, vol. 11, no. 1, pp. 24–29, 2022.

[15] G. Gartlehner et al., "Data Extraction for Evidence Synthesis Using a Large Language Model: A Proof-of-Concept Study," medRxiv, 2023-10, 2023

[16] Y. Zou et al., "ESGReveal: An LLM-based approach for extracting structured data from ESG reports,"arXiv *preprint*, arXiv:2312.17264, 2023

[17] T. Ricketts, *Speech Recognition Application With Tone Analyzer (Doctoral dissertation)*. Alabama Agricultural and Mechanical University, 2023

[18] E. Vidal, "Finite-State Speech-to-Speech Translation," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Munich, Germany, 1997.

[19] S. Ozdemir, *Quick Start Guide to Large Language Models: Strategies and Best Practices for Using ChatGPT and Other LLMs*. Addison-Wesley Professional, October, 2023.

[20] H. Dalianis, "Evaluation Metrics and Evaluation," in *Clinical Text Mining*, Springer, Cham, 2018, doi: 10.1007/978-3-319-78503-5_6.

DOI: https://doi.org/10.34123/jurnalasks.v17i1.793

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika &
Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

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

Quantile Regression with Constrained B-Splines for Modelling Average Years of Schooling and Household Expenditure

Yoga Sasmita[1*], Muhammad Budiman Johra[2], Yogo Aryo Jatmiko[3], Deltha A. Lubis[4], Rizal Rahmad[5], Gama Putra Danu Sohibien[6]

[1]BPS-Statistics Central Kalimantan Province, Palangka Raya, Indonesia, [2]BPS-Statistics South Halmahera Regency, Bacan, Indonesia, [3]BPS-Statistics Indonesia, Jakarta, Indonesia,, [4]BPS-Statistics North Sumatera Province, Medan, Indonesia, [5]BPS-Statistics Pidie Regency, Pidie, Indonesia,[6]Politeknik Statistika STIS, Jakarta, Indonesia

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

ARTICLE INFO

Article history:
Received 30 August, 2024
Revised 29 October, 2024
Accepted 9 November, 2024
Published 24 February, 2025

Keywords:
Average Years of Schooling; B-Splines; COBS; Constrained B-Splines; Quantile Regression; Per Capita Household Expenditure

Abstract

Introduction/Main Objectives: Education serves as a driving force for the transformation of society to break the cycle of poverty. This study examines the relationship between average years of schooling and per capita household expenditure in Kalimantan Tengah Province in 2020. **Background Problems:** The method of estimating a regression model that is assumed to follow a certain form of regression equation such as linear, quadratic and others is called parametric regression. However, researchers often encounter difficulties in determining the model specification through data distribution, so the method used is nonparametric regression. **Novelty:** This research uses a quantile-based approach to explore how the impact of education on per capita expenditure varies across different levels of household education. This provides a more nuanced understanding of the relationship, showing not just whether education matters, but how its influence changes at different levels of educational attainment. **Research Methods:** The relationship between average years of schooling and per capita household expenditure is modeled using a quantile regression model with the constrained B-Splines method. **Finding/Results:** Based on the established classification, it can be concluded that an increase in the average years of schooling among household members tends to have a greater impact on raising per capita expenditure.

1. Introduction

Poverty is seen as an economic inability to fulfill basic food and non-food needs measured in terms of expenditure [1]. Expenditure on food and non-food consumption needs can reflect the level of the community's economic capacity, and the purchasing power of the community can provide an overview of the level of community welfare. The higher the purchasing power of the community, the higher the ability to fulfill their needs, which in turn will lead to an increase in community welfare.

Education serves as a driving force for the transformation of society to break the cycle of poverty. Education helps reduce poverty through its effect on labor productivity and through social benefit channels, so education is an important development goal for the nation [2]. Education is a means to gain insight, knowledge, and skills so that employment opportunities are more open and wages are also higher.

A person's education is one of the determinants of per capita consumption [3]. The average years of schooling, which shows the level of education of the community, can reduce the poverty rate in Indonesia [4]. Highly educated people will have skills and expertise so that they can increase their productivity. Increased productivity will increase company output, increase worker wages, and increase people's purchasing power so that it will reduce poverty. The education, especially an increase in the number of years of learning, is a prerequisite for this stage of economic development [5]. The higher a person's education, the better the quality of human resources and will affect productivity. And of course, higher productivity will increase income and expenditure.

Education is concerned with the development of knowledge as well as the expertise and skills of people and labor in the development process. Due to its enormous contribution to economic development, education is said to be human capital. Education is one of the investments in human resources to get a better life. A person with a higher education usually has greater access to higher-paying jobs, compared to individuals with lower levels of education [6]. Through adequate education, the poor will have a better chance of escaping poverty in the future [7]. This is in line with [8] that if education investment is made evenly, including in low-income communities, poverty will be reduced.

B-splines method has been used in several modelling applications by implementing constraints. The constrained smoothing B-splines (COBS) method nonparametrically estimates interest rate structures while meeting no-arbitrage constraints, such as monotonicity and positive rates, enhancing robustness against outliers. Balancing flexibility and constraint adherence, COBS occupies a middle ground between parametric and nonparametric methods, making it well-suited for markets with varying liquidity [9]. A method for constructing COBS wavelets by [10] using the lifting scheme, enabling multiresolution analysis with control over specific points and derivatives. This approach allows curve smoothing while preserving selected "feature points," seamless representation across different resolutions, and editing under constraints. The algorithm is optimized with linear time and storage complexity in the number of control points, making it highly efficient for large datasets. A method by [11] for designing optimal smoothing splines with derivative constraints, using a linear control system to generate the spline. Constraints on spline derivatives are formulated as controls on the system's input and initial state, useful in applications like trajectory planning and convex shape-preserving splines. The method reduces the problem to convex quadratic programming, effectively handling pointwise constraints.

This study will look at the relationship between education level (average years of schooling) and poverty level (household expenditure per capita) in Kalimantan Tengah Province in 2020. Household expenditure per capita is a proxy for household income per capita, which is difficult to obtain in practice. Furthermore, the data was collected in March 2020 as we know that in that period the COVID-19 outbreak began. Average years of schooling is the number of years of study that the population aged 25 years and over has completed in formal education (excluding years repeated). Concerning household expenditure, the variable that has a significant effect is working/not working status. The reference population aged 18 years and above is used because, at the age point of 18 years and above, the proportion working is greater than those not working. Therefore, the reference population taken is the population aged 18 years and above. Kalimantan Tengah Province has the second lowest poverty rate in Kalimantan Island after Kalimantan Selatan Province. 5.36 percent of the population was recorded as poor in 2016 with an average monthly per capita expenditure of IDR 920,786. The average years of schooling in 2015 was recorded at 8.03 years. Nationally, per capita income was recorded at IDR.868,823 and the average years of schooling was 7.84. So that the higher the average years of schooling, the greater the expenditure/income, so that it will have an impact on poverty status.

At the household level, the relationship between education level (average years of schooling) and poverty level (household expenditure) can be shown based on a regression model. The method of estimating a regression model that is assumed to follow a certain form of regression equation such as linear, quadratic, and others is called parametric regression. However, researchers often encounter difficulties in determining the model specification through data distribution, so the method used is nonparametric regression. One of the estimation techniques in nonparametric regression is B-splines. B-splines is an estimation technique in regression curve fitting that takes smoothing into account. B-Splines are good at handling nonlinear relationships. Through movable knot locations that serve as anchor points where the curve can alter its behavior, they provide flexibility. Because they can describe both linear and complex nonlinear interactions, this flexibility is useful in situations that need for both smoothness and precision [12].

Furthermore, [13] proposes Constrained B-Splines to accomodates the constraines which can be monoton, convec or periodic based on the assumed of the form of curve regression so the regression curve will be more smooth by facilitates the addition of smoothing parameters. The addition of monotone constraints is often applied to estimate parameters where the relationship between the response variable and the predictor variables is assumed to be monotone [14]. The addition of monotone constraints has a smoothing effect on the estimated regression model [15].

2. Materials and Methods

2.1. Materials

The data in this study uses data sourced from the results of the National Socio-Economic Survey (Susenas) semester I 2020 in Kalimantan Tengah Province. The variables used are the Per Capita Expenditure variable as the response variable and the Average Years of Schooling per capita variable as the predictor variable. Household expenditure according to [1] is the cost incurred for consumption by all household members during the month, which consists of food and non-food consumption, regardless of the origin of the goods and is limited to consumption for business purposes or given to other parties. Per capita household expenditure is household expenditure divided by the number of household members in a household or in other words the average household expenditure for each household member.

Average Years of Schooling (RLS) is the number of years spent in formal education. The population included in the calculation of RLS is the population aged 25 years and over. However, based on the background discussed earlier, this study uses the limitation of RLS calculation on the population aged 18 years and above. Average Years of Schooling per capita is the average years of schooling of all household members aged 18 years and above in a household divided by the number of household members. RLS is calculated using the following formula [1]:

$$RLS = \frac{1}{P_{18+}} \sum_{i=1}^{P_{18+}} (LS_i) \tag{1}$$

P_{18+} : Total population aged 18 years and over

LS_i : years of schooling of the i-th population.

Years of schooling of the population aged 18 years and over at the last completed level of education using the following conversion [16]:

Table 1. Conversion highest education completed

No.	Highest education completed	Years
1.	No/never been to school	0
2.	Primary school/equivalent	6
3.	Junior high school/equivalent	9
4.	High school/equivalent	12
5.	Diploma I	13
6.	Diploma II	14
7.	Academy/ Diploma III	15
8.	Diploma IV/ Bachelor (S1)	16
9.	Magister (S2)	18
10.	Doctor (S3)	22

Source: BPS, 2011

2.2. Methods

2.2.1. Nonparametric Regression

Suppose Y is the response variable and X is the predictor variable with $\{(x_i, y_i), i = 1,2, \dots n\}$, $x_i \in X, y_i \in Y$. The relationship between x_i and y_i can be assumed to follow the regression model as follows:

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \dots, n \tag{2}$$

where ε_i is the random error and $f(x_i)$ is the regression function.

2.2.2 B-Splines

B-splines is one of the methods used to estimate nonparametric regression functions. B-splines are defined as polynomial functions that have segmented properties at the interval x formed by knot points (piecewise polynomial) which are then locally estimated at these intervals for a certain polynomial degree [17]. To obtain B-splines of degree v with u knot points, additional knots of 2v are first defined so that a knot row $T = (t_1, \dots, t_v, t_{v+1}, \dots, t_{u+v}, t_{u+v+1}, \dots, t_{u+2v})$ with $t_1 = \cdots = t_v < t_{v+1} < \cdots < t_{u+v} < t_{u+v+1} = \cdots = t_{u+2v}$. Furthermore, the jth B-splines with $j = 1, \dots, u + v = m$ are recursively denoted by the following formula:

$$B_j(x;\upsilon) = \frac{x - t_j}{t_{j+\upsilon-1} - t_j} B_j(x;\upsilon-1) - \frac{x - t_{j+\upsilon}}{t_{j+\upsilon} - t_{j+1}} B_{j+1}(x;\upsilon-1) \tag{3}$$

where:

$$B_j(x;1) = \begin{cases} 1, \text{ if } t_j \le x \le t_{j+1} \\ 0, \text{ others} \end{cases} \tag{4}$$

From equations (3) and (4) it is obtained that on the interval $[t_v, t_{u+v+1}]$, $\sum_{j=1}^{m} B_j(x; v) = 1$ holds for every x.

The regression model (2) is a regression function of unknown shapes that will be approximated by the B-splines function. The B-splines function is formulated:

$$f(x) \approx \sum_{j=1}^{m} \alpha_j B_j(x;\upsilon) \tag{5}$$

From equation (5) above, the regression model (2) becomes:

$$y_i = \sum_{j=1}^{m} \alpha_j B_j(x_i;\upsilon) + \varepsilon_i, \quad i = 1, 2, \dots, n \tag{6}$$

or we can denote it as

$$\mathbf{Y} = \mathbf{B}\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{7}$$

In general, the objective function of B-spline regression is as follows:

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m} \alpha_j B_j(x_i;\upsilon) \right)^2 \right\} \tag{8}$$

So that by using the matrix form (8), the estimator of the B-splines parameter is obtained as follows

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha})^T (\mathbf{Y} - \mathbf{B}\boldsymbol{\alpha}) = \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\alpha}^T\mathbf{B}^T\mathbf{Y} + \boldsymbol{\alpha}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\alpha} \tag{9}$$

The minimum value of $\boldsymbol{\varepsilon}^\mathbf{T}\boldsymbol{\varepsilon}$ is obtained if $\frac{\partial(\boldsymbol{\varepsilon}^\mathbf{T}\boldsymbol{\varepsilon})}{\partial\boldsymbol{\alpha}} = 0$, so $-\mathbf{2B^TY} + \mathbf{2B^TB}\hat{\boldsymbol{\alpha}} = 0$. So, the estimator of B-splines is

$$\hat{\boldsymbol{\alpha}} = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{Y} \tag{10}$$

Based on the results in (10), the estimator for the regression model (7) in matrix form is

$$\widehat{\mathbf{Y}} = \mathbf{AY} \tag{11}$$

where:

$$\mathbf{A} = \mathbf{B}(\mathbf{B^TB})^{-1}\mathbf{B^T}$$

2.2.3. Derivation and Monotonicity of B-Splines (Constrained B-Splines)

The first derivative of v-order B-splines with v>1 in equation (3) is

$$\frac{\partial B_j(x;v)}{\partial x} = \frac{v-1}{t_{j+v-1}-t_j} B_j(x;v-1) - \frac{v-1}{t_{j+v}-t_{j+1}} B_{j+1}(x;v-1), \tag{12}$$

while the first derivative for v = 1 is equal to 0 [17]. After obtaining the first derivative of B-splines (12) then the first derivative for the B-splines function (5) is

$$\frac{\partial f(x)}{\partial x} = (v-1)\left(\sum_{j=2}^{m} \frac{\alpha_j - \alpha_{j-1}}{t_{j+v-1} - t_j} B_j(x;v-1)\right). \tag{13}$$

From the above results, it is obtained that the value of $\frac{\partial f(x)}{\partial x}$ is affected by the value of $\alpha_j - \alpha_{j-1} = \delta_j$ for j = 2,…,m. So, it can be concluded that for

$$\alpha_j \geq \alpha_{j-1} \ (\delta_j \geq 0), \quad j = 2, \dots, m \tag{14}$$

the value of $\frac{\partial f(x)}{\partial x}$ is non-negative so the B-splines function is a monotonically increasing function. While for

$$\alpha_j \leq \alpha_{j-1} \ (\delta_j \leq 0), \quad j = 2, \dots, m \tag{15}$$

the value of $\frac{\partial f(x)}{\partial x}$ is non-positive so the B-splines function is a monotone-decreasing function.

The addition of monotone constraints as in (14) and (15) is often applied to estimate parameters where the relationship between response variables and predictor variables is assumed to be monotone [14]. The addition of monotone constraints provides a smoothing effect on the estimated regression model [15]. From the first derivative of B-splines in equation (12), the second derivative is

$$\begin{aligned}\frac{\partial^2 B_j(x;v)}{(\partial x)^2} = {} & (v-1)(v-2)\left[\frac{1}{(t_{j+v-1}-t_j)(t_{j+v-2}-t_j)} B_j(x;v-2)\right. \\ & - \left(\frac{1}{(t_{j+v-1}-t_j)} + \frac{1}{(t_{j+v}-t_{j+1})}\right)\frac{1}{(t_{j+v-1}-t_{j+1})} B_{j+1}(x;v-2) \\ & \left. + \frac{1}{(t_{j+v}-t_{j+1})(t_{j+v}-t_{j+2})} B_{j+2}(x;v-2)\right].\end{aligned} \tag{16}$$

While the second derivative of the B-splines function is

$$\frac{\partial^2 f(x)}{(\partial x)^2} = (v-1)(v-2)\left(\sum_{j=3}^{m} \frac{\frac{\alpha_j - \alpha_{j-1}}{t_{j+v-1} - t_j} - \frac{\alpha_{j-1} - \alpha_{j-2}}{t_{j+v-2} - t_{j-1}}}{t_{j+v-2} - t_j} B_j(x;v-2)\right) \tag{17}$$

From the above results, the second derivative is obtained if the order of the B-splines is $v > 2$.

2.2.4. Quantile Regression

Quantile regression introduced by [18] is an extension of median regression, where quantile regression allows to estimate of quantile functions at various desired quantile values [19]. Suppose Y is a random variable that has a distribution center, denoted c, then the cumulative distribution function $F_Y(.)$ of c is written:

$$F_Y(c) = P(Y \leq c). \tag{18}$$

For $\tau \in [0,1]$, the τ-th quantile of Y which is based on the objective function L_1 (loss-function), indicates the specific locations of a distribution. The function L_1 is defined

$$q_\tau(Y) = {F_Y}^{-1}(\tau) = \inf\{c : F_Y(c) \geq \tau\}. \tag{19}$$

In general, the τ-th quantile of Y can be expressed by minimizing

$$q_\tau(Y) = \operatorname{argmin}_c \mathrm{E}[\rho_\tau(Y-c)], \tag{20}$$

with the function $\rho_\tau(.)$ referred to as the defined 'check-function':

$$\rho_t(z) = \begin{cases} \tau z, & if\ z > 0 \\ -(1-\tau)z, & others \end{cases} \tag{21}$$

Furthermore, from equations (20) and (21) obtained

$$\mathrm{E}[\,\rho_\tau(Y-c)] = (\tau-1)\int_{-\infty}^{c}(Y-c)dF_Y(y) + \tau\int_{c}^{\infty}(Y-c)dF_Y(y). \tag{22}$$

By minimising the first derivative of the function in equation (22) is obtained:

$$0 = (1-\tau)\int_{-\infty}^{c} dF_Y(y) - \tau\int_{c}^{\infty} dF_Y(y) = F_Y(c) - \tau \tag{23}$$

The function $F_Y(.)$ is monotone, so every element of $\{y: F_Y(y) = \tau\}$ minimizes the function (23). From equation (19), it is obtained that $c = {F_Y}^{-1}(\tau)$ is a unique solution. Suppose $Y_1, \dots, Y_n$ are random samples from Y such that $Y_1, \dots, Y_n$ are independently and identically distributed (*i.i.d*) with *Y*. The empirical cumulative distribution function of $Y_1, \dots, Y_n$ is written:

$$F_n(Y) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} I(Y_i \le y), \tag{24}$$

where I(A) is the indicator of the set A that satisfies the conditions:

$$I(A) = \begin{cases} 1, A\ fulfilled \\ 0, A\ fulfilled \end{cases}$$

The function $F_Y(.)$ can be replaced by $F_n(Y)$ and ${\hat{F}_Y}^{-1}(\tau)$ which is the estimator of ${F_Y}^{-1}(\tau)$ can be obtained by minimizing

$$\mathrm{argmin}_c \int \rho_\tau\,(Y-c)dF_n(Y) = \mathrm{argmin}_c \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\,(Y_i - c). \tag{25}$$

$$Y = \beta_0 + \beta_1 X^{(1)} + \dots + \beta_p X^{(p)} + \varepsilon = \mathbf{X^T}\beta + \varepsilon, \tag{26}$$

with $\beta = (\beta_0, \dots, \beta_p)^T$, $\mathbf{X} = (1, X^{(1)}, \dots, X^{(p)})^T$ and ε is assumed to have a distribution with the notation F. In general, the τ-th quartile of the error (ε) which is

$$F^{-1}(\tau) = \inf\{u: P\{\varepsilon \le u\} \ge \tau\}, \tag{27}$$

with *u* being the error of the regression model (27). The quantile curve equation for *Y* conditional on *X* can be written

$$q_\tau(Y|\mathbf{X}) = [\beta_0 + F^{-1}(\tau)] + \beta_1 X^{(1)} + \dots + \beta_p X^{(p)} = \mathbf{X^T}\beta(\tau) \tag{28}$$

with $\beta(\tau) = \left(\left(\beta_0 + F^{-1}(\tau)\right), \dots, \beta_p\right)^T$. As in the previous discussion, the estimator of the parameter $\hat{\beta}(\tau)$ is obtained by minimizing

$$\min_\beta\ \mathrm{E}\left[\rho_\tau\left(Y - \mathbf{X^T}\beta(\tau)\right)\right]. \tag{29}$$

Let $\left({X_1}^{(1)}, \dots, {X_1}^{(p)}, Y_1\right), \dots, \left({X_n}^{(1)}, \dots, {X_n}^{(p)}, Y_n\right)$ be random samples from $\left(X^{(1)}, \dots, X^{(p)}, Y\right)$ that are independently and identically distributed (*i.i.d*) so that the conditional quantile objective function in equation (28) becomes

$$\min_\beta \sum_{i=1}^{n} \rho_\tau\left(Y_i - \mathbf{X_i}^{\boldsymbol{T}}\beta(\tau)\right), \tag{30}$$

with $\mathbf{X_i} = \left(1, {X_i}^{(1)}, \dots, {X_i}^{(p)}\right)^T$ being the *i*th observation of *X*.

2.2.5. Confidence Interval for Quantile Regression

One system of estimating population parameters based on samples is the confidence interval, which produces more representative parameter estimators than the point estimator system [20]. A confidence interval is an interval between two numbers, where the parameter value of the population lies within the interval. Since quantile regression was introduced, various methods have been used to estimate confidence intervals on quantile regression curves. One of the methods used is the direct method. The direct method is more efficient in estimating confidence intervals than other methods [21].

For $\tau \in [0,1]$ and $\alpha \in (0,1)$, the $(1-\alpha)$ percent confidence interval for the quantile regression curve equation (28) is

$$I_n = \left(\left(\mathbf{X^T}\hat{\beta}(\tau - b_n)\right), \left(\mathbf{X^T}\hat{\beta}(\tau + b_n)\right)\right) \tag{31}$$

with,

$$b_n = z_\alpha\sqrt{\mathbf{XQ^{-1}X^T}\tau(\tau-1)}/\sqrt{n},$$

where z_α is the (1-α) standard normal percentile point and

$$\mathbf{Q} = \mathbf{n^{-1}}\left(\mathbf{X}_i{\mathbf{X}_i}^T\right), \tag{32}$$

Where **Q** is a positive definite matrix of size ((*p*+1)×(*p*+1)).

2.2.6. Quantile Regression Smoothing B-Splines

The quantile objective function for smoothing B-Splines in the form of a linear equation is:

$$\min\left\{\widehat{\mathbf{W}}^T\mathbf{u}+\widehat{\mathbf{W}}^T\boldsymbol{\upsilon}\mid \widehat{\mathbf{X}}\boldsymbol{\alpha}+\mathbf{u}-\boldsymbol{\upsilon}=\widehat{\mathbf{Y}},\left(\mathbf{u},\boldsymbol{\upsilon}\in\square_{+}^{(\mathbf{n+u})}\right)\right\} \qquad (33)$$

Where **u** and **υ** are vectors of positive and negative parts of the regression residuals.

$$\widehat{\mathbf{W}}_{(n+u)\times1}=\begin{pmatrix}\mathbf{W}\\ \mathbf{1}_{u\times1}\end{pmatrix} \qquad (34)$$

with $\mathbf{W}=\left(\rho_\tau\left(z_1\right),...,\rho_\tau\left(z_n\right)\right)^T$ is the weight vector

$$\widehat{\mathbf{Y}}=\begin{pmatrix}\mathbf{Y}\\ \mathbf{0}_{u\times1}\end{pmatrix} \qquad (35)$$

$\widehat{\mathbf{Y}}_{(n+u)\times1}$ is a pseudo response vector with $\mathbf{Y}=\left(y_1,...,y_n\right)^T$

$$\widehat{\mathbf{X}}=\begin{pmatrix}\mathbf{B}\\ \boldsymbol{\lambda}\mathbf{C}\end{pmatrix} \qquad (36)$$

$\widehat{\mathbf{X}}_{((n+u)\times m)}$ is a pseudo matrix design with:

$$\mathbf{B}_{n\times m}=\begin{bmatrix}B_1\left(x_1;\upsilon\right) & B_2\left(x_1;\upsilon\right) & \dots & B_m\left(x_1;\upsilon\right)\\ B_1\left(x_2;\upsilon\right) & B_2\left(x_2;\upsilon\right) & \dots & B_m\left(x_2;\upsilon\right)\\ \vdots & \vdots & \ddots & \vdots\\ B_1\left(x_n;\upsilon\right) & B_2\left(x_n;\upsilon\right) & \dots & B_m\left(x_n;\upsilon\right)\end{bmatrix}$$

$$\mathbf{C}_{u\times m}=\begin{bmatrix}B_1'\left(t_{\upsilon+1};\upsilon\right)-B_1'\left(t_\upsilon;\upsilon\right) & \dots & B_m'\left(t_{\upsilon+1};\upsilon\right)-B_m'\left(t_\upsilon;\upsilon\right)\\ \vdots & \ddots & \vdots\\ B_1'\left(t_m;\upsilon\right)-B_1'\left(t_{m-1};\upsilon\right) & \dots & B_m'\left(t_m;\upsilon\right)-B_m'\left(t_{m-1};\upsilon\right)\end{bmatrix}$$

The objective function:

$$\widehat{\mathbf{W}}^T\mathbf{u}+\widehat{\mathbf{W}}^T\boldsymbol{\upsilon}$$

The control function:

$$\widehat{\mathbf{X}}\boldsymbol{\alpha}+\mathbf{u}-\boldsymbol{\upsilon}=\widehat{\mathbf{Y}}$$

2.2.7. Selection of Smoothing Parameters and Knots

The criterion for selecting the most optimum smoothing parameter (λ) uses the smallest Schawrz Information Criterion (SIC) value [13], with the formulation:

$$SIC(\lambda)=\log\left(\frac{1}{n}\sum_{i=1}^{n}\rho_\tau\left(y_i-\sum_{j=1}^{m}\hat{\alpha}_jB_j(x_i;v)\right)\right)+\frac{1}{2}p_\lambda\frac{\log(n)}{n} \qquad (37)$$

Where p_λ is the sum of the zero residuals for the fitted model.

The number of knots for B-Splines smoothing quantile regression is 20 knots where the location is chosen based on the unique value of the variable X [13]. The u-th knot point (tu) is obtained from:

$$t_u = \text{quantile of } -\left(\frac{u}{20}\right) \text{of the value for variable X;} \quad u = 1, 2, ..., 20 \qquad (38)$$

2.2.8. Monotone Constraint Function on Linear Programmes for Quantile Regression

The addition of an increasing or decreasing monotone constraint function when estimating parameters provides a smoothing effect on the curve of a regression model. The criteria for checking monotone constraints on the objective function of quantile regression on Smoothing B-Splines are:

$\mathbf{H}\boldsymbol{\alpha} > \mathbf{0}$, for monotone increasing function

$\mathbf{H}\boldsymbol{\alpha} < \mathbf{0}$, for monotone decreasing function

With:

$$\mathbf{H} = \begin{bmatrix} B_1^{'}(t_\upsilon;\upsilon) & \cdots & B_m^{'}(t_\upsilon;\upsilon) \\ \vdots & \ddots & \vdots \\ B_1^{'}(t_m;\upsilon) & \cdots & B_m^{'}(t_{m-1};\upsilon) \end{bmatrix}$$

$\widehat{\boldsymbol{H}}\boldsymbol{x} > 0$, for monotone increasing function

$\widehat{\boldsymbol{H}}\boldsymbol{x} < 0$, for monotone decreasing function

With:

$$\widehat{\mathbf{H}} = \left(\mathbf{H} \quad \mathbf{0}_{((u+2)\times 2(n+u))}\right)$$

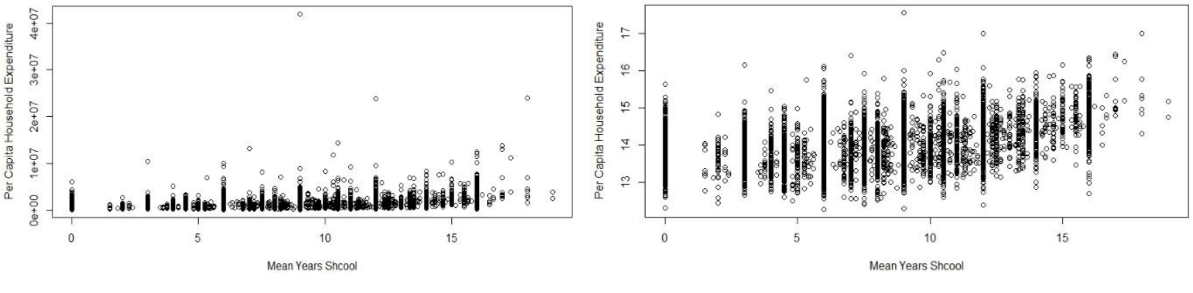
2.2.9. Research Steps

The steps in this study are:

1. Creating a scatter plot between the response variable and the independent variables
2. Performing model specification based on the scatter plot, in this case a B-Splines function approach is used.
3. Checking for outliers in the scatter plot results and if there are outliers then quantiles are used. Checking for outliers can also be done by looking at the distribution of errors with the mean as a measure of data concentration in the B-Splines function.
4. Determine the constraints of the relationship between the two variables, whether it is monotonically increasing or monotonically decreasing.
5. Determine the number of knots and smoothing parameter (λ). In this paper, the B-Splines smoothing function with the number of knots used is 20 knots, and the smoothing parameter (λ) is determined based on the smallest Schawrz Information Criterion (SIC) value.
6. Estimate the quantile regression curve based on the optimal value of smoothing parameter (λ) at several quantile points, namely at τ = 0.2; 0.4; 0.6; 0.8.
7. Estimating the confidence interval for the quantile regression curve by the direct method

3. Results and Discussions

This chapter will explain the relationship between the variables of average years of schooling and average per capita household expenditure in Central Kalimantan Province in 2020 modelled by the COBS method. The reason for using the COBS method is that the data plot (Figure 1) shows a pattern that cannot be clearly specified but has an increasing trend, so it would be better to do the modelling nonparametrically. What is meant by ‘Constrained’ here is the assumption that the relationship between the two data is an increasing pattern which is further referred to as ‘Increase Constrained’. This can be interpreted that the average household expenditure per capita increases along with the average years of schooling of household members.



(a) (b)

Figure 1. Data plots of household expenditure per capita and average years of schooling: (a) original plot; (b) transformation plot

Addition to the irregularity of the data pattern, the next phenomenon is the presence of outlier data (Figure 2) from the residuals of the B-splines model presented in the mean regression. Thus, to capture the phenomenon of the existence of outlier residuals, quantile regression analysis is applied in estimating model parameters.

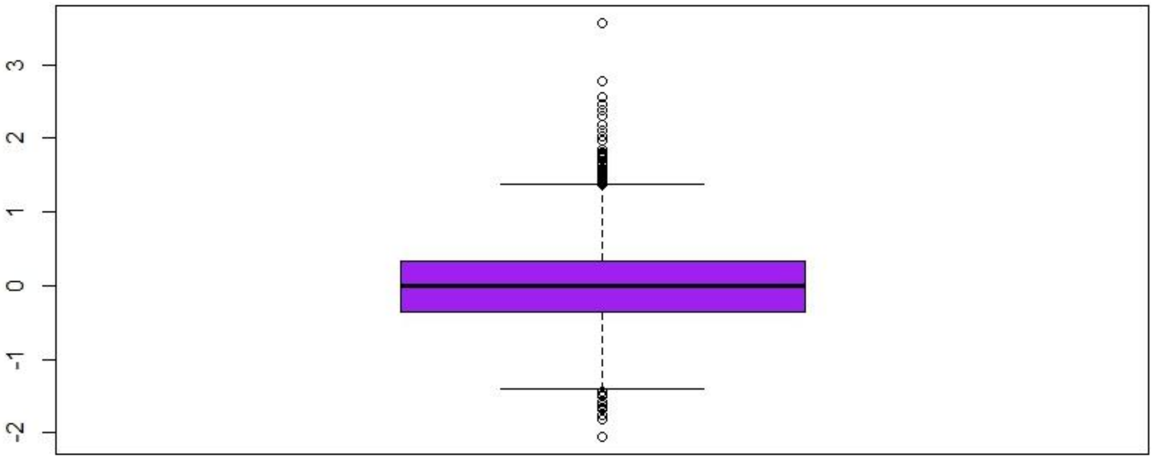


Figure 2. Boxplot of residuals from the b-splines model

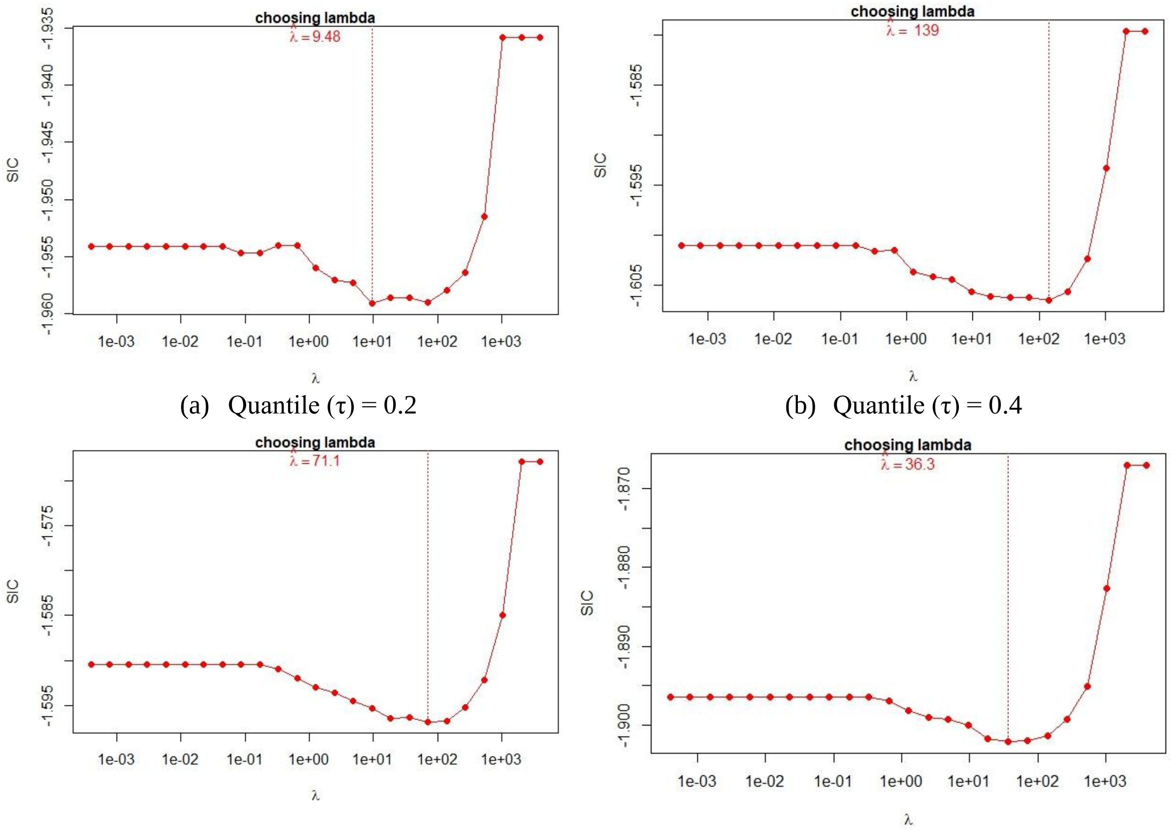
Furthermore, to divide households into groups with similar characteristics based on the average years of schooling of household members and per capita household expenditure, four quantile regression modelling will be applied with the boundaries of the 0.2nd quantile, 0.4th quantile, 0.6th quantile and 0.8th quantile. In quantile regression modelling, each quantile has the same number and location of knot points as presented in Table 2.

