Jurnal Matematika dan Statistika serta Aplikasinya, Vol. 13, No. 2, 98-106, 2025
DOI: https://doi.org/10.24252/msa.v13i2.60319

Nonparametric Model For Poverty Data: The Effect of Internal Factors Using Multi-Predictor Spline Regression in Indonesia

Ruliana[1]*, Rahmat Hidayat[2], Hardianti Hafid[3], Sudarmin[4]
[1,2,3,4] *Statistics Study Program, Universitas Negeri Makassar, Indonesia*
***Corresponding author**: ruliana.t@unm.ac.id

***Submission date**: 30 July 2025, **Revision**: 15 October 2025, **Accepted**: 02 November 2025

ABSTRACT
Poverty, as a multidimensional issue affecting national welfare and development, is the main focus of this research. This study investigates the impact of demographic and educational factors on the percentage of the poor population in Indonesia using a nonparametric Spline regression approach. The variables studied include the average population growth rate, the availability of schools in villages, and school enrollment rates. The best model, selected based on the lowest Generalized Cross Validation (GCV) value (0.204) and a high coefficient of determination (94.67%) is a nonparametric Spline regression model with an optimal combination of knot points. The analysis shows that all three predictor variables significantly influence the poverty rate. The model also meets standard statistical assumptions. These findings highlight the vital role of education and demographic factors in addressing poverty, thus strengthening education and controlling population growth should be a priority in poverty alleviation policies in Indonesia.

KEYWORDS
Poverty, GCV, nonparametric, Spline

1. INTRODUCTION

Every individual faces various life challenges, both personal and sociocultural. Among these challenges, poverty is a crucial issue. Socially, poverty has significant implications for national development. [1] even identifies poverty as a fundamental problem that must be addressed within the context of development. From an individual perspective, poverty is often defined as the inability to meet basic needs, and can even extend to the failure to achieve personal aspirations or desires.

The percentage of the poor population is an essential measure that shows the portion of individuals with expenditure levels below the poverty line relative to the total population in an area [2], [3]. This indicator plays a vital role in assessing the effectiveness of poverty mitigation programs and in assessing the level of social welfare. Referring to the criteria of the Central Statistics Agency (BPS), an individual is classified as poor if their average per capita expenditure per month does not reach the poverty line, which is calculated based on minimum needs standards for both food and non-food consumption. The high percentage of the poor population indicates that the community still has limited access to education, health services, decent work, and social protection, thus impacting the overall low quality of life. Therefore, measuring and analyzing the percentage of the poor population is very relevant in various studies of social and economic development.

The increase in poverty is caused by various factors. These factors need to be understood to address poverty in Indonesia and to inform the government's decision-making in future policies. To identify the factors contributing to the high Poverty Severity Index in Indonesia, a relevant statistical method was applied.

The average population growth rate has the potential to significantly impact the percentage of poor people in a region. High

population growth, if not accompanied by increased employment opportunities, resource availability, and access to public services, can increase economic pressures, particularly for low-income groups. Population growth tends to lead to fiercer competition for jobs, education, and healthcare, potentially trapping some people in disadvantaged economic conditions [4], [5]. Therefore, regions with high population growth but not accompanied by inclusive economic development have the potential to experience higher poverty rates.

The availability of educational facilities that are evenly distributed throughout villages is a crucial factor in poverty alleviation efforts. The more villages that have school facilities, the greater the opportunities for rural communities to access formal education [6]. More accessible education can improve the quality of human resources, open access to more decent jobs, and reduce the dropout rate, which is one of the causes of structural poverty. Therefore, the number of villages with school facilities acts as a driving factor for improving welfare, which can indirectly reduce the percentage of the poor population.

School participation rates reflect the extent to which children and adolescents participate in formal education in a region. High school participation rates indicate that a large portion of the population has access to education, which is a crucial foundation for social and economic mobility [7]. A good education opens up opportunities for individuals to obtain more productive and higher-paying jobs, thus enabling them to escape the cycle of poverty. Therefore, high school participation rates in a region are usually negatively correlated with the percentage of the poor population, as education is key to breaking the intergenerational cycle of poverty.

