Res. Math. Nat. Sci. vol. 6, no. 1, pp. 1–11, 2026, doi:10.55657/rmns.v6i1.375

Research in the Mathematical and Natural Sciences
Journal Homepage: https://journal.scimadly.com/index.php/rmns
eISSN: 2828-6804

Original Research

Modeling the Human Development Index of Regencies/Cities in South Sulawesi in 2024 Using Truncated Spline Nonparametric Regression

Ahmad Zaki[1], Irwan[1,*], and Nurhalisa Ishak[1]

[1]*Department of Mathematics, Universitas Negeri Makassar, Makassar 90224, Indonesia*

Article information

***Corresponding author:**
Irwan
irwanthaha@unm.ac.id

Received: 26 September 2026
Revised: 19 November 2026
Accepted: 23 November 2026
Published: 29 November 2026

Licensed under CC BY-NC-SA 4.0

Published by:
scimadly PUBLISHING

Abstract

The Human Development Index (HDI) of regencies/cities in South Sulawesi ranged from 66.38 to 83.90 in 2024. This study models the relationship between the HDI and three factors that are not HDI components, namely population density, number of poor people, and labor force size, for 24 regencies/cities using Statistics Indonesia data. A linear truncated spline nonparametric regression on log-transformed predictors was applied, with knots selected by minimum Generalized Cross Validation (GCV) subject to at least four observations per segment. The best model is a one-knot linear spline (GCV = 7.80; R^2 = 71.72%; adjusted R^2 = 61.74%). Only population density is significantly associated with the HDI (p = 0.028 from a bootstrap that repeats knot selection). Above about 525 persons/km^2, a 10% increase in density is associated with an HDI increase of about 0.53 points (95% CI: 0.29–0.78), whereas below this threshold the evidence of an association is insufficient. These findings persist when Makassar City is excluded and when count predictors are replaced by rates. Residual normality is not rejected, but the residuals show spatial autocorrelation (Moran's I = 0.196; p = 0.026), so inference should be interpreted with caution.

Keywords: Nonparametric Regression; Truncated Spline; Generalized Cross Validation; Human Development Index; South Sulawesi

1. Introduction

The Human Development Index (HDI) measures human development achievement along three basic dimensions: a long and healthy life, knowledge, and a decent standard of living. Statistics Indonesia (BPS) measures these dimensions using life expectancy, expected and mean years of schooling, and real per capita expenditure [1]. The HDI is also one of the main indicators of community welfare at the regional level [2]. In South Sulawesi Province, the HDI of regencies/cities in 2024 ranged from 66.38 (Jeneponto) to 83.90 (Makassar City). This disparity calls for an understanding of the regional factors associated with human development achievement.

Regression analysis is a common tool for this purpose. Parametric regression requires the form of the regression curve to be known in advance. In contrast, nonparametric regression only assumes that the regression

curve is smooth, that is, it belongs to a certain function space, so that its shape is determined by the data [3, 4]. Approaches developed for this purpose include kernel and local polynomial estimators [5] and spline estimators [6, 7].

A truncated spline consists of polynomial pieces joined at knots, that is, points at which the pattern of the relationship changes [8, 9]. The coefficients of a linear spline can be read directly as segment slopes, which makes it suitable for detecting threshold effects. The optimal knot locations are usually selected by the Generalized Cross Validation (GCV) criterion, which is asymptotically optimal and does not require knowledge of the error variance [10].

Spline regression has been applied to various types of data, for example crime rates [9] and rainfall [11], including for modeling the HDI. Ramadhan et al. [12] modeled the HDI across Indonesian provinces with expected years of schooling, life expectancy, and population size as predictors (three knots; R^2 = 83.07%). Safitri et al. [13] compared truncated spline and Gaussian kernel regression for the provincial HDI with expected years of schooling, mean years of schooling, and the percentage of poor people as predictors (spline R^2 = 92.79%). Permatasari et al. [14] applied it to the HDI on Kalimantan Island (one knot; R^2 = 42.98%).

For South Sulawesi, three studies are the most relevant. Asmira et al. [15] modeled the HDI of 24 regencies/cities in 2015 with five predictors, namely the labor force participation rate, the pupil ratio, population density, health facilities, and GRDP; their best model used three knots with R^2 = 80.29%. Meilinda et al. [16] used life expectancy, mean years of schooling, population percentage, morbidity rate, and GRDP (three knots; R^2 = 90.94%) and found that only GRDP was significant. Amarta et al. [17] used life expectancy, mean years of schooling, and GRDP (three knots; R^2 = 93.07%) and found that only life expectancy was significant.

This comparison reveals two methodological weaknesses. First, most studies use HDI components (life expectancy, mean years of schooling, or expected years of schooling) as predictors, so the relationships obtained are partly definitional and R^2 is artificially high. This is evident from R^2 values above 90% in [13, 16, 17], all of which include HDI components, compared with 80.29% in [15], which does not. Second, with only 24 observations, three-knot models for three to five predictors contain 13 to 21 parameters and are therefore prone to placing knots in segments containing only one or two observations. Their inference also does not account for the fact that the knots are selected from the same data. This study addresses these gaps by (i) using only predictors outside the HDI components, namely population density, number of poor people, and labor force size; (ii) selecting knots by GCV subject to a minimum number of observations per segment; and (iii) complementing inference with a bootstrap that repeats knot selection, influence diagnostics, and a test for spatial autocorrelation. The objectives are to obtain the best truncated spline model for the HDI of regencies/cities in South Sulawesi in 2024, to identify the factors associated with the HDI, and to assess the uncertainty of the threshold at which the association changes.

