Modeling of Total Fertility Rate (TFR) in East Java Province using Mixed Semiparametric Regression Spline Truncated and Kernel Approach

Arip Ramadan[1], I Nyoman Budiantara[1] and Ismaini Zain[2]
[1]*Institut Teknologi Sepuluh Nopember, Jl. Raya ITS Sukolilo Surabaya, Jawa Timur, Indonesia*
[2]*Department of Statistics, Faculty of Mathematics, Computing and Data Science, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*

Keywords: Total Fertility Rate, Mix Estimator, Spline Truncated, Kernel, Semiparametric Regression and Knot Point.

Abstract: The problem of population growth in Indonesia year by year is relatively very high. Based on data from Indonesia Demographic and Health Survey (SDKI), there is a very high increase of TFR and uncontrollable increase of population in Indonesia. TFR in East Java is very high in 2012 reached 20%. The behavior of TFR patterns is associated with the variables that are suspected to affect the Unmet Need, the Age Specific Fertility Rates (ASFR), the Human Development Index (HDI) and the Infant Mortality Rate (IMR) has a very special pattern. The relationship pattern between TFR and Unmet Need tends to be linear. While the pattern of relationship between TFR with ASFR and HDI, in contrast to the Unmet Need that tends to be nonlinear especially change as the increase or decrease value of these variables. The pattern of relationship between TFR and IMR appears not to follow a certain pattern. Taking into account this description, this study modeled using mixed semiparametric regression spline truncated and kernel spline models. The best model is a mixed semiparametric regression spline truncated and kernel model with a combination of knots that has the smallest GCV of 0.003964 and the coefficient of determination of 97.04%.

1 INTRODUCTION

Population is the most important thing in sustaining the development of an area because it is both a subject and an object of development. As the subject of population development will play a role in achieving the achievement of economic and social development that can affect the increase in social welfare, while as an object of population development is the party that gets results from the development of a region.

The Total Fertility Rate (TFR) can affect the development of the population in the future. In 2012, there was a national TFR increase from 2.41 in 2008 to 2.6 in 2012. Based on the report, there were only 10 provinces that experienced a decline in their fertility levels, while the rest were observed to increase. TFR increases experienced by other provinces ranged from 31 percent to 63 percent (BKKBN, 2107).

East Java Province is a province that experienced a very significant decline in TFR since the introduction of family planning (KB) policies, namely the TFR of East Java Province had reached below 2.1 in 2002. However the 2012 East Java Province TFR experienced a significant increase since 2002 which had an effect on increasing the return of TFR Indonesia. TFR East Java experienced a TFR increase of more than 20 percent, from 2.1 in 2002 to 2.6 in 2012.

To see the shape of the relationship between TFR and the variables that influence it can use the regression method. But sometimes the form of relationships that occur can vary, such as parametric or nonparametric. Therefore obtaining a mixed semiparametric spline truncated and kernel regression model is important because it includes parametric and nonparametric.

Ramadan, A., Budiantara, I. and Zain, I.
Modeling of Total Fertility Rate (TFR) in East Java Province using Mixed Semiparametric Regression Spline Truncated and Kernel Approach.
DOI: 10.5220/0008520602710277
In *Proceedings of the International Conference on Mathematics and Islam (ICMIs 2018)*, pages 271-277
ISBN: 978-989-758-407-7

2 LITERATURE REVIEW

2.1 Mixed Semiparametric Regression Truncated Spline

Semiparametric regression is a combination of parametric components and nonparametric components (Budiantara, 2009). In some cases, it can be found the relationship between response variables with one predictor variable is linear, but the relationship with other predictor variables is unknown. Variables that have known data patterns or previous information about their data patterns are classified in parametric components. While the unknown variable data pattern is classified on nonparametric components (Ruppert, Wand, Carrol, 2003).