Most studies examining the determinants of poverty are still dominated by parametric regression approaches, such as multiple linear regression, which assumes a specific relationship between predictor and response variables. However, in socioeconomic contexts, relationships between variables are often nonlinear and complex, so parametric approaches can produce inaccurate or biased results if the model assumptions are not met. The aim of this study is to identify factors that influence poverty levels.

The nonparametric approach used in this study offers the advantage of capturing flexible relationship patterns without requiring a specific functional form [8],[9]. Thus, this model allows for a more realistic exploration of the relationship between the percentage of the poor population and its predictor variables, such as population growth rate, the number of villages with school facilities, and school enrollment rates. The use of this method has also not been widely adopted in empirical studies on poverty in Indonesia, making it a significant methodological contribution to the literature.

Furthermore, another gap this research addresses is the limitation of previous studies in simultaneously integrating demographic and educational factors within a single, flexible analysis model. By utilizing a multivariate nonparametric regression approach, this research is able to evaluate the contribution of each factor to poverty levels more comprehensively and adapt to actual data patterns. This provides a more accurate picture and provides a stronger basis for formulating data-driven poverty alleviation policies.

2. METHODOLOGY

This study involves several phases of analysis, starting with identifying data patterns through dispersion plots. The variables studied include the average population growth rate, the availability of schools in villages, and school enrollment rates. Next, data modeling is performed using a Truncated Spline model. This method examines models with one, two, or three knots, as well as combinations of knots. To find the best knot point, the minimum Generalized Cross Validation (GCV) value is used [10], [11], [12]. In the next stage, parameter testing is carried out both simultaneously and partially. Next, a Spline regression model is used to test residual assumptions. Finally, the goodness-of-fit of the obtained model is examined based on the coefficient of determination value.

3. RESULT & DISCUSSION

3.1 Data Exploration

Before modeling is carried out, data exploration is first performed by creating a scatter plot. A scatter plot illustrates the pattern of the relationship between two variables. After obtaining a general overview of the percentage of poor population, an analysis is conducted between the response variable and each predictor variable to identify the relationship patterns among variables.

The figure above shows that the relationship between the percentage of poor population and its predictor variables does not form any specific pattern. Based on this observation, modeling the percentage of poor population is therefore more appropriate

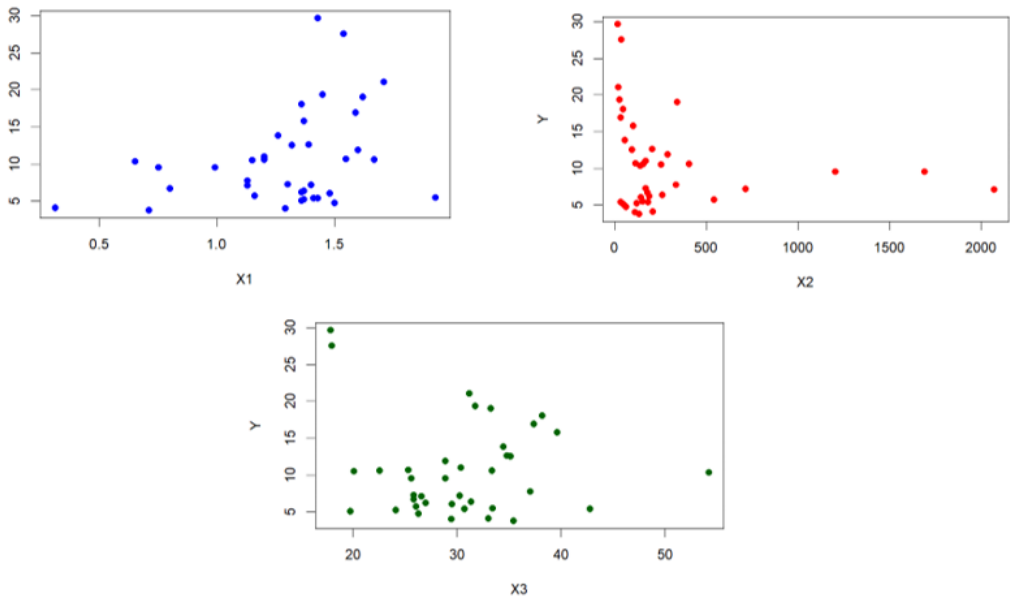

Figure 1. Scatter Plot

using a nonparametric regression approach.