2. Methods

2.1 Data and Research Variables

This study uses secondary data for 2024 from BPS-Statistics of South Sulawesi Province [1]. The observation units are 24 regencies/cities, so $n = 24$. The research variables are presented in Table 1. The HDI used is the new-method HDI published by BPS-Statistics of South Sulawesi Province. Life expectancy, mean years of schooling, and real per capita expenditure are not used as predictors because they are components of the HDI. For the sensitivity analysis, the percentage of poor people and the labor force participation rate (LFPR) in 2024 from the same source are also used.

Table 1: Research variables.

Symbol	Variable	Unit
Y	Human Development Index (HDI)	index
X_1	Population density	persons/km^2
X_2	Number of poor people	thousand persons
X_3	Labor force size	persons

All three predictors are strongly right-skewed because of the extreme values of Makassar City (Table 2). Therefore, the modeling is carried out on $z_{ji} = \ln x_{ji}$. This transformation spreads the candidate knots evenly and allows the slopes to be interpreted as the change in HDI per relative (percentage) increase in a predictor.

2.2 Truncated Spline Nonparametric Regression

Given the data $(z_{1i}, z_{2i}, z_{3i}, y_i)$, $i = 1, \ldots, n$, the additive nonparametric regression model is

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(z_{ji}) + \varepsilon_i, \qquad \varepsilon_i \sim \text{iid } N(0, \sigma^2). \tag{1}$$

If f_j is approximated by a truncated spline of order m with r_j knots $K_{j1}, \ldots, K_{jr_j}$, then [8, 9]

$$f_j(z_{ji}) = \sum_{k=1}^{m} \beta_{jk} z_{ji}^k + \sum_{u=1}^{r_j} \beta_{j(m+u)} (z_{ji} - K_{ju})_+^m. \tag{2}$$

$$(z_{ji} - K_{ju})_+^m = \begin{cases} (z_{ji} - K_{ju})^m, & z_{ji} \geq K_{ju}, \\ 0, & z_{ji} < K_{ju}. \end{cases} \tag{3}$$

This study uses a linear spline ($m = 1$). Thus, β_{j1} is the slope of the first segment, and $\beta_{j(1+u)}$ is the change in slope after knot K_{ju}.

2.3 Parameter Estimation

In matrix notation, the model in equations (1) and (2) is written as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The matrix $\mathbf{X}$ has size $n \times (q + 1)$ with $q = \sum_j (m + r_j)$, and its columns contain the constant, z_{ji}, and $(z_{ji} - K_{ju})_+$. The least squares estimator is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \qquad \mathbf{A}(K) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{4}$$

2.4 Selection of Optimal Knots

The knots are selected by minimizing [9, 10]

$$\text{GCV}(K) = \frac{n^{-1} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2}{\left[n^{-1} \operatorname{tr}\left(\mathbf{I} - \mathbf{A}(K)\right)\right]^2}. \tag{5}$$

The candidate knots for each predictor are the 10th, 15th, …, 90th percentiles of z_j (17 candidates). Models with one, two, and three knots are evaluated over all candidate combinations, using the same percentile level for all three predictors. Combinations with different numbers of knots across predictors are then also evaluated ($r_j \in \{1, 2, 3\}$, giving $3^3 = 27$ combinations). To prevent knots from being determined by only one or two observations, only configurations in which every segment contains at least four observations ($\approx n/6$) are considered. For comparison, the GCV of parametric linear regression is also computed on both the original and the logarithmic scale. The number of observations in the smallest segment of each model is reported in Table 3. All computations were performed in Python 3 (NumPy, SciPy, and statsmodels).

2.5 Parameter Testing

Simultaneous test. The hypotheses are $H_0 : \beta_1 = \cdots = \beta_q = 0$ versus H_1: at least one $\beta_l \neq 0$, with the test statistic

$$F = \frac{SSR/q}{SSE/(n - q - 1)}. \tag{6}$$

H_0 is rejected if $F > F_{\alpha; q, n-q-1}$.

Partial test. Each parameter is tested with $t = \widehat{\beta}_l / SE(\widehat{\beta}_l)$, and $H_0 : \beta_l = 0$ is rejected if $|t| > t_{\alpha/2; n-q-1}$. Because one predictor is represented by several parameters, the effect of the j-th predictor is tested with the partial F test $H_0 : \beta_{j1} = \cdots = \beta_{j(1+r_j)} = 0$:

$$F_j = \frac{\left(SSE_{(-j)} - SSE\right)/(1 + r_j)}{SSE/(n - q - 1)}, \tag{7}$$

where $SSE_{(-j)}$ is the residual sum of squares of the model without the j-th predictor.

Inference after knot selection. The t and F tests above treat the knots as fixed, whereas the knots are selected from the same data through GCV, so the p-values tend to be too small [18]. Therefore, a residual bootstrap with $B = 2{,}000$ replicates was used, repeating the entire knot-selection procedure in every replicate. First, the p-value for the effect of population density that accounts for knot selection is computed as the proportion of replicates under H_0 (data generated from the model without population density) that yield a partial F-test p-value as small as or smaller than the observed value. Second, a 95% percentile interval for the slope of the density function is computed from replicates under the best model.