Table 2. Knot points at each quantile

Point	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
Knot	0	3.00	4.80	5.60	6.60	7.33	8.14	8.60	9.33	10.12

Point	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}	t_{17}	t_{18}	t_{19}	t_{20}
Knot	10.50	11.25	11.67	12.33	12.83	13.50	14.40	15.00	16.00	19.00

Table 2 shows the knot points calculated by the quantile method from the unique values of the average years of schooling variable as many as 126 values. Meanwhile, the optimum curve smoothing parameter (λ) in each quantile has different values as presented in Figure 3.



(a) Quantile (τ) = 0.2

(b) Quantile (τ) = 0.4

(c) Quantile (τ) = 0.6

(d) Quantile (τ) = 0.8

Figure 3. Optimum smoothing parameter (λ) based on the smallest SIC value in each quantile

Figure 3 (a) shows that at the 0.2th quantile, the optimum smoothing parameter (λ) is 9.48 with a minimum SIC of -1.9591. This indicates that, at the lower end of the distribution, a relatively smaller *λ* provides the best smoothing effect, resulting in a more accurate model with minimized information loss. Figure 3 (b) shows that at the 0.4th quantile, the optimum smoothing parameter (λ) is 139 with a minimum SIC of -1.6065. This suggests that, as we move towards the median of the data distribution, a much larger λ is required to achieve optimal smoothing, potentially due to increased variability in this middle range that requires more significant smoothing to reduce the SIC. Figure 3 (c) shows that at the 0.6th quantile, the optimum smoothing parameter (λ) is 71.1 with a minimum SIC of -1.5969. This result indicates a moderate level of smoothing is ideal for the upper-middle quantile, which is lower than that required for the 0.4th quantile but higher than at the 0.2th quantile.

This pattern might reflect changes in data variability or distribution characteristics that affect the model's performance at this quantile level. Figure 3 (d) shows that at the 0.8th quantile, the optimum smoothing parameter (λ) is 36.3 with a minimum SIC of -1.9021. Compared to the lower quantiles, the decrease in the optimum λ suggests less need for aggressive smoothing, possibly due to reduced variability or a different distribution pattern in the upper quantiles. Armed with the optimum knot points and smoothing parameters (λ) that have been obtained at each quantile, the quantile regression curve based on the COBS method for linear B-Splines smoothing with the assumption of monotonous increase (Increase Constrain) is presented in Figure 4.

Figure 4 shows the estimated household expenditure per capita based on the average years of schooling of household members at the 0.2 quantile, 0.4 quantile, 0.6 quantile and 0.8 quantile. At the 0.6 and 0.8 quantiles, an average year of schooling of around seven years and above drastically increases per capita expenditure. The estimated per capita household expenditure is obtained from the quantile regression model which has coefficients (α) as presented in Table 3.

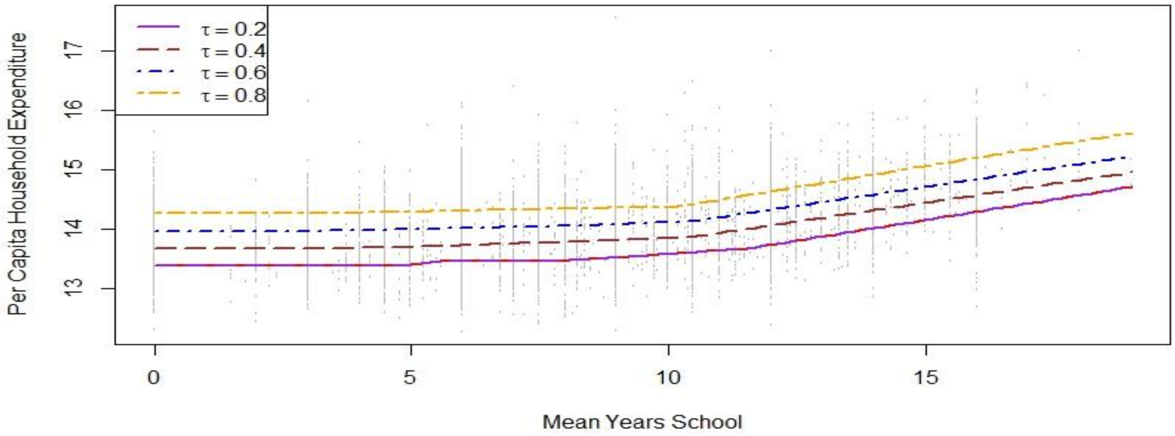


Figure 4. Quantile regression curve with COBS

Table 3 shows that the coefficient value in each quantile tends to increase, this is what makes the quantile regression curve in Figure 4 tend to rise. Thus, the relationship between the average years of schooling of household members and household expenditure per capita tends to increase in line with the assumption (Increase Constrain).

Table 3. Linear b-splines quantile regression coefficients at the 0.2nd, 0.4th, 0.6th and 0.8th quantiles

Knot i^{th} (t_i)	Coefficient (α)			
	τ = 0.2	τ = 0.4	τ = 0.6	τ = 0.8
1	13.3890	13.6773	13.9650	14.2611
2	13.3890	13.6773	13.9650	14.2611
3	13.3890	13.6921	13.9970	14.2906
4	13.4570	13.7163	14.0112	14.3037
5	13.4640	13.7466	14.0289	14.3201
6	13.4692	13.7688	14.0420	14.3321
7	13.4749	13.7933	14.0563	14.3454
8	13.5016	13.8071	14.0720	14.3529
9	13.5444	13.8293	14.0972	14.3649
10	13.5907	13.8533	14.1243	14.3779
11	13.6126	13.8753	14.1372	14.4274
12	13.6565	13.9705	14.2328	14.5329
13	13.6934	14.0234	14.2860	14.5916
14	13.7859	14.1081	14.3710	14.6854
15	13.8553	14.1716	14.4348	14.7558
16	13.9478	14.2563	14.5198	14.8497
17	14.0728	14.3706	14.6346	14.9764
18	14.1560	14.4468	14.7111	15.0608
19	14.2948	14.5738	14.8386	15.2016
20	14.7112	14.9549	15.2213	15.6239

Source: Susenas March 2020 BPS, processed

The quantile regression model with the coefficients presented in Table 3 is as follows:

a) Quantile Regression Model for the 0.2nd Quantile:

$$\hat{y} = \sum_{j=1}^{20} \alpha_j B_j(x; v = 2)$$

$$= 13.39\left(\frac{x}{3.00}B_1(x;1) - \frac{x-4.80}{1.80}B_2(x;1)\right) + 13.39\left(\frac{x-3.00}{1.80}B_2(x;1) - \frac{x-5.60}{0.80}B_3(x;1)\right) +$$

$$13.39\left(\frac{x-4.80}{0.80}B_3(x;1) - \frac{x-6.60}{1.00}B_4(x;1)\right) + \cdots + 14.16\left(\frac{x-15.00}{1.00}B_{18}(x;1) - \frac{x-19}{3.00}B_{19}(x;1)\right) +$$

$$14.29\left(\frac{x-16.00}{3.00}B_{19}(x;1)\right)$$

$$= 4.46xB_1(x;1) - (7.44x - 35.71)B_2(x;1) + \cdots + (4.76x - 76.21)B_{19}(x;1)$$

b) Quantile Regression Model for the 0,4th Quantile:

$$\hat{y} = \sum_{j=1}^{20} \alpha_j B_j(x; v = 2)$$

$$= 13.68\left(\frac{x}{3.00}B_1(x;1) - \frac{x-4.80}{1.80}B_2(x;1)\right) + 13.68\left(\frac{x-3.00}{1.80}B_2(x;1) - \frac{x-5.60}{0.80}B_3(x;1)\right) +$$

$$13.69\left(\frac{x-4.80}{0.80}B_3(x;1) - \frac{x-6.60}{1.00}B_4(x;1)\right) + \cdots + 14.45\left(\frac{x-15.00}{1.00}B_{18}(x;1) - \frac{x-19}{3.00}B_{19}(x;1)\right) +$$

$$14{,}57\left(\frac{x-16{,}00}{3{,}00}B_{19}(x;1)\right)$$

$$= 4.56xB_1(x;1) - (7.60x - 36.48)B_2(x;1) + \cdots + (4.86x - 70.71)B_{19}(x;1)$$

c) Quantile Regression Model for the 0,6th Quantile:

$$\hat{y} = \sum_{j=1}^{20} \alpha_j B_j(x; v = 2)$$

$$= 13.96\left(\frac{x}{3.00}B_1(x;1) - \frac{x-4.80}{1.80}B_2(x;1)\right) + 13.96\left(\frac{x-3.00}{1.80}B_2(x;1) - \frac{x-5.60}{0.80}B_3(x;1)\right) +$$

$$13.99\left(\frac{x-4.80}{0.80}B_3(x;1) - \frac{x-6.60}{1.00}B_4(x;1)\right) + \cdots + 14.71\left(\frac{x-15.00}{1.00}B_{18}(x;1) - \frac{x-19}{3.00}B_{19}(x;1)\right) +$$

$$14.84\left(\frac{x-16.00}{3.00}B_{19}(x;1)\right)$$

$$= 4.65xB_1(x;1) - (7.76x - 37.23)B_2(x;1) + \cdots + (4.95x - 79.15)B_{19}(x;1)$$

d) Quantile Regression Model for the 0,8th Quantile:

$$\hat{y} = \sum_{j=1}^{20} \alpha_j B_j(x; v = 2)$$

$$= 14.26\left(\frac{x}{3.00}B_1(x;1) - \frac{x-4.80}{1.80}B_2(x;1)\right) + 14.26\left(\frac{x-3.00}{1.80}B_2(x;1) - \frac{x-5.60}{0.80}B_3(x;1)\right) +$$

$$14.29\left(\frac{x-4.80}{0.80}B_3(x;1) - \frac{x-6.60}{1.00}B_4(x;1)\right) + \cdots + 15.06\left(\frac{x-15.00}{1.00}B_{18}(x;1) - \frac{x-19}{3.00}B_{19}(x;1)\right) +$$

$$15{,}20\left(\frac{x-16{,}00}{3{,}00}B_{19}(x;1)\right)$$

$$= 4.75xB_1(x;1) - (7.92x - 38.03)B_2(x;1) + \cdots + (5.07x - 81.07)B_{19}(x;1)$$

Next, we calculate the estimated per capita expenditure at several average years of schooling of a person indicating a certain level of education. Some of the education levels used in the estimation of per capita expenditure include 0 years (no/never been to school), 6 years (elementary school/equivalent), 9 years (junior high school/equivalent), 12 years (senior high school/equivalent), 13 years (Diploma I/II), 15 years (Academy/Diploma III), 16 years (Diploma IV/Bachelor's degree), 18 years (Master's/Secondary degree) and 22 years (Doctoral degree).

Table 4. Estimated value of household expenditure by average years of schooling at the 0.2, 0.4, 0.6 and 0.8 quantiles

Mean Years School (Years)	Estimated Expenditure Per Capita (IDR)			
	τ = 0.2	τ = 0.4	τ = 0.6	τ = 0.8
0	652,763	870,933	1,161,264	1,561,410
6	700,668	916,594	1,224,790	1,640,104
9	747,844	1,003,746	1,310,239	1,722,746
12	926,936	1,284,359	1,670,244	2,277,320
13	1,064,952	1,458,297	1,897,447	2,621,560
15	1,405,692	1,880,033	2,448,779	3,473,975
16	1,614,991	2,134,664	2,781,887	3,999,060
18	2,131,721	2,752,004	3,590,205	5,299,431
19	2,449,122	3,124,704	4,078,582	6,100,431

Source: Susenas March 2020 BPS

Table 4 informs that in the 0.2 quantile, household members with an average year of schooling of 0 years have a per capita expenditure of IDR 652,763, while household members with an average year of schooling of 6 years have a per capita expenditure of IDR 700,668, and so on until household members with an average year of schooling of 19 years have a per capita expenditure of IDR 2,449,122. In the 0.4th quantile shows members who have never been to school have per capita expenditure of IDR 870,933, household members with an average length of schooling of 6 years will have per capita expenditure of IDR 916,594, and so on until household members with an average length of schooling of 19 years have per capita expenditure of IDR 3,124,704. In the 0.6th quantile shows members who do not / have never been to school have per capita expenditure of IDR 1,161,264, household members with an average length of schooling of 6 years will have per capita expenditure of IDR 1,224,790, and so on until household members with an average length of schooling of 19 years have per capita expenditure of IDR 4,078,582. At the 0.8 quantile, members who have never been to school have a per capita expenditure of IDR 1,561,410, household members with an average year of schooling of 6 years will have a per capita expenditure of IDR 5,299,431, and so on until household members with an average year of schooling of 19 years have a per capita expenditure of IDR 6,100,431.

Based on the estimated per capita expenditure in each quantile, the classification of households based on average years of schooling and per capita household expenditure in Kalimantan Tengah Province in 2020 will be determined. The classification for each average years of schooling are as follows:

- ‘Very poor’ if the per capita expenditure is less than IDR 652,763.
- ‘Poor’ if the per capita expenditure is between IDR 652,763 and IDR 870,933.
- ‘Middle’ if the per capita expenditure is between IDR 870,933 and IDR 878,788.91.
- ‘Rich’ if the per capita expenditure is between IDR 878,788.91 and IDR 1,161,264; and
- ‘Very rich’ if the per capita expenditure is above IDR 1,161,264.

The complete classification of households with an average year of schooling of 6 years, 9 years, 12 years, 13 years, 15 years, 16 years, 18 years and 19 years is presented in Table 5.

Figure 5 illustrates the relationship between the average years of schooling on the x-axis and household expenditure on the y-axis at 95 percent significance level. The solid red line in Figure 5 (a) represents the estimated trend of household expenditure, while the dashed lines likely represent confidence intervals around the trend estimate of household espenditure in quantile = 0.2. The dashed yellow line in Figure 5 (b) represents the estimated trend of household expenditure, while the dashed lines likely represent confidence intervals around the trend estimate of household espenditure in quantile = 0.4. The dashed purple line in Figure 5 (c) represents the estimated trend of household expenditure, while the dashed lines likely represent confidence intervals around the trend estimate of household espenditure in quantile = 0.6. The dashed black line in Figure 5 (d) represents the estimated trend of household expenditure, while the dashed lines likely represent confidence intervals around the trend estimate of household espenditure in quantile = 0.8. The detail of convidence intervals each quantile presented in Table 6.

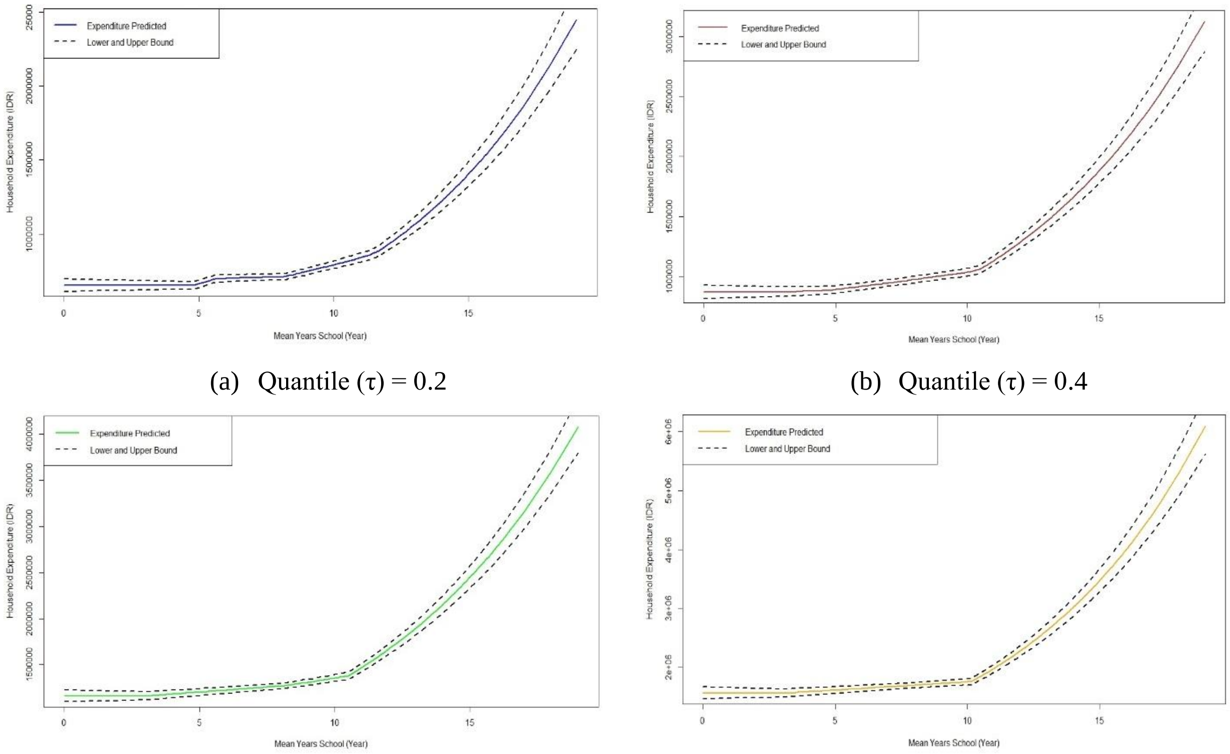
Table 5. Classification of households by average years of schooling and per capita expenditure

Mean Years School (Years)	Estimated Expenditure per Capita (IDR)				
	Very Poor	Poor	Midle	Rich	Very Rich
0	< 652,763	652,763 - 870,933	870,933 - 1,161,264	1,161,264 - 1,561,410	> 1,561,410
6	< 700,668	700,668 - 916,594	916,594 - 1,224,790	1,224,790 - 1,640,104	> 1,640,104
9	< 747,844	747,844 - 1,003,746	1,003,746 - 1,310,239	1,310,239 - 1,722,746	> 1,722,746
12	< 926,936	926,936 - 1,284,359	1,284,359 - 1,670,244	1,670,244 - 2,277,320	> 2,277,320
13	< 1,064,952	1,064,952 - 1,458,297	1,458,297 - 1,897,447	1,897,447 - 2,621,560	> 2,621,560
15	< 1,405,692	1,405,692 - 1,880,033	1,880,033 - 2,448,779	2,448,779 - 3,473,975	> 3,473,975
16	< 1,614,991	1,614,991 - 2,134,664	2,134,664 - 2,781,887	2,781,887 - 3,999,060	> 3,999,060
18	< 2,131,721	2,131,721 - 2,752,004	2,752,004 - 3,590,205	3,590,205 - 5,299,431	> 5,299,431
19	< 2,449,122	2,449,122 – 3,124,704	3,124,704 – 4,078,582	4,078,582 – 6,100,431	> 6,100,431

Source: Susenas 2020 March

The four graphs collectively illustrate a distinct pattern in the expenditure distribution of households based on their average years of schooling. At lower levels of education, particularly when the average year of schooling is between 0 to 6 years, the range of household expenditures is relatively broad. This indicates a high variability in spending among households with minimal education; some households may have very low expenditures, possibly due to limited income-earning opportunities, while others might still maintain moderate levels of expenditure despite lower education levels. This variability suggests diverse economic situations even among households with similarly low education.

Overall, this pattern demonstrates that both low and high education levels are associated with greater variability in household expenditure, while households with moderate levels of education (6-13 years) exhibit more consistent expenditure levels. The findings imply that education significantly influences economic stability and expenditure behavior, with moderate education levels fostering a more uniform economic condition among households, while very low or very high education levels lead to a wider range of economic outcomes.



(a) Quantile (τ) = 0.2

(b) Quantile (τ) = 0.4

(c) Quantile (τ) = 0,6

(d) Quantile (τ) = 0,8

Figure 5. Interval of quantile regression curves with COBS for each quantile: (a) 0.2nd quantile; (b) 0.4th quantile; (c) 0.6th quantile; (d) 0.8th quantile

Table 6 presents the lower and upper bounds of household expenditure by average years of schooling in each quantile. All lower and upper bounds show an increasing trend both from the lowest average years of schooling (0 years) to the highest average years of schooling (19 years) and an increasing trend from the 0.2 to the 0.8 quantile. This indicates that there is an Increase Constrained assumption in the average years of schooling data that significantly affects the increase in per capita household expenditure.

Table 6. Lower and upper bound of per capita household expenditure by average years of schooling and quantiles

Mean	Estimated Expenditure Per Capita (IDR)							
Years	Quantile (τ) = 0.2		Quantile (τ) = 0.4		Quantile (τ) = 0.6		Quantile (τ) = 0.8	
School	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(Years)	Bound	Bound	Bound	Bound	Bound	Bound	Bound	Bound
0	610,644	697,788	816,621	928,857	1,098,394	1,227,733	1,465,592	1,663,509
6	677,720	724,393	887,594	946,532	1,191,222	1,259,304	1,589,055	1,692,792
9	725,474	770,905	974,744	1,033,612	1,277,455	1,343,878	1,673,771	1,773,153
12	889,345	966,117	1,234,048	1,336,734	1,613,522	1,728,942	2,189,536	2,368,625
13	1,016,179	1,116,077	1,393,766	1,525,831	1,824,634	1,973,146	2,507,408	2,740,909
15	1,324,864	1,491,451	1,775,584	1,990,646	2,330,702	2,572,812	3,284,024	3,674,913
16	1,512,100	1,724,883	2,003,227	2,274,703	2,633,173	2,938,971	3,756,770	4,257,020
18	1,968,613	2,308,365	2,548,435	2,971,805	3,359,457	3,836,765	4,913,613	5,715,487
19	2,245,772	2,670,884	2,873,870	3,397,397	3,793,957	4,384,517	5,618,414	6,623,801

Source: Susenas 2020 March

4. Conclusion

This research draws several conclusions regarding household expenditure and schooling through the quantile regression model formed by the COBS method. The model is divided into four quantiles—0.2, 0.4, 0.6, and 0.8—with all coefficients displaying an increasing trend both from the smallest knot to the largest knot and across quantiles from smallest to largest. Meanwhile, the estimated per capita household expenditure shows a similar upward trend along these quantiles and knots. Additionally, based on the estimated values of household expenditure per capita and average years of schooling, households can be classified economically as very poor, poor, middle class, rich, and very rich. This classification suggests that higher average years of schooling among household members significantly influence increased per capita expenditure. In other words, household members with a higher level of education have a higher level of welfare. Quantile regression modelling with COBS is still limited to using only one predictor variable, it is better to use other methods such as P-Spline which can accommodate the use of more than one predictor variable.

Acknowledgments

The authors would like to thank BPS Kalimantan Tengah Province, BPS North Sumatra Province, BPS South Halmahera District, Central Bureau of Statistics, Jakarta, BPS Pidie District, and Politeknik STIS.

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 at BPS-Statistics Kalimantan Tengah Province.

Credit Authorship

Yoga Sasmita: Conceptualization, Data Collection, Formal Analysis, Writing–Original Draft. **Muhammad Budiman Johra:** Transformation Data. **Yogo Aryo Jatmiko:** Methodology. **Deltha A. Lubis:** Writing–Review. **Rizal Rahmad:** Editing**.** **Gama Putra Danu Sohibien:** Supervision.

References

[1] BPS, *Data and Poverty Information of Regencies/Cities in 2014*, 2014th ed., no. 112. Jakarta: Subdirectorate of Social Vulnerability Statistics, 2014.

[2] A. D. Steer, *The New Era of Poverty Reduction in Indonesia*. Jakarta: The World Bank, 2007.

[3] M. Andersson, A. Engvall, and A. Kokko, *Determinants of poverty in Lao DPR*, no. March, 2006.

[4] Suparno, *Analysis of Economic Growth and Poverty Reduction: A Study of Pro Poor Growth Policy in Indonesia*. Bogor: Institut Pertanian Bogor, 2010.

[5] E. A. Hanushek and L. Woessmann, "The Role of Cognitive Skills in Economic Development," *J. Econ. Lit.*, vol. 46, no. 3, pp. 607–668, 2008.

[6] BPS, *Expenditure and Consumption Pattern of Indonesian People 2013*. Jakarta: Household Statistics Subdirectorat, 2013.

[7] M. R. Montecel, "Education as Pathway Out of Poverty," *Idra Newsl.*, vol. XL, pp. 1–8, 2013.

[8] G. Psacharopoulos and H. Patrinos, "Returns to Investment in Education Result," *Economics*, no. April, p. 25, 2018.

[9] M. Poletti Laurini and M. Moura, "Constrained smoothing B-splines for the term structure of interest rates," *Insur. Math. Econ.*, vol. 46, no. 2, pp. 339–350, 2010, doi: 10.1016/j.insmatheco.2009.11.008.

[10] D. Li, K. Qin, and H. Sun, "Curve modeling with constrained B-spline wavelets," *Comput. Aided Geom. Des.*, vol. 22, no. 1, pp. 45–56, 2005, doi: 10.1016/j.cagd.2004.08.004.

[11] H. Fujioka and H. Kano, "Control theoretic B-spline smoothing with constraints on derivatives," *Proc. IEEE Conf. Decis. Control*, no. 1, pp. 2115–2120, 2013, doi: 10.1109/CDC.2013.6760194.

[12] Y. Zhao, M. Zhang, Q. Ni, and X. Wang, "Adaptive Nonparametric Density Estimation with B-Spline Bases," *Mathematics*, vol. 11, no. 2, pp. 1–12, 2023, doi: 10.3390/math11020291.

[13] X. He and P. Ng, "Cobs: Qualitatively constrained smoothing via linear programming," *Comput. Stat.*, vol. 14, no. 3, pp. 315–337, 1999, doi: 10.1007/s001800050019.

[14] K. Kagerer, "A Short Introduction to Splines in Least Squares Regression Analysis," no. 472, p. 51, 2013.

[15] P. Dierckx, "Curve and Surface Fitting with Splines," *Numer. Math. Sci. Comput.*, Mar. 1993, doi: 10.1093/oso/9780198534419.001.0001.

[16] BPS, *Encyclopedia of Social Economic Indicators*. Jakarta: BPS, 2011.

[17] J. R. and C. de Boor, "A Practical Guide to Splines.," *Math. Comput.*, vol. 34, no. 149, p. 325, 1980, doi: 10.2307/2006241.

[18] R. Koenker and G. Bassett, "Regression Quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[19] L. Hudoyo, "Modelling the Relationship between Average Years of Schooling and Household Expenditure Using Constrained B-Splines (COBS) in Quantile Regression," Padjadjaran University, 2017.

[20] E. Walpole, *Introduction to Statistics*, Third. London: MacMilan, 1982.

[21] K. Q. Zhou and S. L. Portnoy, "Direct Use of Regression Quantiles To Construct," *Ann. Stat.*, vol. 24, no. March 1994, pp. 287–306, 1996.

DOI: https://doi.org/10.34123/jurnalasks.v17i1.795

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

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

Comparison of Binary and Traditional Partial Least Squares Structural Equation Modeling: A Study on The Role of Multidimensional Poverty Dimension to Social Protection in Java Island

Diana Bhakti[1*], Ardi Adji[2], Endang Saefuddin Mubarok[3], Renny Sukmono[4], Rudi Salam[5]

[1]BPS-Statistics Badung Regency, Badung, Indonesia, [2]National Team for the Acceleration of Poverty Reduction, Jakarta, Indonesia, [3]Economics Faculty of Jakarta Islamic University, Jakarta, Indonesia, [4]State Treasury Diploma Program of Polytechnic of State Finance STAN, South Tangerang, Indonesia, [5]Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address:* dianabhakti@bps.go.id

ARTICLE INFO

Article history:
Received 31 August, 2024
Revised 26 November, 2024
Accepted 2 December, 2024
Published 24 February, 2025

Keywords:
Binary PLS-SEM; Item Response Theory Model; Multidimensional Poverty; Social Protection; Tetrachoric Correlation

Abstract

Introduction/Main Objectives: The traditional Partial Least Squares Structural Equation Modeling (PLS-SEM) method uses an ordinary least squares regression approach that assumes that indicators must have a continuous scale. When the indicators are categorical, the use of traditional PLS-SEM becomes less appropriate. **Background Problems:** Multidimensional poverty consists of dimensions that are measured by a binary scale. The use of binary PLS-SEM is better than traditional PLS-SEM in modeling the effect of dimensions on social protection on Java Island. **Novelty:** The use of binary PLS-SEM with factor scores from the item response theory model applied to the role of dimensions of multidimensional poverty to social protection has not been carried out yet. **Research Methods:** This study introduces binary PLS-SEM, which is modified from traditional PLS-SEM by changing the data input using a tetrachoric correlation matrix. **Finding/Results:** Empirical results show that the binary PLS-SEM measurement model is better than traditional PLS-SEM. Evaluation of the structural model shows that the path coefficients of binary PLS-SEM are better than traditional PLS-SEM. Both approaches have an overall model fit. The order of multidimensional poverty dimensions that affect social protection are education, living standard, and health.

1. Introduction

Partial least squares structural equation modeling (PLS–SEM) is one of the methods commonly used in estimating complex relationships between indicators and their latent variables [1]. PLS-SEM is formed from two models, namely the measurement model and the structural model [2]. The measurement model is used to describe how well the observed indicators are able to be a measuring tool for the latent variables while the structural model is used to see the relationship between the latent variables [3]. Latent variables can be measured in a reflective and formative way. The reflective method assumes that the indicator is caused by the latent variable while the formative method assumes that the latent variable is formed by its indicators [4].

PLS–SEM uses an iterative algorithm in finding a linear combination of indicators to form factor scores as latent variables and based on these factor scores the parameters of the model are estimated [5]. The algorithm in PLS-SEM also uses the ordinary least squares regression approach so that the data type of the indicator is required to have a continuous scale [6]. In other words, the use of PLS-SEM if the indicator data is categorical is less appropriate [7]. Forcing indicators with a category type to be treated as continuous can produce biased estimates [6]. Currently, there is a PLS-SEM approach that can be used to overcome the problem of ordinal scale category data, namely ordinal PLS (OrdPLS) [6], [8], [9], [10]. This OrdPLS approach is still based on the traditional PLS-SEM algorithm with the main modification being in the input data. With ordinal scale data, the input data used is a polychoric correlation matrix [6]. With slight modifications to the input data, this approach can also be used for indicators with binary scale data, namely using a tetrachoric correlation matrix.

PLS-SEM requires the use of factor scores as a proxy for each latent variable in the structural model [11]. Factor scores in traditional PLS-SEM are obtained from the sum of the multiplication of weights and indicators. This process cannot be done simply when the indicators are categorical [12]. In the OrdPLS method, factor scores are obtained using the mean, median and mode approaches [8]. This approach does not consider opportunities so that another approach is needed that is still related in terms of obtaining factor scores when the data is categorical. One approach that can be used is the item response theory (IRT) model which was originally developed for categorical data, especially binary data [13].

