Application of Semiparametric Spline Regression Model in Analyzing Factors that Influence Population Density in Central Java

Y D Sumantari 1, I Slamet 2, and Sugiyanto 2
1 Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia
2 Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia
E-mail: isnandarslamet@staff.uns.ac.id

Abstract. Semiparametric regression is a statistical analysis method that consists of parametric and nonparametric regression. There are various approach techniques in nonparametric regression. One of the approach techniques is spline. Central Java is one of the most densely populated province in Indonesia. Population density in this province can be modeled by semiparametric regression because it consists of parametric and nonparametric component. Therefore, the purpose of this paper is to determine the factors that influence population density in Central Java using the semiparametric spline regression model. The result shows that the factors which influence population density in Central Java is Family Planning (FP) active participants and district minimum wage.

1. Introduction
Indonesia is an archipelago state which is the fourth most populated country in the world. The number of population is 255 million people which separate in about 2.342 islands (BPS, 2015). Due to the number of both population and island, each island in Indonesia has different population density. Java is an island which has the most densely populated in Indonesia. Central Java is the third most densely populated province in both Java and Indonesia after West Java and East Java. Population density is influenced by many factors. According to Sugiharyanto [8], the factors that influence population density are birth, climate and strategic places, economic, and social.

Semiparametric regression is a statistical analysis method that consists of parametric and nonparametric regression (Ruppert [7]). Parametric regression is used when the curve can be represented in terms of the parametric model (Hardle [4]). By contrast, nonparametric regression refers to flexible functional form of the regression curve (Eubank [2]). There are various approach techniques in nonparametric regression. One of the approach techniques is spline. Spline is work quite well in terms of giving a visually pleasing fit and has a definite computational advantage over the other methods (Eubank [2]).
The semiparametric spline regression has been applied to many cases. We may refer to Engle [1], who used semiparametric regression to estimates the relation between weather and electricity sales. Gimenez [3] applied semiparametric regression in capture-recapture modeling. Pambudi [6] also applied semiparametric spline regression to modeling Dengue Fever sickness in Madiun. Similarly, Marina [5] also applied semiparametric spline regression to determine the factors that influence criminality percentage in East Java. All of that papers provide a good result. Therefore, in this paper we use semiparametric spline regression to determine the factors that influence population density in Central Java.

2. Semiparametric Spline Regression
Semiparametric spline regression is a semiparametric regression that use spline approach. Assumed that a pairwise data set \((x_i, y_i, t_i)\) is following semiparametric regression, then the semiparametric regression equation can be modeled by

\[ y_i = x_i' \beta_i + f(t_i) + \varepsilon_i, i = 1, 2, \ldots, n \]

where \(y_i\) is the \(i\)-th response variable, \(\beta_i\) is parameter of parametric function, \(x_i'\) is predictor, \(f(t_i)\) is nonparametric function, and \(\varepsilon_i\) is residual which has \(N(0, \sigma^2)\) distribution. Nonparametric function in this model can be formed by spline approach. Spline function with \(p\) order and \(k\) knot can be written as

\[ f(t) = \sum_{j=0}^{p} \beta_j t_i^j + \sum_{k=1}^{K} \beta_{p+k}(t - K_k)^p_{+} \]

where \(p\) is the order or degree of spline, and \((t - K_k)^p_{+}\) is piecewise of polynom with

\[(t - K_k)^p_{+} = \begin{cases} (t - K_k)^p, & t \geq K_k; \\ 0, & t < K_k. \end{cases} \]

3. Research of Methodology
Data used in this research is secondary data of population density in Central Java. The factors are density population per \(km^2\) \((y)\) as response variable, and the average age of first marriage \((x_1)\), the number of population \((t_1)\), the number of family planning (FP) active participants \((t_2)\), distance to province capital \((t_3)\), and district minimum wage \((t_4)\) as predictor variables. The steps to solve the problem of this paper are

(1) Draw the scatter plot between response variable and each predictor variables to determine the patterns in data.
(2) Determine the variable belongs to parametric or nonparametric component.
(3) Make the model of population density in Central Java using 1 knot, 2 knots, 3 knots of spline, and its combination.
(4) Choose the optimal knots by the minimum value of Generalized Cross Validation (GCV).
(5) Determine the best model by the optimal knots.
(6) Verify the significance of semiparametric spline regression parameter by \(F\)-test and \(t\)-test.
(7) Verify the residual assumption in the spline model.
(8) Interpret the model and make the conclusion.
4. Results

4.1. Scatter Plot and Knot

First step is to apply the semiparametric spline regression in estimating the relationships among variables. It can be shown by plotting the data.