2.6 Residual Assumption Tests

Identical. Homoscedasticity is tested with the Glejser test [19], that is, by regressing $|e_i|$ on the same matrix $\mathbf{X}$:

$$F = \frac{\sum_{i=1}^{n} \left(\widehat{|e_i|} - \overline{|e|}\right)^2 / q}{\sum_{i=1}^{n} \left(|e_i| - \widehat{|e_i|}\right)^2 / (n - q - 1)}. \tag{8}$$

Independence. Because the data are cross-sectional, the ordering of the observations is not meaningful, so an ACF plot is not relevant. Dependence between regions is examined with Moran's index [20] on the residuals,

$$I = \frac{n}{S_0} \frac{\sum_i \sum_j w_{ij} e_i e_j}{\sum_i e_i^2},$$

where w_{ij} are row-standardized k-nearest-neighbor weights ($k = 4$) based on the distance between the capitals of the regencies/cities and $S_0 = \sum_i \sum_j w_{ij}$. Significance is assessed with a permutation test (9,999 permutations).

Normality. Because the parameters of the normal distribution are estimated from the residuals, the Kolmogorov–Smirnov test is used with the Lilliefors correction [21] and supplemented by the Shapiro–Wilk test. A non-significant result is interpreted as insufficient evidence to reject normality, not as evidence that the residuals are normally distributed.

Influence diagnostics. Influential observations are identified with the leverage h_{ii} (the diagonal of matrix $\mathbf{A}(K)$), Cook's distance $D_i = (r_i^2/p)\, h_{ii}/(1 - h_{ii})$, where r_i is the standardized residual and $p = q + 1$ [22], and the externally studentized residuals. The reference thresholds are $h_{ii} > 2p/n$ and $D_i > 4/n$. The influence of Makassar City is also assessed by re-estimating the model without Makassar.

2.7 Goodness of Fit

The goodness of fit of the selected model is evaluated using the coefficient of determination (R^2) and the adjusted coefficient of determination (R^2_{adj}). The R^2 measures the proportion of total variation in the HDI explained by the model, whereas R^2_{adj} adjusts this measure for the number of parameters included in the model. These measures are calculated as follows:

$$R^2 = \frac{SSR}{SST}, \qquad R^2_{\text{adj}} = 1 - \frac{SSE/(n - q - 1)}{SST/(n - 1)}. \tag{9}$$

Multicollinearity among predictors is examined with the variance inflation factor (VIF). Because the number of poor people and the labor force size both depend on regional population size, a sensitivity analysis is also carried out by replacing both with rates, namely the percentage of poor people and the LFPR (without logarithmic transformation), using the same knot-selection procedure.

3. Results and Discussion

3.1 Descriptive Statistics

Descriptive statistics are presented in Table 2. The lowest HDI is in Jeneponto Regency (66.38) and the highest is in Makassar City (83.90). Makassar City also has the highest values of all three predictors. Its density

reaches 8,282 persons/km^2, far above the median of 219.5 persons/km^2, so the standard deviation of density is more than twice its mean.

Table 2: Descriptive statistics of the research variables (n = 24).

Variable	Mean	Standard deviation	Minimum	Maximum
HDI (Y)	72.79	3.80	66.38 (Jeneponto)	83.90 (Makassar)
Population density (X_1)	673.21	1,659.44	45 (North Luwu)	8,282 (Makassar)
Poor people (X_2)	30.69	18.27	7.94 (Parepare)	79.53 (Makassar)
Labor force (X_3)	203,570	138,540	74,971 (Selayar Islands)	696,851 (Makassar)

The scatter plots in Figure 1 show that the relationship between the HDI and population density is not linear. The HDI is relatively flat in low-density regencies and then jumps in three cities (Makassar, Parepare, and Palopo). The relationships of the HDI with the number of poor people and labor force size show no clear pattern.

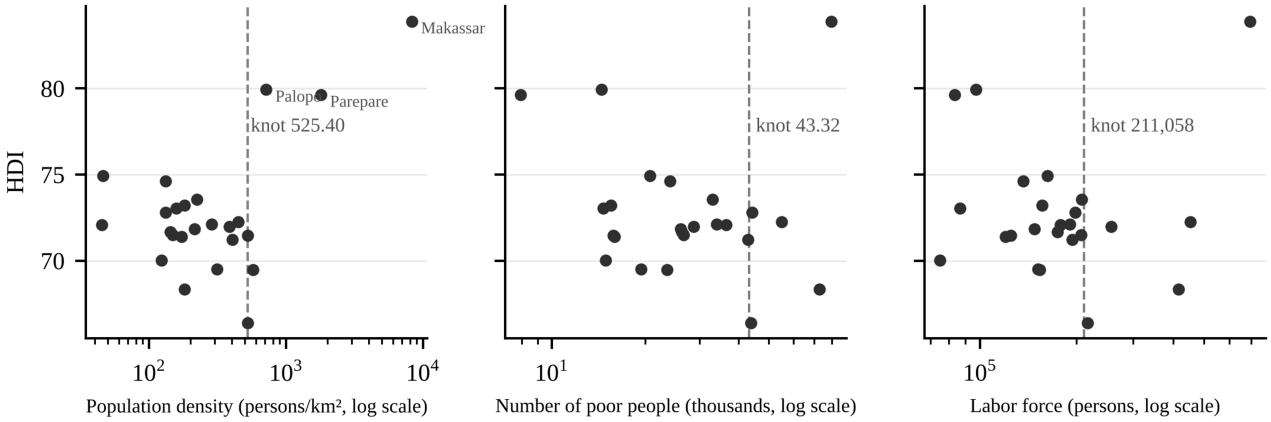


Figure 1: Scatter plots of the HDI against each predictor (x-axis on a logarithmic scale); dashed lines indicate the optimal knots.