3.2 One Point Knot Trial

The Spline model with one knot point using three variables that are suspected of influencing the response variable is presented as follows.

$$\hat{y} = \hat{\beta}_{00} + \hat{\beta}_{11}X_2 + \hat{\beta}_{12}(X_2 - K_{11})_+^1 + \hat{\beta}_{21}X_3 + \hat{\beta}_{22}(X_3 - K_{21})_+^1 + \hat{\beta}_{31}X_4 + \hat{\beta}_{32}(X_4 - K_{31})_+^1 \quad (1)$$

The results of the analysis using statistical software are presented in **Table 1**:

Table 1. GCV Values with Three Variables and One Knot Point

No.	x_1	x_2	x_3	GCV
1	13.91	9.30	6.37	0.590
2	12.75	8.88	9.12	0.512
3	11.46	9.62	9.66	0.992
4	12.85	8.60	8.13	0.928
5	11.09	8.50	9.64	0.968
6	12.81	8.97	7.30	0.666
7	11.96	9.50	8.67	0.841
8	13.18	8.49	8.61	0.985
9	13.67	8.77	7.95	**0.443***
10	12.25	9.70	6.77	0.539

Based on **Table 1**, the Nonparametric Spline Regression model with one knot point produces a minimum GCV value of 0.443. This optimal value is achieved by determining one optimal knot point for each predictor variable: 13.67 for the average population growth rate (x_1), 8.77 for the number of villages with school facilities (x_2), and 7.95 for the school enrollment rate (x_3).

3.3 Two-Point Knot Trial

To estimate the percentage of the poor population, a spline model with two knot points is used. The model estimation is as follows:

$$\hat{y} = \hat{\beta}_{00} + \hat{\beta}_{11}X_2 + \hat{\beta}_{12}(X_2 - K_{11})_+^1 + \hat{\beta}_{13}(X_2 - K_{12})_+^1 + \hat{\beta}_{21}X_3 + \hat{\beta}_{22}(X_3 - K_{21})_+^1 + \hat{\beta}_{23}(X_3 - K_{22})_+^1 + \hat{\beta}_{31}X_4 + \hat{\beta}_{32}(X_4 - K_{31})_+^1 + \hat{\beta}_{33}(X_4 - K_{32})_+^1 \quad (2)$$

The GCV values for the Spline model with two knots are around the minimum GCV value, as shown in Table 3.1. For the Nonparametric Spline Regression model using two knots on all three predictor variables, the minimum GCV value is 0.419. For each variable, we find two optimal knot points. For variable x_1, the average population growth rate is 1.32 and 1.05; for variable x_2, the number of villages with school facilities is 254.49 and 224.34; and for variable x_3, the number of school enrollments is 24.17 and 39.84.

Table 2. Two-Point Knot GCV Values

No.	x_1	x_2	x_3	x_1	x_2	GCV
1	1.79	1.18	180.27	202.18	25.59, 39.59	0.504
2	1.69	2.00	487.00	484.00	39.85, 28.67	0.847
3	1.83	1.84	20.16	196.45	39.10, 39.31	0.451
4	1.85	1.74	83.29	96.63	30.48, 34.69	0.474
5	1.22	1.21	390.56	421.23	26.18, 29.86	0.458
6	1.60	1.02	177.48	165.28	27.55, 21.16	0.589
7	1.32	1.05	254.49	224.34	24.17, 39.84	**0.419***
8	1.23	1.15	414.63	344.77	21.23, 38.58	0.618
9	1.36	1.05	230.51	465.87	34.94, 20.55	0.492
10	1.46	1.84	235.17	364.25	21.79, 25.98	0.466

Table 2 presents the results of the two-knot Nonparametric Spline Regression modeling, which shows a minimum GCV value of 0.419.

