Proximal: Jurnal Penelitian Matematika dan Pendidikan Matematika
ISSN 26158132 (cetak)
ISSN 26157667 (online)

PRINCIPAL COMPONENT–TRUNCATED SPLINE REGRESSION MODELING OF POPULATION WELFARE IN SOUTH SULAWESI

Irwan[1], Ayu Rahayu[1], Nur Fadilah[1], Fettyana[1]
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Negeri Makassar, Makassar, Indonesia[1,2,3,4]
Email: irwanthaha@unm.ac.id[1], ayu.rahayu@unm.ac.id[2], nur.fadilah@unm.ac.id[3], fettyana@unm.ac.id[4]

Corresponding Author: Irwan, **Email:** irwanthaha@unm.ac.id

Abstract. Population welfare modeling with socioeconomic predictors faces two problems: multicollinearity and an unknown functional form. This study combines Principal Component Analysis (PCA) with nonparametric truncated spline regression to model the Human Development Index (HDI), used as a proxy for the human development dimension of population welfare, across the 24 regencies/cities of South Sulawesi in 2024, using six predictors: population density, number of poor people, labor force size, real per-capita expenditure, life expectancy, and mean years of schooling. The KMO measure (0.591) and Bartlett's test ($\chi^2 = 97.72$; $p < .001$) supported dimension reduction, and two components with eigenvalues > 1 were retained (78.83% of variance). The polynomial order (m = 1, 2, 3), number of knots per component (r = 1, 2, 3), and knot locations were selected jointly by minimizing Generalized Cross-Validation (GCV), with knots searched exhaustively within the 15th–85th percentile range of each component. The optimal model, a linear truncated spline with two knots per component (GCV = 0.453), achieved $R^2 = 0.9836$ and AIC = 46.54, outperforming linear principal component regression ($R^2 = 0.9571$; AIC = 61.59); its leave-one-out cross-validated R^2 was 0.9625. The first component, a development-scale dimension, has an increasing effect over three segments (slopes 1.80, 4.22, and 1.43); the second, contrasting poverty–labor pressure with human-capital quality, decreases overall with a local slope change between two closely spaced knots. The estimated knots group regencies/cities into segments with different marginal associations with the HDI, offering an empirical basis for prioritizing regional development.

Keywords: **Population Welfare, Principal Component Analysis, Truncated Spline, Nonparametric Regression, South Sulawesi**

A. Introduction

Population welfare is a multidimensional indicator and a principal benchmark of regional development. In South Sulawesi Province, disparities in socioeconomic indicators across regencies and cities remain pronounced, so mapping and improving welfare is an important agenda that cuts across several Sustainable Development Goals (SDGs). One composite measure commonly used by the government is the Human Development Index (HDI), published regularly by Statistics Indonesia (Badan Pusat Statistik, 2025). The HDI aggregates three dimensions (a long and healthy life, knowledge, and a decent standard of living) and is therefore related to several goals simultaneously, notably SDG 3 (good health and well-being), SDG 4 (quality education), and SDG 1 and SDG 8 (no poverty; decent work and economic growth), rather than to SDG 3 alone. In this study, the HDI is used as a proxy for the human development dimension of population welfare, not as an exhaustive measure of welfare. Accurate statistical modeling of the determinants of the HDI is needed as a basis for *evidence-based policy*, so that local government interventions are well targeted and efficient.

Modeling welfare data in practice faces two structural problems. First, socioeconomic predictors such as population density, poverty, employment, expenditure, health, and education tend to be strongly correlated (multicollinearity). Multicollinearity does not bias the ordinary least squares estimator, but it inflates the sampling variances and hence the standard errors of

CC BY SA

the individual coefficient estimates. As a result, these estimates become imprecise and sensitive to small changes in the data, and inference on the separate contribution of each predictor becomes unreliable. Principal Component Analysis (PCA) is widely used to reduce data dimension by transforming correlated original variables into a smaller set of principal components, that is, linear combinations with orthogonal coefficient vectors whose scores are mutually uncorrelated, that retain most of the variance of the original data (Islamiyati et al., 2022). Because the retained component scores are linearly uncorrelated with one another, using them as regressors removes the linear dependence among predictors that causes the loss of precision described above.