3.2 Selection of the Best Model

The minimum GCV value for each model is presented in Table 3.

Table 3: Comparison of minimum GCV values and smallest segment sizes (requirement: every segment ≥ 4 observations).

Model	Number of parameters	Minimum GCV	Smallest segment (observations)
Parametric linear regression (original scale)	4	8.00	–
Parametric linear regression (log scale)	4	13.05	–
Linear spline, 1 knot (original scale)	7	8.85	6
Linear spline, 1 knot (log scale)	7	7.80	5 (19; 5)
Linear spline, 2 knots (log scale)	10	11.19	4
Linear spline, 3 knots (log scale)	13	16.32	4
Next best knot combination, (1, 1, 2)	8	8.60	4

The one-knot linear spline on the logarithmic scale yields the smallest GCV (7.80). This value is smaller than the GCV of parametric linear regression (8.00) and of the spline model on the original scale. Adding knots increases the GCV, because the additional parameters are not offset by the reduction in error at n = 24. The optimal knot lies at the 80th percentile of each predictor, so each predictor is divided into a lower segment with 19 observations and an upper segment with 5 observations. On the original scale, the knots are 525.40 persons/km^2 for population density, 43.32 thousand persons for the number of poor people, and 211,057.62 persons for labor force size.

3.3 Parameter Estimation and Testing

The results of the simultaneous test are presented in Table 4. The value $F = 7.19$ is greater than $F_{0.05;\,6;\,17} = 2.70$ with p-value $= 0.0006$, so H_0 is rejected. That is, at least one parameter is significantly associated with the HDI.

Table 4: Results of the simultaneous test.

Source	df	SS	MS	F	p-value
Regression	6	238.220	39.703	7.185	0.0006
Error	17	93.934	5.526		
Total	23	332.154			

The parameter estimates and t tests are presented in Table 5 ($t_{0.025;\,17} = 2.110$). Of the six parameters other than the constant, only β_{12} is significant. This parameter is the change in the slope of $\ln X_1$ after the knot.

Table 5: Parameter estimates and partial tests.

Variable	Parameter	Estimate	SE	t	p-value
Constant	β_0	81.450	31.465	2.589	0.019
$\ln X_1$	β_{11}	−1.010	0.746	−1.354	0.193
	β_{12}	6.622	1.542	4.295	< 0.001
$\ln X_2$	β_{21}	−1.839	2.338	−0.787	0.442
	β_{22}	−11.724	10.287	−1.140	0.270
$\ln X_3$	β_{31}	0.151	3.124	0.048	0.962
	β_{32}	5.820	6.572	0.886	0.388

The results of the partial F test for each predictor (Table 6) are consistent with the t tests. Only population density is significant ($F = 10.82 > F_{0.05;\,2;\,17} = 3.59$).

Table 6: Partial F test for each predictor.

Predictor	df	F_j	p-value	Decision
Population density	(2; 17)	10.815	< 0.001	Significant
Number of poor people	(2; 17)	1.418	0.269	Not significant
Labor force size	(2; 17)	0.590	0.565	Not significant

With $z_j = \ln x_j$, the best model obtained is

$$\widehat{y} = 81.450 - 1.010\, z_1 + 6.622\, (z_1 - 6.264)_+ - 1.839\, z_2 - 11.724\, (z_2 - 3.769)_+ + 0.151\, z_3 + 5.820\, (z_3 - 12.260)_+. \tag{10}$$

3.4 Residual Assumption Tests

Identical. The Glejser test yields $F = 0.998$ with $df = (6;\, 17)$ and p-value $= 0.458 > 0.05$. Thus, H_0 is not rejected and there is no evidence of heteroscedasticity.

Independence. Moran's index of the residuals is $I = 0.196$ (its expected value under H_0 is $-1/(n-1) = -0.043$) with a permutation p-value of 0.026. Similar results are obtained for $k = 3$, 5, and 6 neighbors ($I = 0.170$–0.258; all p-values < 0.03). Thus, the residuals show positive spatial autocorrelation: negative residuals cluster in the southern regencies (Selayar Islands, Jeneponto, Takalar, Sinjai), whereas positive residuals cluster in the central and northern parts (Pinrang, Enrekang, Luwu, Palopo). The independence assumption is not fully met, so the standard errors and p-values in Tables 5 and 6 are likely too small and should be interpreted with caution.

Normality. The Kolmogorov–Smirnov test with the Lilliefors correction yields $D = 0.148$ (p-value > 0.20), and the Shapiro–Wilk test yields $W = 0.922$ (p-value $= 0.064$). Neither is significant at $\alpha = 0.05$, so there is insufficient evidence to reject the normality of the residuals. However, the Shapiro–Wilk p-value close to 0.05 and the small sample size make this conclusion weak. The Q-Q plot (Figure 2) shows that most residuals lie close to the reference line, except for Palopo City, which has the largest residual (6.26).