3.4 Optimal Knot Point with Three Knot Points

The following presents a Nonparametric Spline Regression model that estimates the percentage of poor people using three knot points and three predictor variables.

$$\hat{y} = \hat{\beta}_{00} + \hat{\beta}_{11}X_2 + \hat{\beta}_{12}(X_2 - K_{11})_+^1 + \hat{\beta}_{13}(X_2 - K_{12})_+^1 + \hat{\beta}_{14}(X_2 - K_{13})_+^1 + \hat{\beta}_{21}X_3 + \hat{\beta}_{22}(X_3 - K_{21})_+^1 + \hat{\beta}_{23}(X_3 - K_{22})_+^1 + \hat{\beta}_{24}(X_3 - K_{23})_+^1 + \hat{\beta}_{31}X_4 + \hat{\beta}_{32}(X_4 - K_{31})_+^1 + \hat{\beta}_{33}(X_4 - K_{32})_+^1 + \hat{\beta}_{34}(X_4 - K_{33})_+^1 \quad (3)$$

The five GCV values for the three-point knot Spline model are around the minimum GCV value, as shown in **Table 3**.

Table 3. Three-Point Knot GCV Values (Three Variables)

No.	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{3,1}$	$x_{3,2}$	GCV
1	1.74	1.98	1.20	54.89	323.47	38.20	29.48	38.94, 26.52	0.863
2	1.52	1.78	1.32	244.29	7.15	351.77	30.81	32.24, 38.46	0.432
3	1.45	1.08	1.24	488.20	325.51	111.07	29.13	21.83, 31.80	**0.256***
4	1.05	1.22	1.69	182.66	240.22	409.43	30.19	27.00, 24.21	0.270
5	1.56	1.34	1.49	146.88	398.64	39.37	30.80	29.46, 20.24	0.751

3.5 Optimal Knot Point with Knot Point Combination

After conducting trials with one, two and three knot points, another possibility is a model with variations or combinations of knots for each predictor variable.

Table 4. GCV Values of Three Variable Knot Point Combinations

No.	x_1	x_2	x_3	GCV
1	1.66	257.88	21.35	0.278
	1.09	444.69	28.24	
2	1.75	117.22	21.47	0.233
	1.89	431.99	21.85	
			21.06	
3	1.52	265.91	34.21	0.432
	1.72	213.94		
		192.76		
4	1.54	496.73	28.80	**0.204***
	1.35	114.49	32.29	
		251.30	25.22	
5	1.57	95.34	27.76	0.467
	1.74	337.08	34.71	
		139.82		
6	1.75	100.21	31.16	0.386
	1.05			
	1.36			
7	1.55	2.10	24.03	0.369
	1.41		31.66	
	1.09			

The table indicates a minimum GCV of 0.204 resulting from the 2-3-3 knot point combination. The optimal knot point configurations for each predictor variable are: 1.54 and 1.35 for the average population growth rate (x_1); 496.73, 114.49, and 251.30 for the number of villages with school facilities (x_2); and 28.80, 32.29, and 25.22 for the school enrollment rate (x_3).

The GCV values of the Nonparametric Spline Regression model with the observed knot point combinations are close to the minimum GCV values, as shown in Table 3.4. The minimum GCV value of 0.204 results from the 2-3-3 knot point combination, and the optimal knot point configurations for each predictor variable are: 1.54 and 1.35 for the average population growth rate (x_1); 496.73, 114.49, and 251.30 for the number of villages with school facilities (x_2); and 28.80, 32.29, and 25.22 for the school participation rate (x_3).