The second problem concerns the functional form of the relationship between predictors and the response, which is often unknown, that is, nonparametric. Imposing linearity on data with sharp changes in pattern can produce substantial estimation bias. Nonparametric regression with spline estimators addresses this problem because splines capture local changes in pattern flexibly without being bound to a particular global curve (Islamiyanti et al., 2019). Among spline estimators, the truncated spline stands out for its interpretability. The regression curve is represented as polynomial segments joined at knot points, so changes in the relationship over particular ranges of a predictor can be traced explicitly from the estimated knot locations ((Islamiyati et al., 2017)(Ell Goldameir et al., 2015). Unlike the penalized spline, which regularizes smoothness through a penalty parameter on a high-dimensional basis (Latief et al., 2022), the truncated spline controls model complexity through the polynomial order and the number and location of knots. The optimal combination is generally chosen by minimizing Generalized Cross-Validation (GCV) (Utami et al., 2025)(Pratama et al., 2024).

The integration of PCA and truncated spline nonparametric regression has been developed in several settings. Examples include a biresponse model with a truncated spline estimator on principal components (Yulianti et al., 2024), a principal component linear spline quantile regression model for statistical downscaling of rainfall (Ell Goldameir et al., 2015), disease-risk modeling with Multivariate Adaptive Regression Splines (Chamidah et al., 2024), and multi-predictor penalized spline estimators for longitudinal data (Lestari et al., 2018)(Tenggara et al., 2020). However, the specific application of PCA combined with a single-response truncated spline to model population welfare at the regency/city level, particularly in South Sulawesi, is rarely found in the literature.

The contribution of this study is applied and empirical rather than the proposal of a new estimator, and it lies in three elements:

1. combining PCA with an additive truncated spline in which the polynomial order, the number of knots, and the knot locations are selected jointly. The GCV-based joint grid search is a standard model-selection strategy rather than a new method; its role here is to make the choice of order data-driven instead of fixed a priori and to report all candidate models transparently. For each pair of order and number of knots, the knot locations are obtained by exhaustive search over candidate points within the 15th–85th percentile range of each component score and the knot set with minimum GCV is retained; the resulting nine optimal models are then compared by GCV, with AIC and R^2 as supporting criteria (Section B.3);
2. applying it to six official Statistics Indonesia socioeconomic indicators; and
3. using leave-one-out cross-validation to assess generalization for a limited number of regencies/cities.

The objectives of this study are:

1. to identify the principal component structure of six socioeconomic indicators of the regencies/cities of South Sulawesi in 2024; and

CC BY SA

2. to build and evaluate a truncated spline regression model on those components, including the selection of the optimal order and knots, to explain regional variation in the HDI; and
3. to use the estimated knots and segment-specific slopes to group the regencies/cities into segments that can inform the prioritization of welfare development.

Methodologically, the study shows how a PCA–truncated spline model with jointly selected order and knots can be applied to small-area welfare data with strongly correlated predictors. Practically, the estimated model supports priority mapping in three ways: (i) the component loadings identify which original indicators drive each component; (ii) the estimated knots partition the regencies/cities into score segments, and the segment-specific slopes show where along each component the HDI changes most rapidly; and (iii) the membership of each regency/city in these segments indicates which regions lie in ranges where differences in the underlying indicators are associated with the largest differences in the HDI. Because the data are cross-sectional, these associations are descriptive and are not interpreted as causal effects of policy interventions.

B. Research Methods

1. Data and Research Variables

This study uses secondary regency/city-level data from the official publication of Statistics Indonesia (BPS) of South Sulawesi Province for 2024 (Badan Pusat Statistik Provinsi Sulawesi Selatan, 2025). The data cover all 24 regencies/cities ($n = 24$) as observation units. The response variable y is the HDI as a proxy for population welfare. The $p = 6$ predictors $x_1, \dots, x_6$ are population density (x_1, persons/km²), number of poor people (x_2, thousand persons), labor force size (x_3, persons), real per-capita expenditure (x_4, thousand rupiah/capita/year), life expectancy (x_5, years), and mean years of schooling (x_6, years).

2. Principal Component Analysis

Before extraction, the six predictors are standardized to remove differences in measurement units:

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \qquad s_j = \sqrt{n^{-1} \sum_{i=1}^{n} \left(x_{ij} - \bar{x}_j\right)^2} \tag{1}$$

with $i = 1,2, \dots, n$ and $j = 1,2, \dots, p$. The correlation matrix $\mathbf{R} = n^{-1}\mathbf{Z}^T\mathbf{Z}$ is decomposed into eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and the corresponding eigenvectors $\mathbf{a}_k = \left(a_{1k}, \dots, a_{pk}\right)^T$, such that $\mathbf{R}\mathbf{a}_k = \lambda_k \mathbf{a}_k$. The k-th principal component score of the i-th observation is the linear combination

$$t_{ik} = \sum_{j=1}^{p} a_{jk}\, z_{ij}, \qquad k = 1,2, \dots, q \tag{2}$$

The adequacy of dimension reduction is examined with the Kaiser–Meyer–Olkin measure of sampling adequacy (KMO) (Islamiyati et al., 2023) and Bartlett's test of sphericity (Yulianti et al., 2024). Bartlett's test examines the null hypothesis H_0: $\mathbf{R} = \mathbf{I}$ using the statistic

$$\chi^2 = -\left[(n-1) - \frac{2p+5}{6}\right] \ln|\mathbf{R}|, \qquad df = \frac{p(p-1)}{2} \tag{3}$$