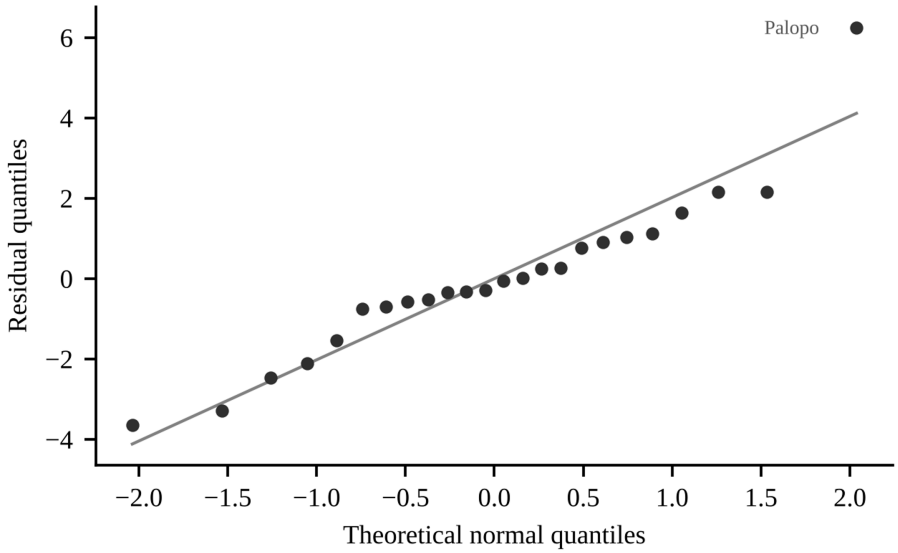


Figure 2: Q-Q plot of the residuals of the best model.

Influence diagnostics. Figure 3 presents the leverage and Cook's distance of each observation. Makassar City has a very high leverage ($h_{ii} = 0.906 > 2p/n = 0.583$) and the largest Cook's distance ($D_i = 0.693$), although it is still below 1. Bone Regency ($h_{ii} = 0.919$) and Gowa Regency ($h_{ii} = 0.778$) also have high leverage because they belong to the upper segments of the number of poor people and labor force size, which contain only five observations. Palopo City is an outlier in the response direction (studentized residual 4.16; Bonferroni p-value = 0.018), whereas the Selayar Islands have a Cook's distance of 0.384. When Makassar City is excluded and the knots are kept, the slope of ln population density above the knot increases to 7.203 ($SE = 2.585$) and the density effect remains significant (partial F test, p-value = 0.038), although R^2 decreases to 55.18% (Table 7). If the knots are reselected without Makassar, the density knot is almost the same (525.70 persons/km^2), but the coefficients of the other predictors become unstable. Hence, the direction and significance of the density association do not depend on Makassar, but the magnitude of the upper-segment slope is sensitive to influential observations.

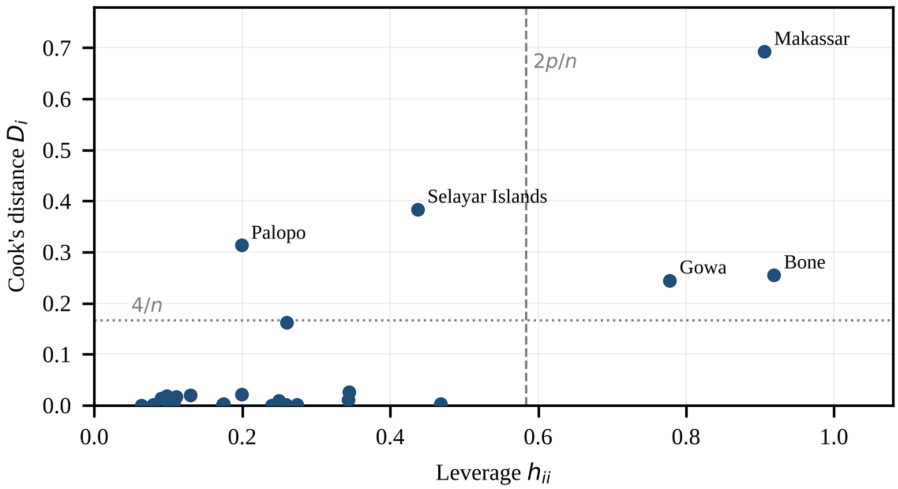


Figure 3: Leverage and Cook's distance of each regency/city; the dashed line is the $2p/n$ threshold and the dotted line is the $4/n$ threshold.

3.5 Goodness of Fit

The values are $R^2 = 238.220/332.154 = 71.72\%$ and $R^2_{\text{adj}} = 61.74\%$. That is, the model explains 71.72% of the variation in the HDI across regencies/cities in South Sulawesi in 2024. This value is lower than those of models that include HDI components as predictors, namely 93.07% in Amarta et al. [17], 92.79% in Safitri et

al. [13], and 90.94% in Meilinda et al. [16], and lower than the 80.29% in Asmira et al. [15], who used five predictors with 21 parameters. This is expected because the present model uses only three factors outside the definition of the HDI with seven parameters. The VIF values (on the log scale) are 1.32 for density, 6.19 for the number of poor people, and 6.79 for labor force size. The correlation between ln number of poor people and ln labor force size reaches 0.90 because both depend on regional population size.