3.6 Selecting the Best Knot Point

The best models obtained from the previous step were then compared to determine the best model. This table clearly shows that the model with the knot point combination provided the lowest GCV value, at 0.204. Based on the criteria for selecting the

Table 5. GCV Values for Each Model Test

Model	GCV
1 knot point	0.443
2 knot points	0.419
3 knot points	0.256
Combination of knot points	**0.204***

best model, the Nonparametric Spline Regression model with a combination of knot points produces the minimum GCV value, namely 0.204.

3.7 Simultaneous Testing of Three Predictor Variables

To determine the collective influence of the predictor variables on the response variable, simultaneous testing was conducted. The statistical test results showed an F-value of 493 and a p-value of 3×10^{25}. With a significance level of 5%, a p-value smaller than α strengthens the rejection of H_0. This proves that the three predictor variables simultaneously have a significant influence on the percentage of the poor population in Indonesia.

3.8 Partial Testing of Three Predictor Variables

The results of the simultaneous tests indicate that at least one parameter in the Spline model is significant. The results of each test are depicted in **Table 6**.

Table 6. Three-Variable Regression Parameter Estimates

Var.	Par.	Est.	p-value	Ket.
	β_{00}	-1.00	0.84	Not significant
x_1	β_{11}	0.21	0.00	Significant
	β_{12}	1.55	0.01	Significant
	β_{13}	-2.31	0.02	Significant
x_2	β_{21}	4.12	0.76	Not significant
	β_{22}	55.19	0.00	Significant
	β_{23}	-135.51	0.00	Significant
	β_{24}	23.76	0.00	Significant
x_3	β_{31}	0.64	0.91	Not significant
	β_{32}	44.81	0.00	Significant
	β_{33}	-71.53	0.00	Significant
	β_{34}	53.67	0.00	Significant

According to **Table 6**, all predictor variables, include the average population growth rate, the availability of schools in villages, and school enrollment rates, demonstrated statistical significance in the model. A p-value lower than five percent indicates that all these variables significantly influence the percentage of the poor population.

The results indicate that the following nonparametric spline regression model is the most suitable for estimating parameters using Ordinary Least Squares (OLS).

$$\begin{aligned}\hat{y} = &-1 + 0,21X_2 + 1,55\,(X_2 - 1,54)^1_+ - 2,31\,(X_2 - 1,74)^1_+ + 4,12X_3 + 55,19\,(X_3 - 496,73)^1_+ - 135,51\,(X_3 - 114,49)^1_+ \\ &+ 23,76\,(X_3 - 251,30)^1_+ + 0,64X_4 + 44,81\,(X_4 - 28,80)^1_+ - 71,53\,(X_4 - 32,29)^1_+ + 53,67\,(X_4 - 25,22)^1_+\end{aligned} \tag{4}$$

3.9 Residual Assumption Testing

To validate the model, residual assumption testing was performed to ensure that the residuals were identical, independent, and normally distributed.

3.10 Identical Assumption Test (Homoscedasticity)

The results of the Glejser test show a p-value of 0.183. Therefore, it can be concluded that the residuals generated by the model meet the identical assumption.

3.11 Testing the Independent Residual Assumption

To meet the validity of the model, the residual independence assumption needs to be verified. This test aims to ensure there is no correlation between residuals. Run Test was applied to perform this test. The calculation results show a p-value of 0.631. At a significance level (α) of 0.05, a p-value greater than α results in failure to reject H_0. Therefore, it can be confirmed that the residual independence assumption is met, indicating that the residuals are random and free from autocorrelation.

3.12 Testing the Assumption of Normal Distribution Residuals

The normality of the residual distribution is the third essential assumption. The Kolmogorov-Smirnov test is applied to evaluate this assumption. A visual representation of the normality of the model's residuals can be seen through the Normal Probability Plot shown in the **Figure 2**.