The number of retained components q is determined by the Kaiser criterion $\lambda_k > 1$, cross-checked against a cumulative proportion of variance $\sum_{k=1}^{q} \lambda_k / p \geq 80\%$.

3. Truncated Spline Regression

The selected component scores $t_{i1}, \dots, t_{iq}$ are used as predictors in the additive nonparametric regression model

$$y_i = \beta_0 + \sum_{k=1}^{q} f_k\left(t_{ik}\right) + \varepsilon_i, \qquad i = 1,2, \dots, n \tag{4}$$

CC BY SA

where ε_i are independent and identically distributed with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$, and $q = 2$ is the number of retained components. Following Islamiyati et al. (2022) and Budiantara (2009), each regression curve f_k is approximated by a truncated spline of order m with r knots $K_{k1} < K_{k2} < \cdots < K_{kr}$:

$$f_k(t_{ik}) = \sum_{h=1}^{m} \beta_{kh}\ t_{ik}^{h} + \sum_{l=1}^{r} \beta_{k(m+l)}\ (t_{ik} - K_{kl})_{+}^{m} \tag{5}$$

with the truncated function

$$(t_{ik} - K_{kl})_{+}^{m} = \begin{cases} (t_{ik} - K_{kl})^{m}, & t_{ik} \geq K_{kl} \\ 0, & t_{ik} < K_{kl} \end{cases} \tag{6}$$

For a linear spline ($m = 1$), β_{k1} is the slope of the first segment and $\beta_{k(1+l)}$ is the change in slope after knot K_{kl}. For a given combination of (m, r) and knot locations, model (4)–(6) is linear in the parameters and can be written in matrix form as $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. The ordinary least squares (OLS) estimator is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \qquad \hat{\mathbf{y}} = \mathbf{A}(\mathbf{K})\,\mathbf{y}, \qquad \mathbf{A}(\mathbf{K}) = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \tag{7}$$

where $\mathbf{X}$ is the $n \times s$ design matrix whose columns are 1, t_{ik}^{h} $(h = 1, \ldots, m)$ and $(t_{ik} - K_{kl})_{+}^{m}$ $(l = 1, \ldots, r)$ for $k = 1,2$, and $s = 1 + q(m + r)$ is the number of parameters including the intercept.

The order m, the number of knots r, and the knot locations $\mathbf{K} = (K_{11}, \ldots, K_{qr})$ are selected by a joint grid search. For every combination $m \in \{1,2,3\}$ and $r \in \{1,2,3\}$, candidate knots located between the 15th and 85th percentiles of each component score are searched exhaustively over all combinations of knots for PC1 and PC2. The combination $(m, r, \mathbf{K})$ that minimizes GCV (Craven & Wahba, 1979; Wahba, 1990) is selected:

$$GCV(\mathbf{K}) = \frac{MSE(\mathbf{K})}{\left[n^{-1}\,\mathrm{tr}(\mathbf{I} - \mathbf{A}(\mathbf{K}))\right]^2} = \frac{n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(1 - s/n)^2} \tag{8}$$

The second equality holds because $\mathrm{tr}\,\mathbf{A}(\mathbf{K}) = \mathrm{rank}(\mathbf{X}) = s$ for the OLS projection matrix. Because the sample is small ($n = 24$), knot candidates are restricted to the 15th–85th percentile range of each component to avoid knots in sparse tails, which would produce unstable estimates.

The selection therefore proceeds in two stages. In the first stage, for each fixed pair (m, r), every admissible knot vector K, that is, every ordered set of r candidate points for PC1 combined with every ordered set of r candidate points for PC2, is fitted by OLS, and the knot vector with the smallest GCV is retained as the optimal knot vector for that pair. The candidate points are lying within the 15th–85th percentile range of each component. In the second stage, the nine resulting models are compared on a common basis using GCV as the primary criterion, because it penalizes the number of parameters s = 1 + q(m + r); in-sample R^2, MSE, and AIC are reported as supporting criteria. The selected model is then compared with the linear PCR baseline using the same criteria and LOOCV (Section B.4).