3.6 Inference after Knot Selection and Sensitivity Analysis

The results of the bootstrap that repeats knot selection are presented in Table 7. After accounting for knot selection, the p-value for the effect of population density increases from 0.001 to 0.028, but it remains significant at $\alpha = 0.05$. The slope of the ln density function at a density of 716 persons/km^2 (Palopo) is positive in 99.8% of replicates, with a median of 5.46 and a 95% percentile interval of [2.12; 8.93]. In contrast, the knot location is unstable: the one-knot configuration for all three predictors is selected in only 14.1% of replicates, and a density knot around 525–550 persons/km^2 is selected in 30.8% of replicates. Hence, the positive association of density in densely populated areas is fairly robust, but the threshold location of about 525 persons/km^2 should be regarded as approximate.

Table 7: Inference after knot selection and sensitivity analysis of the population density effect.

Analysis	Density knot (persons/km^2)	Slope of ln X_1 above the knot	p-value for density	R^2
Main model (n = 24)	525.40	5.613 (SE 1.213)	0.001[a]	71.72%
Bootstrap with knot reselection (B = 2,000)	varies	5.46 [2.12; 8.93][b]	0.028[c]	–
Without Makassar City (n = 23)	525.40	7.203 (SE 2.585)	0.038[a]	55.18%
Rate predictors (% poor people, LFPR)	466.90	4.448 (SE 1.010)	0.002[a]	66.87%

Note. [a] Partial F test; [b] median and 95% percentile interval of the slope at a density of 716 persons/km^2; [c] p-value of the partial F test calibrated against the bootstrap distribution under H_0 with knot selection repeated in every replicate.

The sensitivity analysis with rate predictors is also presented in Table 7. Replacing the number of poor people and labor force size with the percentage of poor people and the LFPR reduces the VIF to 1.10–1.31 and the correlation between the two predictors to 0.30. A one-knot spline model is again selected (GCV = 9.14, smaller than that of linear regression, GCV = 12.33) with R^2 = 66.87%. Population density remains the only significant predictor (partial F test, p-value = 0.002), with a knot at 466.90 persons/km^2 and a slope above the knot of 4.448 (SE = 1.010), whereas the percentage of poor people (p-value = 0.310) and the LFPR (p-value = 0.800) are not significant. Thus, the main finding is not caused by collinearity due to population size.

3.7 Discussion

The slope of each segment, together with its standard error and 95% confidence interval (CI), is presented in Table 8. The standard error of the upper-segment slope is computed from

$$\mathrm{Var}(\widehat{\beta}_{j1} + \widehat{\beta}_{j2}) = \mathrm{Var}(\widehat{\beta}_{j1}) + \mathrm{Var}(\widehat{\beta}_{j2}) + 2\ \mathrm{Cov}(\widehat{\beta}_{j1}, \widehat{\beta}_{j2}).$$

On the logarithmic scale, a slope b means that a 10% increase in a predictor is associated with a change in the HDI of $b \times \ln(1.1) \approx 0.095b$ points, holding the other predictors constant.

Table 8: Slopes of the relationship between each predictor (ln scale) and the HDI, standard errors (SE), and 95% CI.

Predictor	Knot (original scale)	Lower segment (n = 19): slope (SE) [95% CI]	Upper segment (n = 5): slope (SE) [95% CI]	Regencies/cities above the knot
Population density	525.40	−1.010 (0.746) [−2.58; 0.56]	5.613 (1.213) [3.05; 8.17]	Jeneponto, Takalar, Makassar, Parepare, Palopo
Poor people	43.32	−1.839 (2.338) [−6.77; 3.09]	−13.563 (9.756) [−34.15; 7.02]	Jeneponto, Gowa, Bone, Luwu, Makassar
Labor force	211,057.62	0.151 (3.124) [−6.44; 6.74]	5.971 (5.537) [−5.71; 17.65]	Bulukumba, Jeneponto, Gowa, Bone, Makassar

3.7.1 Population Density

Population density is the only factor significantly associated with the HDI, and the pattern is of a threshold type. Below 525.40 persons/km^2, the slope is small and not significant (−1.010; 95% CI [−2.58; 0.56]; p-value = 0.193). This means that there is insufficient evidence of an association in this segment, not evidence that

no association exists; the confidence interval still includes changes in the HDI from −0.25 to 0.05 points per 10% increase in density. Above this threshold, the slope is 5.613 (95% CI [3.05; 8.17]), so a 10% increase in density is associated with an HDI increase of about 0.53 points (95% CI: 0.29–0.78 points). The five regions above the threshold consist of three cities (Makassar, Parepare, Palopo) and two regencies (Jeneponto, Takalar). Because this study does not include an urban/rural status variable, this pattern is only consistent with, but does not prove, the role of the concentration of education, health, and economic facilities in urban areas. Jeneponto and Takalar also exceed the density threshold, but their HDI is relatively low, so density alone does not explain HDI achievement. This finding is in line with Asmira et al. [15], who also found population density to be associated with the HDI of South Sulawesi in 2015, and complements Meilinda et al. [16], who found GRDP to be the only significant factor; density and GRDP both reflect the concentration of economic activity.

3.7.2 Number of Poor People and Labor Force Size

Neither predictor is significantly associated with the HDI, either per parameter or jointly. There are two explanations. First, both are measured as absolute counts and therefore reflect regional population size more than the poverty level or labor participation. Second, the high correlation between them ($r = 0.90$; VIF > 6) inflates the standard errors, making their individual associations difficult to separate. The sensitivity analysis with the percentage of poor people and the LFPR (Table 7) removes this collinearity problem, but both rates remain non-significant. The large slopes above the knot (−13.563 and 5.971) are supported by only five observations and have very wide confidence intervals (Table 8), so they are not interpreted substantively.