One application that uses binary data is multidimensional poverty measurement. Multidimensional poverty is another approach to poverty measurement that has been used so far, namely the monetary approach. If the monetary approach uses the concept of the ability to meet basic needs for food and non-food from an economic perspective, while multidimensional poverty arises when people do not have resources so that they do not have adequate education, or have poor health conditions, or feel insecure, or low self-confidence, or a sense of helplessness, or the absence of the right to freedom of speech [14].

The measurement of poverty using the World Bank's monetary approach is still the most commonly used measure of poverty worldwide [15]. However, since 2010, the United Nation Development Program (UNDP) and the Oxford Poverty and Human Development Initiative (OPHI) have agreed on a new poverty measurement initiative through the Multidimensional Poverty Index (MPI) based on the multidimensional measure of Alkire-Foster [16]. The MPI approach to poverty measurement uses ten indicators divided into 3 dimensions, namely health (2 indicators), education (2 indicators), and standard of living (6 indicators). Indonesia is one of the countries that still uses a monetary approach in measuring its poverty and has not officially used multidimensional poverty. However, there have been many articles that present the MPI in Indonesia. Some fairly recent articles related to measuring the MPI in Indonesia include [12] [15] [17]. All of these articles use the Alkire-Foster (AF) method approach in calculating multidimensional poverty and utilize national socio-economic survey (Susenas) data. Although using the same data and methods, the indicators used are not exactly the same. For example, the standard of living dimension in [18] was replaced by the expenditure dimension in [15]. The use of these different indicators shows that measuring multidimensional poverty in Indonesia is still under development [12].

This paper attempts to utilize OrdPLS from [8] which uses a polychoric matrix as input data so that it can also be used for a tetrachoric matrix for binary data. In addition, the use of the ability parameters from the IRT model as factor scores for the values of the latent variables is also utilized. Furthermore, a comparison of traditional PLS-SEM with binary PLS-SEM is carried out. As an application for this binary PLS-PM, household data from the 2021 Java Island National Socio-Economic Survey (Susenas) is used. This data has been processed in such a way that it becomes indicators involving dimensions that form multidimensional poverty based on the multidimensional measure of Alkire-Foster [16]. By using the writing of [17], the role of which multidimensional poverty dimensions have a greater influence on social protection will be determined.

2. Material and Methods

2.1. Partial Least Squares Structural Equation Modeling

PLS-SEM with latent variables are formed from two models, namely measurement models and structural models [2]. The measurement model describes how well the observed indicators function as measurement instruments for latent variables [3] while the structural model describes the relationship paths between latent variables. The structural model of PLS-SEM is represented by a linear relationship (Rademaker, 2020)

$$\boldsymbol{l}_{endo} = \mathbf{B}\,\boldsymbol{l}_{endo} + \boldsymbol{\Gamma}\,\boldsymbol{l}_{exo} + \boldsymbol{\zeta} \tag{1}$$

Meanwhile, the linear relationship between the measurement model and the reflective type is stated by [19]

$$\mathbf{x} = \boldsymbol{\Lambda}_x \boldsymbol{l}_{exo} + \boldsymbol{\varepsilon}_x \tag{2}$$

$$\mathbf{y} = \boldsymbol{\Lambda}_y \boldsymbol{l}_{endo} + \boldsymbol{\varepsilon}_y \tag{3}$$

Note that the structural relationship in (1) can be rewritten in matrix notation as follows [9]

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_{exo} \\ \eta_{endo} \end{bmatrix} = \begin{bmatrix} l_{exo} \\ l_{endo} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \boldsymbol{\Gamma} & \mathbf{B} \end{bmatrix} \begin{bmatrix} l_{exo} \\ l_{endo} \end{bmatrix} + \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\zeta} \end{bmatrix} = \mathbf{D}\boldsymbol{\eta} + \mathbf{v} \tag{4}$$

where $\boldsymbol{\eta}_{\text{exo}}$ and $\boldsymbol{\eta}_{\text{endo}}$ are vector of n exogenous and m endogenous latent random variables which defining vector $\boldsymbol{\eta} = [\eta_1, \dots, \eta_n, \eta_{n+1}, \dots, \eta_{n+m}]^T$, $\boldsymbol{\zeta}$ a is the vector of m error components. $\boldsymbol{\Gamma}$ dan $\mathbf{B}$ are $(m \times n)$ and $(m \times m)$ matrices containing the structural parameters. $(\mathbf{I} - \mathbf{B})$ is nonsingular. $\mathbf{0}$ of size $(n \times 1)$ is vector zero, $\boldsymbol{\zeta}$ $(m \times 1)$ is vector of error component assumed to have zero expected value, $E(\boldsymbol{\zeta}) = \mathbf{0}$, and is uncorrelated with $\boldsymbol{\eta}_{\text{exo}}$. $\boldsymbol{v}$ of size $(n + m \times 1)$ is the error vector.

The reflective measurement model can be written in as

$$\boldsymbol{\xi} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_x & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_y \end{bmatrix} \begin{bmatrix} l_{exo} \\ l_{endo} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_x \\ \boldsymbol{\varepsilon}_y \end{bmatrix} = \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{5}$$

where the random variable vector $\mathbf{y}$ of size $(p \times 1)$ and $\mathbf{x}$ of size $(q \times 1)$ are the observed variables, $\boldsymbol{\Lambda}_y$ of size $(p \times m)$ and $\boldsymbol{\Lambda}_x$ of size $(q \times n)$ are coefficient matrices indicating the relationship of $\mathbf{y}$ to $\boldsymbol{l}_{endo}$ and $\mathbf{x}$ to l_{exo}, respectively, and $\boldsymbol{\varepsilon}_y$ of size $(p \times 1)$ and $\boldsymbol{\varepsilon}_x$ of size $(q \times 1)$ are the measurement errors in $\mathbf{y}$ and $\mathbf{x}$, respectively. This measurement model describes the relationship between each latent variable η_j in $\boldsymbol{\eta}$ and a block K_j of manifest indicators, ξ_{jk}; $k = 1,2, \dots, K_j$, elements of the random variable vector ξ of size $(q + p \times 1)$.

Once the model is available, the next step is to estimate the parameters. PLS-SEM parameter estimation uses an algorithm consisting of three sequential stages. In the first stage, the latent variable scores are estimated iteratively for each observation in the sample. In the second stage, the scores obtained from stage 1 are used to calculate the parameters of the measurement model (called outer coefficients and outer loadings). Similarly, in the third stage the structural parameters (also called path coefficients) are finally estimated. The first stage is what makes PLS-PM a novel method while the second and third stages are about performing a series of traditional ordinary least squares (OLS) regressions. For this task, the algorithm needs to determine the construct scores that are used as inputs for the partial regression models (single and multiple) in the path model. After the algorithm calculates the construct scores, they are used to estimate each partial regression model in the path model. As a result, estimates are obtained for all the relationships in the measurement model such as the indicator weights or loadings and the structural model, namely the path coefficients.

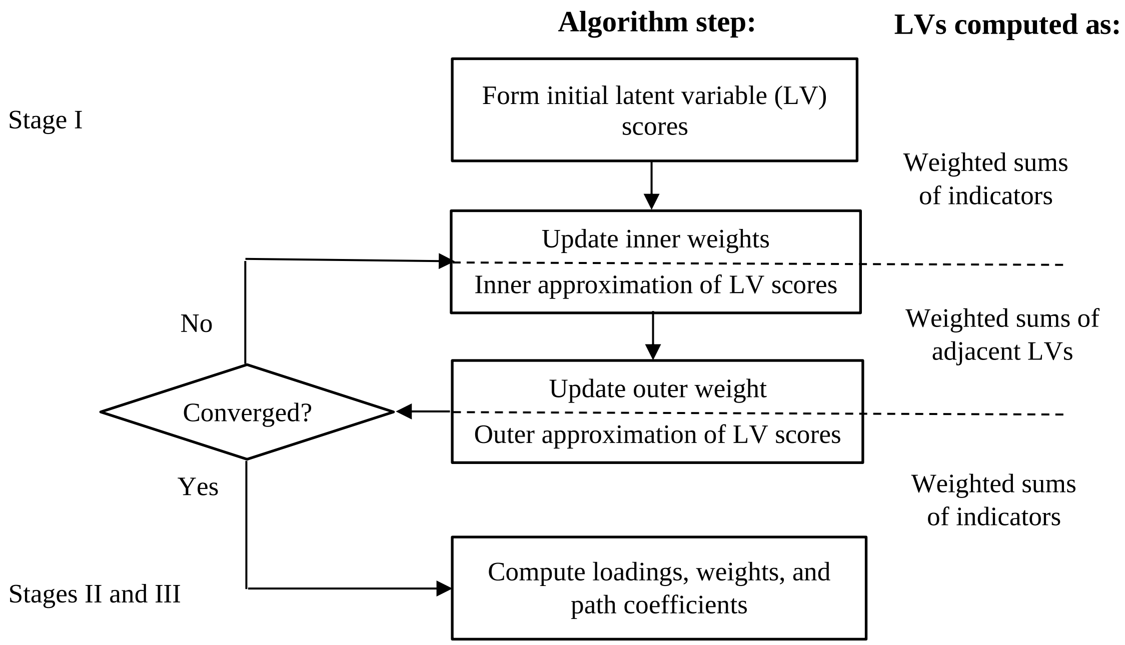

Algorithm step:
LVs computed as:
Stage I
Form initial latent variable (LV) scores
Weighted sums of indicators
Update inner weights
Inner approximation of LV scores
No
Weighted sums of adjacent LVs
Converged?
Update outer weight
Outer approximation of LV scores
Yes
Weighted sums of indicators
Stages II and III
Compute loadings, weights, and path coefficients

Figure 1. Algorithm steps of PLS-SEM (adapted from [20])

The estimation procedure in PLS-SEM involving a series of iterative stages and steps has the consequence that the path coefficient estimates obtained at the end of the procedure cannot be expressed as an explicit function of the indicator data. Therefore, it is impossible to obtain the exact sampling distribution of the estimator in question. Therefore, the only feasible way to perform inferences such as calculating p-values and confidence intervals for the PLS-PM model is through bootstrapping [21].

2.2. *Binary Partial Least Squares Structural Equation Modeling*

PLS-SEM measurement model describes the relationship between each latent variable η_j in $\boldsymbol{\eta}$ and one construct K_j of manifest indicators, Y_{jk}; $k = 1, \ldots, K_j$, elements of the random variable vector $\boldsymbol{Y}$ of size $(p \times 1)$. Suppose there is a measurement model of the reflective type as follows

$$\mathbf{Y} = \Lambda\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{6}$$

If the indicators are binary, then it is assumed that for the set of binary variables $\boldsymbol{Y}$ there are K-dimensional unobserved continuous indicators $\boldsymbol{Y}^*$ represented on an interval scale by a multinormal distribution function [22] [23]. Each observed binary indicator Y_{jk} can assume an existing category that is related to the corresponding continuous indicator Y_{jk}^* through a nonlinear monotone function. From this function, a tetrachoric correlation matrix is obtained which will be used in the PLS-SEM algorithm.

With the presence of indicators with a binary scale, models (1) and (2) need to be modified where the observed variable Y in (2) is replaced with the underlying unobserved continuous indicator Y^*.

$$\mathbf{Y} \leftarrow \mathbf{Y}^* = \Lambda\boldsymbol{\eta} + \boldsymbol{\varepsilon} \tag{7}$$

The dependency relationship between Y and Y^* is not explicitly written because for subject $s = 1,2, \ldots, N$ the actual score of y_{ks}^* for each indicator y_k^* cannot be identified, it is only assumed that the value belongs to the interval determined by the threshold value of the nonlinear function owned as a description of the observed category y_{ks}. The PLS algorithm with binary data (binary PLS) and traditional PLS are not much different apart from changes in the input data and the calculation process that adjusts due to the use of tetrachoric matrices as input data. The occurrence of categorical data being treated as continuous often occurs in applications that cause the resulting Pearson correlation estimate to be biased [24].

Handling of binary category indicators in PLS-SEM based on OrdPLS (ordinal PLS) from [8] uses the same algorithm as in traditional PLS only making modifications to the input data where binary PLS-SEM enters the tetrachoric correlation matrix as input to its algorithm [12]. In the context of binary categories, it can be assumed that the relationship between two dichotomous variables representing continuous variables that are categorized is the tetrachoric correlation coefficient. Tetrachoric correlation is obtained by hypothesizing the existence of a continuous latent variable underlying the true and false dichotomy imposed in scoring a dichotomous item so that it can classify variables into a frequency distribution [25]. The tetrachoric correlation algorithm used is the method from [26].

```
Compute $\mathbf{\Sigma}_{XX}$ (in the case of binary data is a tetrachoric correlation matrix)
Determine the initial weights $\mathbf{W} = \left[\mathbf{w}_1, \dots, \mathbf{w}_j, \dots, \mathbf{w}_{n+m}\right]$
Iterative phase
        Set $\mathbf{W}_{TEMP} = \mathbf{W}$
    Compute:
        $\mathbf{\Sigma}_{\hat{Y}\hat{Y}} = \mathbf{W}^T \mathbf{\Sigma}_{XX} \mathbf{W}$
        ${}_S\mathbf{W} = \mathbf{W}\left\{\left[\mathbf{W}^T \mathbf{\Sigma}_{XX} \mathbf{W}\right] * \mathbf{I}\right\}^{-1/2} = \mathbf{W}\left[\mathbf{\Sigma}_{\hat{Y}\hat{Y}} * \mathbf{I}\right]^{-1/2}$
        $\mathbf{\Sigma}_{\hat{Y}\hat{Y}} = \mathbf{P}_{\hat{Y}\hat{Y}} = {}_S\mathbf{W}^T \mathbf{\Sigma}_{XX}\, {}_S\mathbf{W}$
        $\mathbf{\Upsilon} = \left(\mathbf{T} + \mathbf{T}^T\right) * sign\left(\mathbf{\Sigma}_{\hat{Y}\hat{Y}}\right)$
        $\mathbf{\Sigma}_{XZ} = \mathbf{\Sigma}_{XX}\, {}_S\mathbf{W}\mathbf{\Upsilon}$
        $\mathbf{\Sigma}_{X\hat{Y}} = \mathbf{\Sigma}_{XX}\, {}_S\mathbf{W}$
        $\mathbf{C} = \chi_{\mathbf{w}} * \mathbf{\Sigma}_{XZ}$
        $\pm = sign\left\{\mathbf{1}_{p+q}^T \left[sign\left(\chi \mathbf{w} \mathbf{\Sigma}_{X\hat{Y}}\right)\right]\right\}$
    Update weight $\mathbf{W} = \mathbf{C}\left[diag\left(1_{p+q}^T \mathbf{C}\right)\right]^{-1} diag\left(\pm\right)$
    Obtain ${}_S\mathbf{W}$
Check if $\left\|\mathbf{W} - \mathbf{W}_{TEMP}\right\| < \varepsilon$
```

Figure 2. PLS algorithm using matrix as input data (adapted from [8])

Suppose there is a 2 × 2 contingency table with frequencies given by a, b, c, and d. Any continuous random variable Y can be transformed into a standard normal variable Z_Y by the formula $Z_y = \Phi^{-1}[\Phi_Y(Y)]$ where Φ_Y is the cumulative density function (cdf) of Y, Φ is the cdf of the standard normal distribution, and N is the total frequency. The variable Z_Y is called the standard normal deviate (SND) corresponding to Y. Let z_1 and z_2 be the standard normal deviations corresponding to the marginal probabilities $(a + c)/N$ and $(a + b)/N$, respectively, that is

$$\Phi(z_1) = (a + c)/N, \quad z_1 = \Phi^{-1}\{(a + c)/N\}$$

$$\Phi(z_2) = (a + b)/N, \quad z_2 = \Phi^{-1}\{(a + b)/N\}$$

then the tetrachoric correlation ρ_{tet} is a correlation coefficient that satisfies

$$\frac{a}{N} = \int_{-\infty}^{z_2}\int_{-\infty}^{z_1} \Phi\left(y_1, y_2; \rho_{tet}\right) dy_1 dy_2$$

where $\Phi(y_1, y_2; \rho_{tet})$ is the bivariate normal probability density function (pdf) with mean zero and variance one.

$$\Phi\left(y_1, y_2; \rho\right) = \frac{1}{2\pi\left(1 - \rho^2\right)^{1/2}} \exp\left[-\frac{1}{2\left(1 - \rho^2\right)}\left(y_1^2 - 2\rho y_1 y_2 + y_2^2\right)\right]$$

The probability of the four quadrants formed by performing a dichotomy of variables with the line $y_1 = z_1$ and $y_2 = z_2$ is the same as a/N, b/N, c/N and d/N. This correlation value will be formed into a matrix form called the tetrachoric correlation matrix which will later be used as input data in the PLS-SEM algorithm.

2.3. Score Factor of Item Response Theory Model

In item response theory (IRT) model, factor scores are known as ability parameters (θ). [27] states that there are three popular methods for estimating ability in IRT model, namely Maximum likelihood (ML), Bayesian Maximum a Posteriori (MAP), and Bayesian Expectation a Posteriori (EAP). These methods use a single function called the likelihood function (LF). The ML method has problems when the answer pattern is all true or all false. While the use of MAP has problems with asymmetric LF and iterative arithmetic operations. So that, the ability parameter estimation used in this paper is the EAP approach with the Markov Chain Monte Carlo (MCMC) algorithm because it is generally more robust for complex models [28] [29] [30].

The IRT model used in this study is a two-parameter logistic (2PL) model. The 2PL model use the parameters of the difficulty level of the question b and the discriminatory power a. Parameter a shows the slope and item characteristic curve (ICC) at point b on a certain ability scale. The discriminatory power a functions to determine whether or not a question item can distinguish a group in the aspect being measured, according to the differences in the group. The value of a ranges from -∞ to ∞, but the value of a can be categorized as good if it is in the range of 0 to 2 [31]. The formula for the 2PL model is as follows [32]

$$p\left(y_j = 1 \mid \theta, a_j, b_j\right) = \frac{e^{a_j\left(\theta - b_j\right)}}{1 + e^{a_j\left(\theta - b_j\right)}} \tag{8}$$

In addition, the estimated factor scores with the EAP approach with R quadrature points are [33]

$$\hat{\theta}_i = \frac{\sum_{r=1}^{R} Y_r L\left(Y_r\right) A\left(Y_r\right)}{\sum_{r=1}^{R} L\left(Y_r\right) A\left(Y_r\right)} \tag{9}$$

where Y_r is a node or quadrature point, $L(Y_r)$ is the likelihood function when Y_r is approximating quadrature, $A(Y_r)$ is the quadrature weight corresponding to Y_r which reflects the height of the function g(θ|υ) around Y_r, g(θ|υ) is the continuous population distribution of individuals and υ represents a vector containing the location and scale parameters of the population which have values 0 and 1 respectively [34].

2.4. Multidimensional Poverty

The building of multidimensional poverty in this paper is using M_0 of the method proposed by [35] which is also known as the adjustment headcount ratio. Suppose $x_{ij} \epsilon \mathbb{R}_+$ is the achievement of each individual $i = 1, \dots, n$ on each indicator $j = 1, \dots, d$, and suppose z_j is the deprivation cutoff of the indicator j. Individual deprivation i on the indicator j is defined as $g_{ij}^0 = 1$ when $x_{ij} < z_j$ and $g_{ij}^0 = 0$ otherwise. Then, the deprivation of each individual is weighted by the indicator weight w_j such that $\sum_{j=1}^{d} w_j = 1$. Furthermore, a deprivation score is calculated for each individual, which is then defined as the weighted sum of deprivations $c_i = \sum_{j=1}^{d} w_j g_{ij}^0$. With this score, poor individuals are identified using the second cutoff or poverty cutoff symbolized by , which represents the minimum proportion of deprivation that an individual must experience in order to be identified as a poor individual. In other words, an individual is poor if $c_i \geq k$.

The deprivation of those not identified as poor is then ignored or technically they are censored. Formally, censored deprivation is defined as $g_{ij}^0(k) = g_{ij}^0$ if $c_i \geq k$ and $g_{ij}^0(k) = 0$ otherwise. Analogously, the censored deprivation score is defined as

$c_i = \sum_{j=1}^{d} w_j g_{ij}^0\,(k)$.

Once multidimensionally poor individuals are identified, the measure M_0 combines two fundamental sub-indices, namely the proportion of multidimensionally poor individuals (also called poverty incidence) and the poverty intensity, which is the weighted average of deprivation among poor individuals. Formally, the proportion of poor individuals is given by $H = q/N$, where q is the number of individuals identified as poor. The poverty intensity is given by $A = \sum_{i=1}^{n} c_i(k)/q$. The MPI as M_0 is the product of these two sub-indices

$$MPI = M_0 = H \times A = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} w_j g_{ij}^{0}\left(k\right) \tag{10}$$

By adjusting the incidence of multidimensional poverty based on its intensity, M_0 satisfies dimensional monotonicity [35]. That is, if poor individuals become deprived on additional indicators, then M_0 will increase.

Because of its additive structure, the M_0 measure allows for two types of decompositions that are useful for policy information. First, M_0 can be decomposed into population subgroups. This is because the overall M_0 is the population-weighted sum of the subgroup poverty rates. Then, the percentage contribution of the subgroup to overall poverty can be calculated from the subgroup M_0 weighted by its population contribution compared to the overall M_0. Second, after identification, M_0 can be divided by indicators. The overall M_0 can be expressed as the weighted sum of the proportion of the total population that has been identified as poor and deprived in each indicator (weights refer to the relative weight of each indicator). This proportion is the so-called censored headcount ratio. The percentage contribution of an indicator to overall poverty is calculated as the censored headcount ratio multiplied by its relative weight, divided by the overall M_0 measure.

2.5. Social Protection

Social protection is one of the concepts that has developed in relation to solving multidimensional poverty problems. According to [36], social protection is aimed at addressing the root causes of poverty and is not limited to actions that only solve poverty problems at the symptom level. More broadly, social protection is based on the view that the causes of poverty are related to various social risks faced by the poor and their vulnerability to the impacts of emerging social risks. The emphasis on risk and vulnerability, which are the main causes of poverty, indicates that social protection should have a forward-looking vision and focus on the importance of developing holistic strategies and policies to reduce risks and vulnerabilities for poor groups before they actually occur. Because the concept of social protection is aimed at addressing poverty and vulnerability, the concept of social protection includes two dimensions of social security, namely basic social security for all (horizontal dimension) and the gradual implementation of social security with higher standards (vertical dimension). These two dimensions have been mandated in the ILO Convention Number 102 of 1952 concerning Minimum Standards for Social Security. Therefore, the concept of social protection is not only related to social assistance and social security. Even according to[37], social protection traditionally has a broader concept than social security, social insurance, and social safety nets. Furthermore, [38] stated that social protection is a collection of public efforts to face and overcome vulnerability, risk and poverty that has exceeded the limit. This means that the focus of social protection is on preventing poverty and providing assistance to the poorest people.

Furthermore, the concept of social protection has developed. For example, [39] stated that the concept of social protection traditionally focuses more on short-term protection programs, such as protection mechanisms for people from the impact of shocks caused by natural disasters, unemployment, and death. In contrast, [40] views that social protection has broader components, including protection, prevention, and promotion components to reduce the vulnerability of each individual in the future. Meanwhile, [41] view social protection as having a transformative role, where social protection is aimed at improving status and opening up more livelihood opportunities for marginalized groups in society.

Basically, the framework of social protection refers to the fundamental principle of social justice and the fulfillment of specific universal rights for every person. Everyone should receive social security and an adequate standard of living in obtaining health and welfare services for themselves and their families. [42] states that social protection aimed at overcoming poverty, underdevelopment, and inequality must be complemented by other strategies, such as strengthening labor institutions and social institutions and promoting a pro-worker microeconomic environment. These elements have been included by several countries in their social protection systems. Furthermore, [42] emphasizes that countries with lower middle incomes should create social protection programs that are in line with efforts to reduce poverty, inequality and other social transformations. Furthermore, [36] suggests that social protection should also be aimed at overcoming the root causes of poverty and not limited to actions to resolve symptoms of poverty. This means that social protection must be "forward looking" to avoid various persistent risks that may be faced by poor and vulnerable communities, so that social protection is a way out of the poverty trap.

2.6. Data

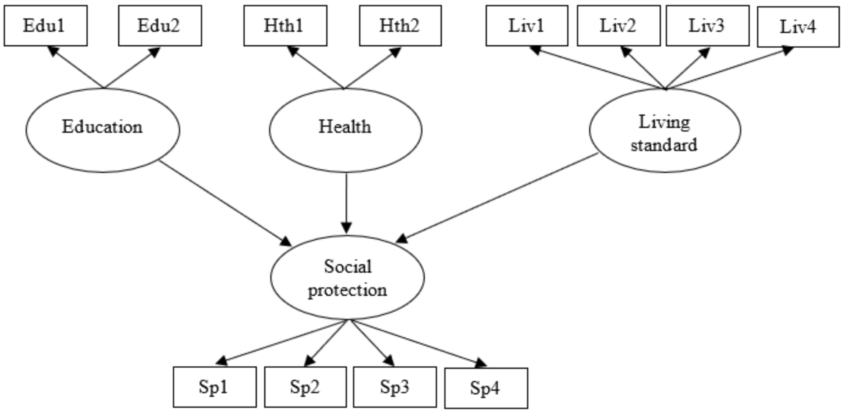


Figure 3. The conceptual framework of the research

The data used in this research is cross-sectional data taken from the 2021 Susenas of Java Island with the unit of analysis being 105200 households. Table 1 shows details of the latent variables and indicators used. Of these latent variables, the dimensions of health, education and living standards are the dimensions that form multidimensional poverty. Figure 3 shows the indicators and latent variables as well as the relationships related to the model studied.

Table 1. Dimensions, indicators, and observed variables

Dimensions	Indicators	Observed Variables
Education	Years of Schooling (EDU1)	- There are household members who do not graduate from junior high school
	School attendance (EDU2)	- There are household members of school age (7-15 years) who do not attend school
Health	Vaccination (HTH1)	- There are births that are not assisted by medical personnel
	Health Insurance (HTH2)	- Households are not covered by health insurance
Living standards	Durability (LIV1)	- Households with non-durable houses
	Sanitation (LIV2)	- Households with inadequate sanitation
	Electricity (LIV3)	- Households with non-Electric Lighting Sources
	Cooking fuel (LIV2)	- Households with biomass/solid cooking fuel
Social protection	Non-cash food assistance (SP1)	- There are household members who receive non-cash food assistance
	Routine assistance (SP2)	- Households receive assistance/social assistance/ subsidies from the local government in the form of routine assistance
	Family hope program (SP3)	- Households receive assistance from the Family Hope Program
	Smart Indonesian program (SP4)	- There are household members who receive assistance from the Smart Indonesia Program

As an empirical study of PLS-SEM, we applied social protection model from [17]. This model is used to show how work affects multidimensional poverty and how education (edu), health (health) and standard of living which are dimensions of multidimensional poverty, affect social protection (SP). All measures are scored for each item on a binary scale. Figure 3 shows the research model used.

3. Results and Discussions

3.1. Evaluation of Measurement Model

The PLS-SEM reflective measurement model was evaluated using reliability and validity measures. Reliability was measured at the indicator level and the latent variable level (internal consistency reliability). Validity assessment focused on convergent validity measured using the average variance extracted (AVE). In addition, the heterotrait-monotrait correlation ratio (HTMT) can also be used to assess the discriminant validity of the construct measured reflectively compared to other construct measures in the same model. Table 2 displays the reliability of the indicator using the loading value. The recommended loading value is above 0.708 because it indicates that the latent variable explains more than 50 percent of the indicator's variance, thus providing acceptable indicator reliability [43]. According to [44], indicators with very low loading (below 0.40) should always be removed from the measurement model. Indicator loading between 0.40 and 0.708 can be considered to be included in the model. Table 2 also shows that in traditional PLS-SEM there are still 5 indicators that are very low (below 0.04) while in binary PLS-SEM only one indicator that is below 0.40. However, because this article aims to compare the method, these indicators are maintained. So that it can be said that the binary PLS-SEM latent variable is better at explaining the variance of each indicator.

The internal consistency reliability and convergent validity is shown in Table 3. There are theree measures to assess internal consistency reliability, namely Cronbach alpha, composite reliability (ρC), and reliability coefficient (rho-A). The recommended value for internal consistency reliability is greater than 0.7 but a value greater than 0.6 is often considered acceptable [45]. Table 3 shows that all measures from binary PLS-SEM for each latent variable is greater than traditional PLS-SEM. So that it can be said that reliability of binary PLS-SEM is better than traditional PLS-SEM.

Table 2. Loading value by dimension and indicator

Dimension	Indicators	Traditional PLS	Binary PLS
Education	Edu1	0.99	0.94
	Edu2	0.18	0.54
Health	Hth1	0.02	0.32
	Hth2	0.99	0.99
Living Standard	Liv1	0.20	0.69
	Liv1	0.71	0.72
	Liv1	0.12	0.72
	Liv1	0.82	0.78
Social protection	Sp1	0.84	0.90
	Sp2	0.31	0.48
	Sp3	0.83	0.93
	Sp4	0.53	0.69

Table 3. Validity and reliability value by dimensions

Dimension	Traditional PLS				Binary PLS			
	Cronbach's α	ρC	ρA	AVE	Cronbach's α	ρC	ρA	AVE
Education	0.08	0.58	0.29	0.51	0.76	0.73	0.58	0.59
Health	0.10	0.51	1.66	0.50	0.83	0.66	1.60	0.55
Living Standard	0.23	0.55	0.37	0.31	0.82	0.82	0.72	0.53
Social protection	0.56	0.74	0.71	0.44	0.85	0.89	0.89	0.64

The next step is to assess the convergent validity of each latent variable. Convergent validity is the extent to which the construct converges to explain the variance of its indicators. The measure used to evaluate the convergent validity of a construct is the AVE for all indicators in each construct. AVE is defined as the overall average value of the squared loadings of the indicators associated with the construct (i.e., the sum of the squared loadings divided by the number of indicators). An AVE of less than 0.5 is considered inadequate, because more variance is due to error variance than indicator variance [44]. Therefore, AVE is equivalent to the communality of a construct. The minimum acceptable AVE is 0.50 where an AVE of 0.50 or higher indicates that the construct explains 50 percent or more of the variance of the indicators that make up the construct [44]. In traditional PLS-SEM, there are two latent variables that have an AVE value of less than 0.5 or invalid, namely living standard and social protection. While the education and health variables are valid. In addition, in the binary PLS-SEM, all latent variables are valid because they have an AVE value greater than 0.5.

The next measure is to assess discriminant validity. This measure indicates the extent to which a latent variable captures the variance of related indicators relative to indicators related to other latent variables in the measurement model. The higher the correlation between a latent variable and its indicators compared to its correlation with other indicators in the model, the clearer the latent variable is. Measures to measure discriminant validity include the heterotrait-monotrait ratio correlation (HTMT) from [46]. HTMT is the ratio of the correlation between traits to the correlation within traits. The HTMT is the average of all indicator correlations across constructs measuring different constructs (i.e., heterotrait-heteromethods correlations) relative to the (geometric) average of the average of indicator correlations measuring the same construct (i.e., monotrait-heteromethods correlations [44].

Table 4. HTMT value by relationships

Relationships	Traditional PLS	Binary PLS
Education – health	0.89	0.54
Education – living standard	0.65	0.49
Education – social protection	0.68	0.50
Health – living standard	0.52	0.31
Health – social protection	0.33	0.21
Living standar – social protection	0.34	0.35

The heterotrait-monotrait ratio (HTMT) of correlations is the average of the heterotrait-heteromethod correlations (i.e., indicator correlations across constructs measuring different phenomena), relative to the average of the monotrait-heteromethod correlations (i.e., indicator correlations within the same construct). [46] proposed a cutoff value of 0.90 for structural models with conceptually very similar constructs. An HTMT value above 0.90 indicates no discriminant validity. However, when the constructs are conceptually more dissimilar, a lower, more conservative cutoff value such as 0.85 is suggested [46] [47]. From Table 4 it can be seen that in both traditional PLS-SEM and binary PLS-SEM the HTMT value for each relationship is less than the recommended 0.85 except for the education-health relationship in traditional PLS-SEM which is 0.89. It can be concluded that binary PLS-SEM is better when viewed from the HTMT value.

3.2. Evaluation of Structural Model

Assessment of model structural can be seen from the significance of the path coefficient and the relevance of the path coefficient are evaluated. The path coefficient is significant at the 5% level if the zero value is not included in the 95% confidence interval. In general, the percentile method should be used to construct the confidence interval [48]. The path coefficient is usually between −1 and +1, with coefficients approaching −1 indicating a strong negative relationship and those approaching +1 indicating a strong positive relationship. Table 5 shows the results of the path coefficients for each dimension of multidimensional poverty. It appears that for each dimension, binary PLS-SEM has a larger path coefficient than traditional PLS-SEM. In other words, the dimensions of multidimensional poverty with the binary PLS-SEM approach show a stronger relationship than traditional PLS-SEM. Meanwhile, if we look at the magnitude, it can be seen that both binary and traditional PLS-SEM have the same order with the Education dimension being the strongest in relation to social protection, followed by the dimensions of standard of living and health.