4. Model Evaluation and Validation

Goodness of fit is evaluated with the coefficient of determination R^2, MSE, GCV, and the Akaike Information Criterion

$$AIC = -2\ln\hat{L} + 2s \tag{9}$$

where $\hat{L}$ is the maximized Gaussian likelihood. The model is compared with a linear principal component regression (PCR) as the baseline. To assess generalization on a small sample, leave-one-out cross-validation (LOOCV) is also carried out. Each observation i is held out in turn, and the model is re-estimated on the remaining $n - 1$ observations with m, r, and $\mathbf{K}$ fixed at their optimal full-data values, so only $\boldsymbol{\beta}$ is re-estimated in each fold. The held-out prediction $\hat{y}_{(-i)}$ is then compared with the observed value:

CC BY SA

$$R^2_{LOOCV} = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(-i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad MSE_{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(-i)}\right)^2 \tag{10}$$

5. Analysis Procedure

The analysis proceeds in the following steps:

1. collecting and cleaning the 2024 BPS South Sulawesi data;
2. standardizing the predictors and testing PCA adequacy (KMO and Bartlett);
3. extracting the principal components and determining q by the criterion $\lambda_k > 1$;
4. fitting the truncated spline on the retained components, with the optimal order, number of knots, and knot locations selected by the GCV-based grid search; and
5. evaluating the model and validating it by leave-one-out cross-validation.

All analyses were performed in Python 3. PCA and the adequacy tests used scikit-learn and factor_analyzer, the truncated spline basis and least squares estimation used NumPy, and the linear PCR baseline used statsmodels.

C. Results and Discussion

1. Descriptive Statistics and Multicollinearity

Descriptive statistics of the research variables for the 24 regencies/cities of South Sulawesi in 2024 are presented in Table 1.

Table 1 Descriptive statistics of the research variables, BPS South Sulawesi 2024 ($n = 24$)

Variable	Unit	Mean	Std. Dev.	Min	Max
y: Human Development Index	index	72.79	3.80	66.38	83.90
x_1: Population density	persons/km²	673.21	1,659.44	45.00	8,282.00
x_2: Number of poor people	thousand persons	30.69	18.27	7.94	79.53
x_3: Labor force size	persons	203,570.33	138,540.37	74,971	696,851
x_4: Real per-capita expenditure	thousand IDR/capita/year	12,109.04	2,111.45	8,319	18,386
x_5: Life expectancy	years	70.35	1.92	67.16	74.23
x_6: Mean years of schooling	years	8.63	1.14	7.01	11.57

Makassar City is the most distinctive observation, with a population density of 8,282 persons/km² and an HDI of 83.90, far above all other regencies/cities. This is the *urban primacy* pattern typical of Indonesian regional data. The correlation matrix of the predictors indicates fairly strong multicollinearity, for example between the number of poor people and labor force size ($r = 0.903$) and between population density and labor force size ($r = 0.717$). This confirms the need for PCA before regression modeling.

PCA adequacy was examined with the KMO and Bartlett tests. The overall KMO was 0.591, which is marginal but acceptable by Kaiser's criterion. Individual KMO values ranged from 0.378 (life expectancy) to 0.756 (population density). Bartlett's test yielded $\chi^2(15) = 97.72$ with $p < 0.001$, rejecting $H_0: \mathbf{R} = \mathbf{I}$. Together, the two tests indicate that the data are adequate for dimension reduction by PCA.

2. Principal Component Analysis Results

The eigenvalues and proportions of variance obtained from the six standardized predictors are presented in Table 2 and Figure 1.

CC BY SA

Table 2 Eigenvalues and proportion of variance of the principal components

Component	Eigenvalue (λ_k)	Variance explained (%)	Cumulative (%)
PC1	3.058	50.97	50.97
PC2	1.672	27.87	78.83
PC3	0.833	13.88	92.71
PC4	0.224	3.73	96.44
PC5	0.163	2.71	99.16
PC6	0.051	0.84	100.00

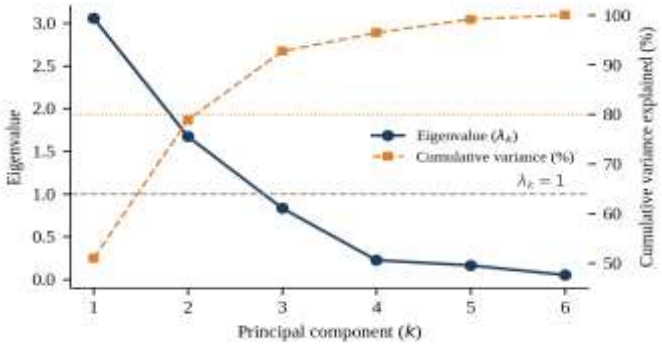


Figure 1 Scree plot of the principal components. The dashed grey line marks the Kaiser criterion $\lambda_k = 1$; the dotted orange line marks the 80% cumulative-variance threshold.