3.7.3 Methodological Implications

Without the minimum requirement of four observations per segment, the GCV on the original scale can fall to 4.07 with $R^2 = 88.5\%$. However, the knots in that model separate only one or two observations (for example, Takalar), so what is modeled are individual points rather than a pattern. In small samples, GCV-based knot selection must be constrained to prevent the spline model from overfitting. In addition, p-values that ignore knot selection are too optimistic (0.001 compared with 0.028 after the bootstrap), so inference from spline models in small samples should be supplemented by procedures that account for knot selection. In line with Ramadhan et al. [12] and Safitri et al. [13], the truncated spline captures changes in the pattern of the relationship. However, the optimal number of knots here is only one, not three as in [12, 15–17], which is consistent with the small sample size.

4. Conclusion

The best model for the HDI of regencies/cities in South Sulawesi in 2024 is a linear truncated spline nonparametric regression with one knot on the logarithms of the three predictors (knot at the 80th percentile). This model has a GCV of 7.80 (smaller than that of parametric linear regression, GCV = 8.00), $R^2 = 71.72\%$, and $R^2_{\text{adj}} = 61.74\%$. Of the three factors studied, only population density is significantly associated with the HDI, including after accounting for knot selection (bootstrap p-value = 0.028), after excluding Makassar City, and after replacing count predictors with rates. This association has a threshold pattern: above about 525 persons/km^2, a 10% increase in density is associated with an HDI increase of about 0.53 points, whereas below this threshold there is insufficient evidence of an association. The number of poor people and labor force size are not significantly associated with the HDI. Hence, higher HDI achievement in South Sulawesi is concentrated in high-density areas, especially the three cities. Extending education and health services to low-density regencies deserves attention.

This study has several limitations. First, the sample size is small ($n = 24$) and the data cover only one year, so the results are associational and cannot be interpreted as causal relationships. Second, the knot location is unstable across bootstrap replicates, so the threshold of 525 persons/km^2 is an approximation. Third, the residuals show positive spatial autocorrelation (Moran's $I = 0.196$; p-value = 0.026), so the standard errors are likely too small; spline models that account for spatial dependence or geographically weighted regression should be explored in future research. Fourth, Makassar City is an influential observation with very high leverage. Fifth, this study does not include an urban/rural status variable, so the mechanism behind the association between population density and the HDI cannot be tested directly.

Declaration of Generative AI Use

The authors used Claude (Anthropic) to assist with language editing, translation into English, and drafting code for the additional diagnostic analyses requested during peer review (post-selection bootstrap, influence diagnostics, Moran's I, and sensitivity analysis). The tool was used under the authors' direction and was not used to formulate the research questions or to draw the conclusions. The authors reviewed and verified all code, results, and text and take full responsibility for the content of this manuscript.

Author Contributions

Ahmad Zaki: Conceptualization, Validation, Supervision, Writing–review and editing. **Irwan**: Conceptualization, Methodology, Formal analysis, Validation, Supervision, Writing–review and editing. **Nurhalisa Ishak**: Data curation, Investigation, Formal analysis, Software, Visualization, Writing–original draft.

Acknowledgments

The authors thank BPS-Statistics of South Sulawesi Province for making the data used in this study publicly available.

Funding

This research received no external funding.

Data Availability

The data used in this study are publicly available from the statistical tables of BPS-Statistics of South Sulawesi Province (https://sulsel.bps.go.id). The compiled dataset and the Python code used for the analysis are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Badan Pusat Statistik Provinsi Sulawesi Selatan, *[Metode baru] Indeks Pembangunan Manusia (IPM) Menurut Kabupaten/Kota di Provinsi Sulawesi Selatan, 2024 [New-Method Human Development Index by Regency/City in South Sulawesi Province, 2024]*, URL: https://sulsel.bps.go.id/id/statistics-table/2/MzAyIzI=/-metode-baru–indeks-pembangunan-manusia–ipm-.html; accessed: Nov. 19, 2026.

[2] I. K. G. Sukarsa and I. G. K. Gandhiadi, "Indikator Kesejahteraan di Provinsi Bali: Suatu Pendekatan Analisis Biplot," *J. Mat.*, vol. 10, no. 1, p. 22, 2020, doi: 10.24843/JMAT.2020.v10.i01.p120.

[3] A. Islamiyati, Fatmawati, and N. Chamidah, "Penalized Spline Estimator with Multi Smoothing Parameters in Bi-Response Multi-Predictor Nonparametric Regression Model for Longitudinal Data," *Songklanakarin J. Sci. Technol.*, vol. 42, no. 4, pp. 897–909, 2020, doi: 10.14456/sjst-psu.2020.115.

[4] M. Setyawati, N. Chamidah, and A. Kurniawan, "Modelling Scholastic Aptitude Test of State Islamic Colleges in Indonesia Using Least Square Spline Estimator in Longitudinal Semiparametric Regression," *J. Phys.: Conf. Ser.*, vol. 1764, no. 1, p. 012 077, 2021, doi: 10.1088/1742-6596/1764/1/012077.

[5] T. W. Utami, N. Chamidah, T. Saifudin, B. Lestari, and D. Aydin, "Estimation of Biresponse Semiparametric Regression Model for Longitudinal Data Using Local Polynomial Kernel Estimator," *Symmetry*, vol. 17, no. 3, p. 392, 2025, doi: 10.3390/sym17030392.

