Tadulako Science and Technology Journal
Vol. 3, No. 1, September 2022, pp. 1-7, E-ISSN : 2776-489

Comparative Analysis Using Epanechnikov and Uniform Regression on Cayenne Chili Production in Central Sulawesi in 2020

Ni Luh Emiliani [a,1,*], Ni Luh Putu Astuti Dewi Ariani [a,2], Afifa [a,3] , Agatha Sora Kedoh [a,4] , Lilies Handayani [a,5]

[a] Department of Statistics, Faculty of Mathematics and Natural Sciences, Tadulako University, Central Sulawesi, Indonesia
[1] nl.emiliani21@gmail.com*; [2] niputudewi2001@gmail.com; [3] afifaa046@gmail.com; [4] soraagatha4@gmail.com; [5] lilies.stath@gmail.com
*corresponding author

ARTICLE INFO

Article history
Received : January 02, 2022
Revised : August 26, 2022
Accepted : September 18, 2022

Keywords
Production of Rawit Chili
Regression Nonparametric
Kernel
Regression
Regression Epanechnikov
Uniform Regression

ABSTRACT

Cayenne pepper (Capsicum frutescent L.) is an annual horticultural crop with high economic value. The production of cayenne pepper in general in Central Sulawesi in 2016 reached 30.18.40 tons with a harvested area of 75 ha with a productivity of 56.74 kW. In 2017 production decreased to 21,299.50 tons with a harvested area of 2,940 ha and productivity of 7,244 kW ha-1 (BPS Central Sulawesi, 2018). The decrease in chili production is due to a decrease in the harvested area that occurs every year. The harvest area is getting smaller because a lot of lands has been built into residential areas. So farmers must be able to take advantage of all existing land even though the land is land that has a low fertility level. To obtain an increase in production and optimal growth of cayenne pepper, adequate plantation land is needed. Based on the description above, it is deemed necessary to conduct comparative analysis research using gaussian kernel regression and epanechnikov kernel regression on cayenne pepper production in Central Sulawesi in 2020.

1. Introduction

Cayenne pepper (Capsicum frutescent L.) is an annual horticultural crop that has high economic value. Because besides being a raw material for the food industry as well as a pharmaceutical raw material, it is not surprising that recently chili production has continued to increase, especially in developing countries, including Indonesia. The spicy taste and distinctive aroma, so that for certain people it can arouse appetite. In general, the nutritional value and vitamins contained in chili plants include; calories, protein, fat, calcium, vitamins A, B1, and vitamin C. Apart from being used for household purposes, cayenne pepper can also be used for industrial purposes, including; in the food industry, the drug industry, or the herbal industry. This cayenne pepper is not only used as vegetables or cooking spices but also serves as industrial raw materials [1].

The Central Statistics Agency (BPS) noted that the production of cayenne pepper in Indonesia reached 1.51 million tons in 2020. This number increased by 9.76% compared to the previous year, which was 1.37 million tons. The production of cayenne pepper in Indonesia has continued to increase over the last five years. During the 2016-2020 period, the average increase in cayenne pepper production was 13.6% per year.

The production of cayenne pepper in general in Central Sulawesi in 2016 reached 30.18.40 tons with a harvested area of 75 ha with a productivity of 56.74 kW ha-1. In 2017 production decreased to 21,299.50 tons with a harvested area of 2,940 ha and productivity of 7,244 kW ha-1 [2].

The decrease in chili production is due to a decrease in the harvested area that occurs every year. The harvest area is getting smaller because a lot of lands has been built into residential areas. So that farmers3=-0[po must be able to take advantage of all existing land even though the land is land that has a low fertility level.

To obtain an increase in production and optimal growth of cayenne pepper, adequate plantation land is needed. Based on the description above, it is deemed necessary to conduct comparative analysis research using gaussian and epanechnikov regression on the production of cayenne pepper in Central Sulawesi in 2020.