Given data in pairs (x_{1i} , x_{2i} , ⋯ , x_{pi} , t_{1i} , t_{2i} , ⋯ , t_{qi} , z_{1i} , z_{2i} , ⋯ , z_{ri} , y_i) then the semiparametric model is formulated as follows:

$$y_i = f\left(x_{1i}, x_{2i}, \cdots, x_{pi}, \mathrm{t}_{1i}, \mathrm{t}_{2i}, \cdots, \mathrm{t}_{qi}, \mathrm{z}_{1i}, \mathrm{z}_{2i}, \cdots, \mathrm{z}_{ri}\right) + \varepsilon_i,$$
$$i = 1, 2, \ldots, n ,$$

with the curve assumed to be additive, it is obtained:

$$y_i = \sum_{j=1}^{p} f(x_{ji}) + \sum_{s=1}^{q} g(t_{si}) + \sum_{k=1}^{r} h(\mathrm{z}_{ki}) + \varepsilon_i ;$$
$$i = 1, 2, \ldots, n ,$$

where y_i is response variable, $\sum_{j=1}^{p} f\left(x_{ji}\right)$ is a parametric component, $\sum_{s=1}^{q} g\left(t_{si}\right)$ is a spline nonparametric component, $\sum_{s=1}^{r} h\left(\mathrm{z}_{ki}\right)$ are kernel nonparametric components, and ε_i is a random error that is assumed to be identical, independent and normally distributed with a zero mean and variant σ^2 (Eubank, 1999).

2.2 Selection of Knot Point and Optimal Bandwith

In nonparametric and semiparametric regression with the Spline approach, the important thing that plays a role in getting the Spline estimator is the optimal selection of knot points. Meanwhile the kernel depends on bandwidth. Bandwidth α is a smoothing parameter that serves to control the smoothness of the estimated curve. The knots and bandwidth points that are too small will produce very rough and fluctuating curves, whereas the point of the knots or bandwidth that is too large or wide will produce a very smooth curve, but not in accordance with the data pattern (Hardle, 1994). The selection of knot points k and optimum bandwidth using GCV is defined as follows:

$$GCV\left(k,\alpha\right) = \frac{MSE\left(k,\alpha\right)}{\left(n^{-1} trace\left[I - M\left(k,\alpha\right)\right]\right)^2} .$$

2.3 Residual Assumption Testing

In the semiparametric regression model linear truncated Spline is assumed to be a random error with an independent normal distribution with zero mean and variance σ^2 (Wahba, 1990). Therefore, before analyzing and making decisions from the modeling results, the residual assumption is tested first. The residual assumption test is an independent test, identical test and normality test (Tupen, 2011).

3 METHODS

3.1 Overview of General Object

This study uses district / city data in East Java Province. The East Java Province consists of 38 regions covering 29 regencies and 9 cities. In detail can be stated in Table 1.

3.2 Research Variable

The response variables used in this study are categorical data, namely TFR by district / city in East Java Province in 2015 and variables that are thought to affect TFR. These variables can be described in Table 2.

3.3 Research Step

Model TFR using spline truncated semiparametric regression and kernel. The steps taken are:

1. Plot the response variable with all predictor variables.

Table 1: List of Regional Names in East Java Province

NO	NAME REG/CITY	NO	NAME REG/CITY	NO	NAME REG/CITY
1	Pacitan	14	Pasuruan	27	Sampang
2	Ponorogo	15	Sidoarjo	28	Pemekasan
3	Trenggalek	16	Mojokerto	29	Sumenep
4	Tulungagung	17	Jombang	30	Kota Kediri
5	Blitar	18	Nganjuk	31	Kota Blitar
6	Kediri	19	Madiun	32	Kota Malang
7	Malang	20	Magetan	33	Probolinggo
8	Lumajang	21	Ngawi	34	Kota Pasuruan
9	Jember	22	Bojonegoro	35	Kota Mojokerto
10	Banyuwangi	23	Tuban	36	Kota Madiun
11	Bondowoso	24	Lamongan	37	Kota Surabaya
12	Situbondo	25	Gresik	38	Kota Batu
13	Probolinggo	26	Bangkalan		