Under the criterion $\lambda_k > 1$, two principal components are retained (PC1 and PC2), jointly explaining 78.83% of the total variance. This is close to the 80% threshold commonly used as an alternative criterion. The loadings a_{jk} of each variable on the two retained components are shown in Table 3.

Table 3 Loading matrix of the retained principal components

Original variable	Loading PC1 (a_{j1})	Loading PC2 (a_{j2})
x_1: Population density	0.536	−0.014
x_2: Number of poor people	0.359	0.554
x_3: Labor force size	0.461	0.415
x_4: Real per-capita expenditure	0.411	−0.230
x_5: Life expectancy	0.200	−0.498
x_6: Mean years of schooling	0.403	−0.470

PC1 has relatively even positive loadings on all original variables. It is therefore interpreted as a *regional development scale* component, in which density, employment, expenditure, and education move together, and it is strongly influenced by the extreme position of Makassar City. PC2 shows a contrast pattern: strong positive loadings on the number of poor people (0.554) and labor force size (0.415), opposed to life expectancy (−0.498) and mean years of schooling (−0.470). It is therefore interpreted as *poverty–labor pressure versus human capital quality*. The distribution of the scores t_{i1} and t_{i2}, colored by observed HDI, is shown in Figure 2.

CC BY SA

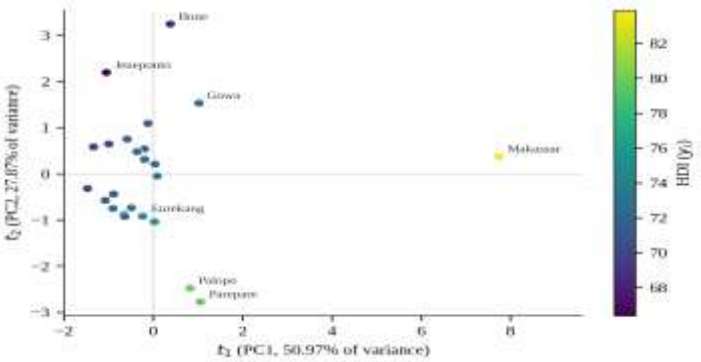


Figure 2 Scores t_{i1} (PC1) and t_{i2} (PC2) of each regency/city in 2024, colored by observed HDI.

Figure 2 shows that the three cities (Makassar, Parepare, and Palopo) consistently have the highest HDI and are clearly separated from the regencies, confirming PC1 as the urban-scale axis. Meanwhile, Bone and Jeneponto have the highest PC2 scores (high poverty pressure) and relatively low HDI, which supports the interpretation of PC2 given above.

3. Selection of the Optimal Order and Knots

The truncated spline model was selected by a joint grid search over nine combinations of polynomial order ($m = 1,2,3$) and number of knots per component ($r = 1,2,3$). For each combination, the knot locations were optimized by exhaustive search over candidate points between the 15th and 85th percentiles of the PC1 and PC2 scores. The search results, sorted by GCV, are summarized in Table 4 and Figure 3.

Table 4 Summary of the GCV grid search over nine combinations of order and number of knots (bold row = minimum GCV)

Order (m)	Knots (r)	Parameters (s)	R^2	AIC	GCV
1	**2**	**7**	**0.9836**	**46.54**	**0.453**
1	1	5	0.9785	49.01	0.475
2	2	9	0.9832	51.06	0.594
1	3	9	0.9830	51.39	0.602
2	3	11	0.9871	48.68	0.606
2	1	7	0.9778	53.79	0.613
3	1	9	0.9810	54.09	0.674
3	2	11	0.9842	53.58	0.744
3	3	13	0.9861	54.55	0.916

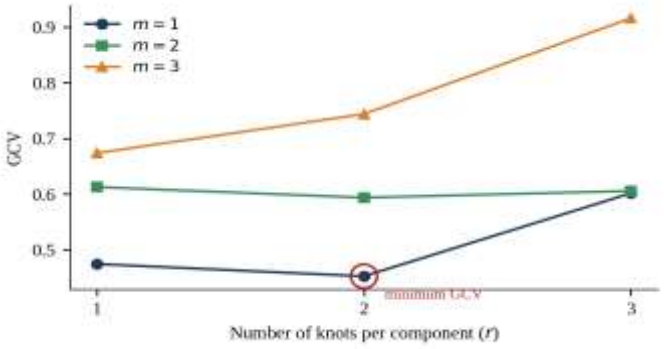


Figure 3 GCV for the nine combinations of order m and number of knots per component r. The red circle marks the minimum GCV ($m = 1, r = 2$).