In the binary PLS-SEM approach, the Education dimension has a path coefficient value of 0.315. This shows that when Education increases by one standard deviation unit, social protection will increase by 0.315. Meanwhile, the dimensions of standard of living and health each have path coefficients of 0.289 and -0.285. Of course, attention is paid to the health dimension because it has a negative path coefficient sign. The reason that can be given is whether there is indeed no longer any deprivation in the health dimension or whether there is an error in the data.

Table 5. Estimates of the parameter

Dimension	Traditional PLS				Binary PLS			
	Coefficient	Standard deviation	Perc. 2.5%	Perc. 97.5%	Coefficient	Standard deviation	Perc. 2.5%	Perc. 97.5%
Education	0.185	0.003	0.179	0.191	0.315	0.020	0.315	0.315
Health	-0.140	0.003	-0.146	-0.135	-0.285	0.018	-0.285	-0.285
Living Standard	0.175	0.003	0.169	0.182	0.289	0.024	0.289	0.289

Table 6. Explanatory power of the model and model fit

Dimension	Traditional PLS	Binary PLS
R-square (R^2)	0.08	0.08
Standardized root mean squares residual (SRMR)	0.07	0.07

The next step in evaluation of structural model involves examining the coefficient of determination (R^2) of the endogenous constructs. R^2 represents the variance explained in each endogenous construct and is a measure of the explanatory power of the model [49], also referred to as in-sample predictive power [50]. R^2 ranges from 0 to 1, with higher values indicating greater explanatory power. As a general rule of thumb, R^2 values of 0.75, 0.50, and 0.25 can be considered substantial, moderate, and weak, respectively, in many social science disciplines [51]. However, acceptable R^2 values are based on the research context, and in some disciplines, R^2 values as low as 0.10 are considered satisfactory [52]. Table 6 shows that the R^2 values in traditional and binary PLS-SEM are not much different but the path coefficient values of binary PLS-SEM are bigger than those of traditional PLS-SEM. Meanwhile, because the SRMR value is belom the recommendation threshold which is 0.08 than that model from the two approaches are fit.

4. Conclusions

In general, from the indicator reliability measures and internal consistency reliability used, namely loading, Cronbach's alpha, composite reliability, and rho-A, the binary PLS-SEM approach shows better performance than traditional PLS-SEM. Likewise, for the validity measures measured using AVE and HTMT, the binary PLS-SEM approach shows better performance. From these results, it can be concluded that if the indicator is a category, then the recommended approach to use is binary PLS-SEM because it produces better measurement model performance.

In the assessment of the structural model, the binary PLS-SEM path coefficient shows a greater value than traditional PLS-SEM. This means that binary PLS-SEM has better performance. Judging from the dimensions of multidimensional poverty, the Education dimension has the largest role in social protection. Followed by the dimensions of standard of living and health. Meanwhile, from the R2 value and SRMR value, the two approaches produce performance that is not much different.

From the applied side, if there are no obstacles in providing social protection, then the three dimensions of multidimensional poverty can be used as a basis for policy. Meanwhile, if there are obstacles such as funds, it is suggested that the education dimension be made the main priority in terms of social protection.

In terms of statistical methods, there are several limitations in this paper. One of them is that the indicators used are only in binary scale. If the indicator is continuous, then the population correlation matrix used is Pearson. Meanwhile, if the indicator is ordinal, then the one used is polychoric correlation.

The next research that can be proposed is to consider using a combination of categorical and continuous indicators as input data. In addition, it can also consider using a high-order model for its dimensions.

Ethics approval

The Ethical approval statement should be provided including the consent. If not appropriate, authors should state: “Not required.”

Acknowledgments

This section contains a form of thanks to individuals or institutions who have provided assistance in carrying out research, preparing the article, providing language help, writing assistance or proof reading the article and others.

Competing interests

A competing interest statement should be provided, even if the authors have no competing interests to declare. If no conflict exists, authors should state: “All the authors declare that there are no conflicts of interest.”

Funding

List funding sources in this standard way to facilitate compliance to funder's requirements. It is not necessary to include detailed descriptions on the program or type of grants and awards. When funding is from a block grant or other resources available to a university, college, or other research institution, submit the name of the institute or organization that provided the funding. If no funding has been provided for the research, please include the following sentence: “This study received no external funding.”

Underlying data

This can be written as: “Derived data supporting the findings of this study are available from the corresponding author on request.”

Credit Authorship

Diana Bhakti: Conceptualization, Data Collection, Formal Analysis, Writing - Original Draft, Visualization. **Ardi Adji:** Methodology, Writing - Review & Editing, Supervision. **Endang Saefuddin Mubarok:** Writing - Review & Editing, Supervision. **Renny Sukmono:** Writing - Review & Editing, Supervision. **Rudi Salam:** Writing - Review & Editing, Supervision

References

[1] M. Sarstedt, J. F. Hair, M. Pick, B. D. Liengaard, L. Radomir, C. M. Ringle. “Progress in partial least squares structural equation modeling use in marketing research in the last decade,” *Psychology and Marketing*, vol. 39, no. 5, pp. 871-873, 2022.

[2] J. Wang & X. Wang, *Struktural Equation Modeling: Applications using MPlus 2nd Edition*. Chichester: John Wiley & Sons Ltd, 2020.

[3] K. G. Jöreskog, U. H. Olsson, & F. Y. Wallentin, *Multivariate Analysis with LISREL*. Springer, 2016.

[4] K. A. Bollen, A. Diamantopoulos. “In defense of causal-formative indicators: A minority report,” *Psychological Methods*, vol. 22, no. 3, pp. 581–596, 2015.

[5] C. Crocetta, L. Antonucci, R. Cataldo, R. Galasso, M. G. Grassia, C. N. Lauro, & M. Marino. “Higher-Order PLS-PM Approach for Different Types of Constructs,” *Social Indicators Research*, vol. 154, no. 2, pp. 725–754, 2021.

[6] F. Schuberth, J. Henseler, & T. K. Dijkstra. “Partial least squares path modeling using ordinal categorical indikators,” *Quality & Quantity: International Journal of Methodology*, vol. 52, no. 1, pp. 9–35, 2018.

[7] V. Savalei. “Improving Fit Indices in Structural Equation Modeling with Categorical Data,” *Multivariate Behavioral Research*, vol. 56, no. 3, pp. 390–407, 2020.

[8] G. Cantaluppi. A partial least squares algorithm handling ordinal variables also in presence of a small number of categories. arXiv: *Methodology*. preprint arXiv:12125049. 2012.

[9] G. Cantaluppi and G. Boari, “A partial least squares algorithm handling ordinal variables,” in *The Multiple Facets of Partial Least Squares and Related Methods: PLS, Paris*, H. Abdi, V. E. Vinzi, G. Russolillo, G. Saporta, and L. Trinchera, Eds. Cham: Springer International Publishing, 2016, pp. 295–306.

[10] F. Schuberth and G. Cantaluppi, “Ordinal consistent partial least squares,” in *Partial Least Squares Path Modeling*, H. Latan and R. Noonan, Eds. Cham: Springer International Publishing, 2017, pp. 109–150.

[11] M. Sarstedt, J. F. Hair, J. Cheah, J-M. Becker, C. M. Ringle. “How to Specify, Estimate, and Validate Higher-Order Constructs in PLS-SEM,” *Australasian Marketing Journal*, vol. 27, no. 3, pp. 197-211, 2019.

[12] R. Salam, I. M. Sumertajaya, H. Wijayanto, A. Kurnia, T. Sirait. “Higher Order Partial Least Squares Path Modeling Using Binary Data: An Application on Multidimensional Poverty and Social Protection in East Java Province,” *Asian Journal of Mathematics and Computer Research*, vol. 30, no. 4, pp. 118–137, 2023.

[13] J. P. Miller, *Essential Statistical Methods for Medical Statistics: A derivative of Handbook of Statistics: Epidemiology and Medical Statistics*, vol. 27. Amsterdam: North Holland, 2011.

[14] S. Alkire, J. E. Foster, S. Seth, M. E. Santos, J. M. Roche, P. Ballon, *Multidimensional Poverty Measurement and Analysis*. Oxford: Oxford University Press, 2015.

[15] W. Hanandita, G. Tampubolon, “Multidimensional Poverty in Indonesia: Trend Over the Last Decade (2003–2013),” *Social Indicators Research*, vol.128, pp.559–587, 2016.

[16] UNDP, *Why is the MPI better than the Human Poverty Index (HPI) which was previously used in the Human Development Reports? Frequently Asked Questions*, 2015. United Nations Development Programme (online). http://hdr.undp.org/en/faq-page/multidimensional-povertyindex-mpi#t295n138. [Accessed: February 27, 2014].

[17] A. Khaliq, B. Uspri, “Kemiskinan Multidimensi dan Perlindungan Sosial [*Multidimensional Poverty and Social Protection*],“ *Jurnal Manajemen*, vol. 13, no. 2, pp. 85-191, 2017.

[18] Prakarsa, *Indeks Kemiskinan Multidimensi Indonesia 2015-2018 [Indonesia Multidimensional Poverty Index 2015-2018]*, Laporan, Perkumpulan Prakarsa, Jakarta, 2020.

[19] M. E. Rademaker, *Composite-based Structural Equation Modeling. Ph.D. Dissertation*, University of Würzburg, Faculty of Economics, 2020.

[20] E. E. Rigdon, “Partial least squares path modeling,” in *Structural Equation Modeling – A Second Course*, 2nd ed., G. R. Hancock and R. O. Mueller, Eds. Information Age Publishing, Inc., 2013, ch. 3.

[21] M. Mehmetoglu, S. Venturini, *Structural Equation Modelling with Partial Least Squares Using Stata and R*, 1st ed. Chapman and Hall/CRC, 2021.

[22] Joreskog, K. G, *Structural Equation Modeling with Ordinal Variables using LISREL*. Scientic Software International Inc., 2005. http://www.ssicentral.com/lisrel/techdocs/ordinal.pdf.

[23] K. A. Bollen, A. Maydeu-Olivares, “A polychoric instrumental variable (piv) estimator for structural equaltion models with categorical variables,” *Psychometrika*, vol. 72, pp. 309-326, 2007.

[24] F. Drasgow, “Polychoric and polyserial correlations,” in *Encyclopedia of Statistical Sciences*, L. Kotz and N. Johnson, Eds. New York: Wiley, 2014, pp. 68–74.

[25] J. Brzezińska, “Item response theory models in the measurement theory,” *Communications in Statistics - Simulation and Computation*, vol. 49, no. 12, pp. 3299–3313, 2018.

[26] E. F. El-Hashash, K. M. El-Absy, “Methods for Determining the Tetrachoric Correlation Coefficient for Binary Variables,”, *Asian J. Prob. Stat.*, vol. 2, no. 3, pp. 1–12, Dec. 2018.

[27] J. Suárez-Cansino, V. López-Morales, L. R. Morales- Manilla, A. Alberto-Rodríguez, J. C. Ramos-Fernández, "Prior Distribution and Entropy in CAT Ability Estimation through MAP or EAP," *Entropy*, vol. 25, no. 1-50, pp. 1-21, 2023.

[28] F. B. Baker, S. Kim, *Item response theory: parameter estimation techniques*. New York: CRC Press, 2004.

[29] C. M. Bishop, *Pattern Recognition and Machine Learning*. Cambridge: Springer, 2006.

[30] J. Fox, *Bayesian item response modeling: theory and applications*. Springer, 2010.

[31] R. D. Bock, R. D. Gibbons, *Item Response Theory*. University of Chicago: John Wiley & Sons, 2021.

[32] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage, 1991.

[33] R. J. de Ayala, *The Theory and Practice of Item Response Theory*. New York: The Guilford Press, 2022.

[34] L. Steinberg and D. Thissen, "Item response theory," in *The Oxford Handbook of Research Strategies for Clinical Psychology*, J. S. Comer and P. C. Kendall, Eds. Madison Avenue, NY: Oxford University Press, 2013, pp. 336–373.

[35] S. Alkire, J. Foster, "Counting and multidimensional poverty measurement," *Journal of Public Economics*, vol. 95, no. 7-8, pp. 476-487, 2011.

[36] [WB] World Bank, *Social protection sector strategy : from safety net to springboard (English)*. Washington, D.C. : World Bank Group, 2001.

[37] A. Barrientos, D. Hulme, "Chronic Poverty and Social Protection: Introduction," *The European Journal of Development Research*, vol. 17, no. 1, pp. 1–7, 2005.

[38] A. Haan, "Social Exclusion: Enriching the Understanding of Deprivation," *Studies in Social and Political Thought*, vol. 2, 2000.

[39] L. Scott, *Social Protection: Improving its Contribution to Preventing Households Falling into Poverty*, 2016.

[40] G. Sanjivi, "Social security options for developing countries," *International labour review*, vol. 133, no. 1, pp. 35-53, 1994.

[41] R. Sabates-Wheeler, S. Devereux, "Social Protection for Transformation," *IDS Bulletin*, vol. 38, no. 3, pp. 23–28, 2007.

[42] International Labour Organization, *Penilaian Landasan Perlindungan Sosial berdasarkan Dialog Nasional di Indonesia: Menuju Landasan Perlindungan Sosial Indonesia* [Assessment of the Social Protection Baseline based on the National Dialogue in Indonesia: Towards the Indonesian Social Protection Baseline], Laporan: Jakarta, 2012.

[43] J. F. Hair Jr., G. T. M. Hult, C. M. Ringle, M. Sarstedt, N. P. Danks, S. Ray, *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*. Springer Cham, 2021.

[44] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed. Thousand Oaks, CA: Sage, 2017.

[45] J. Benitez, J. Henseler, A. Castillo, F. Schuberth, "How to perform and report an impactful analysis using partial least squares: Guidelines for confirmatory and explanatory IS research," *Information and Management*, vol. 57, no. 2, pp. 103168, March 2020.

[46] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015.

[47] G. Franke and M. Sarstedt, "Heuristics versus statistics in discriminant validity testing: A comparison of four procedures," *Internet Research*, vol. 29, no. 3, pp. 430–447, 2019.

[48] M. I. Aguirre-Urreta, M. Rönkkö, "Statistical inference with PLSc using bootstrap confidence intervals," *MIS Quarterly*, vol. 42, no. 3, pp. 1001–1020, 2018.

[49] G. Shmueli, O. R. Koppius, "Predictive analytics in information systems research," *MIS Quarterly*, vol. 35, no. 3, pp. 553–572, 2011.

[50] E. E. Rigdon, "Rethinking Partial Least Squares Path Modeling: In Praise of Simple Methods," *Long Range Planning*, vol. 45, pp. 341-358, 2012.

[51] J. F. Hair, C. M. Ringle, M. Sarstedt, "PLS-SEM: Indeed a silver bullet," *Journal of Marketing Theory and Practice*, vol. 19, no. 2, pp. 139–151, 2011.

[52] S. Raithel, M. Sarstedt, S. Scharf, M. Schwaiger, "On the value relevance of customer satisfaction. Multiple drivers and multiple markets," *Journal of the Academy of Marketing Science*, vol. 40, no. 4, pp. 509–525, 2012.

DOI: https://doi.org/10.34123/jurnalasks.v17i1.797

POLITEKNIK STATISTIKA STIS

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

Jurnal Aplikasi Statistika &
Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

The Implementation of Geospatial Analysis on Hotel Occupancy Rate

Muhammad Fachry Nazuli[1], Satria Bagus Panuntun[2], Addin Maulana[3], Takdir[4], Setia Pramana[5]

[1]BPS-Statistics Lhokseumawe City, Lhokseumawe, Indonesia, [2]Directorate of Statistical Analysis and Development, BPS-Statistics Indonesia, Jakarta, Indonesia, [3]Center for Industrial, Services and Trade Research, National Research and Innovation Agency, Jakarta, Indonesia, [4]University of Tsukuba, Japan, [5]Politeknik Statistika STIS, Jakarta, Indonesia
**Corresponding Author: E-mail address: [1]fachry.nazuli@bps.go.id, [2]satria.bagus@bps.go.id, [3]addi001@brin.go.id, [4]takdir-kde@kde.cs.tsukuba.ac.jp, [5]setia.pramana@stis.ac.id*

ARTICLE INFO

Article history:
Received 2 September, 2024
Revised 5 December, 2024
Accepted 10 December, 2024
Published 24 February, 2025

Keywords:
Big Data; Google Maps; Locations; Spatial Regression; Web-Scraping.

Abstract

Introduction/Main Objectives: One of the main attributes of hotel selection and customer satisfaction is its location. **Background Problems:** Strategic location leads to higher demand for accommodation. Accommodation demand is reflected in hotel occupancy levels, which indicate the percentage of reserved rooms at a specific period. **Novelty:** This study aims to investigate the effect of spatial location on hotel occupancy rates by analyzing data collected in online hotel reservation applications. A study related to the effects of location and hotel occupancy has never been conducted in Indonesia. **Research Methods:** We use data from hotels located in the province of Yogyakarta, which contains 245 hotels spread over three regencies/cities, namely Yogyakarta City, Sleman Regency, and Bantul Regency. We conducted a spatial regression analysis, namely the Spatial Error Model (SEM), with a spatial weight matrix using a radius of 3.2 km. **Finding/Results:** We found that spatial locations affect the occupancy rates of hotels based on the online hotel reservation application that we observed. These spatial locations include the distance from the hotel to the airport, the distance from the hotel to the bus stop, and the number of nearby restaurants, offices, and hotels.

1. Introduction

The geographic position is an important aspect of the economic performance of a hotel. Location significantly affects the probability of a Hotel's survival [1]. As much as 20% of hotel bankruptcy rates are caused by location [2]. Hotels make location their marketing base. For example, many luxury hotels utilize their coastal area as a marketing attraction [3].

Tourists, whether traveling for business or leisure, regard the hotel’s location as a critical factor influencing hotel selection [4]. The geographical location of a hotel can be strongly related to higher room occupancy rates, revenue per available room, and profit from the hotel [5-6]. The hotel must carefully choose its location to avoid the difficulties of relocating.

Customer satisfaction is also closely related to the location of the hotel. The ideal location leads to greater demand for accommodation, better company performance, and higher customer satisfaction [7]. A convenient location is considered one of the main factors influencing hotel selection and the satisfaction of business and leisure tourists. They prefer a location close to the services and facilities available[8]. Satisfaction will be derived not solely from the internal amenities of a hotel but also from the surrounding facilities. Location and proximity to city facilities, such as the central business district, transportation network, sea, lakeside, or tourist attractions, determine the market value of a hotel [9]. Customers usually prefer hotels closer to these facilities, which can increase the hotel's market value. Proximity to the city center can drive increased guest satisfaction and hotel demand [7].

In general, three factors influence hotel selection, which also affects hotel occupancy rates, namely accessibility to the point of interest (POI), transportation convenience, and the surrounding environment [10]. Accessibility to points of interest (POI) includes transportation portals (airports, stations, bus stops, etc.), central business districts, tourist attractions, entertainment venues, and so on. The convenience of transportation is related to the convenience of hotel guests to depart and return to the hotel using public transportation. Environmental factors are closely related to air quality, public safety and security, public infrastructure (such as restaurants, parks, etc.), and culture.

Yogyakarta is a province that has a very strong tourist attraction. As a province with fairly high tourist attractions, it had a total of 108,599 foreign tourists in 2019 and 195,778 in 2020 [11]. This significant decrease is due to the COVID-19 pandemic, which has caused people to reduce their level of mobility. The research conducted by Pramana [12] found an impact of the pandemic on the mobility index, number of flights, occupancy rates, and other big data [12]. The COVID-19 pandemic has significantly affected the accommodation, retail, transportation, and manufacturing sectors, contributing to the current account deficit (CAD) [13]. According to the KBLI (Indonesian Standard Industrial Classification) 2020, the hotel industry categorized as *category I* a sector that provides accommodation, food, and drink [14].

With today's technological advances, hotel reservations are available online. Several applications provide services for online hotel reservations. Online hotel reservations are considered more efficient than making them in person. Customers can make reservations before their departure day. The application is connected to the hotel upon a reservation. Thus, the room occupancy rate can be directly calculated based on the number of rooms available and the number of rooms that have been reserved on a certain date. The data velocity and variation of hotel reservation applications are high, which can be qualified as big data.

Previous studies have known that location influences or impacts a hotel's economic aspects. There are several critical aspects of the hotel economy, including the price per room [9], monthly income [15], profitability [16], hotel distributions [17-18], customer satisfaction [10], and others. However, no study related to the impact of location on occupancy rates has been conducted to the best of our knowledge.

This study aims to analyze the impact of location on the hotel/room occupancy rate using a big data approach. We use data from an online hotel reservation application as the main data source. We obtain locations for urban objects, such as restaurants, train stations, and airports from Google Maps.

2. Material and Methods

Before we look at how location influences hotel occupancy rates in the city of Yogyakarta, we can look at previous research related to hotels and location. Srimulyani [19] analyzed the hotel occupancy rate in West Nusa Tenggara Province. The purpose was to see the impact of COVID-19 on hotel occupancy rates in West Nusa Tenggara using a Big Data approach. Data was collected using the web-scraping method from online hotel reservation websites. They conducted a descriptive analysis of the collected data. They found that the early emergence of COVID-19 reduced the occupancy rate sharply, especially in high-star hotels, and the New Normal policy gradually increased the hotel occupancy rate.

Valentin and O'Neill [9] conducted a study that aims to investigate the significance of the hotel's location to the property market value of the hotel. The market value of a property is represented by the price per room. This research was conducted in Chicago using more than 600 hotels as the unit of observation. They use Ordinary Least Square Regression. The results show that proximity to the city center, namely the Loop, is the most influential factor in market value. One additional mile of the Loop will decrease the property value by 13% on average, for distances under 10 miles from the Loop and be relatively constant at distances above 10 miles from the Loop.

Yang [20] conducted a study that aims to look at the factors that contribute to the potential location of a hotel with a model that combines the characteristics of the hotel and its location. The potential hotel locations are categorized from 1 to 5 based on the ring road in Beijing City. The method used is the Ordered Logit Model. The results of this study indicate that the star level of the hotel, years of construction, service diversification, ownership, agglomeration effects, public service infrastructure, and accessibility (roads, subways, tourist sites) are important determinants in the selection of hotel locations.

Yang [10] conducted a study to determine the determinants of hotel guest satisfaction related to location based on reviews on the TripAdvisor website. The Ordinary Least Square Regression and the Ordered Logit Model were applied. They conclude that the accessibility of hotel properties to the location of tourist attractions, airports, universities, public transportation, green space, water coverage, and surrounding local businesses is vital to hotel reviews. Meanwhile, the variables of proximity to toll roads and crime rates have no significant effect on the model.

Fang [17] researched hotel location choices by developing a GWPR model in the Hong Kong region. The dependent variable used was the number of hotels located in certain sub-districts. They identified nine factors affecting the hotel's location. These include land area, green land, traffic land, commercial land, institutional land, station density, density of tourist attractions, population density, and the average income of residents in the area. The effect of each variable is different for each region. For example, transportation accessibility variables are not significant near the city center.

2.1. Data and variable

This study uses data from online hotel reservation applications. The observed hotels are located in Yogyakarta province. We analyze data from June 2019.

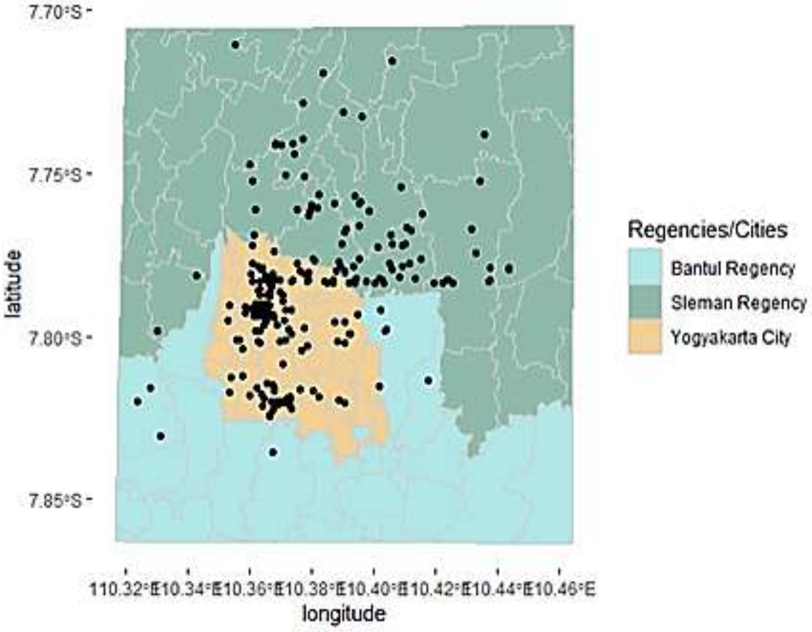


Figure 1. The distribution of hotels in the Special Region of Yogyakarta

In Figure 1, 245 hotels are detected from the application data spread over three regencies/cities in the Province of Yogyakarta, namely Yogyakarta City, Sleman Regency, and Bantul Regency. The dataset contains variables of the number of available, reserved, and occupied rooms. Thus, we can calculate the occupancy rate by using the following formula [15]:

$$occupancy_t = \frac{\Sigma(Total\ rooms - Available\ Rooms)}{\Sigma(Total\ rooms)} x100 \qquad (1)$$

The occupancy rate is the dependent variable in this study There are several variables that are considered to affect the hotel occupancy rate, including:

- hotel count: Number of hotels within a 1 km radius using Euclidean distance. It explains market competition or the agglomeration effect between hotels [17] due to the tendency of hotels to gather in one place.

- airport distance: This variable describes accessibility from the hotel to the airport or vice versa. The market value of a hotel will tend to increase along with its proximity to the airport, which impacts its occupancy rate [9].
- station distance: This variable describes accessibility from the hotel to the train station or vice versa. Access to train stations in urban areas determined hotel customer satisfaction [7].
- bus stop distance: This variable explains accessibility from the hotel to the nearest bus stop or subway [20].
- mall distance: This variable describes accessibility from the hotel to the nearest shopping center [15].
- university distance: This variable explains accessibility from the hotel to the nearest university [10].
- restaurant count: The sum of restaurants within a 1-kilometer radius using the Euclidean distance, excluding the in-house restaurant [10].
- attractions count: The number of attractions within a 1-kilometer radius of the hotel using Euclidean distance. Seven tourist attraction types include museums, theaters, amusement parks, state gardens, stadiums and arenas, beaches, historical sights, and shopping [10].
- office count: Number of offices located within 1 kilometer of the hotel using Euclidean distance. The number of offices has an impact on the type of business guests. The hotel will choose a location close to its potential market, such as shopping and business centers [17].

Of all the variables mentioned above, there are 2 data sources, including the online hotel reservation application to estimate room occupancy rates and Google Maps.

2.2. Analysis Method

a. Spatial Autocorrelation Analysis

Moran's test was applied to see the presence of spatial autocorrelation in hotel data [21]. The initial hypothesis for Moran's test shows the random distribution of the variables in certain areas. The alternative hypothesis is that the variable spreads with no random distribution and has a spatial autocorrelation. Morans' statistical value will be in the range of -1 and +1.

b. Spatial Analysis

Spatial analysis is applied if the Morans test rejects the initial hypothesis or the variables have spatial autocorrelation. The spatial analysis used is the Spatial Autoregressive (SAR) and Spatial Error Model (SEM) analysis. The results were then compared with the classical regression, namely the ordinary least square.

Before doing spatial modeling, it is necessary to test the Lagrange Multiplier. This test focuses on identifying the presence of spatial lag and spatial error in the model. If spatial lag is identified in the model, the Spatial Lag Model or Spatial Autoregressive Model (SAR) is used. Meanwhile, if the identified spatial error is in the model, the Spatial Error Model (SEM) is used. The following is a SAR model that explains the spatial lag (Wy) relationship [22]:

$$\boldsymbol{y} = \rho\boldsymbol{W}\boldsymbol{y} + \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n)$$

Where:

$\boldsymbol{y}$: response variable vector
$\boldsymbol{x}$: predictor variable matrix (n x p)
$\boldsymbol{\rho}$: coefficient of spatial effect parameter of the predictor variable
$\boldsymbol{W}$: spatial weighted matrix
$\boldsymbol{\beta}$: coefficients of predictor variable parameters vector
$\boldsymbol{\epsilon}$: error

Meanwhile, the spatial error model can be explained in the following equation [22]:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{3}$$

$$\boldsymbol{\epsilon} = \lambda \boldsymbol{W\epsilon} + \boldsymbol{u} \tag{4}$$

$$\boldsymbol{u} \sim N(0, \sigma^2 I_n)$$

Where:

$\boldsymbol{y}$	: response variable vector
$\boldsymbol{x}$	: predictor variable matrix (n x p)
$\boldsymbol{W}$	: spatial weighted matrix
$\boldsymbol{\beta}$	: coefficients of predictor variable parameters vector
$\boldsymbol{u}$	: error vector that has autocorrelation
$\boldsymbol{\lambda}$	: spatial effect parameter coefficient error
$\boldsymbol{\epsilon}$	: error

The variable used is the occupancy rate as the dependent variable. As for the independent variables, a number of nine variables were used, as mentioned in the previous section. We performed a significant test on all independent variables.

3. Results and Discussion

In the process of achieving the research objectives, namely understanding the impact of location on hotel occupancy rates, we have undertaken several tasks, including literature review, development of a web-scrapper used to collect data, data preprocessing, the data analysis using descriptive and inferential analysis, model construction using spatial regression, and reviews the previous researches. The resulting spatial regression model shows that there are several location factors that have a significant effect on hotel occupancy rates. This research is useful for determining the location for the appropriate hotel development. The location factor is expected to provide an overview of the hotel’s strategic location to increase the hotel’s occupancy rate.

The use of big data in this study has several advantages, including the fact that data collection is easier to do than conventional methods. In addition, data collection through the scraping process does not incur a high cost. The data collection process is faster. However, using big data presents challenges, namely the need for better computing devices. Additionally, big data exhibits a significant data variance, requiring validation.

This study uses hotel data collected using a web-scraper that has been developed by Adhinugroho [23]. In this study, it was found that the occupancy rate data in the online hotel reservation application has the same pattern as the occupancy rate data issued by the Indonesian statistical BPS. This proves that this big data source can represent the official data effectively with additional location data.

In this study, spatial regression was used to assess the impact of location on hotel occupancy rates. We began with simple linear regression using ordinary least squares before moving to spatial regression (ols). In the modeling phase, it was found that the model did not meet the classical assumptions. Classical assumptions that are not met are normality, autocorrelation, and heteroscedasticity. Therefore, the ols model is not adequate for use with this data. Hence, we use a spatial regression model as an alternative considering that the data is spatial data. The complete results of this study are shown in the explanation below.

3.1. *Data collection*

We gathered data from various sources. The main data in this study is hotel reservation data on online hotel reservation applications. The hotel reservation application used is one of the largest hotel reservation applications in Indonesia, namely Agoda.com. Agoda data is hospitality data obtained using the web-scraping method. The scraper used was based on a previous study conducted by Adhinugroho in 2020 [23].

In addition to utilizing data from hotel reservation applications, this study utilizes urban object data acquired from Google Maps. A web-scraping process is carried out using Python programming to get the location of an object on Google Maps. By using the webdriver-based Selenium library, a robot is formed to retrieve urban objects with certain keywords.

The first step is to collect a list of the names of the villages in the Province of Yogyakarta. However, because the hotel distribution is not in the entire province of Yogyakarta, only all urban villages in Yogyakarta City and some villages in Sleman and Bantul districts are listed. Furthermore, the office of each sub-district was taken as a benchmark point for collecting object location data.

The next step is to do a keyword search for urban objects. There are several keywords used, namely “airport”, “bus stop”, “station”, “restaurant”, “office”, “company”, “mall”, “attractions”, and “university”. For some keywords such as “airport”, “bus stop”, “station”, etc., it only needs to be done at one benchmark point because it has issued object results in all provinces. Meanwhile, keywords such as “restaurant”, “office”, “company”, and “attractions” need to be repeated as many as the number of benchmark points. The results obtained in this step are a list of names of urban objects on Google Maps according to the number of benchmark points. Each page contains a maximum of 20 object names.

Then from the list of objects, we automate click events on the object names one by one. Then the HTML tag containing the name, address, category, number of reviews, rating, and page URL is taken. From the page URL, the longitude and latitude of the object are extracted using a string separation.

The disadvantage of using this method is the search results cannot be limited by a specific region. For example, if we want to retrieve restaurant data in the Gedongkiwo Sub-district, Google Maps cannot limit restaurants in that sub-district but it provides a list of all restaurants close to the benchmark point. Thus, it is necessary to carry out manual selections after retrieving all the data in the village. This process is necessary because the use of the Selenium library takes a long time for a single execution. An overview of Google Maps scraping can be seen in Figure 2.