[6] A. Islamiyati, Fatmawati, and N. Chamidah, "Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with Penalized Spline Regression," *J. Phys.: Conf. Ser.*, vol. 979, no. 1, p. 012 093, 2018, doi: 10.1088/1742-6596/979/1/012093.

[7] N. Chamidah, B. Lestari, H. Susilo, M. Y. Alsagaff, I. N. Budiantara, and D. Aydin, "Spline Estimator in Nonparametric Ordinal Logistic Regression Model for Predicting Heart Attack Risk," *Symmetry*, vol. 16, no. 11, p. 1440, 2024, doi: 10.3390/sym16111440.

[8] A. Islamiyati, "Regresi Spline Polynomial Truncated Biprediktor untuk Identifikasi Perubahan Jumlah Trombosit Pasien Demam Berdarah Dengue," *Al-Khwarizmi: J. Pendidik. Mat. Ilmu Pengetah. Alam*, vol. 7, no. 2, pp. 97–112, 2019, doi: 10.24256/jpmipa.v7i2.799.

[9] M. A. Juniar, A. Fania, D. Ulya, R. Ramadhan, and N. Chamidah, "Modelling Crime Rates in Indonesia Using Truncated Spline Estimator," *BAREKENG: J. Ilmu Mat. Terap.*, vol. 18, no. 2, pp. 1201–1216, 2024, doi: 10.30598/barekengvol18iss2pp1201-1216.

[10] P. Craven and G. Wahba, "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numer. Math.*, vol. 31, no. 4, pp. 377–403, 1979, doi: 10.1007/BF01404567.

[11] A. S. Yulianti, A. Islamiyati, and E. T. Herdiani, "The Principal Component Linear Spline Quantile Regression Model in Statistical Downscaling for Rainfall Data," *J. Sci. Islam. Repub. Iran*, vol. 35, no. 2, pp. 159–166, 2024, doi: 10.22059/jsciences.2024.375343.1007860.

[12] M. R. Ramadhan, Darnah, and S. Wahyuningsih, "Model Regresi Nonparametrik Spline Truncated pada Indeks Pembangunan Manusia di Indonesia," *J. EKSPONENSIAL*, vol. 14, no. 2, pp. 91–98, 2023, doi: 10.30872/eksponensial.v14i2.1090.

[13] U. N. Safitri, I. Sriliana, R. Adelisa, M. Hafiz, and P. Novianti, "Perbandingan Regresi Nonparametrik Spline Truncated dan Kernel Gaussian dalam Menganalisis Faktor-Faktor Penentu Indeks Pembangunan Manusia (IPM) di Indonesia," *J. Gaussian*, vol. 14, no. 2, pp. 554–564, 2025, doi: 10.14710/j.gauss.14.2.554-564.

[14] K. V. Permatasari, F. M. Farid, and T. Anggraini, "Pendekatan Regresi Nonparametrik Spline untuk Pemodelan Indeks Pembangunan Manusia di Pulau Kalimantan," *RAGAM: J. Stat. Its Appl.*, vol. 1, no. 1, pp. 1–13, 2022, url: https://ppjp.ulm.ac.id/journals/index.php/ragam/article/view/6885.

[15] A. Asmira, M. N. Bustan, and M. K. Aidid, "Penerapan Regresi Nonparametrik Spline dalam Memodelkan Faktor-Faktor yang Mempengaruhi Indeks Pembangunan Manusia di Provinsi Sulawesi Selatan Tahun 2015," *VARIANSI: J. Stat. Its Appl. Teach. Res.*, vol. 2, no. 2, pp. 92–109, 2020, doi: 10.35580/variansiunm12902.

[16] S. Meilinda, Sudarmin, and M. Nusrang, "Pendekatan Regresi Nonparametrik Spline Truncated pada Indeks Pembangunan Manusia di Sulawesi Selatan," *VARIANSI: J. Stat. Its Appl. Teach. Res.*, vol. 3, no. 2, pp. 83–89, 2021, url: https://ojs.unm.ac.id/jvariansi/article/view/23860.

[17] M. I. Amarta, R. Ibnas, A. Sauddin, and W. Alwi, "Pemodelan Indeks Pembangunan Manusia (IPM) di Provinsi Sulawesi Selatan Menggunakan Pendekatan Regresi Nonparametrik Spline," *Teknosains: Media Inf. Sains Teknol.*, vol. 19, no. 1, pp. 64–75, 2025, doi: 10.24252/teknosains.v19i1.55454.

[18] B. Efron, "Estimation and Accuracy after Model Selection," *J. Am. Stat. Assoc.*, vol. 109, no. 507, pp. 991–1007, 2014, doi: 10.1080/01621459.2013.823775.

[19] H. Glejser, "A New Test for Heteroskedasticity," *J. Am. Stat. Assoc.*, vol. 64, no. 325, pp. 316–323, 1969, doi: 10.1080/01621459.1969.10500976.

[20] P. A. P. Moran, "Notes on Continuous Stochastic Phenomena," *Biometrika*, vol. 37, no. 1/2, pp. 17–23, 1950, doi: 10.2307/2332142.

[21] H. W. Lilliefors, "On the Kolmogorov–Smirnov Test for Normality with Mean and Variance Unknown," *J. Am. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402, 1967, doi: 10.1080/01621459.1967.10482916.

[22] R. D. Cook, "Detection of Influential Observation in Linear Regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977, doi: 10.1080/00401706.1977.10489493.