Table 2: Variable Operational Definition

Var	Variable Name	Operational definition
Y	TFR (Total Fertility Rate)	The average number of children born to a woman from the beginning of childbearing age to the end of her reproductive period
X_1	Unmet Need	percentage figures that indicate unmet family planning needs or the proportion of women of childbearing age who are married or live together (sexually active) who do not want to have more children or who want to Arrange the next birth within a minimum period of 2 years but do not use contraceptive devices or methods in a district / city
X_2	ASFR (Age Specific Fertility Rate)	The number of births per 1000 women in certain age groups between 15-49 year.
X_3	HDI (Human Development Index)	Measurement of life expectancy, literacy, education and living standards for all countries throughout the world.
X_4	IMR (Infant Mortality Rate)	The number of infant deaths in one particular year per 1000 live births in the same year

2. Determine parametric component variables and truncated spline and nonparametric kernels.
3. Model response variables and predictor variables using spline truncated mix estimators and kernels in semiparametric regression.
4. Model data with a semiparametric regression of mixed truncated spline and kernel with one, two, three, and a combination of knots.
5. Select the optimal point knots, parameters and bandwidth α based on the GCV method.
6. Testing the assumption of independent, identical and normal distribution for residuals.

4 RESULT AND DISCUSSION

4.1 Descriptive Analysis

The results of descriptive statistics can be used to initiate knot points in the next analysis stage.

Table 3: Descriptive Statistics of Response Variables and Predictor Variables.

Var	Mean	Min	Max	Std Dev	Range
Y	2,06	1,52	2,45	0,21	0,93
X_1	69,11	58,18	80,05	5,40	21,87
T_1	16,01	6,87	31,77	5,01	24,9
T_2	36,21	6,40	87,00	18,79	80,60
Z_1	30,92	17,27	60,51	12,09	43,24

From Table 3 above can be seen the characteristics of each variable, both the response variable and predictor variables.

The following is a description of the characteristics for each predictor variable, namely Unmet Need (X_1), ASFR (T_1), HDI (T_2) and IMR (Z_1).

a. The average Unmet Need in East Java Province was 16.01 with a standard deviation of 5.01. The highest Unmet Need in Bangkalan Regency was 31.77 and the lowest was in Bondowoso Regency with 6.87 with a range of 24.90.
b. The average ASFR in East Java Province was 36.21 with a standard deviation of 18.79. The highest ASFR in Bondowoso Regency was 87.00 and the lowest was in Malang City with 6.40.
c. The average HDI in East Java Province was 69.11 with a standard deviation of 5.4. The highest Human Development Index in Malang City was

80.05 and the lowest was in Sampang District with 58.18 with a range of 24.90.

d. The average IMR in East Java Province was 30.92 or 31 people per 1000 live births with a standard deviation of 12.09. The highest IMR in Probolinggo Regency was 60.51 or 61 people and the lowest was in Blitar City with a number of 17.27 or 17 people with a range of 43.24.

4.2 TFR Modeling using Semiparametric Regression Mixture of Truncated Spline and Kernel

4.2.1 Determination of Parametric Component and Nonparametric Component Variables

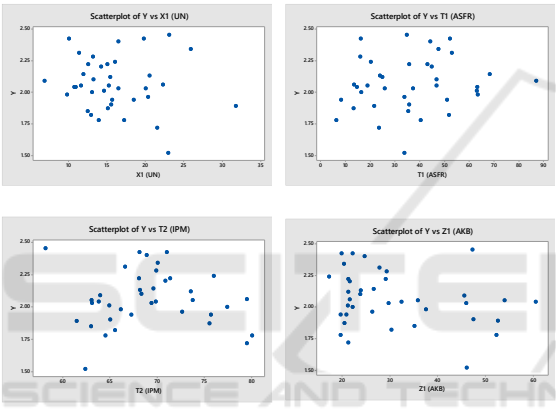

Figure 1. Scatter Plot Response variables and predictors

A summary of the results of the determination of parametric components and nonparametric components is presented in Table 4.