2. Literature Review

2.1 Chili Rawit

Indonesia is known as an agricultural country, so the agricultural sector is the mainstay of the livelihoods of the Indonesian population [3]. One of the leading agricultural crops is cayenne pepper (Capsicum frutescens L.). This chili plant is a multifunctional horticultural plant. It can be used as a cooking spice, sauce, or chili sauce and a mixture of medicines and has a lot of nutritional content [4]. Based on this content, cayenne pepper is a vegetable that is needed by all circles of society. Cayenne pepper production in South Sulawesi Province, with an average of the last five years (2012-2016) is 21,583 tons, with a percentage growth in 2016 during 2015 of 3.66%, while the national percentage can reach 5.29% [2].

2.2 Non-Parametric Regression

In regression analysis, two approach models are used, namely the parametric and nonparametric approaches. Parametric approach model, the shape of the regression curve is assumed to contain certain parameters so that obtaining a regression curve estimator is done by estimating these parameters. In comparison, the nonparametric approach model does not require the distribution of population parameters so that it can be used on data that has a normal distribution or not. Approach nonparametric is a statistical method that can be used to ignore the assumptions that underlie the use of statistical methods parametric, especially those associated with the normal distribution [5].

Model nonparametric regression is mathematically written:

$$y = m(x) + \varepsilon$$

Description:

y = the response variable

( )MX = function nonparametric regression loading predictor variable

ε = Error (error) absolute, formulated with $\varepsilon = y - m(x)$

2.3 Kernel Regression

Kernel regression aims to obtain nonlinear relationships between the variables X and Y. Conditional expectation Y against X is expressed as follows:

$$\text{E (Y|X) = m (X) or } \hat{y} = \text{m(x)} = \int \frac{yf\ (x.y)}{f(x)}\ \text{dy}$$

Where :

f(x,y) = density function together of (X,Y)

f(x) = marginal density function X

According to Musholawati in Indrayanti [6], the weighted regression curve fitting is not done at frequencies *X* but on the response variable *Y* around x. So observation weighting Y_i is determined by the distance X_i against x. So the estimate used is :

$$\widehat{m}\,(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} W\,(x: x_1, \ldots . x_n) y_i$$

One nonparametric regression technique that is often used to estimate the regression function (x) is the use Nadaraya- Watson estimator. This estimator is obtained by using the kernel density function estimation method. Joint opportunities density function (x, y) allegedly by multiplication as follows [6]:

$$\hat{f}\,(\mathrm{x,y}) = \hat{f}_{h1h2}\,(\mathrm{x,y}) = \frac{1}{n}\sum_{i=1}^{n} K_{h1}\,(\mathrm{x}\text{-}x_i\,)\,K_{h2}\,(\mathrm{y}\text{-}y_i)$$

So the equation of the combination conditional probability estimates in the equation Nadaraya-Watson, namely:

$$\widehat{m}\,(\mathrm{x}) = \frac{\frac{1}{n}\sum_{i=1}^{n} K_h\,(\mathrm{x}-x_i\,)\,y_i}{\frac{1}{n}\sum_{i=1}^{n} K_h\,(\mathrm{x}-x_i\,)}$$

Where

$$\sum_{i=1}^{n} K_h\,(\mathrm{x}\text{-}x_i\,) = \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$$

So

$$\widehat{m}\,(\mathrm{x}) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)}$$

with *K* is a kernel function, and h is the bandwidth or smoothing parameter and smoothness controller [7].