CC BY SA

The minimum GCV (0.453) is obtained for the linear order ($m = 1$) with two knots per component ($r = 2$), giving $s = 7$ parameters including the intercept. Linear models consistently yield lower GCV than quadratic or cubic models with the same number of knots (Figure 3). The extra flexibility of a higher polynomial degree therefore does not improve fit enough to offset the complexity penalty at $n = 24$, a pattern common in small-sample nonparametric regression (Eubank, 1999). Notably, the order-2 and order-3 models with $r = 3$ give the highest in-sample R^2 (0.9871 and 0.9861), but both have higher AIC and GCV than the selected model, indicating overfitting caused by too many parameters relative to $n = 24$. The optimal knot locations are given in Table 5.

Table 5 Optimal knot locations of the selected truncated spline model

Component	K_{k1}	K_{k2}
PC1 ($k = 1$)	−0.724	−0.274
PC2 ($k = 2$)	−0.925	−0.756

The two PC1 knots ($K_{11} = -0.724$ and $K_{12} = -0.274$) divide the PC1 scores into three segments with balanced membership. The middle segment contains five regencies: Maros, Wajo, Sidenreng Rappang, Tana Toraja, and Toraja Utara. In contrast, the two PC2 knots ($K_{21} = -0.925$ and $K_{22} = -0.756$) are very close (a difference of only 0.169), and the segment between them contains only two observations (Enrekang and Tana Toraja). The implications of such a narrow segment for estimation stability are discussed in Section C.4.

4. Truncated Spline Regression Model on the Principal Components

With $m = 1$ and $r = 2$, the estimated model is

$$\hat{y}_i = 71.062 + 1.798\, t_{i1} + 2.422(t_{i1} + 0.724)_+ - 2.786(t_{i1} + 0.274)_+ - 2.413\, t_{i2} + 5.284(t_{i2} + 0.925)_+ - 4.626(t_{i2} + 0.756)_+ \quad (11)$$

Each partial effect $\hat{f}_k$ consists of three linear segments. The slope of each segment, obtained as the cumulative sum of the estimated coefficients, is presented in Table 6.

Table 6 Slope of each linear segment of the selected truncated spline model

Component	Segment	Score range	Slope
PC1	1	$t_{i1} < -0.724$	1.798
PC1	2	$-0.724 \le t_{i1} < -0.274$	4.220
PC1	3	$t_{i1} \ge -0.274$	1.434
PC2	1	$t_{i2} < -0.925$	−2.413
PC2	2	$-0.925 \le t_{i2} < -0.756$	2.871
PC2	3	$t_{i2} \ge -0.756$	−1.755

The goodness of fit of the model and its comparison with the linear PCR model are summarized in Table 7.

Table 7 Comparison of the PCA–truncated spline model and the linear PCR model

Criterion	PCA–truncated spline	Linear PCR (baseline)
Number of parameters (s)	7	3
R^2 (in-sample)	0.9836	0.9571
MSE (in-sample)	0.227	0.594
GCV	0.453	0.775
AIC	46.54	61.59
R^2_{LOOCV}	0.9625	—
MSE_{LOOCV}	0.519	—

CC BY SA

The PCA–truncated spline model gives an in-sample R^2 of 0.9836 and a much lower AIC (46.54) than the linear PCR model (61.59), while adding only four parameters ($s = 7$ versus $s = 3$). Its GCV (0.453) is also well below that of the PCR model (0.775). Leave-one-out cross-validation yields $R^2_{LOOCV} = 0.9625$ with $MSE_{LOOCV} = 0.519$. This is higher than even the in-sample R^2 of the linear PCR model, indicating good generalization despite the limited sample size. The estimated partial effect curves are shown in Figure 4, and the agreement between fitted and observed values in Figure 5.

