Modelling rice production in Central Java using semiparametric regression of local polynomial kernel approach

T W Utami\textsuperscript{1}, A Prahutama\textsuperscript{2}, A Karim\textsuperscript{1}, Achmad A R. F\textsuperscript{1}

\textsuperscript{1}Department of Statistics, University of Muhammadiyah Semarang
Jl. Kedungmundu Raya No 22 Tembalang, Semarang 50273, Indonesia

\textsuperscript{2} Department of Statistics, Diponegoro University
Jl. Prof. Soedharto, SH, Tembalang, Semarang 50275, Indonesia

E-mail: tianium@unimus.ac.id

Abstract. Indonesia is an agricultural country with rice as one of the staple foods. Production of rice in the province of Central Java is the highest in Indonesia. The purpose of this study was to model rice production in 31 districts/cities in Central Java Province using semiparametric regression. Semiparametric regression is a combination of parametric and nonparametric regression. Parametric regression curves have a patterned, for example linear, quadratic, and cubic. Nonparametric regression has a smooth curve of the unknown pattern, so in this case required smoothing technique used to smooth curves that one of them is the local polynomial kernel approach and the election of bandwidth the optimal using method Generalized Cross Validation (GCV). Variables used in the study of the production of rice as the response variable, while the predictor variables that harvested area and rainfall. The data used are secondary data from the official website of Central Bureau of Statistics (BPS) of Central Java. Based on the results obtained by applying the model the optimal bandwidth values is 0.43 and polynomial order p =2 when the minimum GCV so the results of the estimation model R$^2$ is 0.968

1. Introduction
Regression analysis has long been developed to investigate the relationship patterns and effect of predictor variables on response variable by estimating the regression curve. In relation to estimating regression curves, there are three regression models that can be used; parametric, nonparametric, and semiparametric regression models.

In some cases, a response variable is recognized for its pattern with one of its predictor variables, but with other predictor variables the pattern of relationship is not recognized. In the state, [10] suggested using a semiparametric regression approach. Some popular semiparametric regression models are kernel, spline, local polynomial, Fourier series and others. Semiparametric regression is a combination of parametric and nonparametric regressions. One of advantage of Semiparametric is able to model the data with two component such as parametric and nonparametric, so that the modelling will be optimal.

According to [7] and [6], one of the advantages of using Local Polynomial estimator is that they are theoretically easy to analyze. Besides, the other advantage is that it is able to reach the level of convergence. The previous study uses a nonparametric approach of Local Polynomial was [4] who
conducted a study on the estimation of confidence interval of nonparametric regression curve with lognormal errors based on Local Polynomial estimator. [12] investigated local polynomial nonparametric regression estimates on longitudinal data, and [13] conducted a research on sea tide modelling using nonparametric regression of Local Polynomial. The previous study using Local Polynomial Semiparametric regression approach was conducted by [11] on longitudinal data.

Paddy is a rice-producing plant used as the main foodstuff for almost 90% of the Indonesian population. The main problem faced in increasing rice production is that still relies on Java Island as the main producer of rice in Indonesia. The rice production in Java reached 56 percent. Java Island serves as a buffer for national rice production [8]. Rice production in Java Island from year to year shows an unstable condition, and it even tends to decline. The decline in rice production was mainly due to a decrease in harvested area and unsupportive weather. In fact, Central Java is one of the largest rice producing provinces in Indonesia with 97 percent of its total rice production as wet land paddy and the rest is dry land paddy. The decrease in harvested area can be explained by population growth every year which causes the demand for housing and infrastructure lands and the changes in the function of agricultural land for industrial development and others [9].

The research on the models and factors that influence the production of rice in Central Java had been widely conducted. [1] conducted a research on rice production modelling in Central Java using an approach of Fourier series.

This study discussed the modelling of wetland rice production in Central Java in 2015. The variables used in this study were: wetland rice production as the response variable and harvested area and rainfall as the predictor variables. The predictor variable with linear relationship with the response variable was harvested area, and the predictor variable with unknown pattern form with the response variable was rainfall.