Table 4: Parametric and Nonparametric Components

Notation	Variable	Component
X_1	Unmet Need	Parametric
T_1	ASFR	Nonparametric
T_2	HDI	Nonparametric
Z_1	IMR	Nonparametric

4.2.2 Modeling TFR in East Java Province Using Semiparametric Mixed Truncated Splines and Kernel One Knots

The semiparametric regression model is linear truncated spline with one parametric component variable and four nonparametric component variables with one knot point are as follows:

$$y_i = \beta_0 + \beta_1 x_i + \theta_{11} t_{1i} + \theta_{21}\left(t_{1i} - K_{11}\right)_+^1 + \theta_{12} t_{2i} + \theta_{22}\left(t_{2i} - K_{12}\right)_+^1 + h(z_i) + \varepsilon_i \;;\; i = 1,2,\ldots,38.$$

GCV values generated using semiparametric regression of truncated spline and kernel mixtures with one knot are presented in Table 5.

Table 5: Comparison of GCV Values using One Knot Point.

Spline		Kernel	GCV	R^2
Knot		Bandwidth		
$t_1 = K_{11}$	$t_2 = K_{21}$	α		
56.30	71.72	0.0485	0.004164	95.06
60.13	72.76	0.0494	0.004268	94.69
52.46	70.68	0.0480	0.004301	94.99
63.97	73.80	0.0501	0.004450	94.28

Based on Table 5 the minimum GCV value produced is equal to 0.004164. Location of knots points on variables (t_1) that is 56.3 (K_{11}) and (t_2) that is 71.72 (K_{12}) while the bandwidth provided is as big as $\alpha = 0,0485$.

4.2.3 Model TFR in East Java Province with Truncated Two Point Knot Spline Components

The truncated spline semiparametric regression model using two knots with one parametric component predictor and five nonparametric components are as follows:

$$y_i = \beta_0 + \beta_1 x_i + \theta_{11} t_{1i} + \theta_{21}\left(t_{1i} - K_{11}\right)_+^1 + \theta_{31}\left(t_{1i} - K_{21}\right)_+^1 + \theta_{12} t_{2i} + \theta_{22}\left(t_{2i} - K_{12}\right)_+^1 + \theta_{32}\left(t_{2i} - K_{22}\right)_+^1 + h(z_i) + \varepsilon_i \quad i = 1,2,\ldots,38.$$

The GCV values produced using semiparametric regression of truncated and kernel spline mixtures with one knot are presented in Table 6. Based on

Table 6 the minimum GCV value produced is equal to 0.004185. Location of knots points on variables (t_1) that is 68.4 (K_{11}), 80.8 (K_{12}), for (t_2) that is 75 (K_{21}) 78.37 (K_{22}) while the bandwidth provided is as big as $\alpha = 0.04472$.

Table 6: Comparison of GCV Values using Two Knot Points.

Spline		Kernel	GCV	R^2
Knot		Bandwidth		
$t_1 = K_{11}$	$t_2 = K_{21}$	α		
$t_1 = K_{11}$	$t_2 = K_{21}$			
68.4	80.8	0.04472	0.004185	96.32
75	78.37			
18.8	56	0.04909	0.004207	95.32
61.54	71.64			
68.4	74.6	0.04553	0.004303	95.91
75	76.69			
12.6	56	0.05001	0.004346	95.09
59.86	71.64			