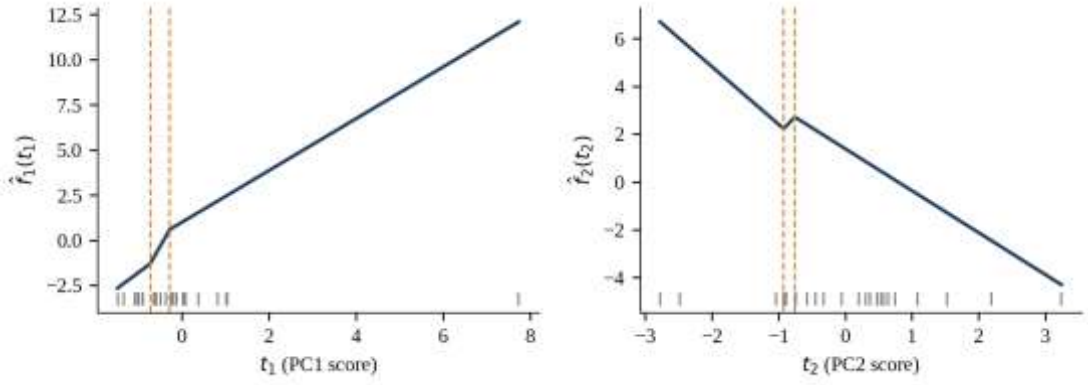


Figure 4 Estimated truncated spline partial effects $\hat{f}_1(t_1)$ (left) and $\hat{f}_2(t_2)$ (right) on the HDI. Dashed orange lines mark the optimal knots; grey ticks show the observed scores.

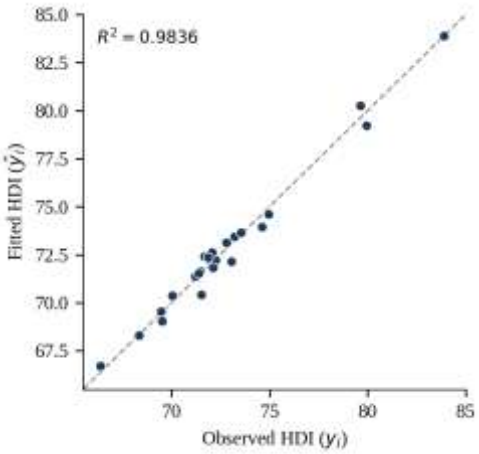


Figure 5 Observed versus fitted HDI of the PCA–truncated spline model. The dashed line is the reference line $\hat{y}_i = y_i$.

The PC1 curve (Figure 4, left) increases on all three segments but at different rates. The steepest slope (4.220) occurs on the narrow middle segment, whereas on the low-score segment and the high-score segment (dominated by Makassar City) the rate of increase is more moderate (1.798 and 1.434). The PC2 curve (Figure 4, right) generally decreases (slopes −2.413 and −1.755 on the first and third segments) but shows a local increase (slope 2.871) on the narrow segment between its two knots.

These segments provide the basis for the priority mapping described in the Introduction. On PC1, the five regencies in the middle segment (Maros, Wajo, Sidenreng Rappang, Tana Toraja, and Toraja Utara) lie in the range where the estimated slope is largest (4.220), so differences in the development-scale indicators that load on PC1 are associated with the largest differences in the HDI there; this group can be regarded as a priority for interventions that strengthen this dimension. Regencies with high PC2 scores, notably Bone and Jeneponto, combine high poverty–labor pressure with relatively low HDI and lie on the decreasing third

CC BY SA

segment (slope −1.755); for these regions, poverty reduction and improvement of health and education outcomes are the dimensions most directly indicated by the model. In contrast, the positive slope on the narrow PC2 segment (2.871) is estimated from only two observations (Enrekang and Tana Toraja), so it has high sampling variability and is largely determined by those two regencies; it is better regarded as a local feature of knot placement than as a substantive pattern and is not used for prioritization. As the analysis is cross-sectional, these groupings describe associations rather than causal effects.

D. Conclusion

This study integrated Principal Component Analysis and truncated spline regression to model population welfare (HDI) in the 24 regencies/cities of South Sulawesi Province in 2024. The KMO (0.591) and Bartlett ($p < 0.001$) tests confirmed the adequacy of dimension reduction for six strongly correlated socioeconomic predictors, yielding two principal components that explain 78.83% of the total variance. The joint grid search over polynomial order and number of knots selected a linear truncated spline ($m = 1$) with two knots per component ($r = 2$) as the optimal model (GCV = 0.453). This model gives $R^2 = 0.9836$ and $AIC = 46.54$, outperforming linear PCR ($R^2 = 0.9571$; $AIC = 61.59$), with a leave-one-out cross-validated $R^2_{LOOCV} = 0.9625$ indicating good generalization despite the limited sample size.

The first principal component represents regional development scale, with an increasing effect on three segments of different slope, steepest in medium-scale regencies. The second represents the contrast between poverty–labor pressure and human capital quality, with a decreasing effect interrupted by a local kink on a segment containing very few observations. The PCA–truncated spline approach thus handles multicollinearity while revealing explicit, policy-relevant piecewise nonlinear patterns, and, through the segment membership of each regency/city, provides an empirical basis for identifying regional welfare development priorities.