2. Local polynomial kernel

The method of nonparametric regression is the regression method used when the curve is relationship between dependent and independent variable, and Independent variable is not known for the form and pattern. The common nonparametric regression model is as follows [5]:

\[ y_i = \eta(x_i) + e_i; i = 1,2,\ldots,n \]  

(1)

\( y_i \) = dependent variable  
\( x_i \) = independent variable

The function curve of \( \eta(x_i) \) is assumed to be smooth in certain function spaces. According to [11], the regression model of Semiparametric Regression Local Polynomial Kernel as in the equation: (2).

\[ \eta(x_i) \] function is a function of unknown shape called regression function. Suppose \( x \) is a predictor variable where the function \( \eta \) be estimated by local polynomial kernel estimators.

\[ y_i = \alpha_0 + t_i \alpha_1 + t_i^2 \alpha_2 + \ldots + t_i^p \alpha_p + \eta(x_i) + e_i; i = 1,2,\ldots,n \]  

(2)

\[ y = T\alpha + \eta + \varepsilon \]

With, \( \eta(t_i) \approx \eta(t) + (t_i - t)\eta^{(1)}(t) + \cdots + (t_i - t)^p\eta^{(p)}(t)/p! \)

\[ \alpha = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_p \end{bmatrix} \]
With, 

\[
C = \begin{bmatrix}
1 & (x_i - x) & (x_i - x)^2 & \ldots & (x_i - x)^p \\
M & M & M & \ldots & M \\
1 & (x_i - x) & (x_i - x)^2 & \ldots & (x_i - x)^p \\
M & M & M & \ldots & M \\
1 & (x_i - x) & (x_i - x)^2 & \ldots & (x_i - x)^p \\
M & M & M & \ldots & M \\
1 & (x_i - x) & (x_i - x)^2 & \ldots & (x_i - x)^p \\
M & M & M & \ldots & M \\
1 & (x_i - x) & (x_i - x)^2 & \ldots & (x_i - x)^p \\
M & M & M & \ldots & M \\
\end{bmatrix}
\]

The kernel function \(K\) with bandwidth \((h)\) is defined as follows: 

\[
K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)
\]

With Gaussian Kernel: 

\[
K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{1}{2} \cdot \frac{-x^2}{h^2}\right), \quad -\infty < x < \infty
\]

3. Rice production

Indonesia Ministry of Agriculture data indicates that from 1981 to 1985 and from 1998 to 1999, the conversion of paddy in Indonesia is an agricultural country with one of the staple food in Indonesia is rice [3]. The majority of rice field conversion took place in Java, which had accounted for 60% of national rice production. Production of rice in the province of Central Java is the highest in Java. In Central Java according to the Central Bureau of Statistics paddy crop production is fluctuating [2].

This results in the need for modelling to predict and know how to change productivity wetland paddy in Central Java. The result of the modelling is expected to help the concerning parties the strategic steps is needed to be done so that not suffering significant losses. Therefore, the appropriate statistical method for modelling productivity wetland paddy in Central Java is using the semiparametric regression approach of Local Polynomial Kernel.

4. Research method
4.1 Data source and research variable
The data used in this study was the secondary data obtained from the Central Bureau of Statistics of Central Java Province in 2015. In this study, a research unit of 31 districts/cities in Central Java Province in 2015 was used. The variables examined in this study consisted of response and predictor variables. The response variable was the production of wetland rice, and the predictor variables were harvested area and rainfall.

4.2 Analysis method
The analysis stages used in this study were:
1) Making a plot of rice production data with the factors deduced to influence it
2) Creating an algorithm and program for determining the polynomial order and optimal bandwidth using the GCV method
3) After obtaining polynomial order and optimal bandwidth, the next step was to create an algorithm and estimation program for semiparametric regression models using Local Polynomial Kernel approach

Modeling rice productivity in Central Java with the factors deduced to influence it using a semiparametric regression approach of Local Polynomial Kernel

5. Results and discussions
4.1 Rice Production Plot and the Influencing Factors
Rice production is influenced by several factors, such as harvested area and rainfall. The relationship of these factors with rice production had a patterned curve, and some curves had unknown patterns. Therefore, the modeling that could be applied was semiparametric regression of kernel local polynomials. The first step to get a semiparametric regression model was to create a scatterplot to determine the parametric and nonparametric components. The following is a scatterplot of rice production with the influencing factors:

![Figure 1. Scatterplot harvested area versus rice production](image1)

![Figure 2. Scatterplot rainfall versus rice production](image2)

Based on the scatterplot above, it can be seen that figure 1 shows the rice production with harvested area included in parametric components, and the variable of rice production with rainfall included in nonparametric components.