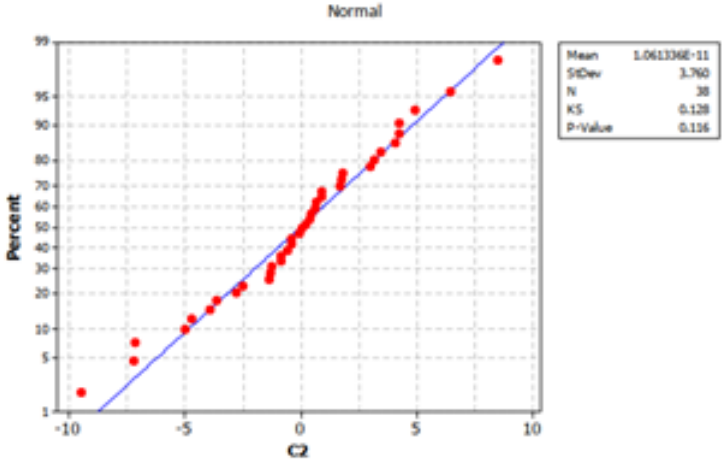

Figure 2. Normal Probability Plot of Residuals

Visual verification through **Figure 2** shows that the residuals are normally distributed, indicated by the conformity of the residual pattern to the normal line. For statistical confirmation, the Kolmogorov-Smirnov test produces a p-value of 0.116. At a significance level of $\alpha = 0.05$, a p-value greater than α fails to reject H_0, thus proving that the residuals of the model are normally distributed. Considering that the three residual assumptions (identical, independent, and normality) have been met, the developed model is valid to explain the relationship between the predictor variables and the response variable.

3.13 Coefficient of Determination

As a measure of model goodness-of-fit, the coefficient of determination reflects the proportion of variability in the percentage of poor people that can be explained by the regression model. The modeling results yielded a coefficient of determination of 94.67%. This means the resulting Nonparametric Spline Regression model significantly explains 94.67% of the variability in the percentage of poor people, with the remaining variability explained by variables outside the model. This value confirms the excellent quality of the model.

This study aims to analyze the factors influencing the percentage of poor people in Indonesia using a nonparametric Spline regression approach. Simultaneously, three predictor variables, namely the average population growth rate, the number of villages with school facilities, and the school enrollment rate, were proven to have a significant effect on the percentage of poor people, as indicated by a high F-value and a p-value ¡0.05. Partial testing also showed that most of the model parameters had a significant effect, indicating that each variable has a contribution in explaining variations in poverty levels. For example, the school enrollment rate variable showed a consistently significant effect, reinforcing previous findings that education has a crucial role in breaking the chain of poverty.

Methodologically, the use of nonparametric spline regression in this study provides an important contribution because it can accommodate nonlinear relationship patterns. This is an advantage over traditional parametric regression approaches, which assume linear relationships and homogeneity of variance. This approach provides more realistic estimation results and is more appropriate for complex socioeconomic phenomena.

In addition, the resulting model has met the three basic assumptions of regression, namely: (1) identical residuals (homoscedasticity), (2) independent residuals (no autocorrelation), and (3) normally distributed residuals. The Glejser Test, Run Test, and Kolmogorov-Smirnov all produced p-values > 0.05, indicating no violation of these assumptions. This confirms that the constructed regression model is suitable for use as a basis for decision-making and scientific interpretation.

4. CONCLUSION

This study aims to analyze the factors influencing the percentage of poor people in Indonesia using a nonparametric spline regression approach. The three predictor variables analyzed are the average population growth rate, the number of villages with school facilities, and the school enrollment rate. Based on the analysis, a nonparametric spline regression model with specific knot point settings for each predictor variable was identified as the most optimal model. This model achieved a minimum Generalized Cross Validation (GCV) value of 0.204. Furthermore, with a coefficient of determination of 94.67%, the model demonstrated high capability in explaining most of the variation in the percentage of the poor population.

Statistically, the three predictor variables were shown to have a significant effect, both simultaneously and partially, on poverty levels. Furthermore, the model met all residual assumptions: residuals were identical, independent, and normally distributed, making it suitable for interpretation and policy recommendations.