![Scatter plot between y vs x1](image1)

![Scatter plot between y vs t1](image2)

![Scatter plot between y vs t2](image3)

![Scatter plot between y vs t3](image4)

![Scatter plot between y vs t4](image5)

**Figure 1.** a) Scatter plot between $y$ vs $x_1$, b) Scatter plot between $y$ vs $t_1$, c) Scatter plot between $y$ vs $t_2$, d) Scatter plot between $y$ vs $t_3$, e) Scatter plot between $y$ vs $t_4$

From the Figure 1, it can be seen that scatter plot between density population per $km^2$ ($y$) and the average of first marriage ($x_1$) represents linear curve and scatter plot between density population per $km^2$ ($y$) and the number of population ($t_1$), density population per $km^2$ ($y$) and the number of family planning (FP) active participants...
\( (t_2) \), density population per km\(^2 \) \( (y) \) and distance to province capital \( (t_3) \), and also density population per km\(^2 \) \( (y) \) and district minimum wage \( (t_4) \) do not represent any formed curve. For the variables that not represent any formed curve, it can be used spline method to approach the nonparametric function. Knot selection is important because it’ll influence the result of model. The best knot is a knot which has minimum value of GCV. The knots and its GCV is shown in Table 1.

### Table 1. Knots and its GCV

| No | Titik Knot | \( t_1 \)  | \( t_2 \)  | \( t_3 \)  | \( t_4 \)  | GCV  |
|----|------------|------------|------------|------------|------------|------|
| 1  | 1          | 576375     | 97622      | 62         | 1051655    | 728305|
| 2  | 2          | 576375     | 97622      | 62         | 1051655    | 559577|
|    |            | 975376     | 170371     | 116        | 1175603    |      |
|    |            | 234373     | 24872      | 39         | 991655     |      |
| 3  | 3          | 348374     | 76836      | 87         | 1101362    | 213616|
|    |            | 633375     | 98999      | 90         | 1161250    |      |
|    | 1,2,3,3    | 576375     | 97622      | 62         | 1051655    | 728305|
|    |            | 170371     | 116        | 1175603    |      |
|    |            | 234373     | 24872      | 39         | 991655     |      |
| 4  | 1,3,2,3    | 576375     | 97622      | 62         | 1051655    | 664548|
|    |            | 76836      | 116        | 1161250    |      |
|    |            | 98999      | 90         | 1161250    |      |
| 5  | 2,2,3,3    | 576375     | 97622      | 62         | 1051655    | 193795|
|    |            | 170371     | 116        | 1175603    |      |
|    |            | 234373     | 24872      | 39         | 991655     |      |
| 6  | 2,3,2,2    | 576375     | 97622      | 62         | 1051655    | 193795|
|    |            | 76836      | 116        | 1175603    |      |
|    |            | 98999      | 90         | 1161250    |      |
| 7  | 2,3,2,3    | 576375     | 97622      | 62         | 1051655    | 193795|
|    |            | 76836      | 116        | 1175603    |      |
|    |            | 98999      | 90         | 1161250    |      |
| 8  | 3,3,2,2    | 234373     | 24872      | 62         | 1051655    | 203049|
|    |            | 348374     | 76836      | 116        | 1175603    |      |
|    |            | 633375     | 98999      | 90         | 1175603    |      |
| 9  | 3,2,3,2    | 234373     | 97622      | 39         | 1051655    | 648211|
|    |            | 348374     | 170371     | 87         | 1175603    |      |
|    |            | 633375     | 98999      | 90         | 1175603    |      |

From Table 1, it can be shown that the minimum value of GCV is 193795. It means that the best knots is the combination of 2,3,2,2 knots.