Future research is recommended to extend the analysis to multi-year panel data, to apply a finer knot search and a sensitivity analysis for high-leverage observations, and to develop biresponse/longitudinal models that capture welfare dynamics over time.

Acknowledgment

The authors thank the Institute for Research and Community Service (LP2M) of Universitas Negeri Makassar for funding this research through the 2026 PNBP Basic Research scheme, and Statistics Indonesia (BPS) of South Sulawesi Province for the open secondary data used in this study.

REFERENCES

Badan Pusat Statistik. (2025). *Indeks Pembangunan Manusia 2024* [Human Development Index 2024]. Badan Pusat Statistik Republik Indonesia. https://www.bps.go.id/id/publication/2025/05/15/dd9a3ce7dfae1c733e46338f/indeks-pembangunan-manusia-2024.html

Badan Pusat Statistik Provinsi Sulawesi Selatan. (2025). *Provinsi Sulawesi Selatan dalam angka 2025* [South Sulawesi Province in figures 2025]. Badan Pusat Statistik Provinsi Sulawesi Selatan.

https://sulsel.bps.go.id/en/publication/2025/02/28/a7beacd81481497cecb44084/provinsi-sulawesi-selatan-dalam-angka-2025.html

Ell Goldameir, N., Djuraidah, A., & Wigena, A. H. (2015). Quantile spline regression on statistical downscaling model to predict extreme rainfall in Indramayu. *Applied Mathematical Sciences*, *9*(126), 6263–6272. https://doi.org/10.12988/ams.2015.57502

Islamiyanti, A., Statistika, D., Matematika, F., & Alam, P. (2019). *Regresi Spline Polynomial Truncated Biprediktor untuk Identifikasi Perubahan Jumlah Trombosit Pasien Demam Berdarah Dengue*. *7*(2), 97–112.

Islamiyati, A., Fatmawati, & Chamidah, N. (2017). Fungsi Goodness of Fit dalam Kriteria Penalized Spline pada Estimasi Regresi Nonparametrik Birespon untuk Data Longitudinal. *Seminar Nasional Matematika Dan Aaplikasinya*, *2007*, 216–221.

Islamiyati, A., Kalondeng, A., Sunusi, N., Zakir, M., & Kamal, A. (2022). Journal of King Saud University – Science Biresponse nonparametric regression model in principal component analysis with truncated spline estimator. *Journal of King Saud University - Science*, *34*(3), 101892. https://doi.org/10.1016/j.jksus.2022.101892

Islamiyati, A., Sahriman, S., & Massalesse, J. (2023). *Model Regresi Kuantil Spline Orde Dua Dalam Menganalisis Perubahan Trombosit Pasien Demam Berdarah*. *5*(1), 38–45.

Latief, A. N., Islamiyati, A., Sunusi, N., & Kuswanto, H. (2022). *The Implementation of Hierarchical Linear Spline Regression Model to The Implementation of Hierarchical Linear Spline Regression Model to Maternal Mortality Rate Data in Indonesia*. *December*. https://doi.org/10.47119/IJRP10011311120214161

Lestari, B., Fatmawati, Budiantara, I. N., & Chamidah, N. (2018). Estimation of Regression Function in Multi-Response Nonparametric Regression Model Using Smoothing Spline and Kernel Estimators. *Journal of Physics: Conference Series*, *1097*(1). https://doi.org/10.1088/1742-6596/1097/1/012091

Pratama, Y. M., Fernandes, A. A. R., Wardhani, N. W. S., & Hamdan, R. (2024). Nonparametric Smoothing Spline Approach in Examining Investor Interest Factors. *JTAM (Jurnal Teori Dan Aplikasi Matematika)*, *8*(2), 425. https://doi.org/10.31764/jtam.v8i2.20192

Tenggara, N., Hidayati, L., Budiantara, I. N., & Chamidah, N. (2020). *Bi-Response Semiparametric Regression Model based on Spline Truncated for Estimating Computer based National Exam in West*. *ICMIs 2018*, 357–361. https://doi.org/10.5220/0008521903570361

Utami, T. W., Chamidah, N., Saifudin, T., Lestari, B., & Aydin, D. (2025). *Estimation of Biresponse Semiparametric Regression Model for Longitudinal Data Using Local Polynomial Kernel Estimator*. 1–22. https://doi.org/https://doi.org/10.3390/sym17030392

Yulianti, A. S., Islamiyati, A., & Herdiani, E. T. (2024). *The Principal Component Linear Spline Quantile Regression Model in Statistical Downscaling for Rainfall Data*. *35*(2), 159–166.