Substantively, the results of this study confirm that demographic and educational factors, particularly access to school facilities and educational participation rates, play a crucial role in poverty reduction. Therefore, strengthening equitable educational development, particularly in rural and densely populated areas, is a strategic key to poverty alleviation. This research still has room for development, particularly in incorporating other relevant variables such as healthcare access, income inequality, and infrastructure quality. Furthermore, testing the model on panel or spatial data could provide a more comprehensive picture of poverty dynamics in Indonesia.

ACKNOWLEDGMENTS

This research was funded by PNBP of Makassar State University with Contract Number 339/UN36.II/TU/2025.

REFERENCES

[1] Ruliana, I. N. Budiantara, B. W. Otok, and W. Wibowo, "Parameter estimation of nonlinear structural model sem using spline approach," *Applied Mathematical Sciences*, vol. 9, no. 149, pp. 7439–7451, 2015.

[2] D. Amaliah, "Pengaruh partisipasi pendidikan terhadap persentase penduduk miskin," *Faktor: Jurnal Ilmiah Kependidikan*, vol. 2, no. 3, 2016.

[3] A. Hadi, "Pengaruh rata-rata lama sekolah kabupaten/kota terhadap persentase penduduk miskin kabupaten/kota di provinsi jawa timur tahun 2017," *Media Trend*, vol. 14, no. 2, pp. 148–153, 2019.

[4] D. Desmawan, F. Fitrianingsih, N. A. Drajat, N. W. Diani, and S. Marlina, "Pengaruh jumlah penduduk terhadap pertumbuhan ekonomi di kabupaten tangerang tahun 2019-2020," *Jurnal Penelitian Ekonomi Manajemen dan Bisnis*, vol. 2, no. 2, pp. 150–157, 2023.

[5] A. Azulaidin, "Pengaruh pertumbuhan penduduk terhadap pertumbuhan ekonomi," *Juripol (Jurnal Institusi Politeknik Ganesha Medan)*, vol. 4, no. 1, pp. 30–34, 2021.

[6] J. Setiawan, R. Jaenudin, and S. Fatimah, "Pengaruh biaya pendidikan dan fasilitas pendidikan terhadap hasil belajar mata pelajaran ekonomi peserta didik sma bukit asam tanjung enim," *Jurnal Profit*, vol. 2, no. 1, pp. 14–27, 2015.

[7] R. M. Shari and J. Abubakar, "Pengaruh pertumbuhan penduduk, angka partisipasi sekolah dan tingkat partisipasi angkatan kerja terhadap pertumbuhan ekonomi pada 5 provinsi di indonesia," *Jurnal Ekonomi Regional Unimal*, vol. 5, no. 2, pp. 20–32, 2022.

[8] R. Hidayat, I. N. Budiantara, B. W. Otok, and V. Ratnasari, "The regression curve estimation by using mixed smoothing spline and kernel (mss-k) model," *Communications in Statistics-Theory and Methods*, vol. 50, no. 17, pp. 3942–3953, 2021.

[9] R. Hidayat, M. Ilyas, and Y. Yuliani, "Spline model in the case of cervical cancer patient resilience," *Trends in Sciences*, vol. 20, no. 8, pp. 6170–6170, 2023.

[10] N. Syamsualam and R. Hidayat, "Application of truncated spline nonparametric regression in modeling traffic accident rate in palopo city," *Journal of Applied Mathematics and Computation*, vol. 7, no. 2, pp. 185–196, 2022.

[11] R. Hidayat, M. Ilyas, Y. Yuliani, S. Sifriyani, and D. Denysia, "Mathematical model of the unemployment rate with multiple spline regression," in *AIP Conference Proceedings*, vol. 3235, no. 1, 2024, p. 020003.

[12] R. Hidayat, I. N. Budiantara, B. W. Otok, and V. Ratnasari, "A reproducing kernel hilbert space approach and smoothing parameters selection in spline-kernel regression," *Journal of Theoretical and Applied Information Technology*, vol. 97, no. 2, pp. 465–475, 2019.