### 4.2. Semiparametric Spline Regression Model

After we achieved the best knots, the next step is to determine the semiparametric spline regression model. Using Minitab, the result can be shown in Table 2.
Table 2. Result of Regression

| Variable | Parameter | Coefficient | $p$-value |
|----------|-----------|-------------|-----------|
|          | $\beta_0$ | 9410        | 0.221     |
| $x_1$    | $\beta_1$ | -88         | 0.762     |
|          | $\beta_2$ | -0.0163     | 0.131     |
| $t_1$    | $\beta_3$ | 0.0184      | 0.089     |
|          | $\beta_4$ | -0.00023    | 0.861     |
| $t_2$    | $\beta_5$ | -0.2123     | 0.001     |
|          | $\beta_6$ | 0.648       | 0.000     |
|          | $\beta_7$ | -1.328      | 0.000     |
|          | $\beta_8$ | 0.8825      | 0.000     |
| $t_3$    | $\beta_9$ | 16.7        | 0.243     |
|          | $\beta_{10}$ | -16.9   | 0.315     |
|          | $\beta_{11}$ | -1.17   | 0.852     |
| $t_4$    | $\beta_{12}$ | 0.00309  | 0.165     |
|          | $\beta_{13}$ | -0.00386 | 0.339     |
|          | $\beta_{14}$ | 0.01432  | 0.006     |

From Table 2, it can be formed the semiparametric spline regression model. The model is $\hat{y} = 9410 - 88x_1 - 0.0163t_1 + 0.0184(t_1 - 576375)_+ - 0.00023(t_1 - 975376)_+ - 0.2123t_2 + 0.648(t_2 - 24872)_+ - 1.328(t_2 - 76836)_+ + 0.8825(t_2 - 98999)_+ + 16.7t_3 - 16.9(t_3 - 62)_+ - 1.17(t_3 - 116)_+ + 0.00309t_4 - 0.00386(t_4 - 1051655)_+ + 0.01432(t_4 - 1175603)_+$, where the value of both $R^2$ dan MSE are 98.78% and 121250, respectively.

4.3. $F$-test and $t$-test

(1) $F$-test

$F$-test is used to verify the significance of models with null hypothesis is no significant regression model. Using $\alpha = 0.05$, the critical region is to reject the null hypothesis if $p-value < \alpha$. The result of $p-value$ is 0.000. Therefore, the null hypothesis is rejected. It means that regression model is significant.

(2) $t$-test

$t$-test is used to verify the significance of parameters with null hypothesis is no significant parameter. Using $\alpha = 0.05$, the critical region is to reject the null hypothesis if $p-value < \alpha$. From Table 2, it can be shown that the significant parameter are $\beta_5$, $\beta_6$, $\beta_7$, $\beta_8$, and $\beta_{14}$.

4.4. Residual Assumption

(1) Identically Distributed

A data is identical if its residual has homoscedasticity assumption. Homoscedasticity can be detected using Glejser test, with null hypothesis is residual has homoscedasticity assumption. Using $\alpha = 0.05$, the critical region is reject the null hypothesis if $p-value < \alpha$. The result of $p-value$ is 0.443. Therefore, the null hypothesis is not rejected. It means that residual has homoscedasticity assumption.

(2) Independent Distributed

Independent test is a test that expected the residuals have no correlation. Independent test can be check using autocorrelation function (ACF) plot. If
in ACF plot there is no lag which out of the margin level, then there is no autocorrelation in residuals. The result can be shown below.

![Figure 2. ACF Plot](image)

From Figure 2, can be shown that there is no lag which out from confidence interval. Therefore, the residuals have no autocorrelation.

(3) Normality Test
Normality test is used to determine the normality of residual. This test is using Anderson-Darling test with null hypothesis is residual has normal distribution. For $\alpha = 0.05$ and $p-value$ is 0.623, residual has normal distribution.

### 4.5. Interpretation of Model
From the analysis above, it necessary to interpret the model. The coefficient of determination for this model is 98.78%$. It means that 98.78% of population density can be explained by predictor variables. The high coefficient of determination shows that the regression line fits to data. The interpretation of semiparametric spline regression model that consists of factors that influence population density in Central Java shown below.

(1) Relationship between FP active participant ($t_2$) and population density per $km^2$ ($y$) with the assumption that other variables are constant is

$$\hat{y} = -0.2123t_2 + 0.648(t_2 - 24872) + 1.328(t_2 - 76836) + 0.8825(t_2 - 98999)$$

where

$$\hat{y} = \begin{cases} 
-0.2123t_2 & ; t_2 < 24872 \\
0.4357t_2 - 16117 & ; 24872 \leq t_2 < 76836 \\
85921 - 0.8923t_2 & ; 76836 \leq t_2 < 98999 \\
-1446 - 0.0098t_2 & ; t_2 \geq 98999 
\end{cases}$$