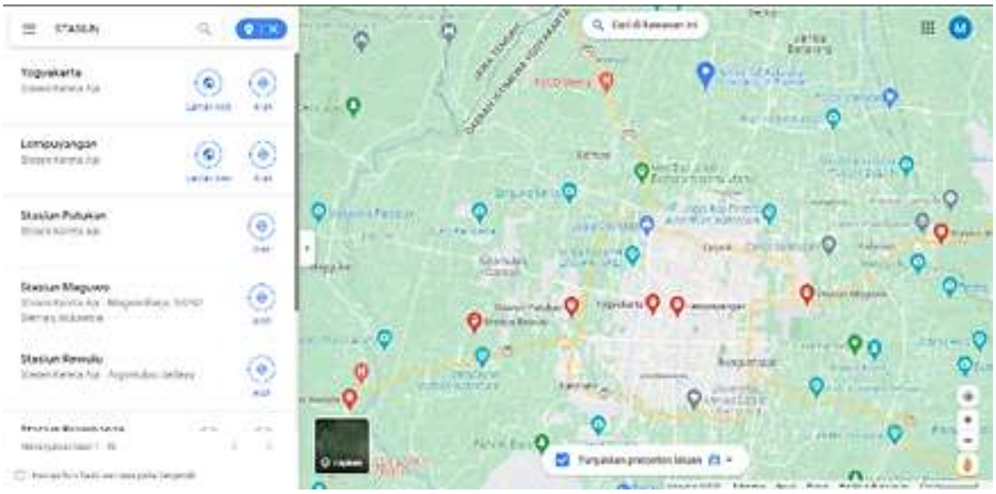


Figure 2. Google Maps scraping process

After the scraping process on Google Maps was carried out, urban objects were obtained from several categories, namely "airport", "bus stop", "station", "restaurant", “attractions”, etc. The number of objects obtained is different for each category. They can be seen in Table I below:

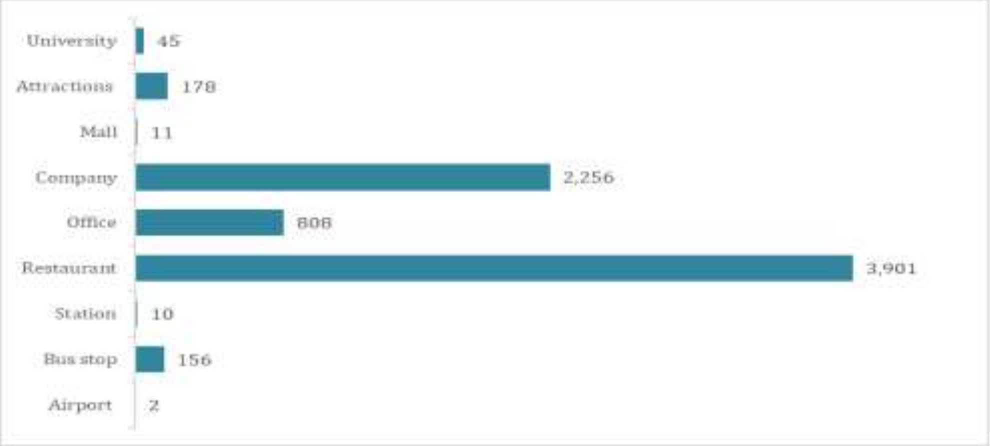


Figure 3. Number of Google Maps scraped objects

In Figure 3, the results from Google Maps data collection are mostly in the restaurant category, totaling 3,901 restaurants, followed by companies and offices, which, when combined, reach 3,064 objects. The lowest are airports and stations, which are 2 and 10, respectively. Figure 4 shows the distribution of each urban object from the scraping results of Google Maps.

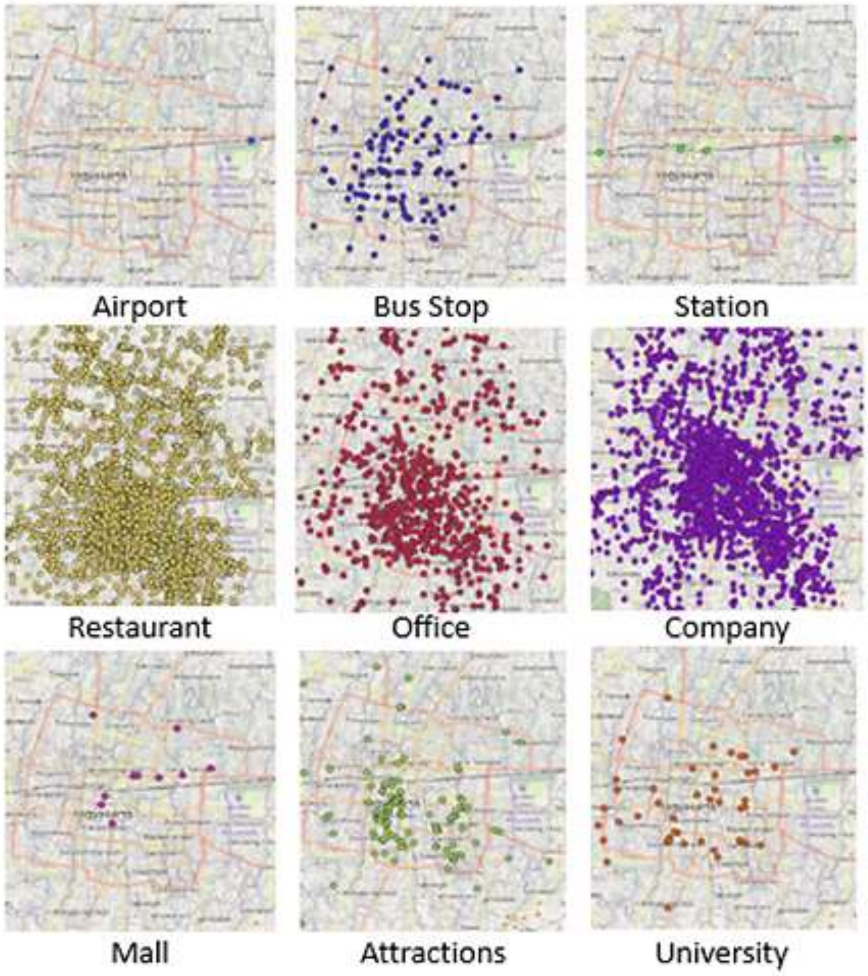


Figure 4. Distribution of Google Maps web-scraping data

To evaluate the web-scraper that has been built, the time needed to run it will be measured. Table 1 shows the time it takes to run the web-scraper for each stage. The average time required for Stage 1 and Stage 2 is relatively the same. The longest stage is in stage three, which is getting details of urban objects. This is because, after retrieving detailed object data, the web-scraper needs to return to the previous page, which takes twice as long as the previous stage. After getting urban objects from Google Maps, each hotel will measure the distance to the nearest object. Each existing hotel is paired with the nearest airport, station, bus stop, mall, and university using the Euclidean distance or the length of a straight line from two points. To calculate the distance traveled by two objects, measurements are made using the API from OpenStreetMap.

Table 1. Running time web-scraper

Stage	Running time	Number of data	Average time
Stage 1	1,251.726 seconds	206	6.08 seconds
Stage 2	6,418.425 seconds	1034	6.21 seconds
Stage 3	11,321.421 seconds	1000 (simulation)	11.36 seconds

The OpenStreetMap API is used because it is open source and offers free API access. However, it has a limit of 2,000 requests per day for each API. For objects that are categorized as restaurants, tourism objects, and offices and companies, the number of objects around the hotel is calculated using a radius of 1 km. For the effect of agglomeration or market competition between hotels, the same calculation method is used.

3.2. *Data Preprocessing*

Data preprocessing is conducted to prepare the data before it is used in the analysis stage. There are several tasks at the data preprocessing stage, such as checking for missing values, removing duplicates, eliminating inappropriate, and joining data. Data preprocessing is carried out on both hotel and urban object data. In the hotel data, several variables have the potential to be used in this study, such as the number of restaurants, the number of bars, and the price per room. However, these variables cannot be used because there are many missing values. Consequently, these variables are removed from our analysis. In hotel data the hotel ID, latitude, longitude, date, number of rooms reserved, and total number of rooms are used.

We initially filtered the hotel data according to the research period, which is June 2019. Additionally, only active hotels are taken. An active hotel on a day is characterized by having at least one hotel room reserved on that day. The hotels that were active during the research period were those that were active for at least 20 days in June 2019. In the Google Maps data, we began the preprocessing stage by eliminating duplicate data. Then, the data is filtered based on the description: *open or close*. The next step is to manually filter by name and category of urban objects. For example, an object named "shelter airport" with the category "airport", which does not have a matching category, is deleted.

3.3. *Hotel Occupancy Rate Estimation*

Based on the literature study, hotel occupancy is strongly influenced by the location of the hotel. Factors that affect hotel occupancy are distance to the airport, station, bus stop, mall, and university, as well as the number of offices, restaurants, tourist attractions, and hotels within a 1 km radius. Before formulating the model using regression, it is necessary to look at the relationship between the dependent and independent variables using the Pearson correlation value. The following graph shows the correlation of each independent variable with the dependent variable of hotel occupancy.

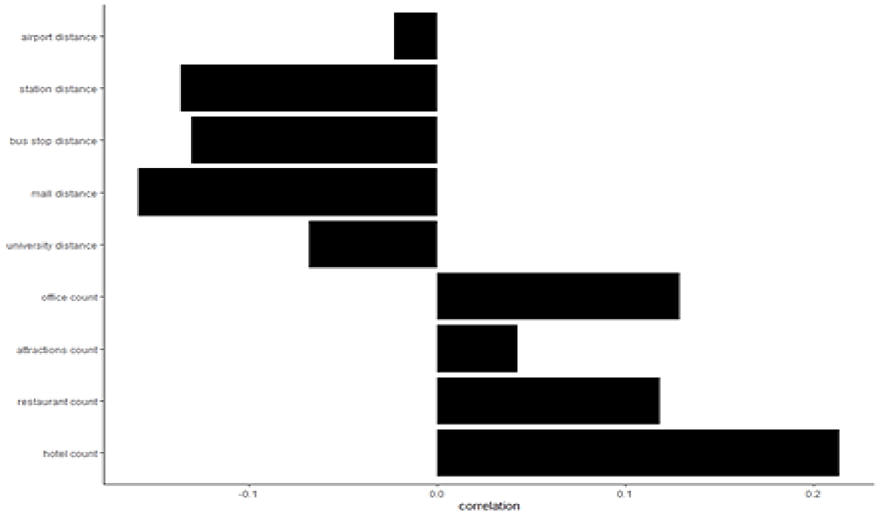


Figure 5. Variable correlation value

Figure 5 shows the pattern of relationships represented by the correlation value between the dependent variable and each independent variable. The variable "count" has a positive relationship with the occupancy rate. This indicates that the greater the number of urban objects within a 1 km radius of the hotel, the higher the hotel occupancy rate. While the variable "distance" has a negative relationship with the level of hotel occupancy. This indicates that the closer the hotel is to certain urban objects, the higher the occupancy rate of the hotel will be.

3.3.1. Spatial Autocorrelation Test

In regression modeling using spatial data, the most suitable autocorrelation test is the Global Moran's I test. This test is used to detect the spatial effect between neighboring observation units. The initial hypothesis or null hypothesis in this test is that the data is randomly distributed in a certain area, while the alternative hypothesis is that the data are not randomly distributed and have spatial autocorrelation. From the results of the Moran I test, the test statistic was 0.078014977 and the p-value was 0.008044. With a significance level of 5%, it can be concluded that the data are not randomly distributed and have spatial autocorrelation. Thus, spatial analysis can be an appropriate analytical step for the data in this study.

3.3.2. Spatial Regression

Before performing the spatial regression modeling, the Lagrange Multiplier was tested. The Lagrange Multiplier test aims to determine certain spatial effects on the model. The purpose of this test is to determine the spatial regression model that fits the data. If the Lagrange Multiplier test results in rejecting H0 in both tests, it can be concluded that spatial lag and spatial error exist in the model. The test will be continued with the Robust Lagrange Multiplier test. It is used because bias is generated by the Lagrange Multiplier test. The following table shows the results of testing the Lagrange Multiplier and Robust Lagrange Multiplier on spatial lag and spatial error.

Table 2. Result lagrange multiplier test

LM-test	statistic	p-value
Lagrange Multiplier (lag)	5.2009	0.02258**
Robust LM (lag)	2.5241	0.1121
Lagrange Multiplier (error)	9.4567	0.002104**
Robust LM (error)	6.78	0.009219**

Notes: **significant at 5%

Based on Table 2, it can be seen that the p-value in the spatial lag test has a value of 2.25%. Thus, at the 5% significance level, there is sufficient evidence of the occurrence of a spatial lag in the model. Meanwhile, when testing the spatial error in the model, the p-value is 0.21%. At the 5% significance level, there is sufficient evidence that there is an error link between regions or there is a spatial error. In the Lagrange multiplier test, significant results on spatial lag and spatial error are obtained. It will be continued with the Robust Lagrange Multiplier test. It can also be seen that the p-value of the spatial error is significant at a significance level of 5%, while the spatial lag is not significant. Therefore, the SEM model with the generalized methods of the moment estimation method is the right model to use. In this study, the autocorrelation test uses weights based on neighbors using the radius method. The distance or radius used is 3.2 km which is obtained from the maximum value of the distance between the hotel and the nearest other hotel. The following table shows the result of modeling using the SEM model with a spatial weighting matrix of the neighboring radius of 3.2 km.

Table 3 shows that there are five significant variables out of 9 independent variables at a significance level of 10%. At the 5% significance level, there are four significant variables. The significant variables are the distance to the airport, the distance to the bus stop, the number of offices, the number of restaurants, and the number of hotels in the vicinity. Therefore, the SEM model obtained is:

$$\begin{aligned}\widehat{Occupancy_i} = {} & 0.80393 + 0.005258 airport\ distance_i^{*} + 0.008216 station\ distance_i \\ & - 0.03822 bus\ stop\ distance_i^{**} - 0.00701 mall\ distance_i \\ & + 0.004436 university\ distance_i - 0.00257 \mathrm{restaurant\ count}_{\mathrm{i}}^{**} + 0.000476 offi \\ & + 0.000075 attractions\ count_i + 0.006901 hotel\ count_i^{**} + u_i^{*}\end{aligned}$$

$$\boldsymbol{u_i^{*} = -1.9734 \Sigma_{j=1, i \neq j}^{n} w_{ij} u_j} \tag{5}$$

Notes: * significant at 10%; **significant at 5 %

Table 3. SEM model result

variable	coefficient	p-value
(intercept)	0.80393	< 0.0001**
airport distance	0.005258	0.087688*
station distance	0.008216	0.171019
bus stop distance	-0.03822	0.003868**
mall distance	-0.00701	0.435993
university distance	0.004436	0.564319
office count	0.000476	0.011438**
attractions count	-7.50E-05	0.946433
restaurant count	-0.00257	0.003601**
hotel count	0.006901	1.8E-05**
λ	-1.9734	4.52e-06**

Notes: *significant at 10%; **significant at 5%

Table 3 also shows that the coefficient has a p-value of less than 0.05. This indicates that spatial error has a significant effect on the model formed. The R-Square value obtained for the SEM model is 17.5%. This value increases when compared to the value in linear regression, which is only 6.7%. When compared with the AIC value, the SEM model has a lower AIC value of -136.14 while the linear regression model has an AIC value of -117.1. Therefore, the SEM model can be considered superior to the multiple linear regression model.

The distance to the airport significantly affects the hotel occupancy rate in the model. This is in line with the research by Valentin and O'Neill, 2018 [9] which explains that accessibility or distance to the airport affects the hotel economy. The agglomeration effect described by the number of hotels within a 1 km radius also affects hotel occupancy. Chung and Kalnins [24] confirmed that the number of hotels in the vicinity will affect high demand to increase the occupancy rate of a hotel. The variable number of restaurants around the hotel significantly affects the hotel's occupancy rate with a negative coefficient value. It is in line with the findings by Yang [20] that restaurants will affect non-star hotels that do not have restaurant facilities, while star hotels tend not to be located in areas with many small-business scale restaurants. It is in line with Fang [17] that a hotel's economy is significantly affected by the number of business units (described by the number of offices). The more business units around the hotel, the more the hotel accommodation will increase. The distance to the bus stop has a negative and significant relationship. It illustrates that the closer the hotel is to the bus stop, the higher the occupancy rate. However, this is not in line with Yang's research [20], which states that accessibility to bus stops will only affect hotels located on the city outskirts.

4. Conclusions

Web-Scraper Google Maps has been developed to generate data on urban objects, ranging from airports, stations, bus stops, restaurants, offices, companies, malls, and universities. The results of this web-scraping are used at the analysis stage as the dependent variable. Distances to airports, stations, bus stops, malls, and universities are used as the distance from the hotel to the nearest object. At the same time, the remaining objects are used to determine the number of urban objects around the hotel.

From the results of the experiment, the model constructed in this study is the Spatial Error Model (SEM). The model concludes that spatial location has an impact on the occupancy level of hotels in the online hotel reservation application in the Province of Yogyakarta. The impact of spatial location that affects the occupancy rate is the distance from the hotel to the airport, the distance from the hotel to the bus stop, the number of nearest restaurants, the number of nearest offices, and the number of closest hotels.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

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

Funding

This study received no external funding.

Underlying data

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

Credit Authorship

Muhammad Fachry Nazuli: Conceptualization, Methodology, Data Collection, Data Analysis, and Manuscript Writing. **Satria Bagus Panuntun**: Data Collection, Software Development. **Addin Maulana**: Manuscript Review and Editing. **Takdir**: Manuscript Review and Editing. **Setia Pramana**: Conceptualization, Methodology, Manuscript Review, Research Advisor.

References

[1] R. Lado-Sestayo, M. Vivel-Búa, and L. Otero-González, “Survival in the lodging sector: An analysis at the firm and location levels”, *International Journal of Hospitality Management*, 59, 19 – 30, 2016. https://doi.org/10.1016/j.ijhm.2016.08.005

[2] R. Lado-Sestayo, and M. Vivel-Búa, “Diagnosis of bankruptcy hospitality and tourist destination; [Diagnosis de quiebra hotelera y destino turístico]”. *Lurralde: Investigacion y Espacio*, 41, 149 – 174, 2018. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85026880026&partnerID=40&md5=ccc920fcfb7dc05ab6e36304ec94ed0f

[3] T. Kajla, S. Raj, S. Sharma, M. Joshi, and A. Kaur, “Key preferences of tourists during COVID-19 pandemic in luxury hotels: Evidence from qualitative data”, *Tourism and Hospitality Research*, 22(4), 473 – 487, 2022. https://doi.org/10.1177/14673584211066742

[4] M. Li, L. Fang, X. Huang, and C. Goh, “A spatial–temporal analysis of hotels in urban tourism destination”, *International Journal of Hospitality Management*, vol. 45 34-43, 2015. https://doi.org/10.1016/j.ijhm.2014.11.005

[5] H. Luo, and Y. Yang, “Intra-metropolitan location choices for star-rated and non-rated budget hotels: The role of agglomeration economies”, *International Journal of Hospitality Management*, 72-83, 2017, https://doi.org/10.1016/j.ijhm.2016.09.007

[6] R. J. G. Calinao, M. V Amores, E. M. A. Rellores, and F. T. G Tabla, “The Determinants of Hotel Room Price In Philippines: A Structural Equation Modeling Analysis”, *Journal of Applied Structural Equation Modeling*, 6(1), 2022. https://doi.org/10.47263/JASEM.6(1)02

[7] Z. Xiang and M. Krawczyk, “What Does Hotel Location Mean for the Online Consumer? Text Analytics Using Online Reviews”, *Information and Communication Technologies in Tourism*, 383-395, 2016. http://dx.doi.org/10.1007/978-3-319-28231-2_28

[8] G. Li, R. Law, H. Q. Vu, J. Rong, and X. R. Zhao, “Identifying emerging hotel preferences using Emerging Pattern Mining technique”, *Tourism Management*, 46, 311-321, 2015. http://dx.doi.org/10.1016/j.tourman.2014.06.015

[9] M. Valentin, and J. W. O'Neill, "The Value of Location for Urban Hotels", *Cornell Hospitality Quarterl*, 5-24, 2018, https://doi.org/10.1177/1938965518777725

[10] Y. Yang, Z. Mao, and J. Tang, "Understanding Guest Satisfaction with Urban Hotel Location", *Journal of Travel Research*, 1-17, 2017. https://doi.org/10.1177%2F0047287517691153

[11] BPS Province D.I. Yogyakarta, Daerah Istimewa Yogyakarta Province in Figures 2021, BPS:2021 [Online]. Available: https://yogyakarta.bps.go.id/en/publication/2021/02/26/3a501d00eaa097f65efc96f9/provinsi-di-yogyakarta-dalam-angka-2021.html

[12] S. Pramana, D. Y. Paramartha, G. Y. Ermawan, N. F. Deli, W. Srimulyani, "Impact of COVID-19 pandemic on tourism in Indonesia", *Current Issues in Tourism*, 2022. https://doi.org/10.1080/13683500.2021.1968803

[13] Susilawati, R. Falefi, and A. Purwoko, "Impact of COVID-19's Pandemic on the Economy of Indonesia", *Budapest International Research and Critics Institute-Journal (BIRCI-Journal),* 1147-1156, 2020. 10.33258/birci.v3i2.954

[14] BPS, *Klasifikasi Baku Lapangan Usaha (KBLI) [Standard Classification of Business Fields 2020]*, (2020), BPS:2020 [Online]. Available: https://ppid.bps.go.id/upload/doc/KBLI_2020_1659511143.pdf

[15] Y. Yang, and Z. Mao, Z. "Location advantages of lodging properties: A comparison between hotels and Airbnb units in an urban environment", *Annals of Tourism Research*, 81, 2020. https://doi.org/10.1016/j.annals.2020.102861

[16] R. Lado-Sestayo, M. Vivel-Búa, and L. Otero-González, "Connection between hotel location and profitability drivers: an analysis of location-specific effects", *Current Issues in Tourism*, 1–18, 2018. https://doi.org/10.1080/13683500.2018.1538203

[17] L. Fang, H. Li, and M. Li, "Does hotel location tell a true story? Evidence from Geographically Weighted Regression Analysis of Hotels in Hong Kong", *Tourism Management*, 72, 78-91, 2019, https://doi.org/10.1016/j.tourman.2018.11.010

[18] L. Fang, Y. Xie, S. Yao, and T. Liu, "Agglomeration and/or differentiation at regional scale? Geographic spatial thinking of hotel distribution – a case study of Guangdong, China", *Current Issues in Tourism*, 1–17, 2020. https://doi.org/10.1080/13683500.2020.1792852

[19] W. Srimulyani, M. Faris, N. F. Deli, and S. Pramana, "Profile of Occupancy Rate Hotel in NTB During Pandemic Covid-19 With Big Data Approach", *Seminar Nasional Official Statistics*, 273-280, 2020, https://doi.org/10.34123/semnasoffstat.v2020i1.503.

[20] Y. Yang, K. K. F. Wong, T. Wang, "How do Hotels Choose Their Location? Evidence from Hotels in Beijing", *International Journal of Hospitality Management*, 675-685, 2012. https://doi.org/10.1016/j.ijhm.2011.09.003

[21] F. Yalcin, M. Mert, "Determination of hedonic hotel room prices with spatial effect in Antalya". *Economía, Sociedad y Territorio*, 18, 697-734, , 2018. Doi: 10.22136/est20181228.

[22] J. P. LeSage. *The Theory and Practice of Spatial Econometrics*. Toledo: University of Toledo, 1999.

[23] Y. Adhinugroho, A. P. Putra, M. Luqman, and E. Ermawan, "Development of online travel Web scraping for tourism statistics in Indonesia", *Information Research*, 25(4), 2020. https://doi.org/10.47989/irpaper885

[24] L. Woo and S. G. Mun, "Types of agglomeration effects and location choices of international hotels in an emerging market", *Tourism Management*, 77, 2019. https://doi.org/10.1016/j.tourman.2019.104034

DOI: https://doi.org/10.34123/jurnalasks.v17i1.798

POLITEKNIK STATISTIKA STIS

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

Jurnal Aplikasi Statistika & Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

The Application of Partial Proportional Odds Model on Determinants Analysis of Household Food Insecurity Level in Papua, Indonesia

Rolyn Abigael[1*], Cucu Sumarni[2], Ray Sastri[3]

[1]BPS-Statistics North Kayong Regency, North Kayong, Indonesia, [2]Politeknik Statistika STIS, Jakarta, Indonesia, [3]School of Finance and Economics, Jiangsu University, Zhenjiang City, Jiangsu Province, China.
**Corresponding Author: E-mail address:* siahaanrolyn@gmail.com

ARTICLE INFO

Article history:
Received 2 September, 2024
Revised 30 November, 2024
Accepted 2 December, 2024
Published 24 February, 2025

Keywords:
Food Insecurity; Logistic Regression; Ordinal Response Model; Parallel Lines Assumption Violation; Partial Proportional Odds

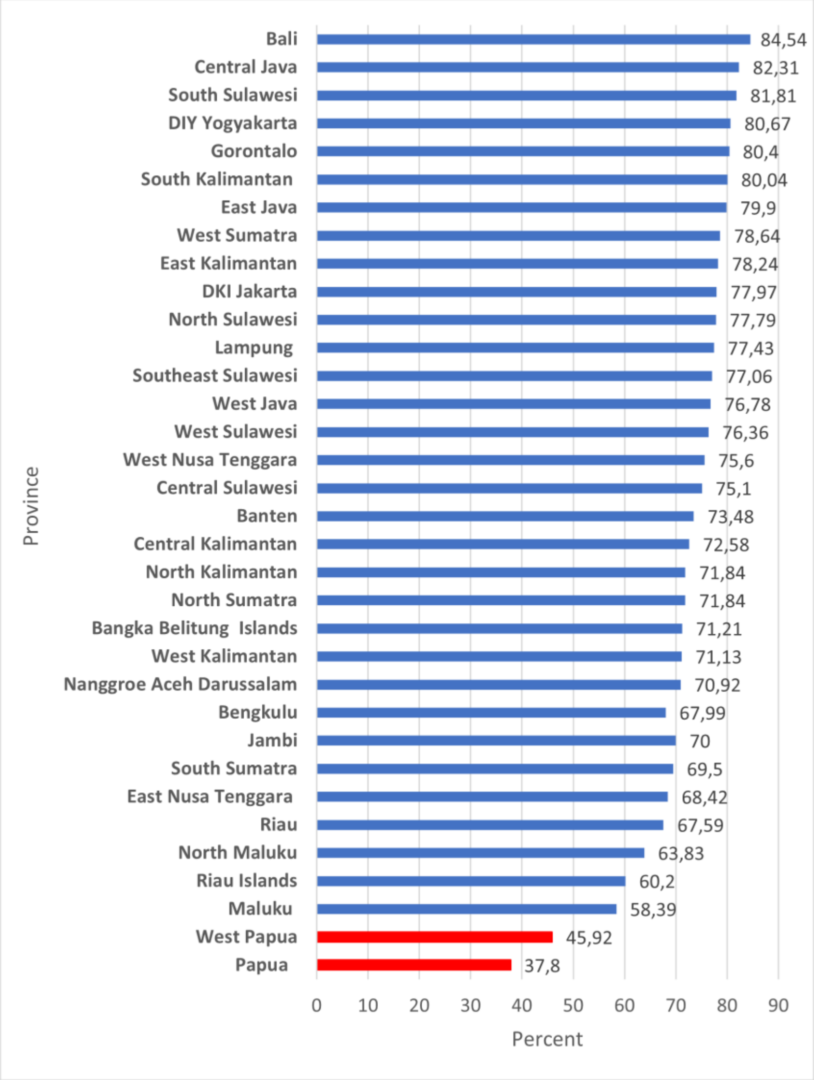
Abstract

Introduction/Main Objectives: Food insecurity in Papua, Indonesia, is still high. However, the study on that issue is limited. This research aims to analyze the determinants of food insecurity in Papua. **Background Problems:** An ordinal logistic regression can be used. However, this model generally requires the parallel lines assumption. However, somehow, the assumption is often violated. **Novelty:** This study used a model that relaxes the assumption of parallel lines. This model can capture the condition that some parameters are assumed to meet parallel lines and some do not. **Research Methods:** In this case, the partial proportional odds model was applied to find the determinant of household food insecurity status by using the National Socioeconomic Survey (SUSENAS) data. **Finding/Results:** The results show that a female head of household, age 60 years and above, junior high school education and below, has a higher tendency to be at least mildly food insecure, and the effect is the same for each level of food insecurity. Household heads who do not work, work in agriculture, and have household drinking water sources that are not feasible can aggravate the food insecurity level. Meanwhile, food assistance provided by the government influences reducing food insecurity levels.

1. Introduction

Food is one of the most basic needs of humanity to sustain life, causing the demand for food needs to increase as the number of people increases with positive population growth. Therefore, food security must be considered by the government. Food security is a condition that describes the sufficiency of food needs from the national level to individuals or households in terms of quantity, quality, safety, distribution, and affordability. The UN is committed to realizing food security through the Sustainable Development Goals(SDGs) in the second goal in target 2.1 regarding the right to food [1]. The government can monitor food security achievements through the Food Security Index (FSI).

Figure 1 shows two provinces with the lowest FSI values: Papua Province at 37.80 percent and West Papua Province at 45.92 percent. According to BPN [2], this figure shows that these provinces are classified as vulnerable to food insecurity. These provinces are located on the Papua Islands. We know that this region, especially Papua Indonesia region, faces various difficulties in achieving, maintaining, and improving the quality of sustainable food security. So, it is common for food insecurity to occur on this island. However, to mitigate more severe food insecurity, it is necessary to find the determining factors of the level of food insecurity at the individual/household level.



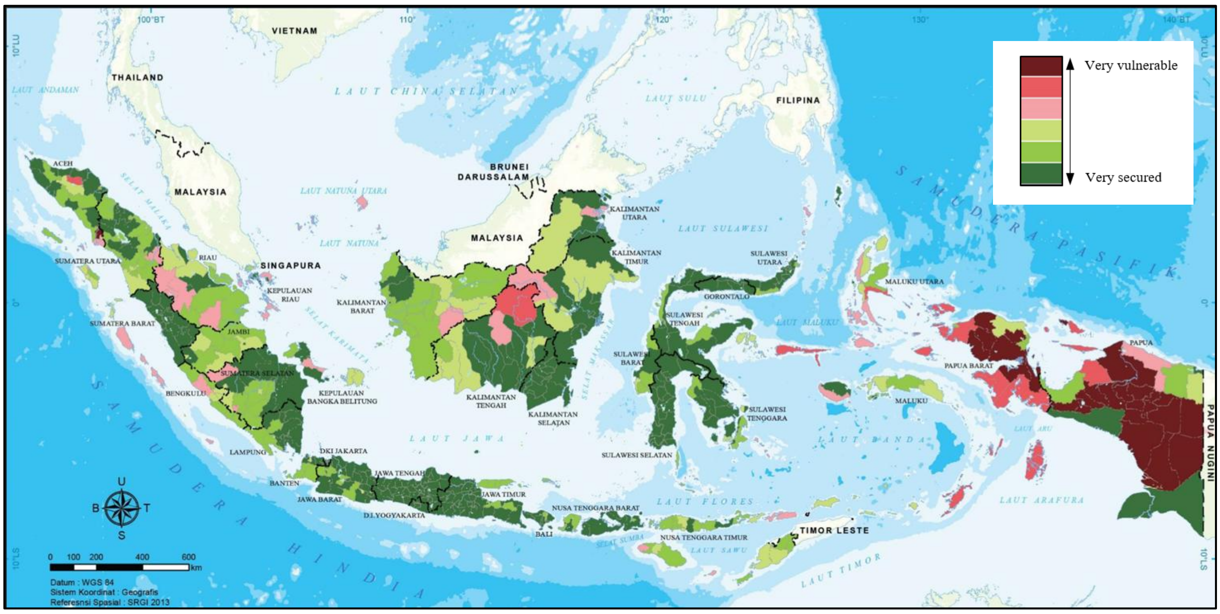
Source: BPN, 2022

Figure 1. Household food security index of 34 provinces in Indonesia 2022

Food insecurity is when a person does not have sufficient physical and economic access to nutritious food on an ongoing basis for normal growth and development and an active and healthy life. In terms of the prevalence of people with moderate or severe food insecurity, based on the 2022 food insecurity experience scale, the provinces of Papua region have a higher prevalence than the national figure of 4.85 percent, namely Papua province at 6.77 percent and West Papua province at 10.31 percent.

The government can monitor the state of food insecurity in Papua through the Food Security and Vulnerability Atlas (FSVA) to alleviate food insecure areas. FSVA categorizes the status of food security into six priority categories: the first is a category of highly vulnerable to food insecurity, and the last is a category of food security [3].

Figure 2 shows that most Papua districts are highly vulnerable to food insecurity. The food condition in this region is different from the achievement of the second goal of the SDGs indicators, making the Papua region an urgency in overcoming the problem of food insecurity at a macro level. In addition, food insecurity issues in this region will affect the status of food insecurity at the household level. Two components cause food insecurity at the household level: inadequate access to nutritious and safe food supplies and inadequate utilization of food by households. This fact makes it important to assess the status of food insecurity at the household level because household food insecurity can affect regional food access uncertainty directly.