4.2.4 Modeling TFR in East Java Province with Truncated Three Point Spots Components

The truncated spline semiparametric regression model using three point knots with one parametric component predictor and five nonparametric components are as follows:

$$y_i = \beta_0 + \beta_1 x_i + \theta_{11} t_{1i} + \theta_{21}(t_{1i} - K_{11})_+^1 + \theta_{31}(t_{1i} - K_{21})_+^1 + \theta_{41}(t_{1i} - K_{21})_+^1 + \theta_{12} t_{2i} + \theta_{22}(t_{2i} - K_{12})_+^1 + \theta_{32}(t_{2i} - K_{22})_+^1 + \theta_{42}(t_{2i} - K_{32})_+^1 + h(z_i) + \varepsilon_i \; ; \; i = 1,2,\ldots,38.$$

GCV values produced using a semiparametric regression of truncated spline and kernel mixture with three knots are presented in Table 7. Based on Table 7, the minimum GCV value produced is equal to 0.004115 Location of the knots on the variable (t_1) that is 57.69 (K_{11}) 72.35 (K_{12}) 79.67 (K_{13}) for (t_2) that is 72.1 (K_{21}) 76.07 (K_{22}) 78.06 (K_{23}) while the bandwidth provided is as big as $\alpha = 0.04277$.

4.2.5 TFR model in East Java Province with a Knot Spots Truncated Combination Component

The selection of a combination of knots is done by combining the optimum knots that have been obtained previously from the calculation of 1 knot, 2 knots and 3 knots. Furthermore, the minimum GCV will be calculated based on the combination obtained and the model chosen with the minimum GCV among the combinations. A truncated spline semiparametric regression model using a combination of point knots with a predictor of one parametric component and three nonparametric components are as follows.

$$y_i = \beta_0 + \beta_1 x_i + \theta_{11} t_{1i} + \theta_{21}(t_{1i} - K_{11})_+^1 + \theta_{31}(t_{1i} - K_{21})_+^1 + \theta_{12} t_{2i} + \theta_{22}(t_{2i} - K_{12})_+^1 + \theta_{32}(t_{2i} - K_{22})_+^1 + \theta_{42}(t_{2i} - K_{32})_+^1 + h(z_i) + \varepsilon_i \; ; \; i = 1,2,\ldots,38.$$

GCV values produced using semiparametric regression of truncated and kernel spline mixtures with a combination of knots are presented in Table 8. Based on Table 8 the minimum GCV value generated is equal to 0.003964 with a combination of knot points, 2.3 and the location of the knot points on the variable (t_1) that is 68.4 (K_{11}), 80.8 (K_{12}) and for (t_2) that is 72,1 (K_{21}) 76.074 (K_{22}), 78.062 (K_{23}) while the bandwidth provided is as big as $\alpha = 0.0427$.

Table 7: Comparison of GCV Values using Three Point Knots

Spline		Kernel	GCV	R^2
Knot		Bandwidth		
$t_1 = K_{11}$	$t_1 = K_{12}$	α		
$t_2 = K_{21}$	$t_2 = K_{22}$			
$t_3 = K_{31}$	$t_3 = K_{32}$			
57.69	72.1			
72.35	76.07	0.04277	0.004115	97.05
79.67	78.06			
65.02	74.09			
72.35	76.07	0.04357	0.00423	96.66
79.67	78.06			
50.36	70.11			
72.35	76.07	0.04242	0.004295	97.05
79.67	78.06			
43.04	68.12			
72.35	76.07	0.04213	0.004412	97.12
79.67	78.06			

4.2.6 Selection of the best model

Based on the GCV value for each knot point that has been calculated previously, the next best model is selected by comparing the GCV values generated by each model show in Table 9.

Table 8: Comparison of GCV Values using Knot Point Combinations.