4.2 Making algorithm and programming the rice production modelling in Central Java using semiparametric regression approach of Kernel Local Polynomial
The algorithm was used to make the program in software R used to obtain the estimator of the semiparametric regression model of population mean based on the estimator of local polynomial kernel as follow as:

A. The algorithm of optimal bandwidth (h) and polynomial order.

The steps to determine the optimal bandwidth (h) and polynomial order with the criteria of GCV were as follows:
1) Defining the response, parametric and nonparametric predictor variables
2) Determining the Gaussian Kernel function
3) Determining the polynomial order of p
4) Determining the set of Bandwidth h Є [bb, ba]
5) For each h Є [bb, ba] and the polynomial order of p, the GCV was calculated using the following steps:
   a. Getting the matrix of $A = C(C^T K_b C)^{-1} C^T K_h$
   b. Getting the matrix of $A_{par} = T[T^T (I - A)^T (I - A) T]^{-1} T^T (I - A)^T (I - A)$.
   c. Getting the matrix of $\hat{y} = [A_{par} + A(I - A_{par})] y$
   d. Getting the matrix of $A^*(h) = A_{par} + A(I - A_{par})$
   e. Determine the value of $GCV(h) = (N^{-1} y^T (I - A^*(h))^T (I - A^*(h)) y) \left( N^{-1} tracel[I - A^*(h)] \right)^2$
6) Based on step 5, the minimal GCV was selected. The value of corresponding bandwidth and polynomial order to minimal GCV is optimal bandwidth and polynomial order

B. Algorithm of semiparametric regression model estimates based on the estimator of local polynomial kernel

The steps to estimate the model used polynomial order and optimal bandwidth obtained from algorithm in part A as follows:
1) Defining the response variable Y and the predictor variable X
2) Determining the Gaussian Kernel function
3) Inputting the optimal bandwidth value and polynomial order obtained from algorithm in part A.
4) Getting the matrix of $A = C(C^T K_b C)^{-1} C^T K_h$
5) Getting the matrix of $A_{par} = T[T^T (I - A)^T (I - A) T]^{-1} T^T (I - A)^T (I - A)$.
6) Getting the matrix of $\hat{y} = [A_{par} + A(I - A_{par})] y$
7) Calculating the error value of $e = y - \hat{y}$
8) Calculating the value of $MSE(h)$, with $MSE(h) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

4.3 Selection of optimal bandwidth (h)

Before making the model, the optimal bandwidth and polynomial order were determined in advance using the GCV method. Table 1 shows the results of GCV values for each bandwidth value and polynomial order.

The selection of optimal bandwidth (h) and polynomial order was seen from the minimum GCV value. Based on Table 1, the minimal GCV value is 1442709112, so the optimal bandwidth is 0.43 and the polynomial order is 2.

| Polynomial Order | Bandwidth | GCV       |
|------------------|-----------|-----------|
| 1                | 0.33      | 1442709725|
| 2                | 0.43      | 1442709112|
| 3                | 0.16      | 1442709725|
| 4                | 0.16      | 1442709725|
4.4 Rice production modelling using the semiparametric regression of local polynomial kernel

To obtain the parameters, both parametric and nonparametric components, can form a semiparametric regression modeling of local polynomial kernel as follows:

\[
\hat{y} = x\hat{\beta} + TM + \varepsilon
\]

The following is the equation of the semiparametric model of local polynomial kernel:

\[
y_t = (-2.42 \times 10^{-1}) + 6.23x_t - 5490.13 - (t_t - 137.6)185.62 + (t_t - 137.6)^2 565.09 / 2!
\]