So :

$$y = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} + \varepsilon$$

2.4 Optimum Bandwidth Selection

There are several methods used in selecting *bandwidth* optimum, one of which is using the criteria *Generalized Cross Validation (GCV)* (Galub et al., 1979), defined by:

$$GCV = \frac{MSE}{\left(\frac{1}{n} tr(I - H(h))\right)^2}$$

With :

n = amount of data

I = identity matrix

h = *bandwidth*

X = data matrix

$H(h)$ $= X(X'X + nhI)^{-1}X'$

MSE $= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - m_h(x_i)\right)^2$

According to Komang and Gusti [8], the goodness of an estimator can be seen from the level of error. There are several criteria to determine the best estimator in nonparametric regression models, including:

a. *Mean Square Error* (MSE)

To measure *error, the* usually used *Mean Square Error* is. The best estimator is selected based on the smallest MSE value. *Mean Square Error* is the average of the squares of errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} {e_i}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

b. *Root Mean Square Error (*RMSE)

$$RMSE = \sqrt{MSE}$$

c. *Mean Absolute Deviation* (MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|e_i| = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Description :

n = amount of data

y_i = actual data

$\hat{y}_i$ = predictive value of variables y_i

3. Method

3.1 Types and Data Sources

This research is a type of quantitative research because the data used are quantitative and prioritize the ability to process data and numbers using formulas. The data used in this study is secondary data obtained from statistical data and published in general by the Central Statistics Agency (BPS) in 2020.

3.2 Identification of Variables

This study uses two variables, namely the response variable and the predictor variable, as follows:
Response Variable (Y): Production of cayenne pepper (tons)
Predictor Variable (X): Harvested area (ha)

3.3 Data Analysis Method Data

Analysis techniques are steps for solving problems to completion, which will be illustrated through the research flow chart in the flow chart image. Based on the pictures flow diagram, it can be described technique of data analysis in this study as follows:

a. Data on the production cayenne pepper and the area harvested in the Central Bureau of Statistics 2020
b. Applying regression kernel regression method *kernel Epanechnikov* and *kernel Uniform* using software *R*
c. Specifies $bandwidth$ used in the estimator
d. Enters $bandwidth$ estimator the *Epanechnikov kernel* and the *Uniform kernel estimator*
e. Calculates the GCV value and $bandwidth$ that has been used in the *kernel*.
f. Selecting the $bandwidth$ optimum based on the minimum GCV value.
g. Comparing the estimation results between the *Epanechnikov kernel* and the *Uniform kernel estimator* using $bandwidth$ is optimum.
h. Comparing *the Epanechnikov kernel* dan *Uniform kernel* with other kernel functions to confirm the selection of the best *kernel estimator* from the two previously selected functions.
i. Determine the best estimator for secondary data.

4. Results and Discussion

An overview of the data processed using software *R* in detail can be seen in Table 1.

Table 1. Descriptive statistics

	N	Min	Max	Mean	Standard Deviation
Cayenne pepper	13	83	165048	19196.54	44445.03
Harvested Area	13	33	763	298.4615	214.1275

The number of observational data is 13, with minimum cayenne pepper production being 83 tons, maximum production being 165048 tons, minimum harvested area being 33 ha, and the maximum area being 763 ha. The average production is 19196.54 tons, and the average harvest area is 298.4615 ha.

The form of the relationship between the predictor variable (Harvest Area) and the response variable (Cabe Cabe Production) is seen from the plot between the two variables (Figure 1.).