Source: BPN, 2022

Figure 2. Food Security and Vulnerability Atlas (FSVA) in Indonesia by District 2022

The categorization in FSVA only provides an overview of the state of vulnerability to food insecurity at the regional level. However, it cannot be used to describe food insecurity at the household level. Meanwhile, household food insecurity is urgent in overcoming food insecurity problems at a regional level. Solving this food insecurity problem needs to consider the prioritization of households affected by food insecurity based on the level of food insecurity. Thus, FAO distinguishes household food insecurity based on levels into four categories. They are food security, mild food insecurity, moderate food insecurity, and severe food insecurity. There are many measurements for food insecurity, but according to Leroy et al. [4], the Food Insecurity Experience Scale (FIES) indicator is more standard than others. The FIES indicator was measured by eight questions about worry about getting food, eating healthily, the kinds of food that are eaten, eating a meal or not, eating less than usual, running out of food, feeling hungry or not, and any condition without food for a day.

Many studies have been conducted to examine the problem of food insecurity. These studies have various kinds of determination of food insecurity status, the methods applied, and the variables used. Borku et al. [5] and Ndhleve et al. [6] determined food insecurity with Household Food Insecurity Access Prevalence (HFIAP). Furthermore, Smith [7], Grimaccia & Naccarato [8], and Sheikomar [9], determine food insecurity status with the Food Insecurity Experience Scale (FIES).

Generally, they use binary logistic regression. Only Grimacia and Naccarato [8] applied ordinal logistic regression among researchers who used FIES. They made FIES into nine categories according to 8 FIES questions plus food security if all 8 FIES questions are answered "no". Because there are too many categories, many cell contents are zero. Of course, this will affect the modeling. Therefore, FAO itself classifies food insecurity into four categories. In addition, they still used the conventional ordinal logistic model with proportional odds so that it is not visible which variables have worsened food insecurity potentially. Besides that, the ordinal logistic regression with proportional odds model requires the assumption of parallel lines to be met. The assumption of parallel lines means that the categories in the dependent variable are parallel to each other so that the model has the same value for each category of different response variables. When the assumption is violated, a partial proportional odds model can be used if only some independent variables violate the parallel lines assumption or the non-proportional odds model if all independent variables violate the parallel lines assumption [10].

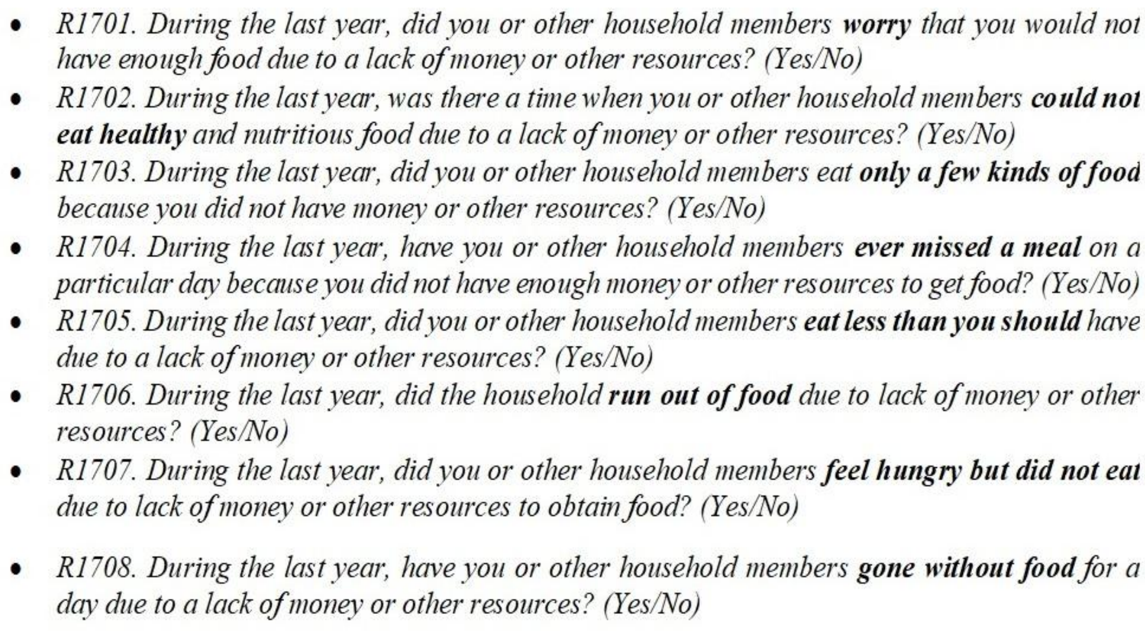
Existing research employing determinant analysis to assess household food insecurity in Papua remains limited. Given the widespread prevalence of severe food insecurity in many regions, further investigation is necessary to develop effective mitigation strategies. Ordinal logistic regression analysis, with categories aligned with the FAO's four-tier classification, can be employed to examine the influence of various factors on the severity of food insecurity. This study proposes a more flexible ordinal logistic model to avoid the restrictive parallel lines assumption inherent in the proportional odds model.

2. Material and Methods

2.1. Data

This study was undertaken in Papua and West Papua Provinces. The household food insecurity status can be measured by Statistics Indonesia (BPS) through the National Socioeconomic Survey (SUSENAS) according to questions R1701 to R1708. We used SUSENAS in March 2022. In this study, we have 20 975 household samples.

The dependent variable of this study is the household food insecurity level. The household food insecurity level is determined based on responses to the FIES questions on household access to food over the past year contained in Block XVII details of questions 1701-1708 in the March 2022 SUSENAS KOR questionnaire (VSEN22.K). The details of these questions are shown in Figure 3.



- *R1701. During the last year, did you or other household members **worry** that you would not have enough food due to a lack of money or other resources? (Yes/No)*
- *R1702. During the last year, was there a time when you or other household members **could not eat healthy** and nutritious food due to a lack of money or other resources? (Yes/No)*
- *R1703. During the last year, did you or other household members eat **only a few kinds of food** because you did not have money or other resources? (Yes/No)*
- *R1704. During the last year, have you or other household members **ever missed a meal** on a particular day because you did not have enough money or other resources to get food? (Yes/No)*
- *R1705. During the last year, did you or other household members **eat less than you should** have due to a lack of money or other resources? (Yes/No)*
- *R1706. During the last year, did the household **run out of food** due to lack of money or other resources? (Yes/No)*
- *R1707. During the last year, did you or other household members **feel hungry but did not eat** due to lack of money or other resources to obtain food? (Yes/No)*
- *R1708. During the last year, have you or other household members **gone without food** for a day due to a lack of money or other resources? (Yes/No)*

Source: BPN, 2022

Figure 3. FIES questions in SUSENAS March 2022 questionnaire

The FIES questions are asked at the household level, represented by the household head/partner/household members aged 15 years and above. Table 1 provides the criteria for categorizing the level of food insecurity.

Table 1. Operational definition of dependent variable

Food Insecurity Levels	Code	Condition
Food Security	1	All question items R1701-R1708 are answered "No."
Mild Food Insecurity	2	At least one question in R1701-R1703 is answered "Yes" and all questions in R1704 - R1708 are "No"
Moderate Food Insecurity	3	At least one question in R1704 - R1706 is answered "Yes," and the questions in R1707 and R1708 are "No".
Severe Food Insecurity	4	There is an answer "Yes" in R1707 and/or in R1708

Meanwhile, the independent variables used in this study are derived from the sociodemographic characteristics of household heads and standard household living conditions included in the March 2022 SUSENAS KOR responses (VSEN22.K). Table 2 explains the operational definitions of the independent variables used in the study.

Table 2. Operational definition of independent variable

Independent Variables	Categories
Head of household's sex	Female Male*
Head of household's age	Less than 60 years* 60 years and above
Head of household's education	Junior high school and below More than junior high school*
Head of household's work	Working at non-agriculture* Working at agriculture Not working
Drinking water source	Yes* No
Food aid recipient	Yes No*

**reference category*

2.2. Ordinal Logistic Regression Analysis

Ordinal logistic regression is a statistical analysis method for modeling the relationship between an ordinal dependent variable and one or more explanatory variables. An ordinal variable is a categorical variable with clear category levels. Meanwhile, the explanatory variables may be either continuous or categorical.

Ordinal logistic regression can be used when the dependent variable has at least three categories and the absolute distance between levels is unknown [11]. Several models are often used and can be distinguished based on how the logit is formed, such as the adjacent-category model, continuation ratio, and cumulative logit [12]. Based on these three models, the cumulative logit model is the easiest to interpret [10]. According to the assumptions that must be met, there are three types of cumulative logit models: proportional odds model, partial proportional odds model, and non-proportional odds model.

Proportional odds model (POM)

The proportional-odds model is an ordinal logistic model in which the intercepts depend on the jth category, but the slopes are all equal. The form of the cumulative logit model of proportional odds property is written as equation (1).

$$logit[P(Y > j)] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x} = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,\ j = 1,\ \ldots,\ J-1 \quad (1)$$

Based on equation (1), it is implied that there are $J-1$ models formed. For example, if there are three categories of response variable/dependent variable, then two cumulative logit models will be formed. The α_j is the unknown parameter's variable estimator, meaning that each cumulative logit has its intercept. $\boldsymbol{\beta}$ is a vector $p \times 1$ of regression coefficient parameter (slopes), where p is the number of the parameters.

$logit[P(Y > j)]$ is the logit of the cumulative probability of an analysis unit belonging to a category higher than j ($j+1$ or $j+2, \ldots$), where j is the dependent variable category. According to Agresti [11], the probability to be at or below the jth category can be defined as follows:

$$P(Y \leq j) = \pi_1 + \pi_2 + \cdots + \pi_j, \qquad j = 1, \ldots, J-1 \quad (2)$$

Where π_j is the probability (odds) of category j. and the probability of being above the jth category is thus the complement of the cumulative probability:

$P(Y > j) = 1 - P(Y \leq j)$
Then, the logit function of the cumulative odds as equation (1) can be expressed by equation (3) [13].

$$logit[P(Y > j)] = log\left[\frac{P(Y\ >\ j)}{1 - P(Y\ >\ j)}\right] =\ log\left[\frac{P(Y\ >\ j)}{P(Y \leq\ j)}\right] = \alpha_j +\ \boldsymbol{\beta}'\boldsymbol{x},\qquad j = 1, \dots, J \tag{3}$$

This model applies simultaneously to all $J - 1$ cumulative probabilities and assumes an identical effect of the predictors (independent variable) for each cumulative probability. The Proportional Odds Model (POM) ensures that the predicted odds for category j are no smaller than that of a category lower than category j and no larger than that of a category higher than category j. This model has assumptions that must be met. Specifically, the slope of the model must be the same for all logits [13]. This assumption is known as cumulative logit parallelity or parallel lines. The parallel lines assumption means that the categories in the dependent variable are parallel to each other so that the model has the same value for each category of different response variables [14]. If these assumptions are violated, the results of ordinal regression may not be valid.

Partial proportional odds model (PPOM)

When a parallel line assumption is not met in the POM model, it can occur because only some are not met. The partial proportional odds model (PPOM) is present to cover this. PPOM can be applied when some independent variables violate the parallel lines assumption. PPOM allows the slopes of some independent variables to violate the parallel lines assumption while others fulfill the parallel lines assumption. The cumulative probability of the partial proportional odds model for a dependent variable with j categories is as follows [15]:

$$logit[P(Y > j)] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x} + \boldsymbol{\gamma}_{\boldsymbol{j}}'\boldsymbol{u},\ \ j = 1, \dots, J - 1 \tag{4}$$

$\boldsymbol{\beta}$ is a vector of regression parameters (slopes) of independent variables that meet the parallel line assumption, and $\boldsymbol{\gamma_i}$ is a vector of parameters of independent variables that violate the parallel line assumption (different slopes for each jth dependent variable category).

Non-proportional odds model (NPOM)

The parallel line assumption violation in the POM model can also occur because all independent variables do not meet the assumption. A Non-proportional Odds Model (NPOM) can be applied when all independent variable coefficients violate the parallel line assumption. This model has varying slopes for each category of the dependent variable. The cumulative probability of the non-proportional odds model is as follows [15]:

$$logit[P(Y > j)] = \alpha_j + \boldsymbol{\gamma}_{\boldsymbol{j}}'\boldsymbol{u},\ \ j = 1, \dots, J - 1 \tag{5}$$

Parallel Lines Assumption Test

The parallel lines assumption means that the association between dependent and independent variables does not change for the categories of dependent variables. The parallel line assumption test can be done to determine whether the proportional odds model can be applied or not. This assumption can be tested through the likelihood ratio test to present an overall test of the parallel lines assumption on each independent variable [16]. The null hypothesis in this test is that the value of the regression coefficient (slope) is the same across all logit models ($j = 1, \dots, J - 1$). A rejection of this null hypothesis thus implies that the assumption is violated, whereas failure to reject this hypothesis supports the assumption.

$$PL =\ -2\ ln\left[\frac{L_0}{L_1}\right] \sim \chi^2_{p(J-2)} \tag{6}$$

Where, L_0 is the maximum likelihood value of the model with independent variables assuming parallel lines and L_1 is the maximum likelihood value of the model with independent variables that does not assume parallel lines. The null hypothesis can be rejected if $PL > \chi^2_{\alpha;p(J-2)}$ or $p - value < \alpha$.

Furthermore, the Brant test of parallel lines can identify the suitable model when the parallel lines assumption is not met. The Brant test compares separate estimates of each predictor. The hypothesis used in the Brant test is as follows,

H_0 : $\boldsymbol{R\beta}^* = \mathbf{0}$ (regression coefficients (slope) of all logit models are the same)
H_1 : $\boldsymbol{R\beta}^* \neq \mathbf{0}$

where

$$\boldsymbol{R} = \begin{bmatrix} I & -I & 0 & \cdots & 0 \\ I & 0 & -I & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I & 0 & 0 & \cdots & -I \end{bmatrix}_{(J-2)p\times(J-1)p} \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{J-1} \end{bmatrix}_{(J-1)p\times 1}$$

The test statistic is as follows:

$$\chi^2_{hit} = \left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}^*\right)^T \left[(\boldsymbol{R})\left(\boldsymbol{Asy.Var}\left(\widehat{\boldsymbol{\beta}}^*\right)\right)\left(\boldsymbol{R^T}\right)\right]^{-\mathbf{1}}\left(\boldsymbol{R}\widehat{\boldsymbol{\beta}}^*\right) \tag{7}$$

where, $\boldsymbol{Asy.Var}\left(\widehat{\boldsymbol{\beta}}^*\right)$ is a covariance estimation matrix of regression coefficient estimate, $\boldsymbol{R}$ is a contrast matrix, and $\boldsymbol{I}$ is the design of matrix $\boldsymbol{R}$, depending on the contrast to be compared in the parameters, the sum of the contrast coefficients in each row is zero. The decision will be obtained when the null hypothesis can be rejected if $\chi^2_{hit} > \chi^2_{\alpha;(J-2)}$ or $p-value < \alpha$. Based on the Brant test results, NPOM can be applied, when it shows that all predictors violate the parallel lines assumption. PPOM can be applied when it shows that only some predictors violate the parallel lines assumption.

Model Fit Test

To ensure which model is suitable, we can compare the models and call it the model fit test [14]. Model fit testing in PPOM for large samples can be done through the likelihood ratio (LR) test. The LR test in this model is carried out by comparing a simpler model with a more complex model. In this case, the LR test is carried out between POM and PPOM, and between PPOM and NPOM. The model fit testing hypothesis is as follows:

a. POM vs PPOM
H_0 : The POM model better fits the data
H_1 : The PPOM model better fits the data

Test statistics:

$$LR_1 = -2\ ln\left[\frac{L_{POM}}{L_{PPOM}}\right] \sim \chi^2_{v_1} \tag{8}$$

L_{POM} is the maximum likelihood of POM, while L_{PPOM} is the maximum likelihood of PPOM and $v_1$is the degree of freedom calculated from the difference in the number of parameters of the POM and PPOM models. The decision will be obtained when the null hypothesis is rejected if $LR_1 > \chi^2_{(\alpha;v_1)}$ or $p-value < \alpha$ so that it can be concluded that PPOM fits the data better than POM.

b. PPOM vs NPOM
H_0 : The PPOM model better fits the data
H_1 : The NPOM model better fits the data

Test Statistics:

$$LR_2 = -2\ ln\left[\frac{L_{PPOM}}{L_{NPOM}}\right] \sim \chi^2_{v_2} \tag{9}$$

L_{NPOM} is the maximum likelihood of NPOM and v_2 is calculated form the difference in the number of parameters of the PPOM and NPOM model. The decision will be obtained when the null hypothesis is rejected if $LR_2 > \chi^2_{(\alpha;v_2)}$ or $p-value < \alpha$ so that it can be concluded that NPOM fits the data better than PPOM.

According to Parry [17], there are many software options for running ordinal logistic regression models, such as SPSS, SAS, R and STATA. Williams [18] proposed the **gologit2** module in STATA. This model can directly check the parallel lines assumption in the POM model and, at the same time, can also test the model fit between the PPOM and NPOM models if the parallel lines assumption is violated.

3. Empirical Result and Discussion

3.1. Households Sample in Papua Indonesia 2022 Overview

Based on the results of the March 2022 SUSENAS presented in Figure 4 regarding the percentage level of household food insecurity in the Papua Indonesia region in 2022, the households sample generally experienced food security, which amounted to 82.31 percent. However, around 17.69 percent of households are still experiencing food insecurity. Of the 17.69 percent of households experiencing food insecurity, there are 6.69 percent of households experiencing mild food insecurity, 4.9 percent of households experiencing moderate food insecurity, and 6.2 percent of households experiencing severe food insecurity. This finding indicates that 6.69 percent of households are worried about uncertainty in food access; 4.9 percent of households experience a decrease in the quality and quantity of food and are not sure they can obtain food due to limited household resources; and 6.2 percent of households experience food shortages and do not eat for one or more days despite being hungry due to limited household resources.

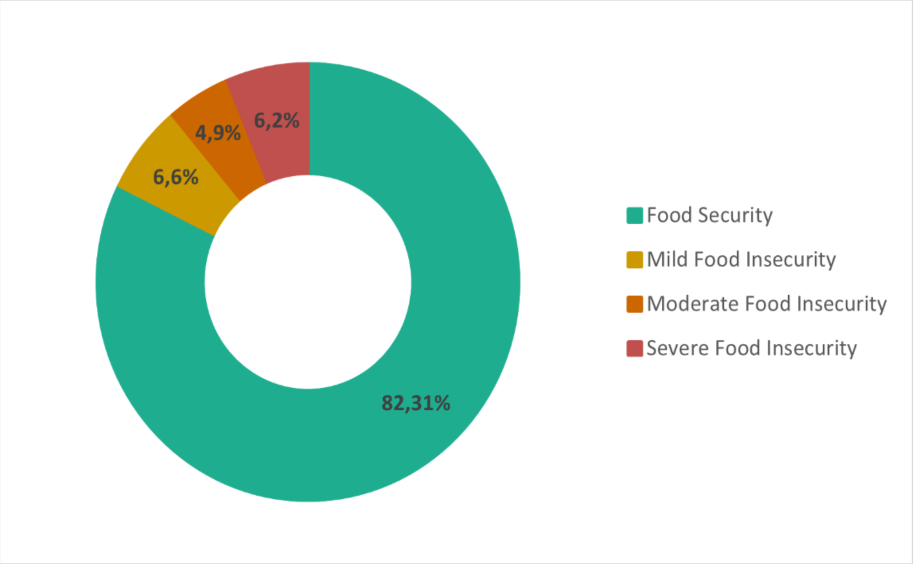


Figure 4. Percentage of households sampled by food insecurity level in Papua Indonesia region in 2022

Table 3 shows that the sample distribution of households experiencing food insecurity has characteristics derived from the gender of the female head of household, the age of the head of household is 60 years and above, the education level of the head of household is junior high school or below, the head of household is not working, work in agriculture, the household receives food assistance, and the household has inadequate drinking water sources. When viewed from the severity level, the most severe level of food insecurity appears to have more household samples for these characteristics. However, the difference is mostly similar to the level below it. The most severe looks were quite different when the head of the household was not working, working in agriculture, the family's source of drinking water was not feasible and not receive food assistance from the government.

Table 3. Percentage of sample households by food insecurity level and household characteristics in Papua Indonesia in 2022

Independent Variables	Categories	Food Security	Food Insecurity		
			Mild	Moderate	Severe
Head of household's sex	Female	78.60%	7.00%	7.06%	7.35%
	Male*	82.64%	6.56%	4.68%	6.12%
Head of household's age	Less than 60 years*	82.66%	6.53%	4.85%	5.97%
	60 years and above	80.08%	7.04%	5.08%	7.80%
Head of household's education	Junior high school and below	81.62%	6.45%	5.06%	6.87%
	More than junior high school*	83.35%	6.82%	4.60%	5.22%
Head of household's work	Working at non-agriculture*	83.18%	7.78%	4.52%	4.53%
	Working at agriculture	82.24%	5.84%	4.96%	6.96%
	Not working	78.26%	7.52%	6.07%	8.15%
Drinking water source	Feasible*	82.61%	7.51%	5.04%	4.85%
	Not feasible	82.01%	5.70%	4.72%	7.57%
Food aid recipient	Yes	61.73%	21.07%	8.47%	8.73%
	No*	83.35%	5.86%	4.70%	6.09%

3.2. Determinants Analysis of Household Food Insecurity Level in Papua Indonesia in 2022

First, we perform ordinal logistic regression modelling with proportional odds, then we check the parallel line assumption. The parallel line statistic test has a Chi-Square value of 148.56, and the p-value is 0.000. The p-value is less than the significance level (α=0,05). This finding indicates that with a significance level of 5 percent, the model does not meet the parallel lines assumption. Then perform a Brant test to know which predictors (independent variables) violate the parallel lines assumption. The Brant test statistic values produced based on formula (6) are listed in Table 4.

Table 4. Parallel lines assumption test results with brant test

Independent variables	Categories	*Chi-Square*	*p-value*	df
Head of household's sex	Female	3.76	0.153	2
Head of household's age	60 years and above	2.14	0.342	2
Head of household's education	Junior high school and below	0.08	0.960	2
Head of household's work	Working at agriculture	12.26	0.002*	2
	Not working	6.23	0.004*	2
Drinking water source	Not feasible	78.41	0.000*	2
Food aid recipient	Yes	19.86	0.000*	2

* Significance at level 0.05

The independent variables that meet the parallel line assumption have a p-value greater than the significance level ($\alpha = 0.05$). From Table 4, it can be seen that of the six variables, only three met the parallel line assumption, such as gender, age, and the education level of the household head. The other three variables, such as the economic activity of the household head, access to feasible drinking water, and the food aid recipient status, violate the parallel line assumption. This result shows that only some variables met the parallel line assumption. This indicates that the ordinal logistic model with proportional odds isn't suitable.

A model-fitting test was done to determine which of the three types of ordinal logistic regression models is the most appropriate. The test results can be seen in Table 5. From these results, we can decide that the partial proportional odds (PPOM) model is the most appropriate.

Table 5. Model fit test result

Model Hypothesis	*Chi-square*	*p-value*	Decision
H_o : $POM\ model\ better\ fits\ the\ data$ H_1 : $PPOM\ model\ better\ fits\ the\ data$	166.07	0.000*	Reject H_o
H_o : $PPOM\ model\ better\ fits\ the\ data$ H_1: $NPOM\ model\ better\ fits\ the\ data$	8.72	0.1899	Do not reject H_o

* Significance at level 0.05

The partial proportional odds model was then applied to determine the household food insecurity level in the Papua Indonesia region. Three logit models were formed because there are four levels of dependent variable categories: Model 1 (at least mild food insecurity versus food security), Model 2 (at least moderate food insecurity versus food security and mild food insecurity), and Model 3 (severe food insecurity versus food security to moderate food insecurity). The results of partial parameter testing are in Table 6.

Table 6. Partial proportional odds model results

		Model 1 (at least mild food insecurity versus food security) $g(y) = \log\frac{P(Y > 1)}{P(Y \leq 1)}$			Model 2 (at least moderate food insecurity versus food security and mild food insecurity) $g(y) = \log\frac{P(Y > 2)}{P(Y \leq 2)}$			Model 3 (severe food insecurity versus food security to moderate food insecurity) $g(y) = \log\frac{P(Y > 3)}{P(Y \leq 3)}$		
Variables	Categories	$\hat{\beta}$	Se($\hat{\beta}$)	*p-value*	$\hat{\beta}$	Se($\hat{\beta}$)	*p-value*	$\hat{\beta}$	Se($\hat{\beta}$)	*p-value*
Intercept	-	-1.632	0.036	0.000	-2.406	0.047	0.000	-3.31	0.066	0.000
Head of household's sex	Female	0.145	0.051	0.004*	0.145	0.051	0.004*	0.145	0.051	0.004*
Head of household's age	60 years and above	0.139	0.052	0.008*	0.139	0.052	0.008*	0.139	0.052	0.008*
Head of household's education	Junior high school and below	0.114	0.039	0.004*	0.114	0.039	0.004*	0.114	0.039	0.004*
Head of household's work	Working at agriculture	0.233	0.044	0.000*	0.363	0.053	0.000*	0.435	0.071	0.000*
	Not working	0.346	0.084	0.000*	0.529	0.096	0.000*	0.665	0.122	0.000*
Drinking water source	Not feasible	0.005	0.038	0.898	0.220	0.044	0.000*	0.458	0.059	0.000*
Food aid recipient	Yes	0.675	0.061	0.000*	0.418	0.073	0.000*	0.408	0.093	0.000*

* Significance at level 0.05

Based on the results presented in Table 6, it can be seen that for all logit models, each independent variable generally produces a p-value smaller than the significance level (α = 0.05). This finding shows that with a significance level of 5 percent, each variable affects to distinguish between households group with food insecurity and no food insecurity, moderate food insecurity from the level below it, and severe food insecurity from the level below it significantly. However, in Model 1 the source of feasible drinking water does not affect significantly. This result means that the variable can not distinguish significant groups of food-secure households from those with food insecurity. Interpretation of the effects per variable is easier if using the odds ratio. The odds ratio values for each variable category are presented in Table 7.

Table 7. Odds Ratio

Variables	Categories	*Odds Ratio* (OR)		
		Model 1 (Y = 2,3,4 vs Y = 1)	Model 2 (Y = 3,4 vs Y = 1,2)	Model 3 (Y = 4 vs Y = 1,2,3)
Intercept		0.195	0.09	0.037
Head of household's sex	Female	1.156	1.156	1.156
Head of household's age	60 years and above	1.149	1.149	1.149
Head of household's education	Junior high school and below	1.121	1.121	1.121
Head of household's work	Working at Agriculture	1.261	1.437	1.545
	Not working	1.413	1.697	1.579
Drinking water source	Not feasible	1.005	1.247	1.581
Food aid recipient	Yes	1.963	1.519	1.503

The head of household's sex, age, and education level fulfill the parallel lines assumption. So, these variables have the same influence on each level of food insecurity. We can see from Table 7. This table depicts the same odds ratio of these variables in Models 1, 2, and 3. However, the other three variables (head of household working status, feasibility of drinking water sources, and food recipient status) violate this assumption. Hence, the effects in these three models are different. More details about each variable effect can be expressed as follows.

Head of Household's Sex

Assuming other influencing variables are constant, the odds of a female head of household experiencing at least mild food insecurity is 1.156 times greater than that of a male. Likewise, her probability of experiencing at least moderate food insecurity and of experiencing severe food insecurity is 1.156 times greater than that of a male head of household. This result is similar to the results of Smith et al. [7], Grimaccia and Naccarato [8], and Nigusu and Shewadinber [19]; households headed by women are more likely to experience food insecurity than households headed by men. Female-headed households have limitations in accessing resources that affect food production and access, so they experience food insecurity compared to male-headed households [6].

Head of Household's Age

Compared to households with a head of household under 60, those aged 60 years and above are 1.149 more likely to experience mild food insecurity (compared to food security), assuming other variables are constant. 1.149 is more likely to experience moderate and severe food insecurity (compared to maximum mild food insecurity), and 1.149 is more likely to experience severe food insecurity (compared to maximum moderate food insecurity). This result is related to Gebre's [20], the older the household head, the more likely they are to experience food insecurity. The older the household head, the more food insecurity they will experience due to decreased productivity and efficiency in doing work [21]. In addition, households with older heads of household are usually multigenerational, with more older people to feed and unable to contribute to income generation, increasing the incidence of food insecurity [7].

Head of Household's Education

Assuming other variables are constant, for households with the education level of head households junior high school and below, the odds of being very or somewhat likely to have food insecurity (severe, moderate, or mild) versus likely to have no food insecurity is 1.121 times that of households whose heads have more than junior high school. This finding is similar to the statement from Ndheleve et al. [6] and Birhane et al. [22] that household heads who have low education are more vulnerable to food insecurity. The higher the level of formal education of the household head, the lower the household food insecurity because the education of the household head is important in improving the quality of life and providing opportunities to obtain decent work so that they have sufficient income to meet food needs [23].

Head of Household's Work

Households in which the head of household did not work (assuming other variables are constant) are 1.413 times more likely to experience mild food insecurity or more (compared to food security) than those working in non-agriculture, 1.697 times more likely to experience moderate or severe food insecurity (versus food security or mild food insecurity), and 1.925 times more likely to experience severe food insecurity (versus food security or no more than moderate food vulnerability). Meanwhile, household heads working in agriculture tend 1.262 times to experience mild food insecurity or more (compared to food security) than those working in non-agriculture, 1.437 times to experience moderate and severe food insecurity (compared to food security or mild food insecurity), and 1.545 times to experience severe food insecurity (compared to food security or no more than moderate food vulnerability).

Thus, in the Papua Indonesia region in 2022, households where the head of household did not work and work in agriculture tend to aggravate food insecurity more than non-agriculture households. This result is in line with Etana and Tolossa [24], that unemployed household heads have a higher potential for food insecurity than employed household heads because unemployed household heads cannot buy food in terms of quality and quantity. In addition, household heads working in the agricultural sector earn smaller salaries and have lower welfare than non-agricultural workers, so they cannot fulfill their food needs. [25].

Drinking water source

The same thing happens with access to infeasible drinking water. Households with infeasible drinking water sources tend to increase food insecurity than households with feasible drinking water sources. The odds of households with infeasible drinking water sources being more likely to experience moderate and severe food insecurity are 1.247 times greater than that of feasible drinking water sources (versus food security or mild food insecurity), and the odds to be more likely severe are 1.581 times.

Similar research results such as Azwardi, et al [26] also prove that households with adequate drinking water sources tend to be food insecure. When drinking water does not come from a proper source, it will increase the risk of individuals getting diseases due to contamination of drinking water. Hence, the food utilization dimension needs to be realized.

Food aid recipient

The provision of food assistance from the Government in Papua appears to have different effects on the household food insecurity level. We see that the higher the severity, the lower the odds ratio. Assuming other variables are constant, households receiving food assistance are 1.963 times more likely to experience mild food insecurity or more (versus food security) than households not receiving food assistance, 1.519 times more likely to experience moderate and severe food insecurity (versus food security or mild food insecurity), and 1.503 times more likely to experience severe food insecurity (versus food security or no more than moderate food vulnerability).

The decrease in the odds ratio shows that this program has successfully reduced the severity of food insecurity. Nonetheless, the odds ratio value is still quite high (more than 1.5), so this program has yet to be able to address food insecurity fully. In addition, Amrullah [27] shows that the provision of food assistance has a small impact, so more than alleviating household food insecurity is needed to depend on food assistance received.

4. Conclusions

This study empirically shows a violation of the parallel lines assumption in the usual ordinal regression model (proportional odds model), and the partial proportional odds model is more appropriate to describe the determinants of household food insecurity levels. Based on this model, households in the Papua region of Indonesia with characteristics of female heads of household, aged 60 years and above, junior high school education or below, have an unsafe drinking water source, do not receive food assistance, farmers, and worse if they do not work have a greater tendency to experience food insecurity at least mild insecurity.

In addition, with the proportional odds model, we can find out which variables can worsen the level of food insecurity or, vice versa, reduce the severity. The employment status of the head of the household, which reflects a household's economic conditions, can aggravate the food insecurity level; if he works as a farmer or does not work at all, the higher the severity level. Thus, access to infeasible drinking water can also worsen food insecurity. However, on the contrary, food assistance from the government can reduce the severity of food insecurity. Thus, the government should continue to run the program accompanied by education on how the family economy improves and socialization of the importance of education.

Ethics approval

Not required.

Acknowledgments

We thank Statistics Indonesia for providing SUSENAS data for this research.

Competing interests

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

Underlying data

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

Credit Authorship

Rolyn Abigael: Writing, Visualization, Software. **Cucu Sumarni**: Conceptualization, Methodology, Supervision. **Ray Sastri:** Reviewing and Editing,

References

[1] BAPPENAS- *Pilar Pembangunan Sosial Metadata Indikator Tujuan Pembangunan Berkelanjutan Indonesia [Indonesian National Development Planning Agency, Pillars of Social Development Metadata Indicators of Sustainable Development Goals of Indonesia]*, Jakarta: Deputy for Maritime Affairs and Natural Resources, Indonesian National Development Planning Agency, 2023.