4.5 The comparison of rice production prediction results with the actual data

Having known the modeling of rice production, the next step was to compare the predicted results of rice production based on the results of semiparametric regression of local polynomial kernel modeling with actual data. The following is the figure of the data plot as the prediction results with the raw data:

![Figure 3. Comparison of rice prediction with raw data](image)

Based on the graph, it can be stated that the actual response data with the predictive response data have the same data pattern. Then, the obtained modelling is suitable for predicting rice production. The region with the lowest prediction value of rice production is in the city of Salatiga. The average rice production in Central Java is 354432.6 kg, while the average prediction of rice production using the local polynomial kernel semiparametric model is 356368.2 kg.

4.6 Model goodness of fit

The model goodness of fit can be seen in the coefficient of determination (R^2) and Mean Square Error (MSE). The higher the coefficient of determination, the better the model, and the smaller the MSE value the better the model. The results of the R program show that the coefficient of determination and MSE are 0.9681 and 1262558130. It means that 96.81% of the predictor variables can explain the variance of the response variable, while the remaining 3.19% is influenced by other factors.

5. Conclusion

The Rice Production Modelling using Semiparametric Regression of Kernel Local Polynomial as follow as:

\[
y_t = (-2.42 \times 10^{-1}) + 6.23x_t - 5490.13 - (t_t - 137.6)185.62 + (t_t - 137.6)^2 565.09 / 2!
\]

The regression model has the coefficient of determination (R^2) of 0.9681 and MSE value of 1262558130. It shows that the predictor variables of harvest area and rainfall have the ability to explain the variance of rice production by 96.81% and the remaining 3.19% is explained by other variables which are not included in this study.

Acknowledgment
Thanks for DRPM Kemenristek Dikti for support in the PDP grant 2017-2018 with contract No: 014/UNIMUS.J/PJ/PG/2018. Thanks for the Statistics Department of University of Muhammadiyah Semarang.

References

[1] Asrini J 2013 Model Regresi Semiparametrik Deret Fourier (Studi Kasus pada Produksi Padi Sawah Irigasi di Jawa Tengah) Tesis ITS Surabaya.
[2] Balai Besar Penelitian Tanaman Padi 2015 BPS: Produksi Padi Tertinggi dalam 10 TahunTerakhir. (http://bbpadi.litbang.pertanian.go.id/index.php/berita/beritaautama/content/106-bps-produksi-padi-tertinggi-dalam-10-tahun-terakhir).
[3] Badan Pusat Statistik 2015 Survei Sosial Ekonomi Nasional BPS Jawa Tengah.
[4] Chamidah N 2004 Selang Kepercayaan Kurva Regresi Nonparametrik Dengan error Berdistribusi Lognormal (Laporan Akhir Research Grant Tahun II Oktober 2004 Unair Surabaya).
[5] Eubank R L 1999 Spline Smoothing and Nonparametric Regression (New York: Marcel Dekker).
[6] Fan J and Gijbels I 1998 Local Polynomial Modelling and its Applications (London Chapman and Hall).
[7] Hong S Y 1999 Statistica Sinica 9 775.
[8] Malian A H Mardianto S and Ariani M 2004 Jurnal Agro Ekonomi 22 119.
[9] Susianto Y 2008 Model Regresi Semiparametrik Kernel Untuk Menduga Produksi Padi Sawah Di Jawa Tengah Tesis ITS Surabaya.
[10] Wahba G 1990 SIAM Philadelphia CBMS-NSF Regional Conference Series in Applied Mathematics 59.
[11] Utami T W 2013 Estimasi Kurva Regresi Semiparametrik pada Data Longitudinal Berdasarkan Estimator Polinomial Lokal Jurnal Statistika Universitas Muhammadiyah Semarang 1.
[12] Utami T W and Nur I M 2014 International Conference on Biomedical Engineering Technology and Application UGM Yogyakarta.
[13] Utami T W and Nur I M 2015 Prosiding Seminar Nasional Matematika dan Pendidikan Matematika UNDIP.