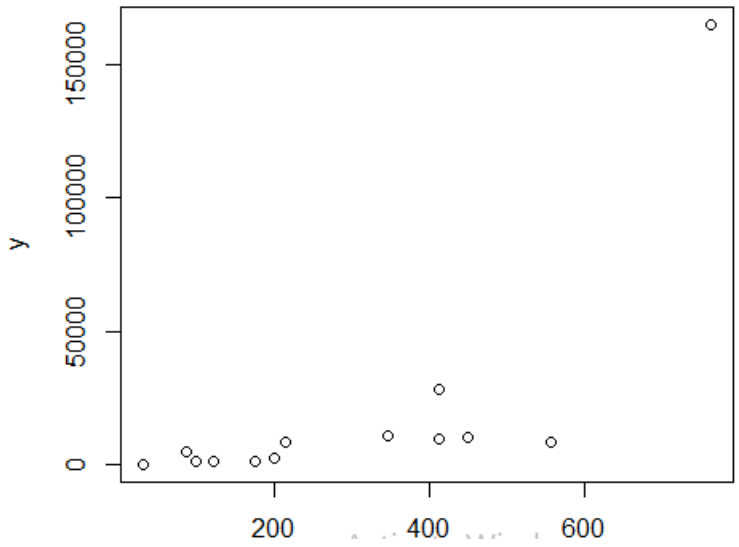
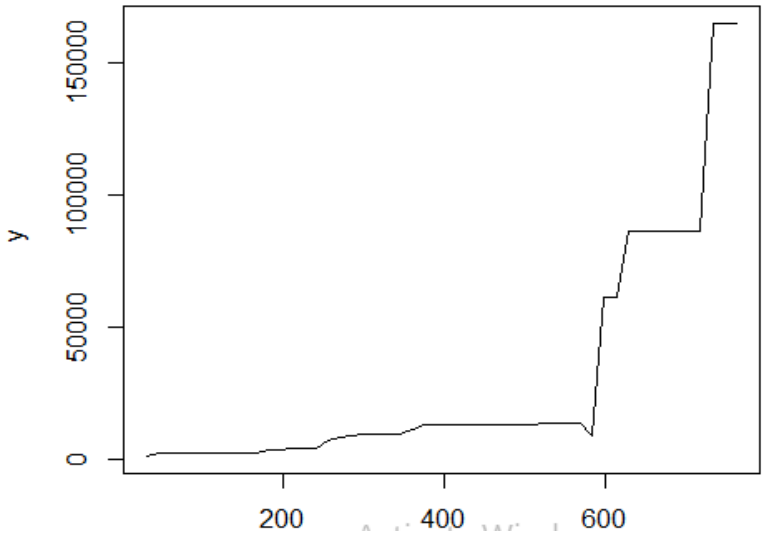


Figure 1. Scatterplot Y to X

Based on the scatterplot above, it can be seen that the plot points do not form a certain pattern, or the pattern formed is unknown. So that the model is continued with nonparametric regression, in this case, the method used is a comparison of the Uniform kernel function and the Epanechnikov kernel function.

4.1 Estimating Cayenne Pepper Production Data with a Uniform Kernel Estimator

Bandwidth selection is the most important step in kernel *smoothing*; if the bandwidth value is too small, a very rough regression curve will be obtained *(under-smoothing)*; otherwise, if the bandwidth value is too large, it will result in a very large curve smooth *(over-smoothing)*.



Gambar 2. The plot of Kernel Estimation *Uniform* with *Bandwidth* = 169.2468.

4.2 Cayenne Pepper Production Data Estimator with Epanechnikov Kernel Estimator

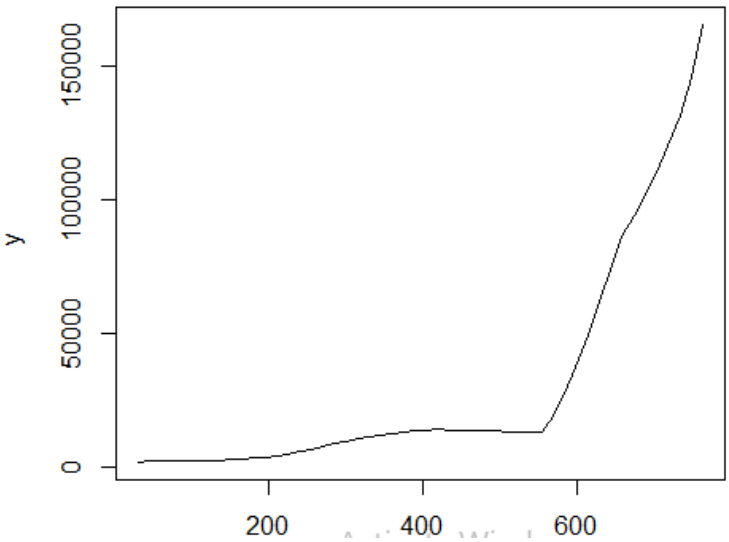


Figure 3. Kernel Estimation Plot *Epanechnikov* with *Bandwidth* =92.12453 compare

If we look at the Uniform and Epanechnikov estimation curves, it can be seen that the Epanechnikov estimation curve has a smoother shape than the Uniform estimation curve.