[2] BPN- *Indeks Ketahanan Pangan [Food Security Index]*, Jakarta: Deputy for Food and Nutrition Vulnerability, National Food Agency, 2022.

[3] BPN- *Food Security and Vulnerability Atlas (FSVA) 2022*, Jakarta: Deputy for Food and Nutrition Vulnerability, National Food Agency, 2022.

[4] J.L. Leroy, M. Ruel, E.A. Frongillo, J. Harris, and T.J. Ballard. "Measuring the food access dimension of food security: A critical review and mapping of indicators," *Food Nutr. Bull.*, vol. 36, no. 2, pp. 167–195, 2015, doi: 10.1177/0379572115587274.

[5] A. W. Borku, A. U. Utallo, and T. T. Tora, "The Level of Food Insecurity among Urban Households in Southern Ethiopia: A Multi-Index-Based Assesment," *Build. Environ.*, p. 107386, 2020, doi: 10.1016/j.jafr.2024.101019.

[6] S. Ndhleve *et al.*, "Household food insecurity status and determinants: The case of botswana and south africa," *Agraris*, vol. 7, no. 2, pp. 207–224, 2021, doi: 10.18196/agraris.v7i2.11451.

[7] M. D. Smith, W. Kassa, and P. Winters, "Assessing food insecurity in Latin America and the Caribbean using FAO's Food Insecurity Experience Scale," *Food Policy*, vol. 71, pp. 48–61, 2017, doi: 10.1016/j.foodpol.2017.07.005.

[8] E. Grimaccia and A. Naccarato, "Food Insecurity in Europe: A Gender Perspective," *Soc. Indic. Res.*, vol. 161, no. 2–3, pp. 649–667, 2022, doi: 10.1007/s11205-020-02387-8.

[9] O. B. Sheikomar, W. Dean, H. Ghattas, and N. R. Sahyoun, "Validity of the Food Insecurity Experience Scale (FIES) for Use in League of Arab States (LAS) and Characteristics of Food Insecure Individuals by the Human Development Index (HDI)," *Curr. Dev. Nutr.*, vol. 5, no. 4, pp. 1–10, 2021, doi: 10.1093/cdn/nzab017.

[10] E. Ari and Z. Yildiz, "Paralel Lines Assumption in Ordinal Logistic Regression and Analysis Approaches," *Int. Interdiscip. J. Sci. Res.*, vol. 1, no. 3, pp. 8–23, 2014.

[11] A. Agresti, *Categorical Data Analysis Third Edition*, New Jersey: John Wiley & Sons, Inc., 2013.

[12] D. W. Hosmer, S. L. Jr., and R. X. Sturdivant, *Applied Logistic Regression.3rd Edition*, New Jersey: John Wiley & Sons, Inc., 2013.

[13] A. Agresti, *An Introduction to Categorical Data Analysis Second* Edition, New Jersey: John Wiley & Sons, Inc., 2007.

[14] D. G. Kleinbaum and M. Klein, *Logistic Regression: A Self Learning Text*, New York: Springer, 2010. doi: 10.1007/978-1-4419-1742-3.

[15] A. Agresti, *Analysis of Ordinal Categorical Data Second Edition*, New Jersey: John Wiley & Sons, Inc., 2010.

[16] R. Azen and C. M. Walker, *Categorical Data Analysis for the Behavioral and Social Sciences*. New York: Taylor and Francis, 2011.

[17] S. Parry, "Ordinal Logistic Regression models and Statistical Software : What you need to know," *Cornell Statistical Consulting Unit Stat News #91*, 2020.

[18] R. Williams, "Generalized ordered logit/partial proportional odds models for ordinal dependent variables," *Stata J.*, vol. 6, no. 1, pp. 58–82, 2006, doi: 10.1177/1536867x0600600104.

[19] A. Nigusu and M. Shewadinber, "The status and determinants of rural household food insecurity in North Shewa Zone, Oromia Region, Ethiopia," *Turkish J. Food Agric. Sci.*, vol. 4, no. 1, pp. 6–12, 2022, doi: 10.53663/turjfas.1020187.

[20] G. G. Gebre, "Determinants of Food Insecurity among Households in Addis Ababa City, Ethiopia," *Interdiscip. Descr. Complex Syst.*, vol. 10, no. 2, pp. 159–173, 2012, doi: 10.7906/indecs.10.2.9.

[21] T. Sekhampu, "Determination Of The Factors Affecting The Food Security Status Of Households In Bophelong, South Africa," *Int. Bus. Econ. Res. J.*, vol. 12, no. 5, p. 543, 2013, doi: 10.19030/iber.v12i5.7829.

[22] T. Birhane, S. Solomon, S. Hagos, and S. M. Katia, "Urban food insecurity in the context of high food prices: a community based cross sectional study in Addis Ababa, Ethiopia," *BMC Public Health*, vol. 14, pp. 1–8, 2014, [Online]. Available: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L605376010%0Ahttp://dx.doi.org/10.1186/1471-2458-14-680

[23] N. F. Ruhyana, W. Y. Essa, and Mardianis, "Sociodemographic factors affecting household food security in Sumedang regency West Java province," *Agraris*, vol. 6, no. 1, pp. 38–51, 2020, doi: 10.18196/agr.6189.

[24] D. Etana and D. Tolossa, "Unemployment and Food Insecurity in Urban Ethiopia," *African Dev. Rev.*, vol. 29, no. 1, pp. 56–68, 2017, doi: 10.1111/1467-8268.12238.

[25] S. Sophia, E. Erwandri, R. Dewi, and F. Varina, "Analisis Tingkat Ketahanan Pangan Keluarga Penerima Manfaat Bantuan Sosial Pangan (Kpm Bansos Pangan) di Kabupaten Batang Hari [Analysis of Food Security Level of Families Receiving Food Social Assistance in Batang Hari Regency]," *JAS (Jurnal Agri Sains)*, vol. 6, no. 2, pp. 113–121, 2022, doi: 10.36355/jas.v6i2.920.

[26] A. Azwardi, H. F. Widyasthika, R. C. Saleh, and N. Adnan, "Household Food Security: Evidence From South Sumatera," *Jejak*, vol. 12, no. 2, pp. 446–465, 2019, doi: 10.15294/jejak.v12i2.20264.

[27] E. R. Amrullah, A. Pullaila, I. Hidayah, and A. Rusyiana, "Dampak Bantuan Langsung Tunai Terhadap Ketahanan Pangan Rumah Tangga Di Indonesia [Impacts of Direct Cash Transfer on Household Food Security in Indonesia]," *J. Agro Ekon.*, vol. 38, no. 2, pp. 77–90, 2020, doi: 10.21082/jae.v38n1.2020.77-90.

DOI: https://doi.org/10.34123/jurnalasks.v17i1.803

POLITEKNIK STATISTIKA STIS

Jurnal Aplikasi Statistika &
Komputasi Statistik
Vol. 17 No. 1, June 2025
e- ISSN: 2615-1367, p-ISSN:2086-4132

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

Estimation of Gross Regional Domestic Product per Capita at the Sub-District Level in Bali, NTB, and NTT Provinces Using Machine Learning Approaches and Geospatial Data

I Made Satria Ambara Putra[1], Rindang Bangun Prasetyo[2*], Candra Adi Wiguna[3]

[1]BPS-Statistics West Kotawaringin Regency, West Kotawaringin, Indonesia, [2]Politeknik Statistika STIS, Jakarta, Indonesia, [3]Nanyang Technological University, Singapore
**Corresponding Author: E-mail address: rindang@stis.ac.id*

ARTICLE INFO

Article history:
Received 8 September, 2024
Revised 5 December, 2024
Accepted 4 February, 2025
Published 24 February, 2025

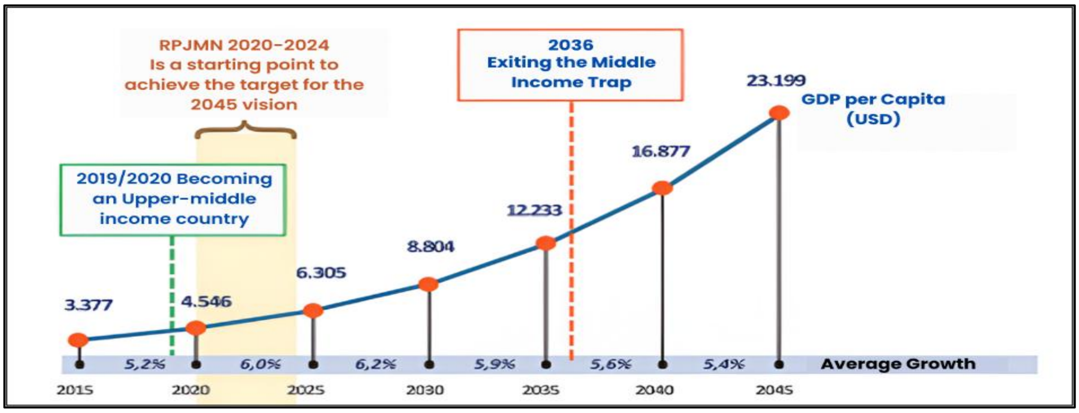
Keywords:
Big Data Geospatial; Gross Regional Domestic Product Per Capita; Machine Learning; Neural Network; Williamson Index

Abstract

Introduction/Main Objectives: This study aims to estimate Gross Regional Domestic Product (GRDP) per capita at the sub-district level. **Background Problems:** Currently, GRDP per capita is calculated only at the district level by BPS. **Novelty:** This study estimates GRDP per capita at the sub-district level using a model developed at the district level, applying machine learning and linear regression methods. **Research Methods:** The model was constructed using geospatial data sourced from satellite imagery, OpenStreetMap, (Village Potential Statistics) PODES, directories of large mining companies, and directories of the manufacturing industry at the district level. Linear regression and machine learning methods, including neural networks, random forest regression, and support vector regression, were used to develop the model. The research focuses on three provinces: Bali, West Nusa Tenggara (NTB), and East Nusa Tenggara (NTT). **Findings/Results:** The best-performing model was support vector regression, with MAE and MAPE evaluations of 10.33 million and 26.11%, respectively. The results indicate that sub-districts with high GRDP per capita are typically urban areas that serve as economic hubs. The Williamson Index results show that districts in the eastern region have higher inequality levels compared to those in the western region.

1. Introduction

The Indonesian government has incorporated the Sustainable Development Goals (SDGs) into its development strategy, as outlined in the National Medium-Term Development Plan (RPJMN), starting from the 2015-2019 period and continuing into the current 2020-2024 period. The RPJMN functions as a strategic framework for the government's initiatives to meet the objectives set forth in Indonesia's Vision 2045, which aims to establish a more Advanced Indonesia. Figure 1 shows the target Gross Domestic Product (GDP) per capita as part of the vision for an Advanced Indonesia. GDP is an economic indicator used to assess the performance of economic development in a country or region. It can be calculated from various metrics that evaluate a country's financial performance [1]. Indonesia's position from 2019 to 2020 has already reached upper-middle-income status, meaning it is above countries with middle-income per capita but still below those with high-income per capita. Indonesia is projected to escape the Middle-Income Trap by 2036. One of the key strategies to achieve this goal is to reduce regional disparities, ensuring that economic growth in each region aligns with national growth. To monitor these conditions and track regional disparities, it is essential to use appropriate indicators. The Gross Regional Domestic Product (GRDP) per capita is one such indicator that can illustrate disparities between regions [2].



Source: Bappenas

Figure 1. Target of GDP per capita growth towards an advanced Indonesia

The Gross Regional Domestic Product (GRDP) is typically calculated at higher administrative levels, such as districts or provinces, and is generally not computed at the sub-district level [3]. This is primarily due to the limited scale and availability of economic data at the sub-district level, where economic activities tend to be smaller and more restricted compared to districts or provinces. The necessary economic data, such as industrial output, agricultural production, trade, and services, may either be unavailable or insufficiently detailed at the sub-district level, leading to challenges in data collection. Additionally, many economic activities in sub-districts are often integrated with surrounding areas, making it difficult to accurately isolate and assess the specific economic contributions of individual sub-districts. For these reasons, GRDP is more frequently calculated at higher levels, such as districts or provinces, where the data is more comprehensive and suitable for evaluating regional economic performance [4], [5].

GRDP at the sub-district level represents added value that can be directly observed by the community [6]. By utilizing GRDP data at the sub-district level, the success of economic development in these areas can be evaluated both directly and indirectly [6]. Regional disparities can be measured using the Williamson Index, which relies on GRDP per capita as a reference metric [7], [8]. Currently, the Williamson Index is typically calculated at the provincial or national level. However, with the availability of GRDP per capita data at the sub-district level, it would become possible to calculate the Williamson Index at the district level. Therefore, estimating GRDP per capita at the sub-district level is crucial for accurately assessing economic disparities and development.

The approach to calculating District GRDP can be done using two methods: the direct method and the indirect method. The direct method involves using primary data from each district to compute GRDP, though in practice, such data is often difficult to obtain. In contrast, the indirect method estimates gross value added by allocating or predicting it using indicators that closely correlate with the respective economic activities. This method leverages proxy indicators to approximate economic output when direct data is unavailable [3], [6]. One of the indirect methods that can be applied to produce GRDP estimates at the sub-district level is through prediction. Various estimation methods, including linear regression, can be used to estimate these values [9]. For instance, Pasaribu et al. [10] utilized a linear regression model to predict GRDP in Jakarta. Similarly, Agu et al. [11] predicted GDP using macroeconomic indicators by applying four regression techniques: Principal Component Regression (PCR), Ridge Regression (RR), Lasso Regression (LR), and Ordinary Least Squares (OLS). These methods help in estimating economic output when direct data is unavailable, providing a robust approach to forecasting regional economic performance.

The linear regression method is one of the simplest and most commonly used techniques to model the relationship between continuous variables [12]. However, its primary limitation is that it can only capture linear associations between the dependent and independent variables. To address this limitation, machine learning techniques can be utilized. Machine learning algorithms can identify and model complex, non-linear relationships, provide more accurate predictions, offer flexibility in model selection and development, address overfitting through regularization techniques, and efficiently process large datasets, making them highly suitable for more advanced statistical modeling and prediction tasks [13].

Several researchers have applied machine learning (ML) techniques to predict GRDP. Muchisha et al. [14] developed and evaluated the performance of six widely used ML algorithms, including Random Forest, LASSO, Ridge, Elastic Net, Neural Networks, and Support Vector Machines, to forecast Indonesia's quarterly GDP growth in real-time. Similarly, Richardson et al. [15] demonstrated that ML algorithms have the potential to significantly outperform simple autoregressive benchmarks and dynamic factor models when predicting New Zealand's GDP. Sa'adah and Wibowo [16] applied two deep learning techniques, LSTM and RNN, to model GDP fluctuations, even during the COVID-19 pandemic. Experimental results from Lai [17] showed that neural networks are reliable in predicting GDP and have practical applications. Meanwhile, Sukono et al. [18] predicted Gross Regional Domestic Product (GRDP) using a genetic algorithm approach based on the Cobb-Douglas model, comparing the results with those obtained from the ordinary least squares (OLS) method. The study conducted by Puttanapong et al. [13] in estimating Provincial GDP in Thailand utilized three methods: neural networks, random forests, and support vector machines. These methods were chosen because neural networks are highly effective in capturing complex and non-linear relationships within data, random forests are known for their ability to produce accurate models by leveraging multiple decision trees, thereby reducing the risk of overfitting, and SVMs perform well with high-dimensional data, making them suitable for analyzing complex datasets [13].

The primary challenge in estimating Gross Regional Domestic Product (GRDP) at the sub-district level lies in the issue of data completeness. Not all economic sectors have data available at such a granular level, necessitating the use of additional data that specifically covers the sub-district level [6]. Furthermore, the data sources traditionally used for GRDP calculations are often costly and time-intensive to collect. Consequently, there is a need for alternative methods and data sources that are both cost-effective and time-efficient while providing comprehensive coverage at the sub-district level.

Recent studies have increasingly explored the use of alternative data sources, such as geospatial data derived from satellite imagery through remote sensing and OpenStreetMap (OSM). These data sources offer the potential to supplement the limited data currently available for calculating GRDP per capita at more granular levels. Remote sensing refers to the use of sensors to detect electromagnetic radiation, which can be processed to generate interpretable images of the earth's surface, thereby yielding valuable information [19]. One significant advantage of remote sensing is its ability to provide extensive spatial coverage, even at very fine scales, coupled with the availability of large volumes of data [19]. Similarly, OpenStreetMap data can accurately represent regional characteristics down to micro levels, and its open-source nature ensures it is freely accessible [20].

Previous research by Fasial and Shakera [21] investigated the use of remote sensing techniques to predict Gross Domestic Product (GDP) through built-up index analysis across nine major cities in Canada. Similarly, Puttanapong et al. [13] demonstrated that the application of geospatial data and machine learning, specifically utilizing the Random Forest algorithm, achieved a prediction accuracy of 97.7% for Provincial Gross Domestic Product. In addition to the Random Forest method, their study also employed two other machine learning algorithms, namely Neural Networks and Support Vector Machines [13]. These findings indicate that the application of machine learning algorithms can significantly enhance the accuracy and comprehensiveness of Provincial GDP predictions, while also facilitating a more detailed integration with geospatial data. Furthermore, Putri et al. applied machine learning techniques to estimate poverty at a granular level, building a model based on sub-district data to estimate poverty at a 1.5 km grid scale [22]. Therefore, this study aims to estimate the GRDP per capita at the sub-district level in the provinces of Bali, NTB, and NTT using linear regression, neural networks, random forest regression, and support vector regression. The models are constructed at the district level to estimate the lower-level sub-districts.

2. Material and Methods

2.1. *Gross Regional Domestic Product Per Capita*

Gross Regional Domestic Product (GRDP) data, both at current prices and constant prices, is one of the key indicators for understanding the economic condition of a region or area within a specific period [23]. In general, GRDP represents the total value added generated by all business units in a region or can be interpreted as the total value of final goods and services produced by all economic units [23]. GRDP at current prices reflects the value added of goods and services calculated based on the prevailing prices in the given year. Meanwhile, GRDP at constant prices represents the value added of goods and services calculated using fixed prices from a specific base year as a reference [23]. Badan Pusat Statistik is an Indonesian government agency responsible for collecting GRDP data.

GRDP per capita is the result of dividing the value added generated by all economic activities by the total population [24]. The size of the population, whether large or small, will affect the GRDP per capita, while the GRDP itself is highly dependent on the potential of natural resources and the production factors available in the region [24]. GRDP per capita at current prices reflects the value of GRDP divided by the total population or per individual [24].

2.2. Geospatial Data

Geospatial data is information about the location, size, and characteristics of objects, both natural and man-made, that are located above or below the Earth's surface [25]. Geospatial information is processed geospatial data that can be utilized as a tool in formulating policies, making decisions, and carrying out activities related to terrestrial space. Geospatial data and information also become very important and beneficial in supporting the development process across various sectors of life [25]. There are five sources of geospatial data: public participatory geographic information systems (PPGIS), participatory geographic information systems (PGIS), volunteered geographic information (VGI), open data, and big data [26]. In this research, geospatial data from satellite imagery sourced from big data and geospatial data from OpenStreetMap sourced from VGI are used.

2.3. Yeo-Johnson Transformation

Yeo-Johnson is an extension of the Box-Cox transformation. Box-Cox Transformation is used to transform positive data to approximate a normal distribution. This transformation can only be applied to positive data. Yeo-Johnson Transformation is a generalization of Box-Cox that can be applied to both positive and negative data. This makes Yeo-Johnson more flexible in handling various types of data [27]. The data transformations are defined by the following equations [27].

$$\psi(x,\lambda)\begin{cases}\dfrac{(x+1)^{\lambda}-1}{\lambda} & x \geq 0;\ \lambda \neq 0\\ \log(x+1) & x \geq 0;\ \lambda = 0\\ \dfrac{(-x+1)^{2-\lambda}-1}{2} & x < 0;\ \lambda \neq 2\\ -\log(-x+1) & x < 0;\ \lambda = 2\end{cases} \tag{1}$$

Where x is the original value and λ is the transformation parameter.

2.4. Pearson Correlation

Pearson correlation is used to determine the strength of the linear relationship between two variables and to identify the direction of the relationship that occurs [28]. The interpretation of the correlation coefficients is presented in Table 1 below.

Table 1. Interpretation of correlation coefficient

Range of Values	Level of Intensity
0.80 – 1.000	Very Strong
0,60 – 0,799	Strong
0,40 – 0,699	Strong Enough
0,20 – 0,399	Weak
0,00 – 0,199	Very Weak

2.5. Mutual Information

Mutual Information has emerged in recent years as an important measure for feature selection criteria, particularly in machine learning [29]. By using Mutual Information, we can uncover unexpected relationships between the measured variables. So far, the Pearson correlation coefficient has been the most popular, but this measure is not sufficient because many phenomena are non-linear in nature [30]. A high Mutual Information value indicates that the association between two variables is stronger [31].

2.6. Variance Inflation Factor

Multicollinearity is the occurrence of a high correlation among factors in a model. Multicollinearity can produce biased results when researchers attempt to determine how effectively each factor can be used to predict or understand the response variable in a statistical model [32]. Overall, multicollinearity can result in wider confidence intervals and less reliable probability values for predictors [32]. This means that the findings from the model with multicollinearity may not be reliable. There are several techniques used to detect multicollinearity, and this study employs the Variance Inflation Factor (VIF). VIF is used to measure the extent to which the variance of the estimated regression coefficients increases when the independent variables are correlated with each other.

2.7. Linear Regression

Linear regression is the simplest and certainly the most common method for measuring the relationship between continuous variables, and this method is easiest to understand through practical examples. Regression models are often used for practical decision-making as well as for more theoretical or scientific investigations [12].

2.8. Machine Learning

Machine learning is a field that lies at the intersection of computer science, statistics, and various other disciplines, focusing on the automatic improvement over time, as well as inference and decision-making under uncertainty [33]. Machine learning requires valid data as learning material (during the training process) before being used in testing to produce optimal output [34]. The ability of networked and cellular computing systems to collect and transfer large amounts of data has grown rapidly, a phenomenon often referred to as "big data." Researchers who gather such data often turn to machine learning to seek solutions in obtaining useful insights, making predictions, and facilitating decision-making from these data sets [33].

2.9. Evaluation Metrics

The evaluation of the optimal model selection is conducted by considering the level of accuracy. The accuracy level is calculated by taking into account the residual values produced by the model that has been built. This research uses several evaluation metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The formulas for these three metrics are as follows [35].

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(F_t - A_t)^2}{n}} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|F_t - A_t| \tag{2}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n}\frac{|F_t - A_t|}{A_t} \tag{3}$$

In this context, F_t represents the forecasted value for time period t, A_t denotes the actual observed value for the same period, and n indicates the total number of time periods. A lower MAPE value signifies a higher accuracy of the forecasting model.

2.10. Williamson Index

The Williamson Index is a tool for measuring the development of a region by comparing it with more advanced areas. In general, the Williamson Index assesses the level of development disparity in a region [8]. The Williamson index describes the relationship between regional disparity and economic levels using economic data from both developed and developing regions [8]. The basis for calculating the Williamson Index involves GDP per capita and the population size in each region. Statistically, the formulation of the Williamson Index can be expressed with the following formula.

$$IW = \frac{\sqrt{\sum_{i}^{n}(y_i - \hat{y})^2 \frac{p_i}{p}}}{\hat{y}} \quad (4)$$

With IW being the Williamson Index, y_i represents the GDP per capita at the district/city level, ŷ represents the GDP per capita of the province, p_i represents the population of each district/city, p is the total population of the province, and n is the number of districts in the province. This study will obtain the GDP per capita at the sub-district level. Therefore, the IW calculation can be conducted at the district level. When the Williamson Index value moves further away from 0, it indicates that income inequality between regions in that area is increasing. Conversely, if the Williamson Index value approaches 0, it shows that income inequality between regions in that area is decreasing [8].

2.11. Analysis Method

This research uses six data sources. First, the data source comes from satellite imagery. Satellite data from remote sensing used in this research includes Night Time Light (NTL) from the NOAA-VIIRS satellite, Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Normalized Difference Built-Up Index (NDBI) from the Sentinel-2 satellite, Land Surface Temperature from the MODIS satellite, and Carbon Monoxide (CO), Nitrogen Dioxide (NO2), and Sulfur Dioxide (SO2) from the Sentinel-5P satellite. This data collection is conducted using a cloud-based platform, namely Google Earth Engine. Secondly, the data source used comes from OpenStreetMap (OSM). The collection of OSM data is done through Overpass Turbo. Overpass Turbo is a web-based mining tool for running queries to access OpenStreetMap data and displaying the results on an interactive map. Tourism data is obtained by entering the query or keyword "tourism," while to get road length data, the keywords "highway=primary," "highway=secondary," and "highway=tertiary" are used.

Thirdly, the data sources used come from the village potential data of 2021 collected by Badan Pusat Statistik (BPS). The researcher proposed data collection through the Statistical Service Information System (Silastik). The data collected includes the number of households using electricity, the number of educational facilities, the number of health facilities, the number of micro and small industries, the number of financial services, the number of villages with kiosks selling agricultural production tools, the number of food and beverage accommodation providers, the number of places of worship, the number of villages with internet access for online gaming cafes and other facilities, the number of village-owned enterprises, and the number of villages with waste banks.

Fourth, the data sources used are from the directory of large mining companies in 2021 and the directory of the manufacturing industry in 2021. Fifth, the data sources used are from the official websites of the BPS of each province to obtain GRDP per capita data at the district administrative level for the year 2021. GRDP per capita is the response variable in this study. The sixth is the population data taken from regional publications in numbers, this data is used for the calculation of the Williamson Index. All data collected is based on the year 2021.

In this study, the model is built at the district level. The resulting model is used to estimate GRDP per capita at the sub-district level. The model development uses data from 9 provinces, namely the Special Region of Yogyakarta, DKI Jakarta, Banten, East Java, Central Java, West Java, West Nusa Tenggara, East Nusa Tenggara, and Bali. However, the predictions were only made in three provinces: Bali, NTB, and NTT. The tools used in this research are Google Earth Engine, Overpass Turbo, QGIS, and Google Colab for data collection, data preprocessing, model development, and mapping. In addition, Google Drive and spreadsheets are used for data storage. The general steps are outlined in the following flowchart.

Data collection and preprocessing is the first stage of this research. Through preprocessing, the data is cleaned, processed, and transformed to be ready for further analysis. Preprocessing steps such as data cleaning help ensure that the data used for training and testing models or algorithms is of high quality and relevant. The collected data consists of data at the district and sub-district administrative levels. The district administrative data is used for model development purposes, while the sub-district administrative data is used for estimating per capita GRDP at the sub-district level. The district administrative data includes 160 districts and cities from 9 provinces, namely Bali Province, West Nusa Tenggara, East Nusa Tenggara, Special Region of Yogyakarta, Jakarta Special Capital Region, Banten, East Java, Central Java, and West Java. Meanwhile, the sub-district administrative data consists of 389 sub-districts from 3 provinces, namely Bali Province, West Nusa Tenggara, and East Nusa Tenggara.

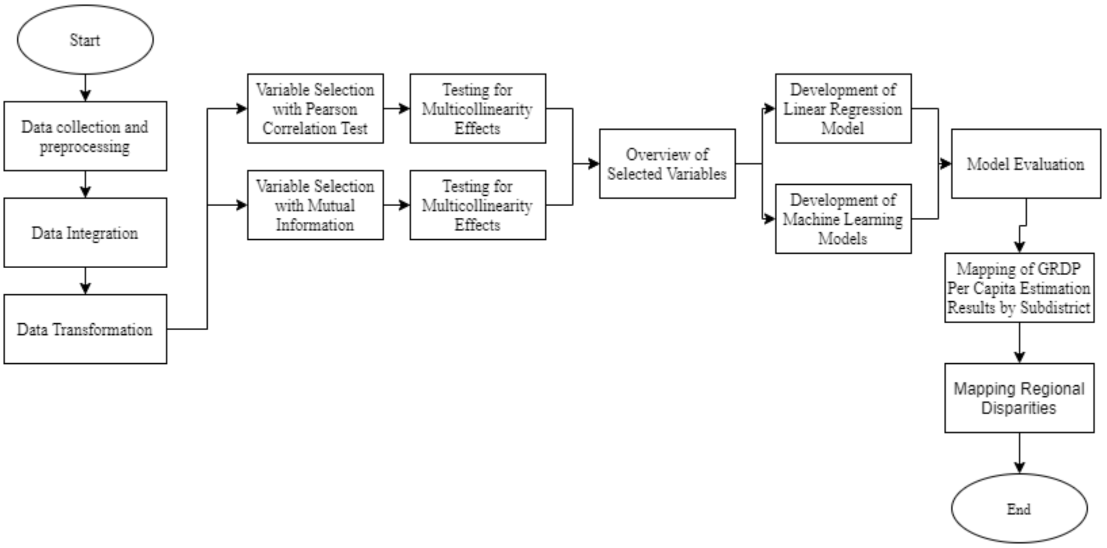

Start
Data collection and preprocessing
Data Integration
Data Transformation
Variable Selection with Pearson Correlation Test
Variable Selection with Mutual Information
Testing for Multicollinearity Effects
Testing for Multicollinearity Effects
Overview of Selected Variables
Development of Linear Regression Model
Development of Machine Learning Models
Model Evaluation
Mapping of GRDP Per Capita Estimation Results by Subdistrict
Mapping Regional Disparities
End

Figure 2. Flowchart of research stages

The expected outcome of the data integration steps is a dataset that encompasses all the necessary data, such as satellite imagery, OpenStreetMap, village potential data, mining, and manufacturing industries. After collecting data and preprocessing it, data from various sources was obtained at the administrative levels of the sub-district and district. At this stage, the data is combined based on its administrative level. Data at the sub-district administrative level is needed to estimate the per capita GRDP at the sub-district level, while data at the district administrative level is necessary for model development. The next step is data transformation. The purpose of data transformation is to improve data quality or to prepare the data for analysis or model development. The transformation used is the Yeo-Johnson transformation. In the Python programming language, the "sklearn.preprocessing" library provides the "PowerTransformer" function for calculating the Yeo-Johnson transformation.

Variable selection means choosing among many variables which ones will be included in the model, that is, selecting the appropriate variables from a complete list by eliminating those that are irrelevant or redundant. Practicality is one of the reasons why variables must be chosen. According to the principle of parsimony, a simple model with fewer variables is preferred over a complex model with many variables. In this study, variable selection for the development of the linear regression model was conducted using Pearson correlation analysis, and variables were chosen based on moderate to strong relationships. After that, the selected variables underwent a multicollinearity test. Variable selection for the development of the machine learning model used mutual information analysis, and 10 variables with the highest values were chosen. Following that, the selected variables underwent a multicollinearity test. Variables that indicate the presence of multicollinearity will be eliminated in the model development. The purpose of this stage is to improve the model's performance.

After the district data set is completed, variable selection is carried out, and the model is built at the district level. The model for estimating GDP per capita is built by applying linear regression and machine learning methods. The machine learning algorithms used are neural network (NN), random forest regression (RFR), and support vector regression (SVR). The data set is divided into 80% training data (128 data points) and 20% testing data (32 data points). The model is built using 80% of the data set. In this research, the GridSearchCV approach is applied to build and select the best model. The evaluation uses the K-Fold Cross Validation method, which is the default evaluation method when using GridSearchCV. Next, evaluate the model using RMSE, MAE, and MAPE values to obtain the best model for estimating per capita GDP at the sub-district level. The model is selected based on the smallest RMSE, MAE, and MAPE values.

3. Result and Discussion

3.1. Candidate Variable Data Exploration

It can be identified in Figure 3 that the relationship between GRDP per capita and the predictor variables has a non-linear pattern, while it can also be identified that the patterns and relationships among other variables also exhibit a non-linear pattern. The variables used are still in different units. Therefore, a transformation is needed to improve data quality. The transformation used in this research is the Yeo-Johnson transformation because it can handle both positive and negative values, as well as manage variables that have different units, such as the variables used in this study. This transformation can produce the best evaluation compared to other transformations. After that, a Pearson correlation test and multicollinearity test were conducted for the development of the linear regression model, while variable selection was done using mutual information and a multicollinearity test for the development of the machine learning model. This is done so that the model can provide accurate and precise predictions.