When the FP active participants less than 24872, if it increases by 1 people then the model predicts that the population density per $km^2$ decreases by 0.2123 people. Area in this segment is Magelang City. When the FP active participants between 24872 and 76836, if it increases by 1 people then the model predicts that the population density per $km^2$ increases by 0.4357 people. Areas in this segment are Surakarta City, Salatiga City, Pekalongan City, and Tegal City. When the FP
active participants between 76836 and 9899, if it increases by 1 people then the model predicts that the population density per $km^2$ decreases by 0.8923 people. Area in this segment is Purworejo Regency. When the FP active participants greater than 98999, if it increases by 1 people then the model predicts that the population density per $km^2$ decreases by 0.0098 people. Areas in this segment are Cilacap Regency, Banyumas Regency, Purbalingga Regency, Banjarnegara Regency, Kebumen Regency, Wonosobo Regency, Magelang Regency, Boyolali Regency, Klaten Regency, Sukoharjo Regency, Wonogiri Regency, Karanganyar Regency, Sragen Regency, Grobogan Regency, Blora Regency, Rembang Regency, Pati Regency, Kudus Regency, Jepara Regency, Demak Regency, Semarang Regency, Temanggung Regency, Kendal Regency, Batang Regency, Pekalongan Regency, Pemalang Regency, Tegal Regency, Brebes Regency, and Semarang City.

(2) Relationship between the district minimum wage ($t_4$) and population density per $km^2$ ($y$) with the assumption that other variables are constant is

$$
\hat{y} = 0.00309t_4 - 0.00386(t_4 - 1051655) + 0.01432(t_4 - 1175603) +
$$

where

$$
\hat{y} = 0.01355t_4 - 12775; t_4 \geq 1175603
$$

When the district minimum wage greater than 1175603, if the district minimum wage increases by Rp 1, then the model predicts that the population density per $km^2$ increases by 0.01355 peoples. Areas in this segment are Demak Regency, Semarang Regency, Kendal Regency, and Semarang City.

5. Conclusion

From the discussion, it can be concluded that

(1) the semiparametric spline regression model is given by

$$
\hat{y} = 9410 - 88x_1 + 0.0163t_1 + 0.0184(t_1 - 576375) + 0.00023(t_1 - 975376) +
-0.2123t_2 + 0.648(t_2 - 24872) + 1.328(t_2 - 76836) + 0.8825(t_2 - 98999) + 16.7t_3 - 16.9(t_3 - 62) + 1.17(t_3 - 116) + 0.00309t_4
-0.00386(t_4 - 1051655) + 0.01432(t_4 - 1175603) +
$$

where $y$ is population density per $km^2$, $x_1$ is the average age of first marriage, $t_1$ is the number of population, $t_2$ is family planning (FP) active participants, $t_3$ is distance to province capital, and $t_4$ is district minimum wage.

(2) The factors that influence population density in Central Java is FP active participants and district minimum wage.

References

[1] Engle, R. F., C. W. J. Granger, John Rice, and Andrew Weiss, Semiparametric Estimates of the Relation Between Weather and Electricity Sales, Journal of the American Statistical Association, Vol. 81, No. 394 (Jun., 1986), pp. 310-320.

[2] Eubank, R. L., Spline Smoothing and Nonparametric Regression, New York: Marcel Dekker.Inc, 1988.

[3] Gimenez, O., C. Crainicenau, C. Barbran, S. Jenouvrier, and B. J. T. Morgan, Semiparametric Regression in CaptureRecapture Modeling, The International Biometric Society, No. 69, (2006) 691-698.

[4] Hardle, W., Applied Nonparametric Regression, New York, Cambridge University Press, 1990.
[5] Marina, S. M. T. dan I Nyoman Budiantara, *Pemodelan Faktor-Faktor yang Mempengaruhi Persentase Kriminalitas di Jawa Timur dengan Pendekatan Regresi Semiparametrik Spline*, Jurnal Sains dan Seni POMITS Vol. 2, No. 2, (2013) 2337-3520.

[6] Pambudi, F. A., *Pemodelan Jumlah Kasus Demam Berdarah Dengue (DBD) di Madiun Dengan Pendekatan Regresi Semiparametrik Spline di Kabupaten/Kota Madiun*, Thesis of Statistics Department of FMIPA ITS, Surabaya, 2010.

[7] Ruppert, M. and R. Carroll, *Semiparametric Regression*, Cambridge University Press, New York, 2003.

[8] Sugiharyanto, *Geografi dan Sosiologi 2*, Yogyakarta: Yudhistira, 2007.

[9] http://www.jateng.bps.go.id