4.3 Comparison of Uniform Kernel Estimator and Ephanechnikov Kernel Estimator

A comparison of Uniform and epanechnikov kernel estimators can be seen in the table below :

Variable	Kernel Function	Bandwidth	R-square	MSE
X	Uniform	169.2468	0.843157	1.364265
	Ephanchnikov	92.12453	0.869083	1.852141

The table above shows the value of the Uniform kernel bandwidth estimator generated and Epanechnikov by using the optimum bandwidth. Optimum bandwidth selection is made by using software *R*.

In this study the aim of this study was to determine the production of cayenne pepper in Central Sulawesi in 2020 using the Uniform kernel nonparametric regression and the Epanechnikov kernel. It can be seen from the bandwidth value that has been obtained that the Uniform kernel function value obtained a bandwidth value of 169.2468. Meanwhile, the Epanechnikov kernel function has a bandwidth value of 92.12453. However, to see the optimum bandwidth value, we must find the smallest MSE value.

From the table above, it can be seen that the Uniform kernel function has a smaller MSE than the Epanechnikov kernel function, so a good prediction result for predicting Cayenne Pepper Production is the Uniform kernel function method with an MSE of 1.364265. A small MSE value is able to provide relatively more consistent results for all input data compared to models that have a large MSE value. In kernel regression, the most important thing is the selection of the optimal bandwidth value, not the selection of the kernel function.

5. Conclusion

Based on the results and discussions that have been carried out, it can be concluded that a good method for predicting Cabe Cabe Production is using the Uniform kernel method with an optimal bandwidth of 169.2468 and an MSE value of 1.364265. A small MSE value is able to provide relatively more consistent results for all input data compared to models that have a large MSE value. In kernel regression, the most important thing is the selection of the optimal bandwidth value, not the selection of the kernel function.

References

[1] Sunarjono, H. 2006. *Bertanam 30 Jenis Sayur. Penebar Swadaya.* Jakarta. 181 hal.
[2] Badan Pusat Statistik Republik Indonesia. (2011). Luas Panen, Produksi dan Produktivitas Cabai, 2009-2010, www.bps.go.id. Diakses pada 10 Januari 2018.
[3] Syamsu Roidah, Ida.: Pemanfaatan Lahan Dengan Menggunakan Sistem Hidroponik. *Jurnal Universitas Tulungagung* 2014, 1, 2, 43-51
[4] Arnanda, F, dan Karim, A. 2016. Pemodelan Produksi Padi di Provinsi Jawa Tengah Dengan Pendekatan Spatial Econometrics. *Jurnal Statistika*, 4(2): 20-27
[5] Eubank, R. L. 1998. Spline Smoothing and Nonparametric Regression Second edition. New York, Marcel Dekker.
[6] Indrayanti, A. (2014). *Estimator Kernel Cosinus dan Kernel Gaussian Dalam Model Regresi Nonparametrik pada Data Butterfly Diagram Siklus ke-23.* Malang: UIN Maulana Malik Ibrahim.
[7] Halim, S. dan Bisono, I. 2006. Fungsi-fungsi Kernel pada Metode Regresi Nonparametrik dan Aplikasinya pada Priest River Experimental Forest's Data. Jurnal Teknik Industri 8 (1).
[8] Komang, G., & Gusti, A. (2012). Estimator Kernel dalam Model regresi Nonparametrik. *Jurnal Matematika 2(1).*
[9] Nisa, S. (2016). Estimator Kernel Epanechnikov dan Kernel Triangel Pada Data Rata-Rata Bulanan Bilangan Sunspot, NOAA. Malang: Diakses pada : http://etheses.uin-malang.ac.id/2895/