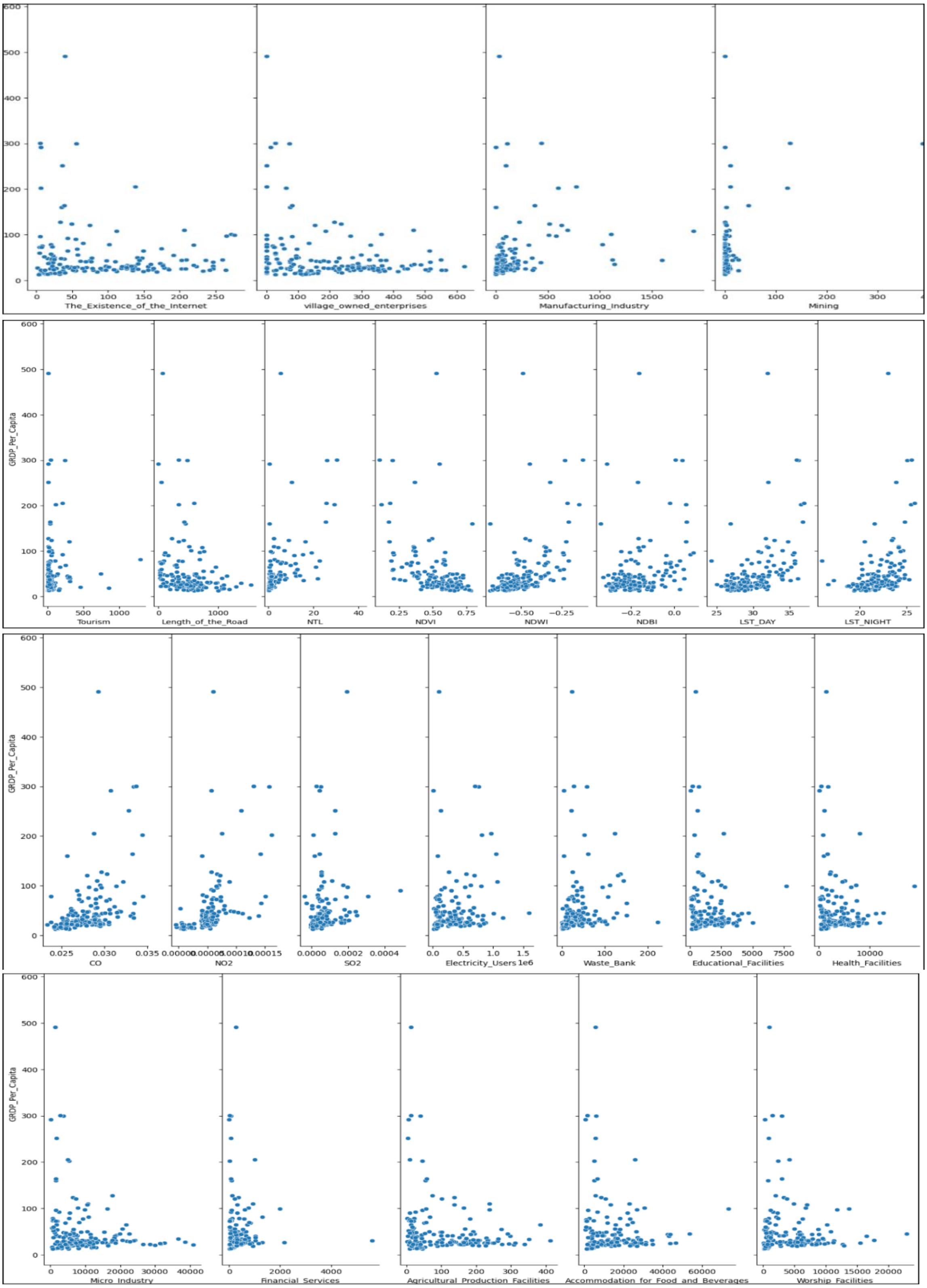


Figure 3. Pairplot

3.2. Variable Selection

In building a linear regression model, selection is carried out using the Pearson correlation test with a moderate to strong relationship. The results of the calculations are shown in Table 2, sorted from the highest value.

Table 2. Results of the pearson correlation calculation

Variable	Pearson	Degree of Closeness	Direction
Night Time Light (NTL)	0.67	Strong	Positive
Normalized Difference Water Index (NDWI)	0.66	Strong	Positive
Nitrogen Dioxide (NO_2)	0.61	Strong	Positive
Day Time Land Surface Temperature (LST Day)	0.57	Strong Enough	Positive
Carbon Monoxide (CO)	0.55	Strong Enough	Positive
Industri Manufaktur	0.48	Strong Enough	Positive
Night Time Land Surface Temperature (LST Night)	0.44	Strong Enough	Positive
Normalized Difference Built-Up Index (NDBI)	0.42	Strong Enough	Positive
Normalized Difference Vegetation Index (NDVI)	-0.61	Strong	Negative

Based on Table 2, it can be seen that the selected variables used in the development of the linear regression model are Night Time Light (NTL), Normalized Difference Water Index (NDWI), Nitrogen Dioxide (NO2), Day Time Land Surface Temperature (LST Day), Carbon Monoxide (CO), Manufacturing Industry, Night Time Land Surface Temperature (LST Night), Normalized Difference Built-Up Index (NDBI), and Normalized Difference Vegetation Index (NDVI). Next, a multicollinearity test was conducted using VIF. The results of the VIF calculations are displayed in Table 3.

Based on Table 3, it can be identified that the variables Normalized Difference Water Index (NDWI), Day Time Land Surface Temperature (LST Day), Normalized Difference Built-Up Index (NDBI), and Normalized Difference Vegetation Index (NDVI) have VIF values of 99.64, 20.48, 12.13, and 122.76, respectively, where a VIF value > 10 indicates that the estimated regression coefficients are weak due to multicollinearity. Variables with VIF values > 10 will be eliminated. Therefore, the selected variables for the development of the linear regression model are Night Time Light (NTL), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Manufacturing Industry, and Night Time Land Surface Temperature (LST Night).

Table 3. Results of VIF calculation after pearson correlation test

Variable	VIF
Night Time Light (NTL)	5.32
Normalized Difference Water Index (NDWI)	99.64
Nitrogen Dioxide (NO_2)	4.74
Day Time Land Surface Temperature (LST Day)	20.48
Carbon Monoxide (CO)	7.32
Manufacturing Industry	0.48
Night Time Land Surface Temperature (LST Night)	0.44
Normalized Difference Built-Up Index (NDBI)	0.42
Normalized Difference Vegetation Index (NDVI)	122.76

Before building the machine learning model, variable selection is carried out by examining the mutual information values. After that, 10 variables are selected based on the highest mutual information values. The results of the calculations can be seen in the following Table 4, which is sorted by the highest mutual information values.

Table 4. Results of mutual information calculation

Variable	Mutual Information Value
Night Time Light (NTL)	0.41
Nitrogen Dioxide (NO_2)	0.32
Normalized Difference Vegetation Index (NDVI)	0.30
Normalized Difference Water Index (NDWI)	0.29
Manufacturing Industry	0.27
Carbon Monoxide (CO)	0.26
The Number of Villages with the Presence of Kiosks Selling Supplies	0.22
Jumlah Fasilitas Kesehatan	0.21
Day Time Land Surface Temperature (LST Day)	0.21
Night Time Land Surface Temperature (LST Night)	0.19

It should be identified that the 10 variables with the highest mutual information values are Night Time Light (NTL), Nitrogen Dioxide (NO2), Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Manufacturing Industry, Carbon Monoxide (CO), Number of Villages with Agricultural Production Supply Kiosks, Number of Health Facilities, Day Time Land Surface Temperature (LST Day), and Night Time Land Surface Temperature. (LST Night). Next, a multicollinearity test was conducted on the selected variables using the Variance Inflation Factor (VIF). The results of the VIF calculation are obtained in the following Table 5.

Table 5. Results of VIF calculation after analyzing autual information values

Variable	VIF
Night Time Light (NTL)	6.07
Nitrogen Dioxide (NO_2)	4.80
Normalized Difference Vegetation Index (NDVI)	95.08
Normalized Difference Water Index (NDWI)	90.93
Manufacturing Industry	3.87
Carbon Monoxide (CO)	2.61
The Number of Villages with the Presence of Kiosks Selling Supplies	2.61
Jumlah Fasilitas Kesehatan	3.04
Day Time Land Surface Temperature (LST Day)	14.32
Night Time Land Surface Temperature (LST Night)	5.25

Based on Table 5, it can be identified that the variables Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), and Day Time Land Surface Temperature (LST Day) have VIF values of 95.08, 90.93, and 14.32, respectively, where a VIF value > 10 indicates that the estimated regression coefficients are weak due to multicollinearity (Shrestha, 2020). Variables with VIF values > 10 will be eliminated. Therefore, the selected variables for the development of the machine learning model are Night Time Light (NTL), Nitrogen Dioxide (NO2), Manufacturing Industry, Carbon Monoxide (CO), the Number of Villages with Agricultural Production Supply Kiosks, the Number of Health Facilities, and Night Time Land Surface Temperature (LST Night).

3.3. Development of GRDP Per Capita Estimation Model

In this research, the creation of linear regression and machine learning models aims to estimate the GDP per capita map based on the spatial characteristics of an area by combining data from satellite imagery, OpenStreetMap, village potential data, directories of large mining companies, and directories of the manufacturing industry that have undergone a previous variable selection stage. This model was built using linear regression and machine learning with three algorithms: neural network, random forest regression, and support vector regression. The GridSearchCV method with 5-fold cross-validation is used for the purpose of tuning parameters in machine learning models.

The first machine learning model built is a neural network. The second machine learning model built is Random Forest Regression. The third machine learning model built is Support Vector Regression. The model with the best combination of hyperparameters is chosen to map GRDP per capita. The specifications of the machine learning model used are presented in Table 6.

Table 6. Hyperparameter specifications

Method	Specifications	Value
Neural Network	Epochs	20
	Batch_Size	32
	Units	128
	Activation	Tanh
	Learning_Rate	0.001
Random Forest Regression	n_estimators	100
	Max_depth	None
	Min_samples_split	10
	Min_samples_leaf	2
Support Vector Regression	C	1
	Gamma	0.1
	kernel	rbf

The selected specifications for the neural network model are an epoch of 20 means the model is trained for 20 full iterations through the entire training dataset, allowing weight updates more than once to improve accuracy. Batch 32 means the training data is divided into groups of 32 samples, and weight updates are performed after each group is processed. The number of neurons in the hidden layer is 128. Activation using hyperbolic tangent (tanh) means the function is used as the activation function, mapping inputs to a range between -1 and 1. Learning rate 0.001 means the model updates its weights in small steps of 0.001 during training to optimize performance.

The second machine learning model built is Random Forest Regression. N_estimators set to 100 in the Random Forest Regression model means the model uses 100 decision trees to make predictions. The final result is the average of the predictions from all the trees. No max_depth means there is no maximum depth limit set for the decision trees in the model. Min samples split 10 means that each decision tree in the model will only split a node if there are at least 10 samples in it. Min_samples_leaf 2 means that each leaf in the decision tree must contain at least 2 samples.

The third machine learning model built is Support Vector Regression. C set to 1 in model represents the regularization value used to control the trade-off between a larger margin and errors in the training data. Gamma set to 0.1 in the Support Vector Regression (SVR) model represents a parameter that controls how far the influence of a single data sample extends to the model. RBF Kernel (Radial Basis Function) means the kernel function used in the model to measure the similarity between data points. To determine the better and selected method for predicting GRDP per capita at the sub-district level, a numerical evaluation was conducted using the three measures outlined in point 2.9.

Table 7. Model evaluation results

Method	RMSE (Million IDR)	MAE (Million IDR)	MAPE
Neural Network	15.26	10.43	28.77%
Random Forest Regression	15.33	10.37	28.63%
Support Vector Regression	15.97	10.33	26.11%
Linear Regression	20.42	13.66	38.60%

The four machine learning models and linear regression produced a MAPE of less than 50%, which means that according to the MAPE significance table, the forecasts are quite accurate. The machine learning model using the SVR algorithm yielded an MAE of 10.33 million IDR and a MAPE of 26.11%. The SVR model achieved the smallest MAE and MAPE evaluations among the other models. Therefore, this study uses SVR to estimate GDP per capita at the administrative district level in the provinces of Bali, NTB, and NTT.

3.4. Mapping of GRDP Per Capita at the Sub-District Level

This research produces a map of GRDP per capita at the administrative district level. This map is expected to assist local governments in monitoring the welfare of their communities. The visualization of GRDP per capita at the sub-district level for the year 2021 in the provinces of Bali, NTB, and NTT, along with direct monitoring or verification in the field using Google Earth. The research locus map of Bali, West Nusa Tenggara, and East Nusa Tenggara provinces can be seen in Figure 4. The mapping was conducted using the natural breaks method and divided the GRDP per capita results into 5 classes. The mapping results are shown in Figure 5.

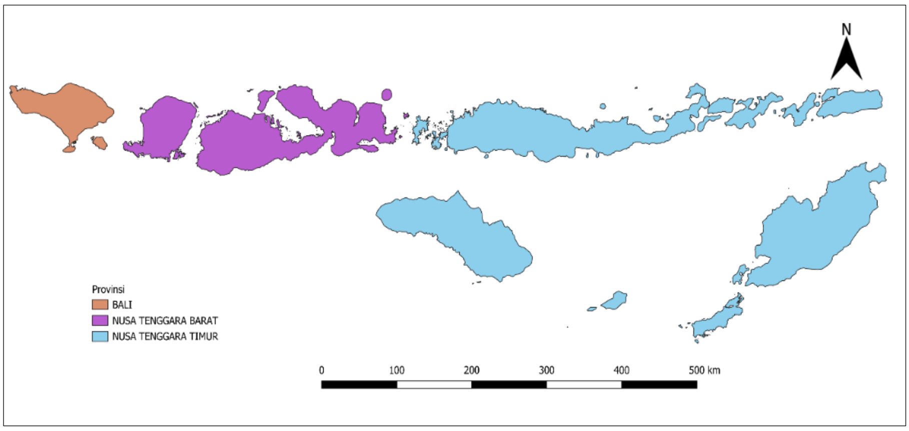


Figure 4. Map of Bali, NTB, And NTT Provinces

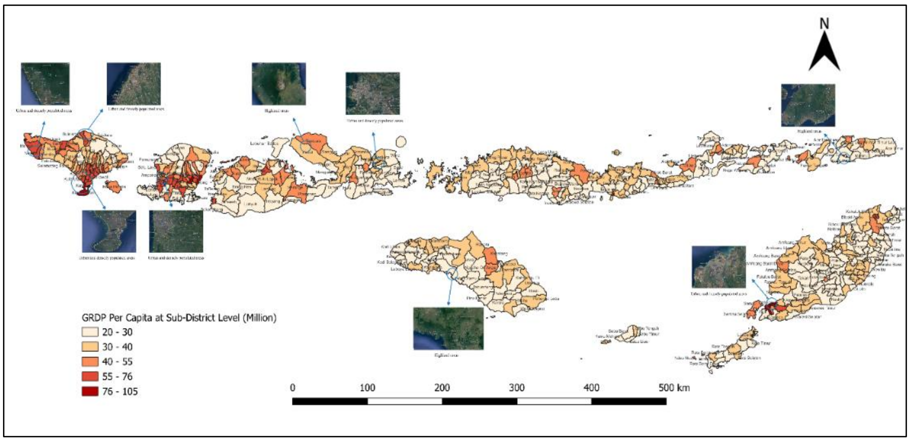


Figure 5. GRDP per capita for Districts in Bali, NTB, and NTT

It can be identified in the image that the districts in the western part have a higher GRDP per capita compared to the GRDP per capita in the eastern part, so the further east the district is, the lower the GRDP per capita value. Urban areas that serve as centers of economic activity have a high GRDP per capita, such as in the southern district of Bali Province, specifically Kuta District, which has a GRDP per capita of 90.31 million. The Cakranegara District in the city of Mataram, NTB Province, which is an urban area, also has a high GDP per capita of 55.05 million. The Kelapa Lima District in the city of Raja, NTT, which is an urban and densely populated area, has a high GRDP per capita of 105.29 million. In addition, districts that are dominated by highlands have a low GRDP per capita, such as Alor Barat Daya District, which has a GDP per capita value of 25.86 million.

Overall, areas dominated by highlands are estimated to be lower than those in the surrounding urban areas. Regions with high GRDP per capita have high values of NTL, CO, and NO_2 as well. The region also has a significant number of manufacturing industries, a number of villages with kiosks selling agricultural production tools, and a large number of healthcare facilities. In addition, the region also has warmer nighttime temperatures compared to areas with low GRDP per capita.

3.5. *Mapping Regional Inequality*

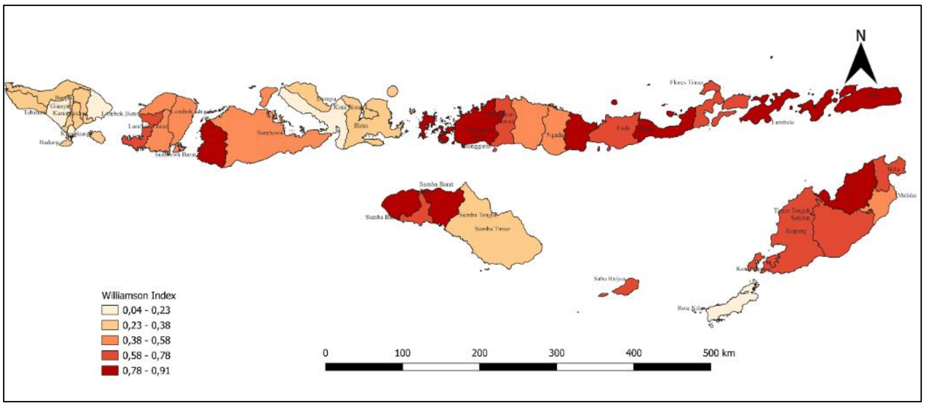


Figure 6. Distribution Map of the Williamson Index in Bali, NTB, and NTT

Figure 6 shows the mapping results of the Williamson Index by district in the provinces of Bali, West Nusa Tenggara (NTB), and East Nusa Tenggara (NTT). It can be seen that the eastern region has a higher disparity between areas compared to the western region. In Bali Province, Denpasar City has the lowest level of regional inequality within the area, with a Williamson Index value of 0.0382, while Klungkung Regency has the highest level of regional inequality with a Williamson Index value of 0.3571. In NTB Province, Mataram City has the lowest level of regional inequality within the area, with a Williamson Index of 0.1859, while West Sumbawa Regency has the highest level of inequality with a Williamson Index value of 0.8076. In NTT Province, Rote Ndao Regency has the lowest level of inequality with a Williamson Index value of 0.22, while West Manggarai Regency has the highest level of inequality with a Williamson Index value of 0.9106.

4. Conclusion

Overall, the estimation model for GRDP per capita at the sub-district level, using linear regression and machine learning, produced a MAPE of less than 50%, indicating that the model forecasts with reasonable accuracy. The support vector regression model was chosen to estimate GRDP per capita at the sub-district level because it resulted in the smallest MAE and MAPE, with values of 10.33 million and 26.11%, respectively. The estimation results indicate that high GRDP per capita is found in urban areas that serve as centers of economic activity, while low GRDP per capita is observed in highland areas with minimal economic activity. Regions with high GRDP per capita also exhibit high levels of NTL, CO, and NO2. These areas are characterized by a large number of manufacturing industries, numerous villages with kiosks selling agricultural production inputs, and a high concentration of healthcare facilities. Additionally, these regions experience warmer nighttime temperatures compared to areas with lower GRDP per capita. Analysis using the Williamson Index reveals that the eastern region has a higher level of inequality compared to the western region.

Ethics approval

Not required.

Acknowledgments

Not required.

Competing interests

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

Funding

This study received no external funding.

Underlying data

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

Credit Authorship

I Made Satria Ambara Putra: Data Curation, Visualization, Writing-Original Draft Preparation. **Rindang Bangun Prasetyo:** Methodology, Investigation, Supervision. **Candra Adi Wiguna:** Validation, Writing-Reviewing and Editing.

References

[1] BPS, Gross Domestic Product of Indonesia by Expenditure 2019-2023. Jakarta: Badan Pusat Statistik, 2024.

[2] M. Nasution, "Ketimpangan Antar Wilayah & Hubungannya dengan Belanja Pemerintah: Studi di Indonesia [Regional Disparities & Their Relationship with Government Spending: A Study in Indonesia]," *J. budg.*, vol. 5, no. 2, pp. 84–102, Nov. 2020, doi: 10.22212/jbudget.v5i2.101.

[3] Dinas Komunikasi dan Informatika Kota Depok, *Indikator Ekonomi Kecamatan Kota Depok 2018 [Economic Indicators of Subdistricts in Depok City, 2018]*, 2019.

[4] R. Capello, "Regional Economics, 0 ed." *Routledge*, 2015. doi: 10.4324/9781315720074.

[5] N. M. Coe, P. F. Kelly, and H. W.-C. Yeung, *Economic geography: a contemporary introduction, Third edition*. Hoboken, NJ: Wiley-Blackwell, 2020.

[6] BPS Kab. Sleman, *Produk Domestik Regional Bruto Kecamatan di Kabupaten Sleman 2016 [Gross Regional Domestic Product of Subdistricts in Sleman Regency, 2016]*, Kabupaten Sleman: Sleman: Badan Pusat Statistik Kabupaten Sleman, 2016.

[7] Sjafrizal, *Ekonomi wilayah dan perkotaan, Cetakan ke-1 [Regional and Urban Economy, 1st Edition]*, Jakarta: PT RajaGrafindo Persada, 2012.

[8] BPS Provinsi Jawa Tengah, *Analisis Indeks Williamson Provinsi Jawa Tengah 2017-2021 [Analysis of the Williamson Index in Central Java Province, 2017-2021]*, Jawa Tengah: Semarang: Badan Pusat Statistik Jawa Tengah, 2021.

[9] S. Pratama, "Prediksi Harga Tanah Menggunakan Algoritma Linear Regression [Land Price Prediction Using the Linear Regression Algorithm]," *Technologia*, vol. 7, no. 2, Jun. 2016, doi: 10.31602/tji.v7i2.624.

[10] V. L. Delimah Pasaribu et al., "Forecast Analysis of Gross Regional Domestic Product based on the Linear Regression Algorithm Technique," *TEM Journal*, pp. 620–626, May 2021, doi: 10.18421/TEM102-17.

[11] S. C. Agu, F. U. Onu, U. K. Ezemagu, and D. Oden, "Predicting gross domestic product to macroeconomic indicators," *Intelligent Systems with Applications*, vol. 14, p. 200082, May 2022, doi: 10.1016/j.iswa.2022.200082.

[12] T. M. H. Hope, "Linear regression, in Machine Learning," *Elsevier*, 2020, pp. 67–81. doi: 10.1016/B978-0-12-815739-8.00004-3.

[13] N. Puttanapong, N. Prasertsoong, and W. Peechapat, "Predicting Provincial Gross Domestic Product Using Satellite Data and Machine Learning Methods: A Case Study of Thailand," *Asian Development Review*, vol. 40, no. 02, pp. 39–85, Sep. 2023, doi: 10.1142/S0116110523400024.

[14] N. D. Muchisha, N. Tamara, A. Andriansyah, and A. M. Soleh, "Nowcasting Indonesia's GDP Growth Using Machine Learning Algorithms," *IJSA*, vol. 5, no. 2, pp. 355–368, Jun. 2021, doi: 10.29244/ijsa.v5i2p355-368.

[15] A. Richardson, T. Van Florenstein Mulder, and T. Vehbi, "Nowcasting GDP using machine-learning algorithms: A real-time assessment," *International Journal of Forecasting*, vol. 37, no. 2, pp. 941–948, Apr. 2021, doi: 10.1016/j.ijforecast.2020.10.005.

[16] S. Sa’adah and M. S. Wibowo, "Prediction of Gross Domestic Product (GDP) in Indonesia Using Deep Learning Algorithm, in 2020 3rd International Seminar," *Research of Information Technology and Intelligent Systems (ISRITI),* Yogyakarta, Indonesia: IEEE, Dec. 2020, pp. 32–36. doi: 10.1109/ISRITI51436.2020.9315519.

[17] H. Lai, "A comparative study of different neural networks in predicting gross domestic product," *Journal of Intelligent Systems*, vol. 31, no. 1, pp. 601–610, May 2022, doi: 10.1515/jisys-2022-0042.

[18] J. Saputra, B. Subartini, J. H. F. Purba, S. Supian, and Y. Hidayat, *An Application of Genetic Algorithm Approach and Cobb-Douglas Model for Predicting the Gross Regional Domestic Product by Expenditure-Based* in Indonesia, 2019.

[19] A. F. Syah, "Penginderaan Jauh dan Aplikasinya di Wilayah Pesisir dan Lautan [Remote Sensing and Its Applications in Coastal and Marine Areas]," *Jurnal Kelautan*, vol. 3, pp. 18–28, 2010.

[20] Z. Wang et al., "Exploring the Potential of OpenStreetMap Data in Regional Economic Development Evaluation Modeling," *Remote Sensing*, vol. 16, no. 2, p. 239, Jan. 2024, doi: 10.3390/rs16020239.

[21] K. Faisal and A. Shaker, "The Use of Remote Sensing Technique to Predict Gross Domestic Product (GDP): An Analysis of Built-Up Index and GDP in Nine Major Cities in Canada," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. XL–7, pp. 85–92, Sep. 2014, doi: 10.5194/ isprsarchives-XL-7-85-2014.

[22] S. R. Putri, A. W. Wijayanto, and S. Pramana, "Multi-source satellite imagery and point of interest data for poverty mapping in East Java, Indonesia: Machine learning and deep learning approaches," *ScienceDirect*, Accessed: Jul. 04, 2024. [Online]. Available: https://www.sciencedirect.com/ science/article/abs/pii/S2352938522001975

[23] BPS Kab. Musi Rawas, *Produk Domestik Regional Bruto Kabupaten Musi Rawas Menurut Pengeluaran 2016-2020 [Gross Regional Domestic Product of Musi Rawas Regency by Expenditure, 2016-2020]*, Kabupaten Musi Rawas: Musi Rawas: Badan Pusat Statistik Kabupaten Musi Rawas, 2021.

[24] BPS Kab. Badung, *Produk Domestik Regional Bruto Kabupaten Badung Menurut Lapangan Usaha [Gross Regional Domestic Product of Badung Regency by Business Sector]*, Kabupaten Badung: Badung: Badan Pusat Statistik Kabupaten Badung, 2022.

[25] M. Urbac, A. Junaidi, M. Syukur, N. Nurhamidah, and R. Ferial, "Kajian Aspek Geospasial Untuk Percepatan Pembangunan dan Pemberdayaan Desa Binaan Kota Padang [Study of Geospatial Aspects for Accelerating Development and Empowering Fostered Villages in Padang City]," *JLBI*, vol. 12, no. 4, pp. 198–204, Dec. 2023, doi: 10.32315/jlbi.v12i4.83.

[26] D. Reynard, "Five Classes of Geospatial Data and The Barriers to Using Them," *Geography Compass*, vol. 12, no. 4, p. e12364, Apr. 2018, doi: 10.1111/gec3.12364.

[27] P. Pan, R. Li, and Y. Zhang, "Predicting punching shear in RC interior flat slabs with steel and FRP reinforcements using Box-Cox and Yeo-Johnson transformations," *Case Studies in Construction Materials*, vol. 19, p. e02409, Dec. 2023, doi: 10.1016/j.cscm.2023.e02409.

[28] B. T. Suryanto, A. A. Imron, and D. A. R. Prasetyo, "The Correlation between Students Vocabulary Mastery and Speaking Skill," *ijoeel*, vol. 3, no. 1, pp. 10–19, Jun. 2021, doi: 10.33650/ijoeel.v3i1, 2042.

[29] G. Zeng, "A Unified Definition of Mutual Information with Applications in Machine Learning," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–12, 2015, doi: 10.1155/2015/201874.

[30] P. Laarne, M. A. Zaidan, and T. Nieminen, "ennemi: Non-linear correlation detection with mutual information," *SoftwareX*, vol. 14, p. 100686, Jun. 2021, doi: 10.1016/j.softx.2021.100686.

[31] J. Zhao, Y. Zhou, X. Zhang, and L. Chen, "Part mutual information for quantifying direct associations in networks," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 113, no. 18, pp. 5130–5135, May 2016, doi: 10.1073/pnas.1522586113.

[32] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *AJAMS*, vol. 8, no. 2, pp. 39–42, Jun. 2020, doi: 10.12691/ajams-8-2-1.

[33] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.

[34] I. Cholissodin, *Buku Ajar AI, Machine Learning & Deep Learning [AI, Machine Learning & Deep Learning textbook]*, Malang: Malang: Filkom Universitas Brawijaya, 2020.

[35] P.-C. Chang, Y.-W. Wang, and C.-H. Liu, "The development of a weighted evolving fuzzy neural network for PCB sales forecasting," *Expert Systems with Applications*, pp. 86–96, 2007.

DOI: https://doi.org/10.34123/jurnalasks.v15i2.442

POLITEKNIK STATISTIKA STIS

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

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

JASKS Jurnal Aplikasi Statistika & Komputasi Statistik

TEMPLATE

Click Here, Type the Title of Your Paper, Capitalize First Letter of Each Word (Times New Romans (TNR), Size 17 pt, exactly spacing at 20 pt, 12 pt spacing for next heading, left alignment))

First Author Name[1*], Second Author Name[2] , Third Author Name[3] (TNR font, size 13 pt, exactly spacing at at 15 pt and 8 pt spacing for the next heading.)

[1]First Affiliation, City, Country, [2,3]Author Affiliation, City, Country [2]Second Affiliation, City, Country, [2,3]Author Affiliation, City, Country
**Corresponding Author: E-mail address: author@institute.xxx*
(TNR font, size 10 pt, with single spacing and 0 pt spacing for the next heading. And for the corresponding author use 10 pt Times New Roman font with single spacing and 8 pt spacing for the next heading.)

ARTICLE INFO

Article history: (TNR, 10pt)
Received dd month, yyyy
Revised dd month, yyyy
Accepted dd month, yyyy
Published dd month, yyyy

Keywords: (TNR, 10 pt)
Type your keywords here, separated by semicolon (;) Capitalize first letter of each word, times new roman, use 10 pt and write alphabetically in 5-10 words

Abstract (Times New Roman, size font 12)

Introduction/Main Objectives: Describe the topic your paper examines. Provide a background to your paper and why is this topic interesting. Avoid unnecessary content. **Background Problems:** State the problem or statistical applied/statistic computing phenomena studied in this paper and specify the research question(s) in one sentence. **Novelty:** Summarize the novelty of this paper. Briefly explain why no one else has adequately researched the question yet. **Research Methods:** Provide an outline of the research method(s) and data used in this paper. Explain how did you go about doing this research. Again, avoid unnecessary content and do not make any speculation(s). **Finding/Results:** List the empirical finding(s) and write a discussion in one or two sentences. Abstract written in English, with a length of 150 - 200 words. Use 10 pt Times New Roman font with justified alignment, single spacing, and 1 pt spacing for the next heading.

1. Main Text (bold, TNR, 14 pt, spacing before- after 12 pt, line spacing 12 pt)

These instructions give you guidelines for preparing papers for Jurnal Aplikasi Statistika & Komputasi Statistik which is published by Politeknik Statistika STIS, effective from the June 2024 edition. Starting from June 2024 Volume 16 No. 1, please use the template available at the following link https://s.stis.ac.id/TemplateJurnalASKS. The paragraphs continue from here and are only separated by headings, subheadings, images and formulae. The section headings are arranged by numbers, bold and 14 pt. Here follow further instructions for authors.

The manuscript was created using Microsoft Office Word only and should be formatted for direct printing. As indicated in the template, manuscript should be prepared in single column format that suitable for direct printing onto paper with A4 paper size (21 x 29.7 cm). All parts of the manuscript are typed in Times New Roman font, size 11, line spacing exactly at 12 pt, with 0.2 line spacing for the next heading and margins of 3 cm of left and 2 cm for top, bottom, and right, the length of header from the top is 1.5 cm and the length of footer from the bottom is 1 cm. For the main text, use justify alignment and special indent for the first line in 0.76 cm. For the purpose of editing the manuscript, all parts of the manuscript (including tables, figures and mathematical equations) are made in a format that can be edited by the editor [1, 2].

The writing style for the Jurnal Aplikasi Statistika & Komputasi Statistik is written in English with a narrative style. Tracing is kept simple and as far as possible avoiding multilevel chronology.

2.1Structure

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.

2.2Tables

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

2.3Construction 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

2.4Section 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).

2.5General 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.

2.6Footnotes

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].

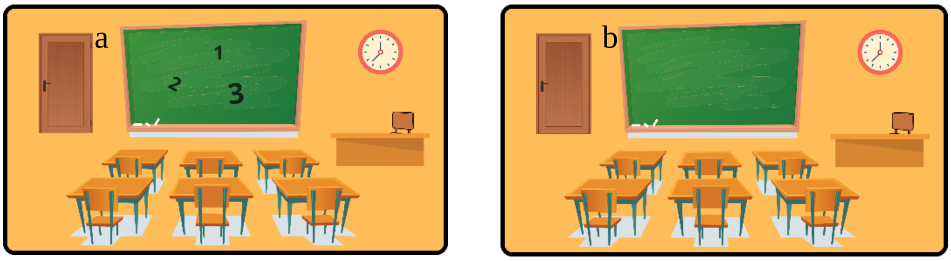


Figure. 1. (a) first picture; (b) second picture.

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.”

Credit Authorship

....

References

[1] W.K. Chen, *Linear Networks and Systems*. Belmont, CA: Wadsworth Press, 2003.

[2] R. Hayes, G. Pisano, and S. Wheelwright, *Operations, Strategy, and Technical Knowledge*. Hoboken, NJ: Wiley, 2007.

[3] K. A. Nelson, R. J. Davis, D. R. Lutz, and W. Smith, “Optical generation of tunable ultrasonic waves,” *J Appl Phys*, vol. 53, no. 2, pp. 1144–1149, Feb. 2002.