Spline		Kernel	GCV	Combination	R^2
Knot		Band width			
$t_1 = K_{11}$	$t_2 = K_{21}$	α			
$t_1 = K_{12}$	$t_2 = K_{22}$				
$t_1 = K_{13}$	$t_2 = K_{23}$				
68.4	72.1				
80.8	76.074	0.0427	0.0039	2.3	97.04
	78.062				
56.3	72.1				
	76.074	0.04275	0.0040	1.3	96.98
	78.062				
68.4	71.72				
80.8		0.04848	0.0040	2.1	95.12
57.69	71.72				
72.345		0.04865	0.0042	3.1	95.11
79.673					

Table 9: Minimum GCV Value on Each Model

Number of Knots	GCV	R-Square
1 Knot Point	0.004164	95.06 %
2 Knot Point	0.004185	96.32 %
3 Knot Point	0.004115	97.05 %
Knot Point Combination	0.003964	97.04 %*

Table 9 shows that the minimum GCV value is found in the combination of knot points. The model chosen is the Spline model with three knots. After obtaining the minimum GCV score for the linear truncated Spline model, the next step calculates the estimate for the linear truncated Spline model. Estimated linear truncated Spline model with a combination of knot points as follows.

$$\hat{y}_i = -0,17519302433 - 0,00012335311x_{1i} - 0,00079110874t_{1i} + 0,00166672135\left(t_{1i} - 68,4\right)_+^1 + 0,00055557378\left(t_{1i} - 80,80\right)_+^1 - 0,00329040416t_{2i} - 0,01911809653\left(t_{2i} - 72,1\right)_+^1 - 0,08784561457\left(t_{2i} - 76,07\right)_+^1 + 0,19710550155\left(t_{2i} - 78,062\right)_+^1 + \sum_{i=1}^{n}\left(\frac{\frac{1}{0,0427}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{z_i - z_{ii}}{0,0427}\right)^2}}{\sum_{i=1}^{n}\frac{1}{0,0427}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{z_i - z_{ii}}{0,0427}\right)^2}}\right)y_i$$

The linear truncated Spline regression model of the combination of these knots has R^2 as big as 97.04%. This means that this model can explain TFR as much 97.04%.

5 CONCLUSION

Applications in TFR data in East Java Province in 2015 got the following results:

1. The best model obtained is by using a combination of knot points with the following equation.

$$\hat{y}_i = -0,17519302433 - 0,00012335311x_{1i} - 0,00079110874t_{1i} + 0,00166672135\left(t_{1i} - 68,4\right)_+^1 + 0,00055557378\left(t_{1i} - 80,80\right)_+^1 - 0,00329040416t_{2i} - 0,01911809653\left(t_{2i} - 72,1\right)_+^1 - 0,08784561457\left(t_{2i} - 76,07\right)_+^1 + 0,19710550155\left(t_{2i} - 78,062\right)_+^1 + \sum_{i=1}^{n}\left(\frac{\frac{1}{0,0427}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{z_i - z_{ii}}{0,0427}\right)^2}}{\sum_{i=1}^{n}\frac{1}{0,0427}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{z_i - z_{ii}}{0,0427}\right)^2}}\right)y_i$$

2. The coefficient of determination (R^2) the amount obtained is equal 97.04 percent, so that the model is suitable for use.

REFERENCES

BKKBN. 2017. *Profil Kependudukan Jawa Timur Tahun 2017*. Jawa Timur: BKKBN.

Budiantara, I. N., 2009., *Spline Dalam Regresi Nonparametrik dan Semiparametrik: Sebuah Pemodelan Statistika Masa Kini dan Masa Mendatang, Pidato Pengukuhan untuk Jabatan Guru Besar*. Institut Teknologi Sepuluh Nopember, ITS Press, Surabaya.

Ruppert, D., Wand, M. P., Carrol, R. J., 2003. *Semiparametric Regression*. Cambridge University, United Kingdom.

Eubank, R., 1999. *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York.

Hardle, W., 1994. *Applied Nonparametric Regression*. Cambridge University Press. New York.

Tupen, S. N., 2011. Uji Hipotesis Dalam Regresi Nonparametrik Spline, Tesis, Jurusan Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh Nopember, Surabaya.

Wahba, G., 1990. *Spline Models for Observational Data*. Philadelphia: Society.